Advancing Text Summarization with Conscience, Comprehension, and Multimodality

A THESIS
TO BE SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

Submitted by
Yash Kumar
Roll No: PhD19017

Advisor: Prof. Vikram Goyal, Dr. Tanmoy Chakraborty

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Department of Computer Science and Engineering
Indraprastha Institute of Information Technology - Delhi
New Delhi-110020 (India)

January 2024

Certificate

This is to certify that the thesis titled **Advancing Text Summarization with Conscience, Comprehension, and Multimodality**, submitted by **Yash Kumar** (PhD19017), to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Vikram Goyal
Professor
Dept. of Computer Science and Engineering
IIIT Delhi, India, 110020

Dr. Tanmoy Chakraborty
Associate Professor
Dept. of Electrical Engineering
IIT Delhi, India, 110016

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisors, Prof. Vikram Goyal and Dr. Tanmoy Chakraborty, for their invaluable guidance, unwavering support, and constant encouragement throughout my PhD journey. Their expertise, insights, and dedication have been instrumental in shaping the direction and quality of my research. Prof. Vikram Goyal's insightful critiques and meticulous attention to detail have refined my analytical skills and enhanced my research methodology. Dr. Tanmoy Chakraborty's innovative thinking and profound understanding of the field have inspired me to pursue ambitious goals and explore new research frontiers. Their mentorship has been a cornerstone of my academic and personal growth.

I am also deeply thankful to my colleagues and friends at the LCS2 lab. The camaraderie, collaboration, and intellectual discussions we have shared have enriched my experience and made my time here truly enjoyable. The friendly and collaborative culture in our lab has fostered an environment of mutual support and continuous learning. I am grateful for the numerous brainstorming sessions, constructive feedback, and the collective effort that has helped me overcome challenges and achieve significant milestones in my research. Each member of the LCS2 lab has contributed to making IIIT Delhi a welcoming and intellectually stimulating place to work.

Special thanks to my yearly reviewers, Dr. Md. Shad Akhtar and Dr. Raghava Mutharaju, for their constructive criticism, thoughtful suggestions, and continuous support. Their feedback has played a significant role in refining my work and pushing me to strive for excellence. Every insightful comment and suggestion from them has helped shape my ideas and improve the quality of my research.

Over the years, I have had the privilege of meeting and collaborating with many knowledgeable researchers and friends at my institute and various conferences and workshops. Their insights and encouragement have been invaluable.

Lastly, I extend my heartfelt gratitude to my family for their unconditional love and support, which has been my source of strength and motivation throughout this journey. It is impossible to express the debt of gratitude I owe them. Their countless sacrifices have ensured my ability to pursue my dreams, and their belief in me has been a constant source of inspiration. I cannot thank them enough for their unwavering support and encouragement. Thank you all for believing in me and helping me achieve this significant milestone.

Abstract

Summarization, an essential technique for efficiently condensing extensive textual content into concise versions, has become increasingly valuable in the face of the information deluge characterizing the modern digital era. This thesis goes beyond the traditional scope by not only pushing the boundaries of summarization techniques, focusing on both unimodal and multimodal approaches for standard and extreme summarization tasks but also addressing the critical issue of biases within summarization systems. In addition to proposing innovative methods for bias identification, the thesis introduces mechanisms to control and mitigate biases, contributing to a more comprehensive and equitable approach in the domain of information condensation and knowledge extraction.

The first section of this thesis investigates the area of unimodal summarization, focusing on the intricate task of transforming extensive textual content into succinct and coherent summaries. Existing datasets and systems often exhibit biases, both intrinsic (stemming from data) and extrinsic (introduced during training), leading to unfaithful and hallucinating summaries. By tackling the challenges of bias in unimodal summarization, this work proposes novel methods to generate coherent, and faithful summaries in both general and extreme summarization settings.

The next section of the thesis explores the area of multimodal summarization, integrating videos, audio, and text to generate comprehensive and coherent summaries. Existing literature in multimodal summarization is still in its early stages, with highly limited datasets and systems available. This is especially true for the task of multimodal summarization of scientific videos. To address this challenge, we introduce the problem of multimodal summarization and extreme multimodal summarization of scientific videos. Multimodal summarization generates concise and coherent summaries that capture the key points of a video in 5-7 lines, while extreme multimodal summarization generates extremely short summaries in 2-3 lines.

In the subsequent section, we meticulously examine biases in existing summarization systems by thoroughly evaluating both datasets and models. Employing diverse intrinsic and extrinsic metrics, we systematically identify biases, gaining a nuanced understanding of the constraints in current summarization datasets and methodologies. Building upon these insights, we introduce a novel method designed to counteract biases and enhance coherence and faithfulness while preserving crucial information. This method represents a significant step forward in advancing the reliability and integrity of summarization systems.

The research findings of this thesis have significant implications for summarization's future applications. By improving unimodal summarization, the proposed methods promise coherent, and faithful models across domains like news aggregation and decision-making. Advancements in multimodal summarization will revolutionize fields rich in multimodal data, like education and entertainment. Enabling automatic generation of comprehensive summaries from various sources empowers users to access and comprehend multimedia content efficiently. Additionally, bias identification and mitigation methods are crucial for ensuring fairness and inclusivity in summarization technologies.

Contents

List of Tables

List of Figures

Publications

Conferences

1. Alvin Dey, Tanya Chowdhury, **Yash Kumar Atri**, and Tanmoy Chakraborty. 2020. Corpora Evaluation and System Bias Detection in Multi-document Summarization. In Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pages 2830–2840.

2. **Yash Kumar Atri**, Vikram Goyal, and Tanmoy Chakraborty. 2023. Fusing Multimodal Signals on Hyper-complex Space for Extreme Abstractive Text Summarization (TL;DR) of Scientific Contents. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), 3724–3736.

3. **Yash Kumar Atri**, Arun Iyer, Tanmoy Chakraborty, and Vikram Goyal. 2023. Promoting Topic Coherence and Inter-Document Consorts in Multi-Document Summarization via Simplicial Complex and Sheaf Graph. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), pages 2154–2166.

4. Simran Kalra, **Yash Kumar Atri**, and Tanmoy Chakraborty. "Fractality-infused Graph Embeddings with FFT Transformer for Extreme Summarization (TL;DR) of Scientific Papers" (Under Review).

Journals

1. **Yash Kumar Atri**, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. Know.-Based Syst. (KBS) 227, C (Sep 2021).

2. **Yash Kumar Atri**, Vikram Goyal and Tanmoy Chakraborty, "Multi-Document Summarization Using Selective Attention Span and Reinforcement Learning," in IEEE/ACM Transactions on Audio, Speech, and Language Processing (IEEE/ACM TASL), vol. 31, pp. 3457-3467, 2023.

3. **Yash Kumar Atri**, Vikram Goyal, and Tanmoy Chakraborty. "Exploiting Representation Bias for Data Distillation in Abstractive Text Summarization." arXiv preprint arXiv:2312.06022 (2023).

Chapter 1

Introduction

1.1 Thesis Overview and Statement

In today's world, the Internet is a big part of our daily lives, helping us get information easily. Whether we're reading news, checking out scientific papers, sharing thoughts, or hanging out on social media, these activities are now second nature. But there's so much information out there that it's impossible to absorb everything. To tackle this problem, summarization has become important. Summarization is the process of condensing the crucial information while keeping the main ideas intact. Research on automatic summarization, using statistical and machine learning algorithms to create short summaries, is gaining attention, showing how technology can help us make sense of a lot of information without getting overwhelmed. This approach not only deals with the challenges of information overload but also makes learning and understanding content more efficient.

Within the scope of this thesis, we extend this narrative by proposing novel text summarization models that enhance the efficiency and effectiveness of information extraction. Additionally, we delve into multimodal text summarization methods, exploring how the integration of diverse modes such as images, audio, and video can contribute to more comprehensive and engaging summaries. Furthermore, our research addresses the critical issues of bias identification and fairness in summarization. We recognize the importance of ensuring that summarization outcomes maintain faithfulness to the original content while being free from biases that may impact comprehension. We centre our primary investigation on six crucial dimensions of summarization: abstractive text summarization, abstractive extreme text summarization, abstractive text summarization using multimodal signals, extreme text summarization using multimodal signals, fingerprinting corpus bias and its effects on systems, and instilling fairness in summarization systems.

Abstractive Text Summarization: Abstractive text summarization simplifies the extraction of vital information from lengthy documents. Unlike traditional methods that merely extract existing content, abstractive summarization generates condensed versions with new, coherent summaries written in a language similar to human-generated content. This method adds a layer of sophistication, providing concise yet comprehensive summaries (cf. Figure 1.1). It effectively addresses the challenges posed by information overload and contributes to enhancing the efficiency and accessibility of knowledge in our data-driven era. In essence, abstractive text summarization

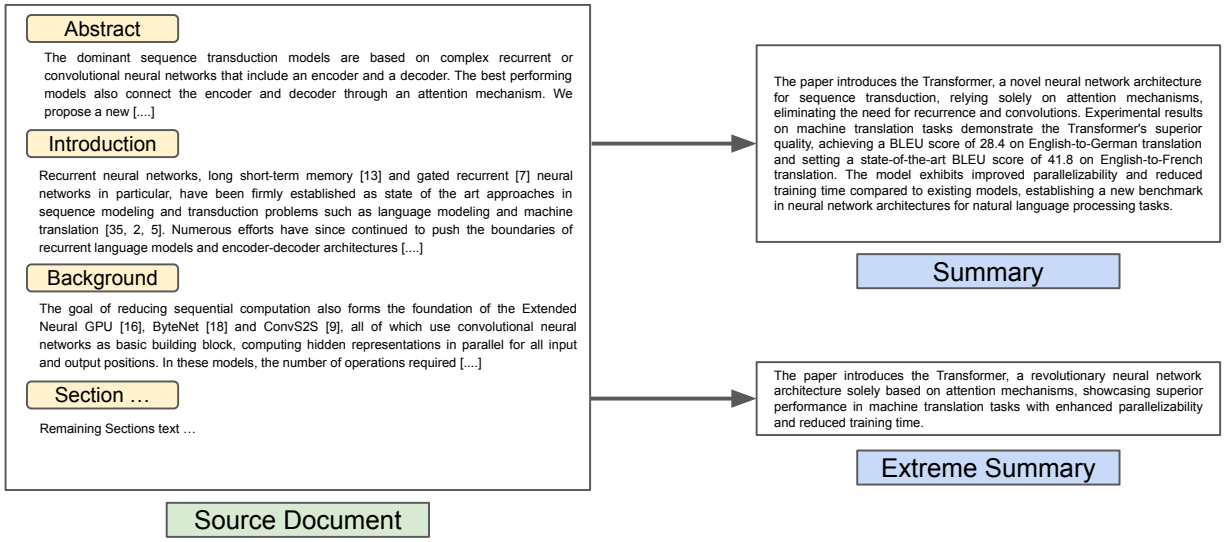


Figure 1.1: Example of abstractive summary and an extreme abstractive summary. Abstractive summarization provides detailed insights in 5-8 lines, while extreme summarization condenses information drastically for a concise overview in 1-2 lines.

acts as a bridge, making intricate content more easily understandable and accessible.

Abstractive Extreme Text Summarization: General summarization, while effective in providing condensed version of information in a few paragraphs (cf. Figure 1.1), may fall short when dealing with time-sensitive scenarios. In situations where users require quick access to critical details, traditional summarization methods might not deliver summaries with the urgency and brevity necessary for rapid decision-making. This is where extreme summarization [5] becomes indispensable. Recognizing the limitations of standard summarization approaches in time-sensitive contexts, extreme summarization aims to distill information into a short, two-line summary, capturing the essence of the content with maximum conciseness (cf. Figure 1.1)). By prioritizing the most vital elements and eliminating superfluous details, extreme summarization ensures that users can rapidly grasp key insights, making it particularly well-suited for scenarios where timely decision-making is paramount.

Abstractive Text Summarization with Multimodal Signals: In this evolving landscape, the exploration of integrating multimodal signals in text summarization holds tremendous promise for elevating the quality of generated summaries (cf. Figure 1.2). Multimodal signals, sourced from diverse channels such as text, image, audio, and video, introduce a wealth of contextual information that significantly enhances the summarization process. Harnessing these signals empowers summarization models not only to preserve the core semantics and overall meaning of the original text but also to capture nuanced tonal-specific details embedded in accompanying audio and visual cues. This capability becomes particularly crucial when considering academic presentation videos, where the convergence of multiple modalities contributes to a more comprehensive understanding of the

content being presented. The escalating volume of multimedia data, further accelerated by events like the COVID-19 outbreak, underscores the imperative of adeptly leveraging multimodal signals to effectively grapple with the intricate challenges associated with summarizing longer academic presentations.

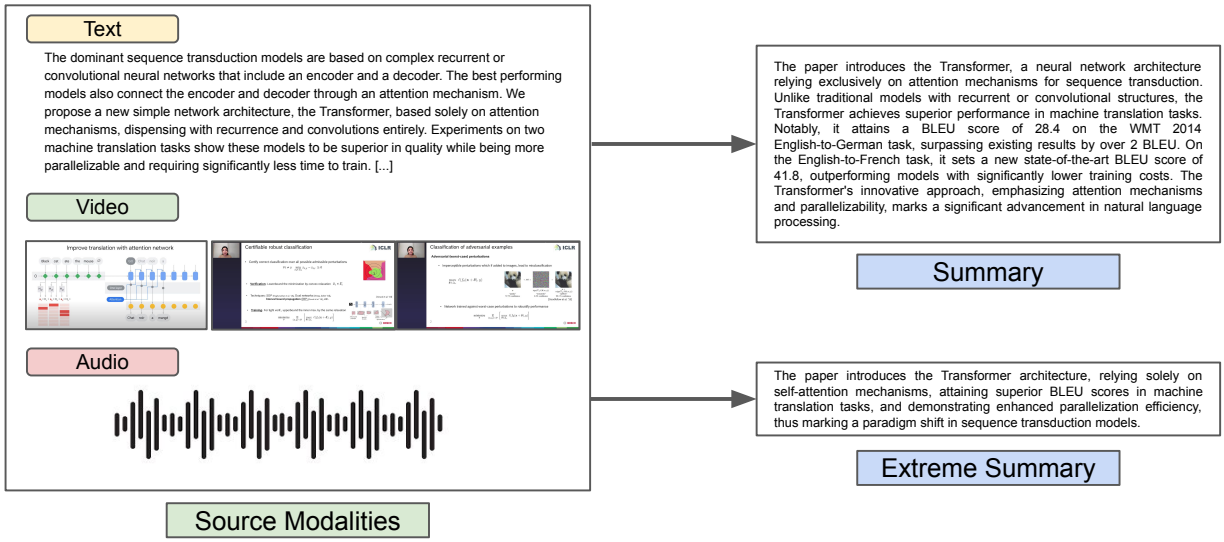


Figure 1.2: Example of abstractive summary and an extreme abstractive summary using multimodal signals (video, audio, and text). Abstractive summarization system utilizes syntactical information from text, important cues from audio, and important keywords mapping from video (OCR) to generate an informative and faithful summary.

Abstractive Extreme Text Summarization using Multimodal Signals: Additionally, the demand for extreme text summarization (cf. Figure 1.2) further accentuates the need to incorporate multimodal signals. The challenges posed by the overwhelming volume of online literature necessitate innovative approaches to distill high-level contributions from extensive textual content. Extreme summarization, aka TL;DR (Too Long; Didn't Read) [5], requires models to efficiently process and synthesize information across different modalities. In scientific content setting, where complex ideas are presented through text, images, audio, and video, a multimodal approach becomes indispensable for generating concise yet informative summaries. Traditional text-only summarization methods fall short in capturing the intricate nuances and comprehensive insights conveyed through diverse modalities. Therefore, the integration of multimodal signals not only enriches the summarization process for longer academic presentations but also becomes a crucial component in addressing the unique challenges posed by extreme text summarization, contributing to a more effective and nuanced approach in distilling essential information from complex scientific literature.

Fingerprinting Corpus Bias and its Effects on Systems: In the midst of these technological advancements, the literature emphasizes a crucial concern: the identification and mitigation of

biases[1] embedded within textual data. Within the context of summarization, biases can manifest in diverse forms [6], exerting influence over the selection and presentation of information [7]. These biases encompass considerations like the relevance of information in multi-document summarization, where certain documents may disproportionately dominate attention. Information redundancy, a prevalent bias, has the potential to overemphasize repeated content in the summary, leading to a skewed overall representation. Position bias[2], reflective of a tendency to prioritize information based on its position (cf. Table 1.3) within the source document, can introduce distortions during the summarization process.

Moreover, biases associated with evaluation metrics [8], such as the widely used ROUGE [9] metric, can impact the assessment of summary quality, potentially favoring specific types of content or structures. Recognizing and addressing these nuanced biases is paramount to ensuring the objectivity and fairness[3] of automatic summarization systems. As the field advances, efforts to mitigate biases become integral to the development of ethical and unbiased automatic summarization methodologies.

Bringing Fairness into Text Summarization: In response to the nuanced biases identified in automatic summarization systems, the imperative to prioritize fairness in terms of ensuring accurate and truthful summaries while minimizing the inclusion of fabricated or misleading information becomes increasingly apparent [10]. Acknowledging the potential impact of biases on the summarization process, there is a growing emphasis on developing models that promote fairness, inclusivity, and unbiased representation of information (cf. Figure 1.3). This pursuit of fairness involves not only detecting and mitigating biases within the corpus but also extends to the design and evaluation of summarization models [11]. This paradigm shift highlights the importance of incorporating robust mechanisms for bias detection, mitigation, and fine-grained control in automatic summarization models. By fostering a more balanced and equitable summarization process, these efforts strive to improve the accuracy and reliability of generated summaries. This research aims to contribute to a more responsible and fair integration of automatic summarization systems within the broader information landscape, ensuring that summaries accurately reflect the source material and avoid the inclusion of incorrect or invented content.

This thesis is motivated by the pressing challenges of information overload, biases, and the imperative for fairness in automatic summarization systems. The contributions of this research are - innovative methodologies are introduced to enhance unimodal summarization, covering standard to extreme summarization, while also exploring multimodal summarization. Additionally, the thesis addresses biases in summarization, providing methodologies for bias identification and assessment of datasets and systems. The research further endeavors to bring fairness into text summarization through models equipped with robust bias detection, mitigation, and control mechanisms. These

[1]In this context, bias refers to partiality or skewness concerning the faithfulness and comprehension of the source document.

[2]Position bias refers to a type of bias in summarization where the initial sentences or paragraphs of a document disproportionately influence the content and structure of the summary. This bias can lead to an overemphasis on the beginning of the text, potentially overlooking essential information located in other sections of the document.

[3]In this context, faithfulness denotes the degree to which the summarization ensures a faithful and accurate coverage of information in alignment with the original document. It emphasizes maintaining the integrity and reliability of the content throughout the summarization process.

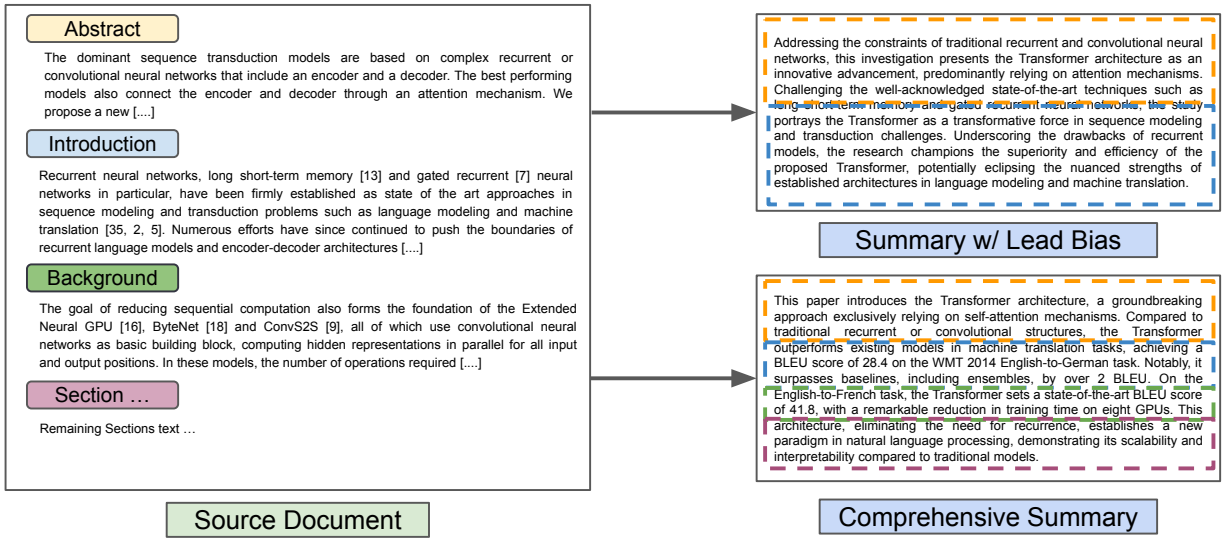


Figure 1.3: Example of abstractive summary with position bias and a summary with reduced bias. The sample on the top right demonstrates a summary with position bias, while the one on the bottom left showcases a summary with significantly reduced bias. Current systems often exhibit position bias, emphasizing the need for improved approaches that foster more balanced and unbiased summarization.

contributions collectively aim to advance the field by providing more accurate, unbiased, and ethically conscious automatic summarization techniques.

This thesis aims to improve summarization methods by focusing on both unimodal and multimodal approaches for standard and extreme summarization tasks, while also addressing the critical issue of biases within existing systems.

1.2 Background

Text summarization is the process of distilling the essential information from a given source text to create a concise and coherent summary while preserving the main ideas and key details. Text summarization comes in two flavors – extractive and abstractive (cf. Figure 1.4). **Extractive summarization** involves selecting and extracting key sentences or phrases directly from the source text to form a concise summary. This method relies on identifying the most significant information and preserving the original wording. In contrast, **abstractive summarization** aims to generate a summary by understanding the content of the source text and expressing it in a new, condensed form. This approach involves paraphrasing and rephrasing the information, often using original language to create a summary that may not be directly present in the source text. While extractive summarization maintains the integrity of the original content, abstractive summarization allows for more flexibility and creativity in conveying the essential meaning.

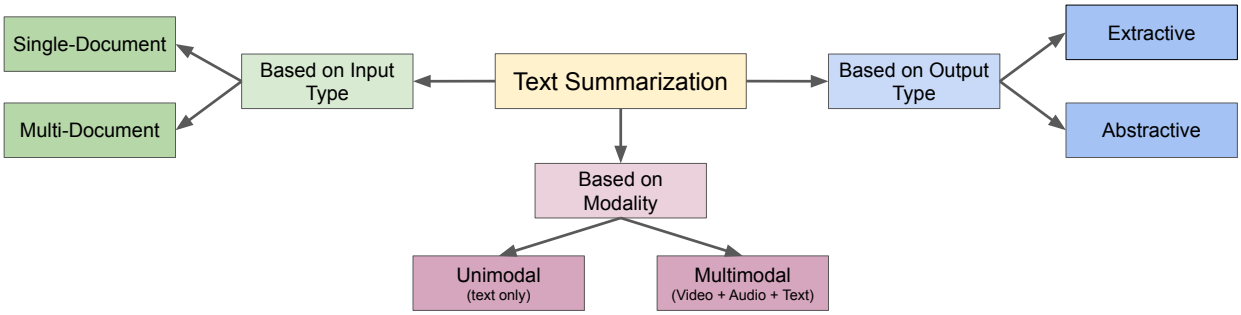


Figure 1.4: Classification of summarization approaches according to input type, output type, and input modalities. Regarding modality, the input can be categorized as either unimodal (text) or multimodal (video, audio, and text), while the output remains consistently in text form.

Single-document summarization and multi-document summarization are two distinct approaches to condensing information but differ in terms of the scope of the input data they handle. **Single-document summarization (SDS)** focuses on summarizing the content of a single document. The goal is to distill the key information, main ideas, and relevant details from a given source document into a concise and coherent summary. Single-document summarization is often used when dealing with individual articles, reports, or pieces of content, providing a condensed version while retaining the essential meaning of the original document. **Multi-document summarization (MDS)** involves the synthesis of information from multiple source documents on a particular topic. The task is to identify and extract the most important information across various documents, offering a comprehensive overview of a subject by considering diverse perspectives. Multi-document summarization is especially useful when dealing with large datasets or collections of documents related to the same theme, such as news articles on a current event or research papers on a specific topic.

Enhancing text summarization through the incorporation of multimodal signals represents a significant avenue for improvement. In this context, multimodal input refers to the utilization of diverse types of data, such as text, audio, and videos, to generate more comprehensive and contextually rich text summaries. By integrating information from multiple modalities, including textual content and associated multimedia elements, the summarization process gains a more nuanced understanding of the input, leading to improved summarization outcomes. This approach acknowledges the complementary nature of various modalities and leverages their combined strength to enhance the overall quality and depth of the generated summaries.

1.3 Challenges and Motivation

Abstractive Text Summarization: Abstractive text summarization, particularly within the context of multi-document summarization, confronts intricate challenges arising from the unique characteristics of handling multiple source documents. One significant challenge stems from the demand for extracting relevant information while considering long context relations and implement-

ing lossy information compression. Existing methods, including document concatenation [12, 13], struggle to maintain layout fidelity and address the varying importance of information across different segments. The state-of-the-art MDS systems [12, 14, 15], utilizing attention-aware [4] encoder-decoder architectures, encounter issues of producing repetitive and unfaithful summaries, resembling extractive systems [7] due to unchecked copying from source documents. The capture of long-range dependencies in MDS poses an additional hurdle, as current approaches often fall short in preserving language semantics, leading to inconsistent and unfaithful summaries. Truncating source documents before encoding introduces complexities related to incomplete information representation, resulting in heightened model hallucinations. Balancing the need for concise yet informative summaries in the face of these complexities remains a fundamental challenge in advancing the field of automatic summarization.

Abstractive Extreme Text Summarization: Summarizing a document in an extremely abstractive manner presents a significant challenge, especially given the impracticality of relying on human annotations to distill crucial information from the massive volume of yearly literature. The available human-annotated datasets are notably small, requiring the development of robust summarization systems capable of extracting meaningful insights from limited data. This challenge is crucial, considering the complex nature of documents, and it involves not only the scarcity of annotated datasets but also the necessity to create systems that can effectively summarize intricate technical documents within constraints.

Abstractive Text Summarization using Multimodal Signals: Generating abstractive text summaries for academic presentation videos using multimodal signals poses substantial challenges. This intricate task becomes even more complex when incorporating information from diverse sources such as text, audio, and video. The goal of preserving the key semantics and overall meaning of the original text while seamlessly integrating tonal-specific details from the speaker's audio and visual cues presents a multifaceted challenge. Integrating automatic speech recognition (ASR) and optical character recognition (OCR) generated text transcripts adds an additional layer of complexity, requiring the model to optimize for semantic understanding, sentiment analysis, and information extraction from various sources. Despite state-of-the-art studies [16, 17, 18, 19] in abstractive text summarization with multimodal signals primarily focusing on short videos and images, these approaches prove inadequate for the distinctive demands of longer academic videos, such as lectures or conference tutorials. The scarcity of benchmark datasets specifically tailored for this task exacerbates the challenges. Existing datasets, as utilized in recent studies, primarily comprise short videos accompanied by brief text summaries, thereby restricting their suitability for longer academic presentations. In essence, the challenges stem from the paucity of relevant datasets and the inherent complexity of integrating multimodal signals for abstractive text summarization in longer academic videos, standing as significant hurdles in advancing research in this domain.

Abstractive Extreme Text Summarization with Multimodal Signals: The introduction of TL;DR, aimed at generating extremely concise summaries from text-only articles, faces challenges when extending this notion to extreme multimodal summarization. The proposed problem statement introduces a novel task of multimodal-input-based TL;DR generation for scientific content, striving

to produce extremely concise and informative text summaries. However, integrating visual elements, tonal-specific details from audio, and aligning them with textual modalities poses substantial challenges, particularly considering the structured and complex vocabulary prevalent in scientific documents. Existing state-of-the-art approaches in related domains, such as image and video captioning, predominantly focus on shorter videos or images with limited vocabulary diversity. This stark contrast underscores the difficulties of the proposed task, which involves handling longer and more intricate scientific documents. Despite recent strides in creating datasets for multimodal text summarization of scientific presentations, current methods fall short in generating coherent summaries for the extreme multimodal summarization (TL;DR) task.

Fingerprinting Corpus Bias and its Effects on Systems: Navigating the challenges of document summarization involves a multifaceted exploration impacting both the quality of existing corpora and the efficacy of summarization systems. The challenges within this landscape span a spectrum, encompassing the variability in quality among crowd-sourced corpora dictated by factors such as community size, genre, and the presence of moderation. Despite the recent introduction of sizable MDS datasets [12, 20], a void persists in comparative studies gauging their relative complexity. Notably, existing MDS systems [4, 12, 15] exhibit divergent behavior across different corpora [12, 13], underscoring the necessity to unravel intrinsic properties influencing their performance. This study lays the foundation for crucial research inquiries concerning the modeling of MDS corpora quality, understanding system performance variations across diverse datasets, addressing biases within generated summaries, and determining the status of the MDS task—whether it is approaching completion or still poised for refinement. In essence, these challenges underscore the imperative need for a holistic understanding of MDS corpora characteristics, insights into system behavior across diverse datasets, and the formulation of strategies to alleviate biases in summarization outputs.

Bringing Fairness into Text Summarization: On the forefront of research lies the endeavor to bring fairness to summarization, particularly in the context of text summarization. Despite the advancements introduced by Language Models (LMs) such as BERT, BART, and T5, ensuring factual accuracy and faithfulness in generated summaries persists as a persistent challenge. Recent studies have explored various avenues to enhance faithfulness and factuality in summarization models, ranging from post-editing models to multi-task learning with question-answering and entailment, and the incorporation of external knowledge. Balancing the need for concise yet informative summaries in the face of these complexities remains a fundamental challenge in advancing the field of automatic summarization. Addressing biases in summarization outputs is crucial, and a comprehensive understanding of the intrinsic properties affecting system performance across diverse datasets is essential.

1.4 Thesis Organization

In this thesis, we identify the causes behind the shortcomings and propose novel methodologies to ameliorate the challenges. In particular, this thesis addresses the below research objectives. Additionally, we list the full forms of major abbreviations used in the thesis in Table 2.1.

1) Abstractive Text Summarization: In the area of text-based summarization, our focus is on introducing novel models tailored for standard abstractive summarization, typically generating concise 5-line summaries. To elevate the quality of conventional summarization, we advocate a reinforcement learning-based methodology that exploits dual rewards, namely Rouge-L and BERTScore. This approach aims to stimulate the generation of syntactically and semantically impactful phrases, enriching the overall summary. Additionally, we incorporate a reinforced attention span mechanism, compelling the model to attentively consider diverse segments of the source document. This comprehensive approach aims to enhance the standard summarization process, providing more nuanced and informative summaries for a variety of textual content.
2) Abstractive Extreme Text Summarization: Venturing into the domain of extreme text summarization, our methodology embraces the confluence of Fractality, Graph Convolutional Networks (GCN), and Fast Fourier Transform (FFT) techniques. This strategic integration enables the extraction of crucial information from two distinct spaces, leading to the creation of highly informative, coherent, and faithful summaries, distilled into 1-2 lines. Leveraging the synergistic potential of these modules, we aim to redefine the boundaries of extreme summarization, offering solutions that are both efficient and effective in capturing the essence of source documents with brevity and precision.
3) Abstractive Text Summarization with Multimodal Signals: Within the domain of multimodal summarization, this thesis underscores the importance of advancing techniques that harness the power of multiple modalities to extract comprehensive information from textual content. A pivotal aspect of this research involves the introduction of a novel multimodal dataset, and a model, tailored for summarizing conference videos. This novel approach integrates video, audio, and transcript inputs, employing a Factorized Multimodal Transformer and Guided Attention to generate summaries that are both faithful and coherent. By exploring the synergies of different modalities, the research aims to enhance traditional summarization processes, providing a more holistic understanding of textual information through the incorporation of diverse signals.
4) Abstractive Extreme Text Summarization with Multimodal Signals: Exploring multimodal extreme summarization, this thesis expands the scope of knowledge extraction from scientific content. The proposed methodology integrates human-annotated summaries with multimodal signals, offering a comprehensive approach to summarizing scientific videos. By harnessing the intricate interplay between different modalities, the research aims to provide nuanced insights and condensed representations of information within the extreme summarization framework. This approach caters to the demand for efficient and effective knowledge extraction from multimedia sources.
5) Fingerprinting Corpus Bias and its Effects on Systems: This thesis sheds light on the critical issue of biases prevalent in the summarization setting, emphasizing the necessity of identifying biases within datasets and models and recognizing their implications during inference.

Our research proposes methodologies to assess the quality of datasets and system-generated summaries using various intrinsic and extrinsic metrics. By analyzing the impact of biases, we contribute to promoting fairness in summarization, ensuring accurate and unbiased representation of information.

It is imperative to highlight that, as of now, the availability of datasets for multimodal summarization remains limited [21], posing a significant impediment to conducting comprehensive studies on biases inherent in multimodal data. Acknowledging this critical gap, we have proactively addressed the issue by introducing two novel multimodal datasets [22, 23] as part of our ongoing research endeavors. These new datasets not only contribute to filling the existing void but also play a pivotal role in advancing research initiatives focused on multimodal summarization. The continued emphasis on fairness and accuracy within summarization underscores the significance of these datasets, marking a substantial stride towards promoting unbiased representation in both multimodal data and the systems designed to process it.

6) Bringing Fairness in Summarization: Armed with insights into the potential limitations and biases within the corpus and existing models, this thesis strives to overcome these challenges by proposing novel models that not only aim for enhanced performance but also prioritize fairness, inclusivity and ethical considerations. By incorporating robust bias detection, mitigation, and fine-grained control mechanisms, our proposed models aim to foster a more balanced and equitable summarization process.

In tandem with the scarcity of datasets, it is crucial to recognize that the landscape of multimodal summarization systems is also constrained [17]. Existing systems often face limitations in addressing the diverse complexities of multimodal data, hindering the exploration of novel approaches for effective summarization. In response to this limitation, we have proposed two novel systems: FLORAL [22] and mTLDRgen [23]. By integrating linguistic, acoustic, and visual modalities, these systems aim to provide a more comprehensive understanding of source content and enhance the summarization process. The introduction of FLORAL and mTLDRgen represents a significant step forward in advancing the capabilities of multimodal summarization systems, offering novel solutions to the evolving challenges in this domain.

In conclusion, this thesis catalyzes significant advancements in text summarization, tackling the complexities posed by information overload, biases, and limitations of existing systems. The envisioned methodologies and models establish a robust bedrock for subsequent research endeavours, fostering equitable, precise, and streamlined knowledge extraction. These advancements will cater to the ever-evolving needs of information management, ensuring that emerging demands are met with fairness, accuracy, and efficiency.

Chapter 2

Related Works

2.1 Abstractive Text Summarization

Numerous advancements have been made in the regime of sequence modelling. With the introduction of seq2seq architectures and attention mechanisms, tasks like machine translation [24, 25, 26, 27], question-answering [28, 29], caption generation [30, 31, 31], dialogue generation [32, 32, 33], abstractive text summarization [34, 35, 36] have all been benefited vastly. Nallapati et al. [37] and See et al. [38] were the first to introduce the attention mechanism in seq2seq network for abstractive summarization and obtained significant improvement on the CNN/Dailymail [39] corpus. Lebanoff et al. [40] exploited the extractive maximum marginal relevance (MMR) approach to improve the pointer-generator (PG) model and showed significant gains on multiple documents. Later, Cohan et al. [41] leveraged the discourse structure of a long document to model the dependencies, while Pasunuru et al. [42] model graph and text encoders to fuse graph-based knowledge into the text-only pre-trained encoder. However, these models are benchmarked on the SDS corpora and fail when ported to the MDS corpora due to its information representation complexity and long-range relations. Fabbri et al. [12] and Chowdhury et al. [43] introduced PG extended hierarchical attention networks to capture sentence-level relations and long context relations in MDS tasks. Recently, Wu et al. [44] introduced the fusion of text encoder and graph encoder to create phrase relations in multi/long documents.

Transformer-based architectures [45] have also been applied for text summarization. Xu et al. [46] used an unsupervised approach by computing the self-attention weights to rank sentences, while Narayan et al. [47] used structured Transformers to generate extractive summaries. Gehrmann et al. [15] extended the big Transformer architecture to enable content selection in the source document and to constrain the model's visibility to generate abstractive summaries. Since the attention computation is computationally expensive, Zaheer et al. [48] proposed attention approximation in a linear time while Beltagy et al. [49] made use of both windowed local attention and task-oriented global attention in linear scalable complexity. Liu et al. [50] extended BERT [51] for both extractive and abstractive summarization by incorporating a document-level encoder.

Recent studies in transfer learning have also produced promising results in SDS. With the introduction of large pretrained language models such as BERT and GPT-2 [51, 52], the finetuning for downstream tasks achieves significant improvement with a limited number of training samples.

The T5 model [53] maps all NLP-based tasks as text-to-text problems; the model was pretrained on the C4 (Colossal) dataset and further fine-tuned on numerous NLP downstream tasks such as summarization and machine translation. However, the T5 model is highly computationally expensive and works on limited task domains. Zhang et al. [36] proposed a new pretraining objective of gap sentences by masking the most relevant sentences and tasking the model to regenerate them, and achieved state-of-the-art performance on various summarization corpora. Fabbri et al. [54] introduced data augmentation and regularization to improve the performance by few-shot transfer. Chen et al. [55] utilized language model and meta-learning to improve low-resource summarization using transfer learning. Goodwin et al. [56] used content selection to generate topic-aware summaries with zero-shot learning.

Keneshloo et al. [57] provided insights into how various RL-based algorithms pave the way for improvement in the seq2seq networks. In an unsupervised setting, Narayan et al. [58], and Dong et al. [59] ranked sentences by tasking the model to optimize the ROUGE-based rewards. In abstractive summarization, [60, 61, 62, 63, 64] introduced the notion of how RL-based models can aid in text summarization. Paulus et al. [61] introduced ROUGE-L-based rewards along with the standard log-likelihood loss to improve SDS. Böhm et al. [64] amalgamate the human feedback to the RL system to generate more human-readable summaries. Pasunuru et al. [63] focused on combining various reward metrics such as ROUGE-L, entailment and ROUGESal to improve readability. Fabbri et al. [65] combined various reward measures similar to [63]. However, all the RL-based models include rewards such as ROUGE-L, entailment, etc., constraining the model to optimize the syntactical information rather than semantics [66], which often leads to high exposure biases. Some of the rewards [65] use pre-trained models trained on a separate dataset that negatively affects the quality of the summarization system by adding extrinsic hallucination effects.

2.2 Abstractive Extreme Text Summarization

The objective of extreme summarization is to condense information within a source document into a succinct summary. Initially, this task was formulated in order to produce TLDRs for social media content [67]. Subsequently, for scientific summary generation, Cachola et al. [5] introduced a novel dataset, SciTLDR. More recently, CiteSum [68], a new corpus, has been introduced using citation information in the source document. The difference between SciTLDR and CiteSum is that CiteSum is produced by an algorithm, whereas SciTLDR is a human-annotated dataset. However, despite these advancements, the challenge of achieving extreme summarization within the realm of scientific documents remains unresolved, and there is a pressing need for a system that can generate succinct summaries without the need for data augmentation.

Scientific document summarization: Scientific document summarization has witnessed significant progress with the use of Transformers [45], and their variations [15, 69, 70, 71, 72]. The availability of datasets like Arxiv, Pubmed [73], and SciSummNet [74] has also contributed to the advancement. More recently, extensions of advanced language models like BERT [75] have further improved performance by using techniques such as sparse attention [76, 77], and directed graphs [78, 79] to compute similarity in vector space. Additionally, utilizing citation context [80] and building connected graphs [81] from citation networks have also been shown to enhance performance

in this task. Topic modelling [82, 83] has also been used to generate topic-aware scientific summaries. However, challenges still persist in summarizing very short documents.

2.3 Abstractive Text Summarization with Multimodal Signals

Abstractive text summarization with multimodal inputs has gained increasing attention in recent years with the surge of multimedia data on the Internet and social media. Unlike unimodal text summarization systems, which are vastly studied, multimodal approaches make use of visual and acoustic modalities in addition to the textual modality, as a valuable source of information for generating a fluent and informative summary. A few existing experiments [84, 85] have shown that compared to unimodal text summarization systems, multimodal summarization can improve the quality of generated summary by using the information in the visual modality.

Abstractive text summarization with multimodality deals with the fusion of textual, acoustic and visual modalities for summarizing a video document with a text precise that outlines the content of the entire video. Multimodal information is very useful in learning human-like meaning representations [86, 87, 88, 89]. Since text rarely occurs in isolation in the real world, it becomes very effective to use all available information to optimize the quality of the summary jointly. The existing literature on multimodal text summarization include multimodal news summarization [85, 90, 91] and summarization of instructional videos [17]. Li et al.[84] were the first to collect a multimodal corpus of 500 English news videos and articles with manually annotated the summaries. However, the size of this dataset is very small. Zhou et al. [92] presented the YouCookII dataset, containing instructional videos for cooking recipes with temporarily localized annotations for the procedures. [93] introduced the notion of multimodal summarization with multimodal output, which takes the news with images as input, and finally outputs a pictorial summary. Most recently, Palaskar et al.[17] studied the task of summarization of instructional videos on the How2 dataset [21], which can be considered as the closest task to ours. However, all existing multimodal text summarization methods focus on summarizing images and/or short videos and generate one- or two-line long summary, and thus, can not be generalized to longer videos.

2.4 Abstractive Extreme Text Summarization with Multimodal Signals

The objective of extreme text summarization is to drastically reduce the size of the source document while preserving its essence in the resulting summary. The concept of extreme summarization was first introduced by voelske et al. [67] with a novel dataset focused on social media summarization. Subsequently, cachola et al. [5] and Mao et al. [68] presented new corpora, namely SciTLDR and CiteSum, respectively, for extreme summarization of scientific documents. However, it remains an open area to explore extreme abstractive text summarization using multimodal signals.

2.5 Fingerprinting Corpus Bias and its Effects on Systems

Previous attempts to evaluate the quality of the benchmark summarization corpora are few in number and mostly from the time when corpora were manually accumulated. [94] primarily used the intrinsic metrics of precision and recall to evaluate corpus quality. In addition, the authors proposed an extrinsic metric, called 'Pseudo Question Answering'. This metric evaluates whether a summary has an answer to a question that is otherwise answerable by reading the documents or not. Although effective, the cost of such an evaluation is enormous and is not scalable to modern day corpora sizes. For such corpora where multiple references are available, [95] used an inter-annotator agreement to model the quality of the corpora. They also used non-redundancy, focus, structure, referential clarity, readability, coherence, length, grammaticality, spelling, layout, and overall quality as quantitative features for an MDS corpus. Recently, [43] proposed an MDS system that used the baseline PG model along with Hierarchical structural attention to take into account long-term dependencies for superior results compared to baseline models.

There have been a series of very recent studies that look into how to strengthen the definition and discover system biases in single-document summarization. Very recently, [96] studied how position, diversity and importance are significant metrics in analyzing the toughness of single-document summarization corpora. Another recent work [97] extensively studied the Layout Bias in news datasets that most single-document summarization systems seem to exploit. Two seminal works, namely [6] and [98], exploited the theoretical complexity of summarization on the ground of importance, analyzing in-depth what makes for a good summary. [6] mathematically modeled the previously intuitive concepts of *Redundancy, Relevance and Informativeness* to define *importance* in single-document summarization. [99] proposed a new single-document summarization corpus and quantified how it compares to other datasets in terms of diversity and difficulty of the data. They introduced metrics such as *extractive fragment density* and *extractive fragment coverage* to plot the quality of SDS corpus. To the best of our knowledge, no comparative work exists for either corpora or systems in MDS, and the current paper is the first in this direction.

2.6 Bringing Fairness into Text Summarization

Existing methods for improving faithfulness include proposing new fact-corrected datasets [100], ensembling language models [101], adversarial techniques [102], target correction [103, 104, 105], Natural Language Inference (NLI) models [106], rejecting noisy tokens [107], controlling irrelevant sentences [108], and contrasting candidate generation [109]. Others use auxiliary information from source documents like additional input [110], entity [111], knowledge graphs [112, 113, 114] or RL-based approaches with rewards as entailment feedback [115], and structured fact extraction [116]. Nevertheless, these methodologies either raise question upon the veracity of the datasets, cull out discordant data points or employ extrinsic data sources to alleviate the factual incongruities. Formulating a new dataset entails substantial costs, and winnowing existing ones might diminish the dataset pool for training a competitive model. In contrast, our proposed approach surmounts these constraints, capitalizing on the entirety of accessible data during training and leveraging the dynamics of simplicial complexes and sheaf graphs to apprehend the inter- and intra-document relationships, thereby engendering exceedingly faithful and factual summaries.

2.7 Abbreviations

Table 2.1: List of abbreviations and their meanings in the thesis.

Abbreviations	**Meaning**
SDS	Single-document Summarization
MDS	Multiple-document Summarization
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
DUC	Document Understanding Conference
TAC	Text Analysis Conference
MMR	Maximal Marginal Relevance
PG	Pointer Generator
BLEU	Bilingual Evaluation Understudy
ILP	Integer Linear Programming
SOTA	State-of-the-Art
SciTLDR	Scientific TLDR
ASR	Automatic Speech Recognition
OCR	Optical Character Recognition
TL;DR/TLDR	too long; didn't read
RL	Reinforcement Learning
GCN	Graph Convolutional Network
FFT	Fast Fourier Transformer
MLE	Maximum Likelihood Estimation
REISA	Reinforced Span Attention Calibration
AVIATE	Audio Video Language Dataset
FLORAL	Factorized Multimodal Transformer based Decoder-only LM
mTLDR	Multimodal TLDR
mTLDRgen	Multimodal TLDR Generator
FABRIC	Fairness using BART, Simplicial Complex, and Sheaf

Chapter 3

Abstractive Text Summarization

Abstractive text summarization systems using recently-improved RNN-based sequence-to-sequence architecture have shown great promise for the task of single-document summarization. However, such neural models fail to perpetuate the performance in the multi-document summarization setting owing to the long-range dependencies within the documents, overlapping/contradicting facts and extrinsic model hallucinations. These shortcomings augment the existing models to generate inconsistent, repetitive and non-factual summaries. In this work, we introduce `REISA`, a sequence-to-sequence model with a novel *reinforced selective attention span* that attends over the input and recalibrates the local attention weights to focus on important segments while generating output at each time step. `REISA` utilizes a reinforcement learning-based policy gradient algorithm to reward the model and formulate attention distributions over the encoder input. We further benchmark `REISA` on two widely-used multi-document summarization corpora – Multinews and CQASumm, and observe an improvement of +2.91 and +6.64 ROUGE-L scores, respectively. The qualitative analyses on semantic similarity by BERTScore, faithfulness by question-answer evaluation and human evaluation show significant improvement over the baseline-generated summaries.

3.1 Introduction

The Internet has become a pre-eminent constituent of our day-to-day life. We utilize the textual formulation of the content to keep us updated by reading news articles, exchanging ideas, and discussing on social media. However, the ever-increasing textual data has curtailed us from going through every bit of information, thus limiting our exposure to wide information. Thankfully, the task of abstractive text summarization has very much helped us quickly grasp the gist of the topic. Early research in text summarization mainly focused on the extractive summarization [117, 118, 119, 120], which lately has been advanced with the abstractive summarization techniques [121, 122, 15, 22].

Abstractive text summarization is broadly classified into two types – single-document summarization (SDS) and multi-document summarization (MDS). Despite apparent similarities w.r.t. the high-level task, MDS is far more challenging than SDS due to its inherent complications. Unlike SDS, MDS deals with n number of source documents to generate a single reference summary. Filtering out the relevant information from each individual document requires long context relations

and lossy information compression. One of the most common ways of representing the source documents is by concatenating them. Multinews [12] concatenates each individual document in a series, while CQASumm [13] arranges each answer of a question using a heuristic to put the most relevant document on the top. As evident from Dey et al. [7], the information layout in a document plays an important role in the performance of the summarization system as the models tend to pick most of the information from the first few segments of the document rather than giving importance to each segment. We aim to address this issue of layout fidelity and long context relation in the summarization models.

The field of abstractive summarization has benefited vastly from the improvements in the attention-aware encoder-decoder architecture [25]. Nallapati et al.[37] and See et al.[38] used attention mechanism for single-document summarization (SDS) and achieved high accuracy; however, the SDS corpora used by them were restrictive in terms of the length of both the source documents and the reference summaries [20]. Our analysis with the baselines on MDS corpora based on quantitative and qualitative parameters reveals that the summaries generated by the sequence-to-sequence (seq2seq) based systems are often repetitive and unfaithful. The coverage mechanism [123] used in traditional seq2seq networks [38, 12] and transformer-based [15, 69] models penalizes the decoder on generating the same set of trigrams, which in turn forces the decoder to again generate inconsistent, non-readable and unfaithful summaries. The pointing mechanism [124] on the vocabulary probabilities also gets skewed and forces the model to copy from the source document rather than generating novel phrases to minimize the training loss. The unchecked copying based only on the probability distribution makes the abstractive model behave more like an extractive system.

The long-range dependencies in the source documents in MDS aid in the complexity of the task. Certain studies for capturing long-range dependencies in MDS corpora are proposed – Huang et al. [125], Jiang et al. [126] and Chowdhury et al. [43] introduced structured attention; Cohan et al. [41], Xu et al. [127] utilizes the discourse information on documents; Manakul et al. [128] and Lebanoff et al. [129] exploited the attention and content selection. However, when these systems are modelled on long or multiple documents, the generated summaries show huge inconsistency in language semantics and high unfaithfulness w.r.t. the source document [130, 131, 132]. This can be attributed to the fact that the source documents are often truncated before feeding them to the encoder. However, the truncated input may not encode the information completely, resulting in high model hallucinations.

Our exhaustive study on 15 baselines for the SDS system unfolds their limitations when used on MDS datasets. We further envision the solution of MDS by incorporating reinforcement learning (RL) based dual reward policies. Where the reward function based on the ROUGE score [133] helps the model to syntactically capture the information, BERTScore [134] based reward aids by capturing the semantic context. Recent studies Li et al. [135], Pasunuru et al. [63] showed promising results by adding reward policies in the RL-based setting on the SDS datasets. However, the attention distribution obtained by the base network architecture is often arbitrary, resulting in redundant summaries. Here, we attempt to mitigate the issue of non-redundancy by recalibrating the attention weights for both the encoder and decoder at each time step. This not only impels the model to look at other source document segments during summary generation but also helps the model to generate more faithful summaries.

Major contributions. Our major contributions are threefold:

1. We propose a novel method, `REISA`[1] that recalibrates the attention weights at the encoded sequence based on the generated token at each time step t. The attention weights are recalibrated using the policy gradient algorithm.

2. We propose *dual loss dual reward functions*. For the RL framework, we devise ROUGE-L and BERTScore-based rewards, which not only minimize the exposure bias introduced due to the classic log-likelihood loss but also produce more faithful summaries.

3. We evaluate `REISA` using three standard metrics – ROUGE, BERTScore and QAEval, to assess the quantitative, and qualitative performances and faithfulness. We demonstrate that `REISA` fares well against five widely-popular extractive and ten abstractive baselines. `REISA` beats the strong Transformer based model [136] (best baseline) by 2.91 ROUGE-L on Multinews and [137] by 6.64 ROUGE-L on CQASumm. On abstractiveness, `REISA` generates highly abstractive summaries compared to all the baselines. Our qualitative analyses on semantic similarity by BERTScore and faithfulness by question-answer evaluation also show considerable gains compared to the baseline-generated summaries. We also perform an exhaustive human evaluations to compare the quality of the system-generated summaries.

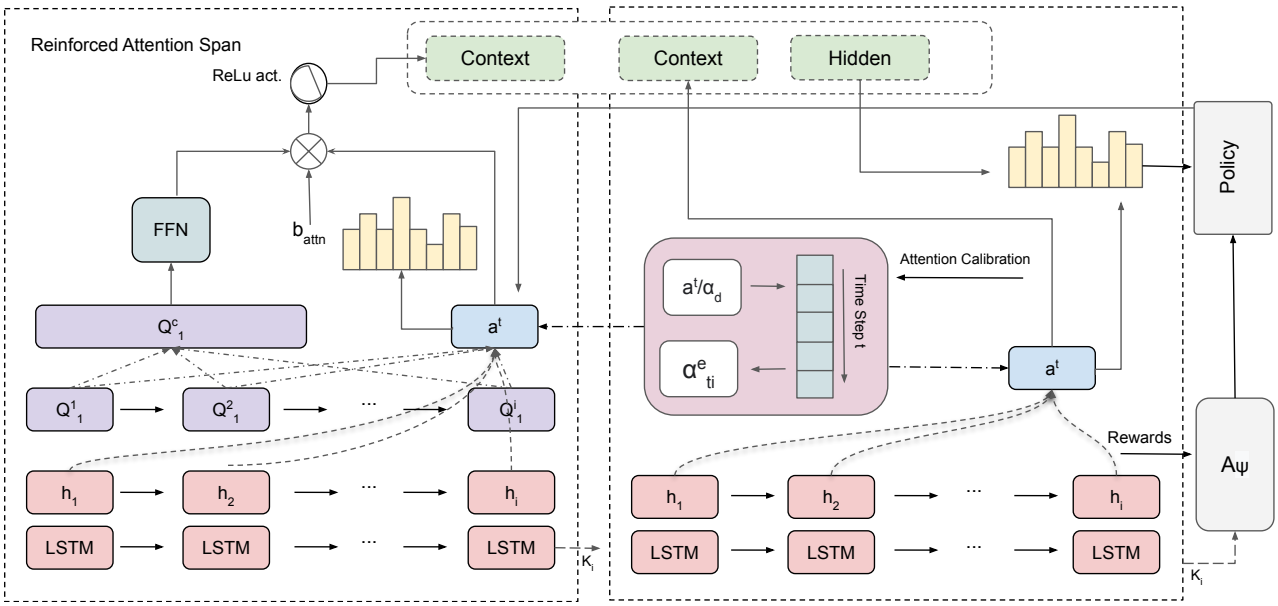


Figure 3.1: Illustration of `REISA`. a^t and α^d are attention distributions of the encoder and decoder, respectively. The context vectors and hidden state are represented by C and H/h, respectively.

[1] `REISA`: **REI**nforced **S**pan **A**ttention calibration

3.2 Proposed Architecture

In this section, we introduce our proposed `REISA` model for multi-document summarization. We first describe the base seq2seq architecture. We use bi-directional LSTMs for both the encoder and decoder. We then elaborate on how the encoder attention weights are recalibrated using the reinforced attention span. The recalibration of attention weights is recomputed at each decoder time step to help the model attain all parts of the source document. Based on the attention scores obtained during a token generation, we penalize the highest attention weights. We then describe the dual reward functions, ROUGE and BERTScore, used in the RL setting. We use both syntactical and semantic level rewards to help the model capture better contextual information. Figure 3.1 shows the schematic architecture of `REISA`.

3.2.1 Base Architecture

We make use of a single bidirectional LSTM layer. Input tokens w_i are fed to the encoder to obtain a hidden representation h_i at each time step t. A single layer unidirectional LSTM obtains the previous word embeddings and the decoder state h_j. The initial attention distribution is obtained by [25]:

$$\begin{aligned} e_i^t &= v^T \tanh(W_h h_i + W_s d_t + b_{attn}); \\ a^t &= \text{softmax}(e^t); h_t^* = \sum_i a_i^t h_i^t \end{aligned} \tag{3.1}$$

We obtain h_t^* as the initial context vector in Equation 3.1. The network parameters are initialized using the above equations.

3.2.2 Reinforced Span Attention

As evident from Equation 3.1, attention distribution [25] is computed over a sequence which is formulated into a context vector. The context vector represents the importance of individual tokens as a weight matrix. However, during the attention weight calculation, the softmax function based on e^t assigns a non-zero score to each token. Therefore, for some tokens, the weights can be attributed to a near-zero value. However, certain tokens with near-zero values can hold importance and may adversely affect the model learning strength. To address these issues, we introduce the notion of *reinforced attention calibration* using the policy gradient training algorithm with ROUGE-L and BERTScore as rewards. The reward functions are discussed later in Section 3.2.3.

For the reinforced attention calibration, we extend the studies of [138] and [139]. We compute attention weights w_i from a given function $f(T, v)$ as follows:

$$w_i = softmax(f(T, v)) = \frac{exp(f(T, v_i))}{\sum_{j=1}^{n} exp(f(T, v_j))} \tag{3.2}$$

Here, T is the target vector, and v_i is the source vector representation. Context vector is computed by $C = \sum_{j=1}^{n} w_j v_j$.

As it can be seen from Equation 3.2 that $\sum_{j=1}^{n} w_j = 1$, the attention mechanism can be thought of as a computation that adaptively learns the distribution over input vectors, accurately representing the context of the problem.

We extend the attention mechanism to adaptively sync the REINFORCE [140] reward settings. Instead of using multi-heads in the attention module of the seq2seq-based network, we make use of RL agents. A total of $m + 1$ critics accompany each agent i. The idea here is that each critic will be able to attend to the information at different document segments, allowing the network to capture long-range dependencies in multiple documents. Finally, the single-layer perceptron encodes the state and state-action information.

The perceptron output is passed on to a feed-forward embedding layer, which in turn formulates the attentions: Key, Value, and Query pairs representing the state-actions and encodings for each agent. Finally, we compute the attention weights using:

$$w_j = softmax\Big(\frac{q_i K_j^T}{\sqrt{d_k}}\Big), \tag{3.3}$$

where d_k represents the key size. The agent critic[2] Q^i is obtained using

$$Q^i = f_i(g_i(o_i, a_i), \sum_{j \neq i} w_j V_j), \tag{3.4}$$

where f_i is a multi-layer perceptron with two layers, o_i is the observation, g_i is an embedding function for agent i, V_j is a state-action encoding, and x_i represents the contribution of other agents to agent i.

The attention vector $W_i(w_i)$ is formulated with all action conditional Q-values $Q_i^k(s, a_i|\vec{a}_{-i}; w_i)$. The Q values represent the critic network, a_i represents the local action of agent i, and a_{-i} represents the joint action. We compute the dot product to calculate $W_i^k(w_i)$:

$$W_i^k(w_i) = \frac{\exp(h_i(w_i) Q_i^k(s, a_i|\vec{a}_{-i}; w_i))}{\sum_{k=1}^{K} \exp(h_i(w_i) Q_i^k(s, a_i|\vec{a}_{-i}; w_i))} \tag{3.5}$$

Here, w_i is the attention weight, and h_i indicates the encoder's hidden state.

The K action conditional Q-value $Q_i^k(s, a_i|\vec{a}_{-i}; w_i)$ is generated for each $\vec{a}_{-i}$ to approximate $Q_i^{\pi_i}(s, a_i, \vec{a}_{-i})$, where w_i is the critic network parameter for i. The $Q_i^k(s, a_i|\vec{a}_{-i}; w_i)$ is generated using a_i on all observations $a_i, \vec{a}_{-i}$. The action conditional Q-values $Q_i^k(s, a_i|\vec{a}_{-i}; w_i)$ are computed using

$$Q_i^{\pi_i|\vec{\pi}_{-i}}(s, a_i) = \Sigma_{\vec{a}_{-i} \in \vec{A}_{-i}} [\vec{\pi}_{-i}(\vec{a}_{-i}|s) Q_i^{\pi_i}(s, a_i, \vec{a}_{-i})] \tag{3.6}$$

Here, in the current state s, π_{-i} represents the joint policy, a_i indicates agent i and $\vec{a}_{-i}$ denotes the other agents.

Lastly, the *contextual Q-value* $Q_i^c(s, a_i, \vec{a}_{-i}; w_i)$ is calculated as a weighted summation of W_i^k and Q_i^k:

$$Q_i^c(s, a_i, \vec{a}_{-i}; w_i) = \sum_{k=1}^{K} W_i^k(w_i) \tag{3.7}$$

[2]In the RL setting, the actor decides which action should be taken, while the critic informs the actor how good the action was and how it should be adjusted.

To align Q_i^c with the encoder-side attention weights, we compute the new contextual Q-value using

$$Q^i = f_i(Q_i^c(s, a_i, \vec{a}_{-i}; w_i)), \tag{3.8}$$

Here, f_i indicates the feed-forward layer.

For the score e_t, the new attention weights are recalibrated by

$$e'_{ti} = exp(e_{ti}) / \sum_{j=1}^{t-1} \exp(e_{ji}) \tag{3.9}$$

where $exp(e_{ti})$ represents the initialized attention weights and Q^i represents the reinforced attention weights. Given Q^i based on rewards obtained, we calibrate $exp(e_{ti})$ by learning a new attention function and assigning a penalty to the high attention weights. The new attention weights at each time step are learned using:

$$\alpha_{ti}^e = v_a tanh(W_a[e'_{ti}; Q^i]) \tag{3.10}$$

where both v_a and W_a are weight matrices that are learned for alignment. The attention weights are recomputed at each decoder time step to update the encoder attention weights.

Finally, the normalized attention scores α_{ti}^e are computed for all inputs:

$$\alpha_{ti}^e = \frac{\alpha_{ti}^e}{\sum_{j=1}^{n} \alpha_{ti}^e} \tag{3.11}$$

and based on α_{ti}^e and the hidden state h_i^e, the context vector c_t^e is formulated as:

$$c_t^e = \sum_{i=1}^{n} \alpha_{ti}^e h_i^e. \tag{3.12}$$

For intra-decoder attention, we use a similar approach as given by [61]. At each decoding step t, the context vector c_t^d is computed. c_1^d is initialized with zeros and at $t > 1$, we use the following equation:

$$e_{tt'}^d = W_a(h_t^d h_{t'}^d) \tag{3.13}$$

Finally, the normalized attention scores α_{tt}^d at the decoder side are computed to formulate the context vector c_t^d as follows:

$$\alpha_{tt'}^d = \frac{exp(e_{tt'}^d)}{\sum_{j=1}^{t-1} exp(e_{tj}^d)}; c_t^d = \sum_{j=1}^{t-1} \alpha_{tj}^d h_j^d \tag{3.14}$$

For word generation, we adopt PG network [38]. The recalibrated attention scores represent the P_{copy} copy probabilities, and the decoder state generates P_{vocab}. P_{gen} acts as a switch to allow the model to copy or generate words from the vocabulary.

$$p(word) = p_{\text{gen}} p_{\text{vocab}}(word) + (1 - p_{\text{gen}}) a_i^t \tag{3.15}$$

3.2.3 Reward and Loss

Paulus et al. [61] used ROUGE-L as a reward measure for the summarization task, which again motivates the model to copy the syntactical information rather than the semantic information. To promote both syntax and semantic capabilities in a model, we introduce the dual reward optimization approach. We use ROUGE-L for syntax-level reward and BERTScore for semantic-level reward.

1. **ROUGE:** Similar to [63], we include the ROUGE-L reward based on the oracle summaries [141]. Our experiments corroborate [63] and [61] that ROUGE-L acts as a better reward measure in support of promoting important phrases than unigram (ROUGE-1) or bigram (ROUGE-2) based rewards. 2. **BERTScore:** Since ROUGE-L alone can be biased towards copying the token/phrases verbatim, we also include BERTScore [134] as another reward for better context-based information optimization. For BERTScore computation, we use "bert-base-uncased" [51] to compute the similarity between the generated summary and the reference summary. The reward scores are combined linearly as $Rw_{mixed} = \nu Rw_{ROUGE} + (1-\nu)Rw_{BertScore}$ Here ν is a tunable parameter.

For the loss function, along with the negative log-likelihood loss, we also incorporate an auxiliary loss function based on the RL policy gradient algorithm [61]. To generate an intermediate summary and optimise the ROUGE-L score during training, we use the oracle summary generation method provided by [7]. The generated oracle summary is extractive in nature and is derived using Integer Linear Programming (ILP) solvers.

$$R_N(R,O) = \frac{\sum_{j=1}^{|U(R)|} \min(C(g_j^N, O), C(g_j^N, R))}{\sum_{j=1}^{U(R)} C(g_j^N, O)} \tag{3.16}$$

Here given the greedily decoded summary from the current probability distribution R_s and an oracle summary as O_s, the ROUGE-L score is maximised. We utilise $R_N(R,O)$ and the output y^s from the sampled decoding probability, $p(y_t^s|y_1^s,...,y_{t-1}^s,x)$ at the current state. The RL loss is computed as:

$$L_{rl}(x,y^*) = (r(\hat{y},x,y^*) - r(y^s,x,y^*)) \sum_{t=1}^{N} \log p(y_t^s|y_1^s,...,y_{t-1}^s,x)$$

We also make use of the standard negative log-likelihood loss:

$$L_{ml} = -\sum_{t=1}^{N} \log p(y_t^*|y_1^*,...,y_{t-1}^*,x)$$

and linearly combine both as $L_{mixed} = \gamma L_{rl} + (1-\gamma)L_{ml}$.

Here γ is a tunable parameter. We use L_{mixed} as a combined loss function. We apply equal weights to both the reward functions.

3.3 Datasets

For comparative evaluation, we run the competing summarization systems on one SDS and two benchmark MDS corpora.

1. **Multinews** [12]: This dataset comprises news articles curated from the `newser.com` site, and summaries are written by professional editors. On average, the source documents consist of $2k$ tokens, while the reference summaries consist of 260 tokens. The dataset is split into $44,972$ (80%) training instances, while the validation and test sets contain 5622 (10%) instances each. With respect to abstractiveness, the Multinews dataset exhibits 32.28% novel unigram, 67.53% novel bigrams, and 80.45% novel trigram formation in reference summaries.

2. **CQASumm** [13]: It makes use of the L6 yahoo! answers dataset for the MDS task. Here, the highest voted, or the accepted answer is treated as the reference summary, while the other answers are treated as individual source documents. The reference summaries are truncated to hundred words. The dataset is split into $80k$ (80%) training instances and $10k$ (10%) validation and test instances, each. In terms of abstractiveness, CQASumm shows the presence of 41.41%, 80.72%, and 88.79% of novel unigram, bigram and trigram formation, respectively.

Table 3.1: Comparative analysis on two datasets – Multinews and CQASumm. We report ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), and ROUGE-L F1 (R-L F1) scores for five extractive (*ext*) and eight abstractive (*abs*) baselines. We also compare the methods based on BERTScore F1 (BS F1) and avg. QAEval F1..

System	Multinews						CQASumm					
	R-1	R-2	R-L	R-L F1	BS F1	QAE F1	R-1	R-2	R-L	R-L F1	BS F1	QAE F1
ext-Lead3	39.41	11.77	14.51	14.84	-	-	8.7	1.01	5.2	5.1	-	-
ext-LexRank [117]	38.27	12.7	13.2	14	83.29	0.0853	28.4	4.7	14.7	14.8	79.28	0.0103
ext-TextRank [118]	38.4	13.1	13.5	14.47	83.82	0.061	27.8	4.8	14.9	14.87	79.64	0.0113
ext-MMR [142]	38.7	11.92	13.11	14.71	84.38	0.0935	29.31	4.97	14.87	14.81	80.34	0.0107
ext-ICSISumm [143]	37.2	13.5	13.87	15.13	84.46	0.0835	28.99	4.24	14.93	14.94	80.67	0.0115
abs-PG [38]	41.85	13.91	16.46	17.14	85.04	0.0525	23.2	4	14.6	14.8	81.45	0.0263
abs-PG-MMR [40]	36.42	9.36	13.23	17.35	85.09	0.0522	16.2	3.2	14.35	11.31	81.47	0.0266
abs-Himap [12]	40.28	13.79	18.81	21.38	83.28	0.0686	30.09	4.69	14.9	14.87	80.57	0.0263
abs-CopyTransformer [15]	42.17	13.03	19.75	20.5	84.57	0.0557	30.12	4.36	14.81	15.32	81.11	0.02
abs-PEGASUS [36]	42.24	13.27	21.44	23.08	84.89	0.0787	32.51	5.14	15.1	15.48	81.68	0.0211
abs-PG+HSA [43]	43.49	14.02	17.21	21.73	84.72	0.0562	31	5	15.2	15.87	81.24	0.0207
abs-rl-RL-ConvS2S [144]	41.17	13.12	16.74	21.47	83.52	0.0527	29.21	4.46	14.83	15.13	81.12	0.0202
abs-rl-ML+RL [61]	42.18	13.21	17.11	21.29	83.58	0.0531	30.72	4.81	15.01	15.26	81.11	0.0205
abs-LongT5 [136]	46.01	**17.37**	23.5	23.31	84.07	0.055	30.29	4.67	14.88	15.03	80.76	0.0202
abs-TGM [137]	46.04	16.43	19.82	22.41	84.29	0.0602	32.18	5.1	17.54	16.72	81.33	0.0224
abs-ChatGPT	44.12	13.7	23.75	24.11	84.12	0.0917	34.23	7.58	22.64	20.32	79.18	0.0231
abs-REISA	**47.28**	17.2	**25.41**	**25.34**	**87.49**	**0.0942**	**36.67**	**7.72**	**24.18**	**22.35**	**83.41**	**0.0286**

3.4 Baseline Systems

We consider 15 baselines for comparative analysis; five of them are extractive, and the remaining ten are abstractive summarization systems. The motivation behind choosing extractive baselines is

to compare their performance against the abstractive baselines. Each dataset possesses its distinct modality – while Multinews favours simple extractive baselines, CQASumm favours complex extractive baselines. We benchmark both datasets across extractive and abstractive baselines to compare the performance.

1. **Lead3** is an extractive summarization system where the first 100 tokens of the source documents form the generated summary.

2. **LexRank** [117] is an unsupervised graph based algorithm. Lexrank represents the sentences as nodes, and the edge represents the sentence importance computed using eigenvector centrality. The similarity is calculated using the frequency of the word in the sentence.

3. **TextRank** [118] is an unsupervised graph-based algorithm that represents the sentences as a fully-connected graph where nodes are the sentences and edges indicate similarities among sentences. TextRank assumes all edge weights between sentences to be unit weights, and finally sentences are ranked based on the derived version of PageRank [145].

4. **Maximal Marginal Relevance (MMR)** [142] reduces redundancy while maintaining relevant sentences. MMR calculates the relevance of a query to sentences using a similarity metric. The similarity value between two sentences is calculated, and similar sentences are removed.

5. **ICSISumm** [143] optimizes the coverage of the summary by adopting linear optimization framework. It extracts sentences containing frequent bigrams from the corpus. Given an upper bound on the summary length, it optimizes the summary using ILP. It generates the globally optimal summary by searching for the most important concepts present in the document.

6. **Pointer Generator (PG)** network [38] is an abstractive summarization framework. The PG architecture uses a bidirectional LSTM based encoder, and a unidirectional decoder architecture along with attention [25]. The context vector and the decoder state produce a vocabulary distribution P_{vocab}, while the attention distribution creates a copy probability P_{copy}. Finally, the pointer variable P_{gen} acts as a switch to allow either to copy or to generate new words from the vocabulary.

7. **Pointer Generator-MMR (PG-MMR)** [40] incorporates the usage of MMR along with the PG network [38]. MMR tries to find K highest-scored source sentences; the weights of all sentences are calculated, and based on the weights, the sentences that relate to the partial summary generated by the model receive low scores. It helps the model to find the important content that might not have been included in the summary.

8. **Hierarchical MMR-Attention PG (Himap)** [12] combines the architectural advantages of PG and MMR. Here, the encoder takes the word tokens for the whole document as input, and the encoder output for the last token is saved to obtain the representation of both source articles and summaries. MMR scores are computed for the sentence representations to determine the optimal sentences based on relevance and redundancy.

9. **Bottom-up Summarization (CopyTransformer)** [15] leverages Transformer [45] along with length and coverage penalty. It uses the same encoder-decoder model but with stacked self-attention and point-wise feed-forward layers. Here, one of the random attention heads acts as a copy pointer.

10. **Pegasus**$_{BASE}$ [36] is a pretrained encoder-decoder based model finetuned over various downstream tasks using the transfer learning approach. The model is based on the big Transformer architecture [45]. The main highlight of Pegasus is its usage of a novel pretraining objective of sentence masking and further tasking the model for regenerating it. We use the Pegasus$_{BASE}$ architecture as our baseline.

11. **PG-HSA** [13] extends the work of PG by incorporating structural attention in a hierarchical encoder setting to capture the long-range contextual dependencies. The authors also proposed a hierarchical encoder with multi-level structural attention to capturing inter-document discourse information and multi-level contextual attention at the word level to generate abstractive summaries.

12. **RL-ConvS2S** [61] is a reinforcement learning (RL) based baseline. This method combines the intra-attention at the encoder as well as the decoder side. It also makes use of a hybrid learning objective along with a policy learning framework to formulate a mixed training function. The proposed method uses ROUGE-L as a reinforcement learning reward.

13. **ML+RL** [144] is an RL-based baseline that uses the greedily decoded sentences and the sampled distribution sentences as a combined RL reward. The ML+RL fuses a topic-aware attention module along with the RL self-critical technique into a CNN architecture to jointly align the text encoding and topic-level information to optimize the overall learning pipeline.

14. **LongT5**$_{BASE}$ [136] scales the T5 [53] architecture over the input length and the model size while combining the Pegasus pretraining objective and TGlobal attention.

15. **TGM** [137] combines topics extracted from multiple documents and represents them as a global knowledge graph. The extracted topics act as Semantic Content Units while generating summaries.

3.5 Experimental Setup

In this section, we discuss the system setup and the evaluation setting of competing systems and the ablation versions of the proposed system. We then elaborate on the experimental results – both quantitative and qualitative analyses.

3.5.1 Evaluation Setup

We provide quantitative analysis of the system-generated summaries using the standard metrics for sequence modelling task: ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) and ROUGE-L F1 [133]. ROUGE measures the unigram, bigram, and longest common sequence overlap between

Table 3.2: Ablation study of REISA on the two datasets. Base Model represents the PG-based seq2seq network with attention. The subsequent systems include RL modules with rewards and attention calibration modules. The final system uses the REISA architecture.

System	#P	Multinews			CQASumm		
		R-1	R-2	R-L	R-1	R-2	R-L
Base Model	60m	34.11	10.17	13.21	17.23	3.64	14.41
+ RW_R	60m	35.57	11.32	14.62	21.87	3.83	14.57
+ RW_{R+BS}	60m	37.75	13.38	15.5	25.8	4.8	14.73
+ RAS	61m	42.43	13.61	19.67	28.62	4.94	15.48
REISA	61m	**47.28**	**17.2**	**25.41**	**36.67**	**7.72**	**24.18**

Table 3.3: Scores for human evaluated metrics - Informativeness (INF), Relevance (REL), Coherence (COH), Fluency (FLU) over extractive and abstractive systems on Multinews and CQASumm datasets.

System	Multinews				CQASumm			
	Inf	Rel	Coh	Flu	Inf	Rel	Coh	Flu
LexRank	2.47	2.51	2.79	3.1	1.74	1.41	2.18	2.07
ICSISumm	2.52	2.64	2.71	2.94	1.87	1.47	2.24	2.17
PG	2.73	2.91	3.21	3.34	2.93	2.54	2.73	2.83
Transformer	2.84	3.14	3.37	3.38	3.38	2.86	2.81	3.17
RL-ConvS2S	2.96	3.21	3.48	3.43	3.47	3.04	3.17	3.29
TGM	3.25	**3.37**	3.54	3.57	3.54	3.12	3.22	3.41
REISA	**3.39**	3.34	**3.59**	**3.64**	**3.61**	**3.15**	**3.31**	**3.48**

the reference and system generated summaries. We also benchmark the performance on the newly introduced BERTScore metric – it utilises the pre-trained BERT model and matches the cosine similarity of tokens between the reference and system summaries. We use the large RoBERTa [146] model for evaluating BERTScore. We further perform a qualitative evaluation using the already proposed QAEval [147] metric. It uses the question-answering task, where a pretrained model formulates questions from the reference summary and then estimates the number of questions that can be answered by the system-generated summary. This metric gives an estimate of the faithfulness of the system-generated summaries.

3.5.2 Human Evaluation Setup

We perform the human evaluation on the system-generated summaries to evaluate the qualitative aspect of the following parameters: Informativeness (Inf), Relevance (Rel), Coherence (Coh) and Fluency (Flu). We randomly select 100 test samples from the system-generated summaries for the assessment. A total of 30 annotators participated in this study. All annotators had strong background in NLP and machine learning. A minimum of two annotators annotated each sample, and inter-annotations of greater than two degrees are discarded. The scores obtained are averaged for each sample.

3.5.3 Ablation Studies

We perform system ablation to understand the impact of individual modules of REISA. Table 3.2 presents the ablation results. We initially run a base model (PG based seq2seq model with the cross-entropy loss). Later, we introduce the reinforcement learning based objective function with ROUGE-L score as the reward metric to capture the syntactical information. Further, we include BERTScore as a reward metric to study the inclusion of both syntactical and semantic-based rewards. When compared with the base model, the inclusion of dual reward helps in attaining an improvement of +2.29 ROUGE-L on Multinews and +0.32 ROUGE-L on CQASumm. Later, we introduce the reinforced attention span, which, when combined with both ROUGE-L and BERTScore as a reward, shows a significant improvement. On Multinews, it gains +13.17 ROUGE-1 and +12.2 ROUGE-L, while on CQASumm, it shows an improvement of +19.44 ROUGE-1 and +9.77 ROUGE-L against the base model. The improvements, when compared with the base PG-based seq2seq model along with the Reinforced Attention Span, gains +8.32 ROUGE-1 on Multinews and +11.39 ROUGE-1 on CQASumm.

As evident from Table 3.2, we infer that the reinforced attention calibration performed at the encoder and decoder side results in the highest gains w.r.t. the model's performance. The improvement is not just consistent with the quantitative score but also performs well with the qualitative metrics.

3.5.4 Quantitative Analysis

We summarise the comparative analysis in Table 3.1. We observe that with the introduction of reinforced attention span, ROUGE-L and BERTScore as reward functions, REISA attains 47.28 ROUGE-1 and 25.41 ROUGE-L scores on the Multinews corpus and 36.67 ROUGE-1 and 24.18 ROUGE-L scores on the CQASumm corpus. Compared to the best baseline (Multinews: ROUGE-1 of TGM and ROUGE-L of LongT5; CQASumm: ROUGE-1 of Pegasus and ROUGE-L of TGM) , REISA performs better over ROUGE-1 and ROUGE-L (R-1 +1.24, R-L 2.91 for Multinews and R-1 +4.16, R-L 6.64 for CQASumm).

A similar trend is observed with the BERTScore as well. REISA attains a BERTScore of 87.49, improving +2.4 over the best baseline (PG-MMR) on Multinews. On CQASumm, REISA shows an improvement of +1.73 by scoring 83.41 BERTScore against the best baseline (Pegasus_{BASE}). We also compare the performance of the proposed model REISA on CNN/Dailymail dataset (Table 3.5). Our model performs better on all three ROUGE metrics, showing an improvement in ROUGE-L of +4.87 over ML+RL on CNN/Dailymail-400 and +3.98 over Transformer on CNN/Dailymail-800. This shows that our approach of recalibrating the attention weights and dual reward not only entitles the model to consider individual segments fairly but also helps in attaining significant improvements over both the quantitative metrics, ROUGE and BERTScore. We present a detailed example of our approach in action, demonstrating its effectiveness in achieving both high faithfulness and minimal hallucinations, in Table 3.4. We find that baseline systems such as PG and Transformer pick sentences from specific source segments forming incoherent phrases. At the same time, REISA takes all individual source segments into account while generating a highly-readable and faithful summary.

Table 3.4: Individual source documents, reference summary and the sample summaries generated by PG, Himap, Transformer, Pegasus, TGM and REISA over a sample of CQASumm corpus. For the summary generation, the baselines use one or a few documents to formulate their summaries; REISA is designed to combine critical information from each document in order to generate a more comprehensive summary. Different colors show positive correlations with the source documents and reference summaries.

Type	Text
Source 1	Probably you have low memories to run the games that means you have to increase the ram. If that doesn't help try to download the latest graphics driver from the manufacturer's website sometimes it helps.
Source 2	This sounds like it could def be a heating issue. "Random" reboots and shutdowns are often as a result of overheating and given that this happens during gaming when you are really pushing your system, this is probably the case. Replace your CPU heatsink/fan and be sure to use Arctic Silver paste between the heatsink and CPU. Also could improve case airflow with additional or bigger fans.
Source 3	its probably ur low RAM or you have a virus. possibly bad power supply but i doubt it.
Source 4	check your video card, it might be over heating. games use alot of proccessing power and can greatly increase the temp. of your computer, open it and blow off the fans. DISCHARGE YOURSELF BEFORE YOU TOUCH ANYTHING. if there is alot of dust goto your neares comp. store and buy a comp. vacume its this really small thing that sucks up dust, ect.. and clean it out.
Source 5	could be something about the games? Are the games labour intensive games for computer? Or are you talking about Solitaire?
Reference Summary	heat is a big issue. when the inside of your computer case is dirty the cooling system ca n't do its job properly. either take it to a reputable shop and have them clean it or go buy a can of compressed air , open the case and blow all the dust out of it taking care to get it all out of the louvres under the cpu fan. DO NOT use a vacuum! it creates a static electricity charge that can corrupt the components on your mother board. a static charge weaker than you can feel can and will
PG	it ' s a gaming card , but it ' s a good time for the first time , the new york times reports . " you need a bigger fan on the mother board. could be something about the control panel , then to system then to hardware then to device manager or here for the rest of the weirdest things that sucks up in the games : click to see a picture of a goto ' s comp. comp. comp. and be sure to be your .
Himap	it ' s a good time for a gaming card , the new york times reports . " sometimes the motherboard will cause it to reboot you need a bigger fan on the mother board. could be something about the about the , the interrupts will cause it the system to reboot. here for the rest of the weirdest things sucks up in the games : you ' ll be able to reports the los angeles times .
Transformer	if you ' re looking for a new ram and buy a comp. possibly bad power supply , you should go to the control panel , then to system your video card , is it a gaming card ? sometimes the interupts will cause it to reboot , reports the los angeles times . it ' s This sounds like it could be over heating. open it and blow off the fans.
Pegasus	you should go to the control panel , then to system your video card , is it a gaming card ? sometimes the interupts will cause it to reboot you need a bigger fan on the mother board. could be something about the control panel blowoff the air weirdest things that sucks up in the games vacume its this really small thing that sucks up dust, ect.. and clean it out.store and buy a comp.
TGM	blowoff the fan. the interrupts will cause it the system to reboot. you should go to the video card and the gaming card. blow the dust. open the power supply, then will cause it to reboot. could be heat and video card. overheating , create a static charge. buy a computer , open it and blow off the fans. could be something about the gaming card.
REISA	buy compressed air, blow all the dust and blow off the fans. heat is a big issue. if you ' re looking for a new ram , you ' ll be able to increase the ram of your computer. it ' s probably your video card or bad power supply , you have low memories. get the latest driver for it. if there is a dust and clean it out. intensive games for computer improve case airflow with heatsink and CPU.

3.5.5 Faithfulness

In analogous to the quantitative evaluation, we also perform human evaluation, Table 3.3 shows that our proposed model REISA generates high quality, redundant, faithful and human-readable summaries. We also use the QAEval [147] metric to evaluate the faithfulness of the system summaries with respect to the reference summaries. QAEval [147] is a question-answer based metric where the question-answer pairs are generated automatically from the reference summary, and the system summaries are evaluated over the presence of information required to answer the

Table 3.5: Comparative analysis on CNN/Dailymail-400 and CNN/Dailymail-800 datasets. We report ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L) scores.

Method	**CNN/Dailymail-400**			**CNN/Dailymail-800**		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead3	40.34	17.7	36.57	40.34	17.7	36.57
LexRank	34.12	13.31	31.93	32.36	11.89	28.12
TextRank	31.38	12.29	30.06	30.24	11.26	27.92
PG	39.53	17.28	36.38	36.81	15.92	32.86
PG-MMR	39.82	17.68	36.72	37.48	16.61	33.49
Transformer	39.94	17.44	36.61	39.50	16.06	36.63
ML+RL	39.87	17.82	36.90	28.42	16.87	36.06
REISA	**43.48**	**20.19**	**41.77**	**42.15**	**19.31**	**40.61**

learned questions.

We compare the faithfulness of summarization systems in Tables 3.1 and 3.3. Here, we observe a considerable gain in terms of the faithfulness of REISA-generated system summaries. Extractive summarization systems on Multinews show significantly better results when compared to other competing abstractive baselines. This can be attributed to the fact that the extractive summaries preserve the sentence syntax that helps the pre-trained language models to formulate better question-answer pairs. The questions generated are robust and require rational answers. However, on CQASumm the trend is inverse. Since CQASumm is extracted from the Yahoo answers and each answer from a thread contains short and unstructured syntax, the pre-trained model finds it hard to extract the necessary information even to formulate the question-answer pairs. Pre-trained QA modal finds it hard to generate insightful questions and often repeats the same questions. The answers to these generated questions show very low correlations with the reference summaries and are often random. This in turn affects the overall faithfulness of the systems. Our analysis shows that the summaries generated by REISA are readable, informative, and faithful. In case of CQASumm, as the summaries are shorter on average compared to Multinews, REISA, with the help of ROUGE-based rewards, is able to learn the short summary representation, showing significant improvements throughout.

3.6 REISA vs ChatGPT

We conducted a comparison between the summaries generated by REISA and ChatGPT-4-turbo, and the results are presented in Table 3.1. For this analysis, we generated summaries from the MultiNews and the CQASumm test set. While ChatGPT's summaries scored well on several quantitative metrics, they did not outperform the proposed method. ChatGPT's summaries generally provide a crisp overview of the source document, starting strong with relevant information. However, as the summaries progress, they tend to deviate from the source material, often expanding based on ChatGPT's internal knowledge graph. This comparison illustrates that while general-purpose

LLMs like ChatGPT excel in linguistic fluency, they often struggle to maintain faithfulness to the factual information and context of the original document, making them less suitable for tasks requiring strict adherence to the source content.

3.7 Analysis and Discussion

This section presents a critical analysis of the proposed model to explain its superiority and limitations. We also discuss further analysis and limitations in this section.

Reward Analysis. We analyze the notion of adding multiple rewards in our architecture. Upon manual inspection, we find that the summaries generated with ROUGE-L as a reward contain more readable and non-repetitive phrases than the baselines; this is in line with the findings reported by [63]. However, the ROUGE-L reward forces the model to directly copy more phrases from the source documents rather than generate them. Therefore, to counter this, we introduce BERTScore as a reward. Compared to ROUGE-L reward, BERTScore is able to generate consistently readable summaries with new novel phrases as it helps the model in structuring the vocabulary with the semantic meaning. We do not consider rewards such as entailment, NLI [148] etc. In the entailment, NLI systems need a separate dataset to train and rank sentences in the summarization corpora; therefore, the performance quality of entailment/NLI rewards will be directly affected by the dataset and the quality of the trained models, thus leading to degradation of the summarization model.

Abstractiveness. We analyse the formation of novel n-gram in the system-generated summaries by three abstractive text summarization systems and `REISA` with only ROUGE-L as a reward and complete `REISA`. Our analysis shows that `REISA` generates the highest abstractive summaries against the most popular systems. Also, when we use `REISA` with only ROUGE-L as a reward, the abstractiveness of the system drops significantly. This supports our earlier intuition that just the ROUGE-based rewards copy the phrases directly from the source documents hurting the abstractness of the system. However, when combined with BERTScore, the model generates more abstractive summaries.

Attention Calibration. We compute the attention distribution using [25]. As the softmax function assigns all the tokens a non-zero value, the computed attention distribution sometimes gets very skewed due to high vocabulary size, which may adversely affect the values of important tokens during summary generation. To counter this issue, we recalibrate all the attention weights at each step by penalizing the highest attention weights after the decoder generates a token. The highest encoder attention weights are penalized and recomputed to help the model attain to other source document fragments and maintain uniformity while generation. The results show that the summaries generated by `REISA` account for all informative source segments fairly and generate a highly faithful summary. The benchmarks performed on the system-generated summaries perform better w.r.t. all evaluation metrics.

Length Analysis of the Generated Summaries. For the baselines and the proposed `REISA` system, we set the generation summary length in analogous to the length of the target summaries. However, we also assess the effects of generating long and short summaries to examine the robustness of the systems. We observe that the generated summaries with above the average length often show redundant information, high hallucinations with respect to the source documents

or copy verbatim phrases from source, while the ones with below the average length fail to capture the entire gist of the source document and are incomplete in information.

Limitations of Traditional Seq2seq Networks. In NLP, deep learning systems tend to optimize the maximum likelihood estimation in the network. At the same time, evaluation metrics such as BLEU and ROUGE are used for assessing the performance of the systems. This, in turn, creates an inconsistency between the metrics used for training and testing the network. The encoder-decoder-based architecture can only work with metrics such as MLE during training since it is differentiable, while ROUGE and BLEU are non-differentiable metrics. However, RL-based models can optimize the non-differentiable metrics and help the models learn similar evaluation metrics during training. Another problem with the seq2seq network is the availability of ground-truth and the previous state during training. However, the ground-truth is not made available during inference, and the model relies on its distribution, thus leading to high exposure biases.

Limitations of Our Proposed Model. Our proposed `REISA` model uses reinforcement learning and the traditional seq2seq network, allowing the model to optimize multiple loss and rewards metrics jointly. During reward computation, `REISA` makes use of oracle summaries generated using eq. 3.16 which is extractive and greedy, adding additional cost and limitation in terms of an extractive-based reward policy. The addition of rewards and recalibration of attention weights also adds to the computing cost of the overall network. The ROUGE-L and BERTScore scores are highly compute-intensive and stall GPU for seconds to get the reward calculated and added to the loss function. This GPU stalling adds a toll on the entire training pipeline and increases the overall training time. We note that our model requires extensive experiments to draw sound conclusions.

3.8 Conclusion

In this work, we introduced `REISA`, a reinforced attention calibration-based model that constraints the encoder into attending at the fixed number of tokens at each step and recalibrates the attention scores to allow the model to attend to other source segments fairly. With the inclusion of dual rewards, ROUGE-L and BERTScore, and dual loss, MLE and RL loss, `REISA` shows significant improvements on ROUGE and BERTScore over both Multinews and CQASumm datasets. We further studied the faithfulness of our generated summaries using the QAEval metric, showing that our model generates more faithful summaries than any other competing baselines. We also perform human evaluations and extensive ablation studies with `REISA` to argue that our model performs better on all fronts. In future, we plan to extend our work by applying it to low-resource summarization tasks and for various other domains.

Chapter 4

Abstractive Extreme Text Summarization

The task of succinctly and accurately encapsulating salient points of long scientific papers, while preserving the integrity of the original content and accentuating novel findings, is a formidable undertaking that necessitates specialized expertise for the extreme text summarization task. The utilization of human annotations for this purpose can be exorbitantly expensive. To address the challenge, we introduce `ExGrapf2`, a novel encoder-decoder-based architecture that utilizes the efficiencies of Fast Fourier Transform (FFT), the expressiveness of Fractal Graph Embeddings, the relational comprehension of Graph Convolutional Networks (GCN), and the quantifiable improvement of contrastive loss. FFT enhances the model's performance for the NLP task, while Fractal Graph Embeddings capture the fundamental concepts. GCN is employed to apprehend the inter-sentential representations, and contrastive loss optimizes the quality of the generated summaries. Our experimentation on the SciTLDR dataset demonstrates that `ExGrapf2` surpasses the state-of-the-art models in terms of +3.97 Rouge-1 and +4.17 Rouge-L points, without resorting to data augmentation. Furthermore, our qualitative evaluations along with human assessments show that `ExGrapf2`'s summaries are more *faithful* to the original content and offer more profound insights than the target summaries themselves.

4.1 Introduction

The challenge of generating extreme summary (*aka* TL;DR/tl;dr/TLDR – too long; didn't read) for scientific papers is a problem of paramount importance, given the sheer volume of literature published annually [149]. The use of human annotations for this task is impractical and onerous, given the expertise and time required to produce an accurate and faithful summary. To date, the only human-annotated benchmark dataset available for extreme summarization of scientific papers is SciTLDR [5]. However, this dataset is relatively small, particularly for a generation task, with only $1,992$ training instances.

This necessitates the development of robust systems capable of extracting more information from a limited-size dataset. In this chapter, we address *extreme abstractive text summarization of long scientific documents* (*aka* TLDR generation). A common approach to address this challenge is to train models for prolonged periods, which may lead to overfitting. However, we posit that by leveraging distinct modules to encompass different perspectives of the data, it is possible to

enhance the performance even with limited data, as opposed to the current paradigms in which reliance on increasing parameter size and larger datasets is the norm.

To this end, we present a novel and unconventional architecture, `ExGrapf2`[1], for the formidable task of extreme abstractive summarization of scientific papers that comprises four key modules.

- The first module is the Fast Fourier Transform (FFT), which ingeniously blends tokens and affords the feed-forward sublayers unencumbered access to all of the Transformer tokens [150]. This approach diverges from FNet [150], in which FFT was substituted with attention, as `ExGrapf2` synergizes it with the attention mechanism. Our experimentation reveals that FFT helps in identifying keywords within the source.

- While FFT is efficacious at identifying keywords, it does not provide context. To surmount this, we introduce the second module – *fractality*. For years, fractals have bewitched the mathematical world because of their ability to create intricate patterns despite their simple formulation [151]. The self-similarity structure of fractals aligns with how text is presented in scientific papers. A scientific paper can be thought of as a structure with "one central concept" repeatedly occurring in various forms. To date, fractality has been primarily used for keyword extraction [152] or extractive summarization [153]. We are the first to apply the concept of fractality for abstractive text summarization.

- The third module, Graph Convolutional Network (GCN), applied on Sentence Relation Graph (SRG) constructed using MPNet [154] and SciBERT [155], provides a sentence-level view of the document and apprehends the inter-sentential relationships within the same document. Both fractality and GCN can be combined to provide word and sentence perspectives.

- Lastly, we adopt the idea from BRIO [14] and use a non-deterministic distribution, unlike one-point target distribution, for summary generation, using contrastive loss. This approach assigns probability mass to different candidate summaries according to their quality.

In short, our contributions are as follows. We propose a novel encoder architecture, `ExGrapf2`, which can be seamlessly unified with existing encoder architectures to enhance their ability to extract more information from smaller data without data augmentation. `ExGrapf2` employs Fractality-infused Graph Embeddings of sentences and FFT to identify salient details at different levels, viewing the task of extreme summarization at the level of words, sentences, and the document.
Word view and its interaction with other words: FFT effectively blends the Transformer tokens, providing the feed-forward sublayers with unimpeded access to all tokens [150]. Additionally, it is capable of capturing keywords, as demonstrated in our experiments.
Individual word view: Fractality also caters to keyword capturing but based on the context and its distribution over the entire document.
Sentence view: GCN derives graph embeddings of sentences from Sentence Relation Graphs, yielding a holistic view of sentences and the relationships between them.
Document view: We employ a Transformer to generate embeddings for the entire document, encapsulating document-level information.

[1] `ExGrapf2` stands for **Ex**treme abstractive summarization utilizing **Grap**h embeddings integrated with **f**ractality and **F**FT Transformer.

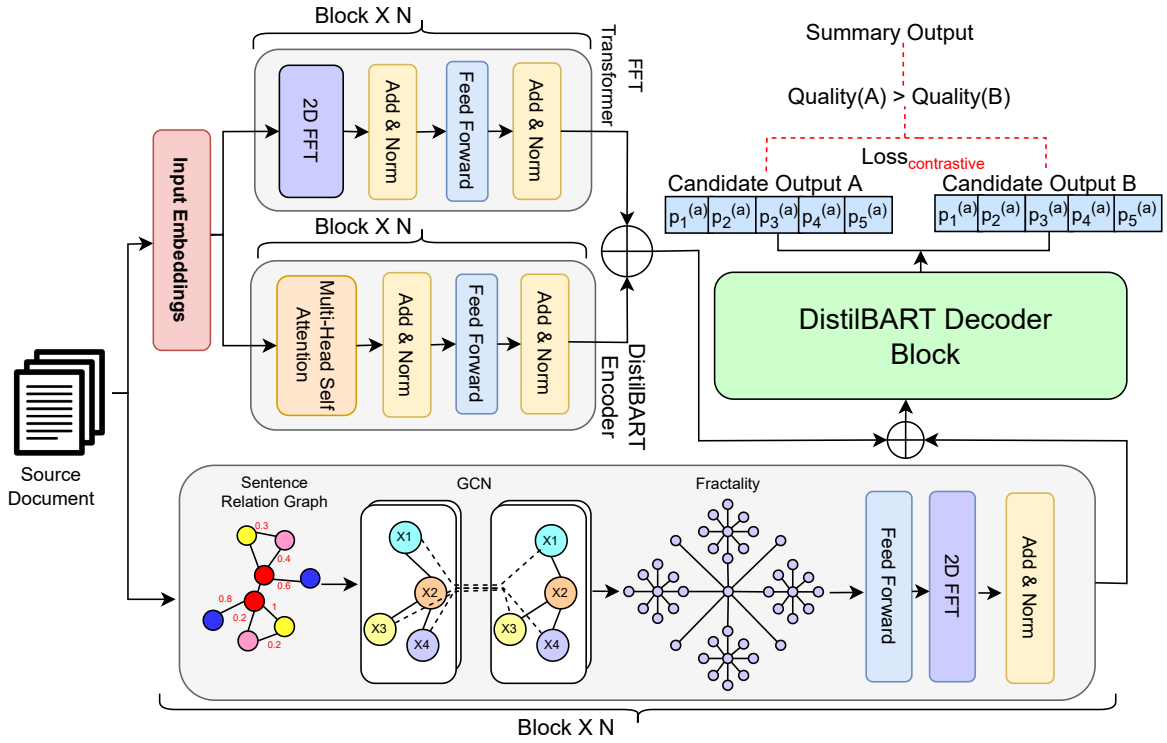


Figure 4.1: A schematic architecture of `ExGrapf2`. The encoder of DistilBART [2] is adopted, and the decoder remains unchanged. Newly-added modules in the encoder are 2D-FFT, Fractality, Sentence Relation Graph (SRG), and Graph Convolution Network (GCN). The contrastive loss is used at the decoder layer. The details for SRG and GCN are given in Figure 4.2.

We conduct extensive experiments and ablation of `ExGrapf2` on the SciTLDR dataset and compare its performance against 15 widely-popular baselines. We evaluate our generated summaries over standard quantitative (Rouge-1, Rouge-2 and Rouge-L) [133] and qualitative (BERTScore [134] and FEQA [66]) metrics. We also perform human evaluations to access the quality of the generated summaries. `ExGrapf2` attains an improvement of +3.97 Rouge-1 and +4.17 Rouge-L; while on qualitative metrics like BERTScore and FEQA, the improvements are +4.02 and +0.04, respectively.

4.2 Datasets

We consider SciTLDR [5] for our experiment. This dataset consists of a source document and a target (TLDR) summary. For the purpose of our experiment, we utilize the AIC subset of SciTLDR, encompassing the Abstract, Introduction, and Conclusion sections of a given paper. On average, the source document comprises $2k$ tokens, while the reference summary comprises 50 tokens. The dataset is split into $1,992$ training, 619 validation, and 618 test samples. The statistics of the dataset are shown in Table 4.1.

Table 4.1: Statistics of SciTLDR: training, validation, and test sets (average token length for source document, target summary and the number of samples).

	Train	**Validation**	**Test**
Source length	2000	2000	2000
Target length	50	50	50
num samples	1992	619	618

The training set features a singular target summary corresponding to each source instance. However, in the validation and test sets, certain samples possess one or more target summaries. In this scenario, the Rouge score is calculated as the maximum Rouge score[2] between the generated summary and the various candidate summaries. The presence of multiple gold summaries per scientific paper is crucial for evaluation, as they account for the natural variability in human-written summaries.

4.3 Proposed Methodology

We present `ExGrapf2`, a novel and perceptive method for extreme abstractive text summarization of scientific papers. Our approach overcomes the hindrance of the restricted size of training data by integrating four components to glean more knowledge. Figure 4.1 shows a schematic diagram of `ExGrapf2`. This section explains each component of `ExGrapf2` in detail.

4.3.1 2D-Fast Fourier Transform

The first component introduces 2D-FFT (two dimensional Fast Fourier Transform), capable of extracting signal properties and features from composite signals. `ExGrapf2` examines the utilization of attention and 2D-FFT in tandem to ascertain if a synergistic effect can be observed. The 2D-FFT technique measures the Discrete Fourier Transform (DFT) twice – one along the hidden dimension and the other along the sequence dimension. We notice that the utilization of FFT extends beyond its application in token mixing and possesses the potential to extract imperative keywords from scientific papers [152]. This approach is inspired by the insight that a scientific paper can be perceived as a composite signal, wherein simplified components correspond to the words or sentences that compose the document. Furthermore, the novel concept of a paper would be reiterated and used more frequently than less crucial terms, which FFT could potentially capture. To maintain consistency, only the real part of FFT is retained while the rest of the architecture remains intact. The following equation explains this:

$$y = \Re(F_{seq}(F_h(x))) \tag{4.1}$$

where $\Re$ represents the real input, F_{seq} represents the sequence dimension, and $F_h(x)$ represents DFT along the hidden dimension.

[2]Furthermore, we disseminated the outcomes of the average Rouge evaluation metrics.

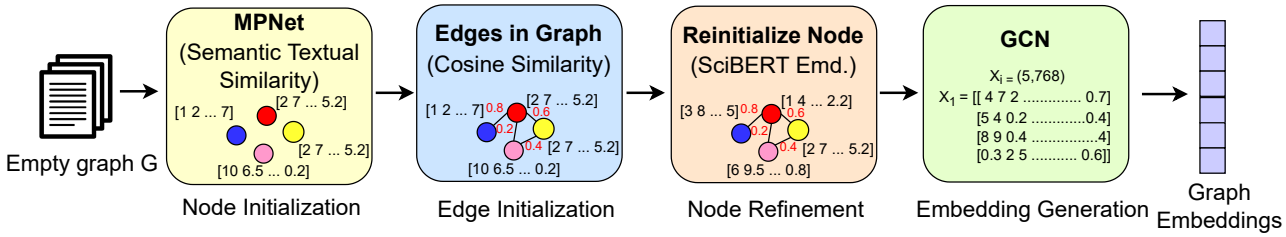


Figure 4.2: SRG and GCN modules of `ExGrapf2` – Sentence Relation Graph (SRG) for sentences in the document and Graph Convolution Network (GCN) to obtain graph embeddings for sentences.

The outcome of the experiments demonstrates that FFT indeed supplies supplementary information, as elaborated in Section 4.5. This is likely owing to the fact that attention concentrates on the relationship of one token to others, which is highly informative; while FFT provides a "word-level" perspective of the document generating a hierarchical representation that is more advantageous and efficient despite being less rich representation than attention.

4.3.2 Fractality

Fractals are not only visually striking but also possess powerful mathematical properties. In simpler terms[3], fractals exhibit self-similarity. This self-similarity can manifest in the form of repeating or iterative patterns. It is worth noting that even slight variations in the repeating unit can result in distinct structures. A term in a scientific paper repeated again and again with different contexts will form a separate unit. For instance, the process of 'backpropagation' may repeatedly appear in the context of 'recurrent neural networks'. In some instances, it may refer to the standard method, while in other cases, it may pertain to the specific application of 'backpropagation' over time. The ability of fractals to inherently identify repeating units with slight variations is momentous.

We employ the box-counting method to calculate the fractal dimension of words [152]. This method allows for the identification of repeating units within the text, as well as any variations or tweaks in those units.

The mathematical formulation for the fractal dimension through box-counting follows power-law and hence transforms to a Linear Regression (Eq. 4.2) on a log-log scale.

$$log\ N = log\ c + D\ log\ (1/s) \tag{4.2}$$

where N denotes the number of boxes that are impacted by the word of concern, D denotes the dimensionality of the fractal, s symbolizes the box size, and c represents a constant. The fractal dimension can be obtained by determining the slope of the line obtained from the mathematical formulation presented in Eq. 4.2. In our specific problem, the window size serves as the equivalent of the box size, and a box is considered to be touched by the document if the word of interest appears within that window.

[3]A little caveat, more precisely, a fractal is by definition a set for which Hausdorff-Besicovitch dimension (D) [156] strictly exceeds the topological dimension.

The fractal dimension of the word can be mathematically determined as:

$$fracDim(word) = \frac{D_s(word)}{D_{ns}(word)} \tag{4.3}$$

where $D_s(word)$ and $D_{ns}(word)$ are the dimensions of shuffled and non-shuffled words, respectively, and D_s is given by

$$D_s = \frac{\sum_{i=1}^{T} D_i}{T} \tag{4.4}$$

where D_i represents the $i-th$ word and T is the number of shuffled experiments.

Note that the sequence of words in a sentence plays a significant role in determining the meaning of the text. Typically, less important words are uniformly distributed throughout the document, while important words are clustered in certain areas. To measure the degree of fractality, we compare a measure calculated before shuffling the text (x) to a measure calculated after shuffling the text (y). As the clustering of important words is disrupted after shuffling, the value of y will be greater than x. However, the distribution of less important words remains unchanged, resulting in the proposed formulation. While this method has been previously applied in the context of keyword extraction [152], to the best of our knowledge, it has not been utilized for text summarization.

4.3.3 Sentence Relation Graph and Graph Convolutional Network

To instil the sentence-level perspective, we construct a Sentence Relation Graph (SRG) on which Graph Convolution Network (GCN) can be applied. The steps to generate the graph are explicated below (c.f. Figure 4.2).

1. For each sentence in a document, MPNet[4] embeddings are computed. The sentences are used as nodes in the graph.
2. An edge is established between two sentences if the cosine similarity between them exceeds a predefined threshold (we set it to 0.2). The weight of the edge is set to the cosine similarity value.
3. Reinitialize node features with the SciBERT embedding of the sentence. We use SciBERT as it is pre-trained on scientific papers and will generate rich embeddings for the task.
4. Apply GCN [78] on this graph to compute the embeddings of the sentence. This methodology astutely embodies the context of neighbouring sentences into the embedding of the sentence of interest, providing a comprehensive "sentence view" of the data.

4.3.4 Combining Graph and Fractality

Our approach employs a synergistic fusion of GCN and fractality to formulate a unified and robust model. By exploiting the fractal dimension of words, our method can identify prominent keywords within a document. By calculating the prevalence of top fractal words in each sentence of the document, we assign weights to the corresponding graph sentence embeddings. The more top

[4]MPNet is selected for this purpose as it is the benchmark for semantic text similarity.

fractal words present in a sentence, the higher the weight of the corresponding graph embedding. This weighting mechanism is applied post-GCN graph embedding computation and pre-encoder layer input, resulting in fractality being overlaid on the GCN-induced embeddings. This results in sentences containing pivotal keywords having a greater influence on the process.

4.3.5 Contrastive Loss

Inspired by BRIO [14], we adopt the notion of contrastive loss and a non-deterministic target distribution, in which the system generates multiple candidate summaries and then learns to rank them. This aligns with our philosophy of extracting more information from the same data. We adopt a similar approach to compute the contrastive loss using Eq. 4.5.

$$Loss_{cstv} = \sum_{i} \sum_{j>i} \max(0, S(C_j) - S(C_i) + \beta_{ij}) \tag{4.5}$$

The decoder employs varied techniques such as beam search [14], to generate two variations of summaries, represented by C_i and C_j in Eq. 4.5. The Rouge score is calculated for both sets of summaries. The difference in rank between candidate summaries is represented by β_{ij}, while $S(C_i)$ denotes the estimated probability values.

4.3.6 Module Intergation

Figure 4.1 presents the fusion of these modules into the encoder architecture. FFT has a parallel branch with multi-head attention, and in each encoder layer, they add up after the application of the feed-forward layer. The embeddings generated by the module combined with GCN and fractality are subsequently added to the encoder layer. We use static embeddings as they are already SciBERT embeddings fused with fractality, and making it a trainable parameter would make it lose its efficacy. Therefore, we add them only with the encoder layer's output. Since the dimensions aren't the same (the dimensions of the graph depend on the number of sentences in the document), we use dense layers to keep the dimension consistent. We also use 2D-FFT to boost the performance further and compensate for not making it a trainable parameter. The decoder remains unaltered. The training regime is contrastive learning.

4.4 Experimental Setting

We prototype our model on Pytorch (1.10) framework with CuDNN 7.2 installed. All experiments are run over A6000 with 48GB of VRAM. Our proposed `ExGrapf2` model is trained over 13 epochs, with 8 gradient steps accumulated. We set the MLE weight as 0.1, max learning rate as $2e-3$, warmup steps at 10000, and ranking loss margin as 0.001. We clip the source document at 1024 tokens and the summary at 35 token length. During decoding, we use a beam size of 8. We use cross-entropy as the loss and ReLU as the activation function for the hidden layers.

Evaluation metrics. We benchmark `ExGrapf2` on the quantitative metrics – Rouge-1 (R1), Rouge-2 (R2) and Rouge-L (RL), over the maximum and average settings, as well as on qualitative

Table 4.2: Evaluation benchmark over five extractive (*ext*) and ten abstractive (*abs*) baselines along with the ablation study of `ExGrapf2`. The evaluations are performed over Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), BERTScore (BS) and FEQA. The Rouge scores are computed over the max and mean settings. The best model (*resp.* best baseline) is highly in bold (*resp.* in italics).

System	**Max Rouge-Score**			**Mean Rouge-Score**			**BS**	**FEQA F1**
	R1	**R2**	**RL**	**R1**	**R2**	**RL**		
ext-Top-K Sentence	22.82	4.61	15.47	17.34	3.00	12.24	62.41	0.10
ext-PACSUM [157]	27.41	6.87	17.43	18.61	4.18	15.94	63.70	0.11
ext-LexRank [158]	29.18	7.82	20.22	22.12	4.63	15.60	63.10	0.13
ext-BertExtractive [159]	30.72	10.91	23.66	21.9	5.78	16.75	64.52	0.15
ext-MatchSum [160]	28.57	9.34	18.39	20.42	5.10	16.20	63.44	0.16
abs-Pointer Generator [38]	24.14	5.43	19.83	21.08	4.33	16.48	59.86	0.15
abs-Himap [12]	24.61	5.48	19.94	21.09	4.36	16.57	59.91	0.17
abs-Bottom-up [15]	25.71	6.08	21.67	23.46	5.26	18.39	60.7	0.19
abs-BERT [75]	26.55	6.68	22.71	24.85	4.79	20.17	62.14	0.21
abs-BART [161]	36.99	11.08	31.71	*29.84*	7.11	21.87	65.87	0.24
abs-LongT5 [136]	35.11	8.71	26.05	26.46	5.33	20.19	63.92	0.22
abs-T5 [162]	37.65	13.42	27.2	29.57	6.75	22.82	65.84	0.22
abs-Pegasus [163]	36.37	13.14	26.66	29.71	7.13	22.91	64.36	0.24
abs-BigBird [48]	35.83	12.47	23.57	27.21	5.57	20.74	64.13	0.22
abs-BRIO [14]	*41.38*	*17.54*	*32.87*	29.79	*8.36*	*23.11*	*66.32*	*0.27*
`ExGrapf2`	**45.35**	**22.49**	**37.04**	**32.37**	**11.62**	**25.53**	**70.34**	**0.31**
– FFT	44.74	21.71	35.97	32.24	11.36	25.11	67.14	0.26
– Fractality	45.09	22.09	36.43	32.18	11.47	25.20	68.92	0.26
– (Fractality + GCN)	44.84	21.53	36.16	31.92	11.21	25.00	68.57	0.28
– (FFT + Fractality + GCN)	44.01	20.72	35.32	31.85	11.04	25.00	68.54	0.29
$\Delta_{\texttt{ExGrapf2}-BEST}$	↑ 3.97	↑ 5.0	↑ 4.17	↑ 2.53	↑ 3.26	↑ 2.42	↑ 4.02	↑ 0.04

Table 4.3: Results on test dataset for three configurations of placing the 2D-FFT module. Config. 1: 2D-FFT placed in parallel with MHA, and they share the same feed-forward networks; Config. 2: Same as Config. 1, except that they have different feed-forward networks; Config. 3: 2D-FFT used after MHA.

Configuration	**Rouge-1**	**Rouge-2**	**Rouge-L**
Transformer	12.39	1.12	10.28
Transformer+ FFT (Config. 1)	13.43	1.12	10.70
Transformer+ FFT (Config. 2)	14.07	1.49	11.21
Transformer+ FFT (Config. 3)	11.04	0.63	9.20

Table 4.4: Rouge scores (Rouge-1 (R1), Rouge-2 (R2), and Rouge-L (RL)) on testing dataset for Transformer and Transformer fused with fractality.

	R1	**R2**	**RL**
Transformer	12.39	1.12	10.28
Transformer + Fractality	13.54	0.81	10.96

metrics – BERTScore and FEQA. Rouge computes the token overlap between the target and the generated summaries. BERTScore computes the similarity between the target and generated summaries in the embedding space, while FEQA evaluates faithfulness using the pre-trained Q/A generation model.

4.5 Ablation of `ExGrapf2`

Our experimentation involves evaluating individual modules of `ExGrapf2` in isolation, devoid of pre-trained models, in order to access their relative importance. The aim is to determine whether the output is solely a product of the novel module and not influenced by pre-trained models. Furthermore, we seek to discern if the modules are mutually complementary. On the basis of these findings, we subsequently integrate all the modules into the final architecture of `ExGrapf2`.

To evaluate the efficacy of the FFT module, we employ a standard Transformer and incorporate FFT in three distinct configurations – (**Config. 1**) placing 2D FFT in parallel to multi-head attention (MHA) to help FFT capture relations among tokens and MHA capture relations between keywords, (**Config. 2**) utilizing feed forward independently in the 2D-FFT and MHA to learn the same components independently, and (**Config. 3**) placing 2D-FFT after MHA to help FFT use the additional context. The performances are presented in Table 4.3, with the results being generated when the difference between consecutive training losses is less than 0.001. These figures, representing the maximum Rouge score, are reported for F-measure.

Analysis of the results: We employ various enumerations of configuring the FFT module in the standard Transformer. The configurations – Config. 1 and Config. 2 of the FFT module outperform the standard Transformer, with the exception of Config. 3. The vanilla Transformer produces higher Rouge scores than Config. 3 (c.f. Table 4.3 and Table 4.5); however, the summaries generated by Config. 3 are found to be more coherent and meaningful than those of the vanilla Transformer. Nevertheless, Config. 1 and Config. 2 also produce superior results, and their summaries amalgamate relevant keywords.

Based on this result, we consider proceeding with Config. 2 – utilizing 2D-FFT in parallel with the multi-head attention mechanism in `ExGrapf2`. This configuration is optimal as multi-head attention and 2D-FFT provide complementary information and yield superior results when employed in parallel.

Subsequently, to the experimentation of the FFT module, we conduct a thorough examination of fractality. Figure 4.3 illustrates the overlap between the most fractal words from the source and target summaries. The curve increases until 100, at which it begins to plateau. As a result,

the GCN embeddings are weighted based on the top 100 fractal words. We add fractal words as inputs, in conjunction with the source, to the Transformer and subsequently compare the results to validate the utility of fractality. Table 4.4 confirms the significance of fractality. The graph for divide + log_{10}(frequency) may appear to be more favourable than "divide only"; however, upon further analysis, we notice that when the frequency term is added, words such as "the" and "and" are also included. Therefore, we ultimately decide to go forward with "divide only". In our final model of `ExGrapf2`, we use 75 top fractal words. we explore different variations for `ExGrapf2`; the results are presented in Section 4.6.

4.6 Performance Comparison

Table 4.2 shows that `ExGrapf2` attains 45.35 R1 and 37.04 RL at the maximum Rouge setting, beating the best-performing baseline BRIO by +3.97 R1 and +4.17 RL. With the average Rouge setting, `ExGrapf2` attains 32.37 R1 and 25.53 RL, beating BART by +2.53 R1 and BRIO by +2.42 RL. We also perform an ablation study to evaluate the effectiveness of the proposed modules of `ExGrapf2`. Table 4.2 showcases the ablation efficacy. Compared to the base Transformer architecture, the inclusion of FFT gives 44.74 R1 and 35.97 RL at max Rouge setting. Further, fractality boosts R1 and RL scores by +0.4 points each. All modules combined push `ExGrapf2` to attain an improvement of +0.6 R1 and +1.0 RL against the only FFT variant. This shows that each module of `ExGrapf2` helps in gathering more context and generating better summaries.

We further assess the quality of the summaries generated by `ExGrapf2` using BERTScore and FEQA. For BERTScore, `ExGrapf2` attains an improvement of +4.02 F1, and for FEQA +0.04 F1 against the best baseline (BRIO). The qualitative improvements are also supported by human evaluation[5]. We randomly sample 100 summaries from extractive baselines – LexRank, BERTExtractive and abstractive baselines – Bottom-up Transformer, Pegasus and BRIO and compare them with `ExGrapf2`-generated summaries. We evaluate the summaries over four parameters – Informativeness, Relevance, Coherence and Fluency. Informativeness assesses the preservation of information w.r.t. the source document. Relevance measures the information overlap between the source document and the summary. Coherence refers to well-structured and organised sentences. Fluency refers to the quality of generated sentences concerning syntaxes and grammar.

Table 4.6 shows that `ExGrapf2` generates highly faithful, relevant and fluent summaries against the baselines. We also compare `ExGrapf2` generated summaries with ChatGPT[6] and find `ExGrapf2` generating significantly better summaries (c.f. Section 4.9).

[5]The human evaluators are NLP experts (i.e., graduate and undergraduate students with sufficient background in ML, DL and NLP), and their age varies between 25 to 35 years old. A total of 17 human evaluators compared our system with the baselines. Each instance is annotated by a minimum of 3 evaluators. Inter-annotations with a difference of more than 3 degrees are discarded.

[6]`https://openai.com/blog/chatgpt/`

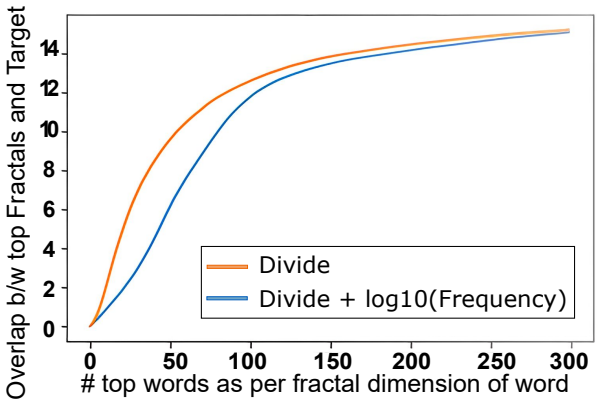


Figure 4.3: Number of top fractal words chosen from source vs avg. number of overlapping words between top fractal words and target summary.

Table 4.5: Generated summaries for three configs of placement of 2D-FFT modules. Config. 1: 2D-FFT placed in parallel with MHA, and they share the same feed-forward networks; Config. 2: Same as Config. 1, except that they have different feed-forward networks; Config. 3: 2D-FFT used after MHA.

Target Summary: the paper presents a multi-view framework for improving sentence representation in nlp tasks using generative and discriminative objective architectures.
Transformer: a general and effective model for avoiding negatve transfer in neural network few shot learning
Transformer + FFT (Config. 1): multi view learning improves the semi supervised learning improves training and dependency relationship
Transformer + FFT (Config. 2): multi view learning improves unsupervised sentence representation learning
Transformer + FFT (Config. 3): multi view learning improves unsupervised sentence representation learning

Table 4.6: Scores for human evaluation metrics - Informativeness (Inf), Relevance (Rel), Coherence (Coh), Fluency (Flu) over two extractive (*ext*) and four abstractive (*abs*) systems on SciTLDR.

System	**SciTLDR**			
	Inf	Rel	Coh	Flu
ext-LexRank	2.78	2.45	2.37	3.11
ext-BERTExt	2.53	2.61	2.65	2.83
abs-Bottom-up	2.82	3.08	3.14	3.23
abs-Pegasus	3.12	3.32	3.38	
abs-BRIO	3.21	3.32	3.44	3.42
`ExGrapf2`	**3.41**	**3.54**	**3.61**	**3.71**

4.7 Error Analysis

As demonstrated in Table 4.6, we observe substantial improvements across all four metrics of human evaluation. The generated summaries are comparable to the target summaries and in many cases,

Table 4.7: ExGrapf2 generated summary. The summary is generated via mathematical formulation of the source document.

Target Summary: An image to image translation method which adds to one image the content of another thereby creating a new image .
ExGrapf2 **Summary:** we study the problem of learning to map , in an unsupervised way , between domains A and B such that the samples vb contain all the information that exists

are superior. For example, in sample #1 in Table 4.8, the generated summary captures the entire gist of the source document and is highly correlated to the target summary. For sample #2, the generated summary is superior to the target summary, highlighting the unique contributions of the scientific work. Similarly, for sample #3, the model effectively generates a summary that is similar to the target summary. However, on further manual inspection, we observe that ExGrapf2 struggles to generate a summary within the token limit or fails to capture the entire gist of the document in summary. Additionally, a few samples (c.f. Table 4.7) also show that ExGrapf2 deviates from the target summary and generates a mathematical formulation of the source document giving an enriched bird's eye view of the document.

These qualitative improvements cannot be quantified using the standard Rouge metric [132], and any qualitative improvements will negatively impact the Rouge score. The target summaries themselves are subjective [164], and therefore, cannot be relied upon as a definitive measure of performance for any system.

4.8 ExGrapf2 vs ChatGPT

In this study, we compare our proposed model, ExGrapf2, with OpenAI's ChatGPT. We generate 50 summaries using the ChatGPT interface and compare them to those generated by ExGrapf2. Our manual analysis reveals that the summaries generated by ChatGPT are overly general and provide only minimal scientific contributions. In contrast, ExGrapf2's summaries are highly informative, providing detailed explanations of the major contributions and the reasoning behind the methodology. Table 4.8 presents a few examples of summaries generated by both ExGrapf2 and ChatGPT, along with the target summary for comparison. This analysis highlights the importance of task-specific systems as a general language model may not be sufficient for all NLP tasks.

4.9 ExGrapf2 vs REISA

While REISA has proven effective for general summarization, its application to extreme summarization is limited due to inherent constraints. Extreme summarization demands a highly focused approach to condense vast amounts of information into concise summaries, posing challenges for REISA in terms of information compression, prioritization, and maintaining coherence within a constrained space. Therefore, a dedicated system tailored specifically for extreme summarization is necessary to address these challenges effectively. Our proposed model, ExGrapf2, integrates FFT, fractality, GCN, and contrastive loss to capture nuanced contextual information, prioritize salient

Table 4.8: Comparative study of summaries produced by ExGrapf2 and ChatGPT under TLDR word limit.

Example 1 (Target Summary) We take face recognition as a breaking point and propose model distillation with knowledge transfer from face classification to alignment and verification .
Example 1 (ExGrapf2 Summary) we take face recognition as a breaking point and propose model distillation with knowledge transfer from face classification to alignment and verification .
Example 1 (ChatGPT Summary) The paper presents a method for compressing large neural networks using knowledge distillation, tested on face recognition and showing good results, even surpassing the teacher network in some cases.
Example 2 (Target Summary) we proposed a novel contextual recurrent convolutional network with robust property of visual learning .
Example 2 (ExGrapf2 Summary) In this paper , we proposed a novel Contextual Recurrent Convolutional Network with contextual recurrent connections with feedback , which widely exists in biological visual system .
Example 2 (ChatGPT Summary) This paper introduces a new CNN architecture with feedback connections, outperforming standard feedforward CNNs in various image classification tasks.
Example 3 (Target Summary) Generatively discover meaningful , novel entity pairs with a certain medical relationship by purely learning from the existing meaningful entity pairs , without the requirement of additional text corpus for discriminative extraction .
Example 3 (ExGrapf2 Summary) a generative model called Conditional Relationship Variational Autoencoder (CRVAE) , which can discover meaningful and novel relational medical entity pairs without the requirement of additional external knowledge .
Example 3 (ChatGPT Summary) The paper introduces a model, CRVAE, that generates novel and meaningful medical entity pairs without external knowledge. Tested and shown to be effective.

details, and ensure coherence and readability in extremely short summaries.

4.10 Conclusion

In this work, we presented `ExGrapf2` an encoder-decoder model that utilizes a combination of Fast Fourier Transform (FFT), fractality, Graph Convolutional Networks (GCN), and contrastive loss to produce informative, coherent, and fluent extreme summaries. The FFT component of `ExGrapf2` provides a word-level perspective, while GCN and fractality offer a blend of sentence-level and word-level perspectives. Additionally, the contrastive loss function contributes to further improvements by generating multiple candidate summaries. Our experimental results demonstrated that `ExGrapf2` outperforms fifteen state-of-the-art models both quantitatively and qualitatively. We also conducted human evaluations to argue that our generated summaries are highly faithful, coherent and fluent against all baselines.

Chapter 5

Abstractive Text Summarization using Multimodal Signals

In recent years, abstractive text summarization with multimodal inputs has started drawing attention due to its ability to accumulate information from different source modalities and generate a fluent textual summary. However, existing methods use short videos as the visual modality and short summary as the ground-truth, therefore, perform poorly on lengthy videos and long ground-truth summary. Additionally, there exists no benchmark dataset to generalize this task on videos of varying lengths.

In this chapter, we introduce AVIATE, the first large-scale dataset for abstractive text summarization with videos of diverse duration, compiled from presentations in well-known academic conferences like NDSS, ICML, NeurIPS, etc. We use the abstract of corresponding research papers as the reference summaries, which ensure adequate quality and uniformity of the ground-truth. We then propose `FLORAL`, a factorized multi-modal Transformer based decoder-only language model, which inherently captures the intra-modal and inter-modal dynamics within various input modalities for the text summarization task. `FLORAL` utilizes an increasing number of self-attentions to capture multimodality and performs significantly better than traditional encoder-decoder based networks. Extensive experiments illustrate that `FLORAL` achieves significant improvement over the baselines in both qualitative and quantitative evaluations on the existing How2 dataset for short videos and newly introduced AVIATE dataset for videos with diverse duration, beating the best baseline on the two datasets by 1.39 and 2.74 ROUGE-L points respectively.

5.1 Introduction

With the emergence of multimedia technology and the rapid growth of social media video-sharing platforms such as Youtube and Vimeo, multimedia data (including text, image, audio, and video) have increased dramatically. Specifically, during the COVID-19 outbreak in the last six months, there has been a steep rise in various e-learning platforms, resulting in a drastic increase of online video tutorials and academic presentation videos. However, such videos often do not have the text meta-data associated with them, or the existing ones fail to capture the subtle differences with related videos [165]. Additionally, different modalities in most of these videos are asynchronous

with each other, leading to the unavailability of subtitles. In this work, we address the task of generating an abstractive text summary of a given academic presentation video so that the viewers can acquire the gist of the presentation in a short time, without watching the video from the beginning to the end. For this purpose, we incorporate automatic speech recognition (ASR) and optical character recognition (OCR) generated text transcripts and capture tonal-specific details of the speaker in addition to extracting semantics and sentics from the video, which are jointly optimized to produce a rich and informative textual summary of the entire presentation. We also show the generalizability of our model on the non-academic dataset (instructional videos).

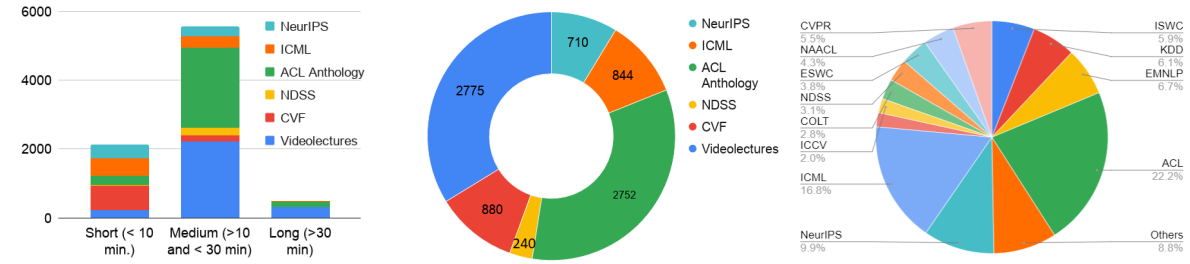
State-of-the-art and Limitations: The existing studies on abstractive text summarization with multimodal signals include multimodal news summarization [85, 91, 90] and summarization of instructional videos [17]. However, all of them use images and/or short videos as the visual modality, which do not generalize on long videos. The generated summaries by these systems are also one or two lines long, and therefore, not suitable for longer academic videos (such as course lecture, conference tutorials). Some other closely related studies include image and video captioning [166, 167, 168, 169, 170], video story generation [171], video title generation [172] and multimodal sentence summarization [173]; but all of them deal with short videos or images which are not appropriate for our application. The lack of previous studies on this task can be attributed to the absence of a suitable benchmark dataset. In a very recent work, [17] studied the task of summarization of instructional videos on the How2 dataset [21], which is the only existing dataset for abstractive text summarization with multimodality. However, the How2 dataset consists of short videos, with an average duration of 90 seconds only. The ground-truth text summaries of this dataset have an average length of 33 words, which are very small as well.

Our Contributions: In this chapter, we explore the role of multimodality in abstractive text summarization for academic presentation videos of diverse duration and introduce a new resource to further enable research in this area. More specifically, our main contributions in this work are as follows:

1. We curate the first large-scale dataset, **A**udio **VI**deo l**A**nguage da**T**as**E**t (AVIATE), for abstractive text summarization using multimodal inputs for academic presentation videos of diverse duration. To collect the videos for this dataset, we scraped 6 publicly available websites and accumulated paper presentation videos from 28 well-known international conferences in computer science and social science. To obtain the transcripts of these videos, we apply Deep Speech [1], a pre-trained end-to-end automatic speech recognition (ASR) system. We use the abstracts of corresponding research papers as the ground-truth summaries, which ensure adequate quality and uniformity. In contrast to How2, AVIATE contains longer videos and larger ground-truth summaries, which help the deep learning models trained on AVIATE to generalize the performance on other datasets.

2. We introduce several baselines to show that multimodal frameworks are substantially more effective when compared to their unimodal variants for abstractive text summarization.

3. We propose **Factorized Multimodal Transformer based decoder-only Language Model (`FLORAL`)**, which uses an increasing number of self-attentions to inherently capture inter-modal and intra-modal dynamics within the asynchronous multimodal input sequences. `FLORAL`



(a) Duration statistics across the dataset.

(b) Source statistics across the dataset.

Figure 5.1: Duration and source statistics of AVIATE.

demonstrates the utility of pre-trained language model (LM) for summary generation in relatively low-resource setups over traditional encoder-decoder based networks.

4. For the videos of AVIATE, we show the importance of OCR generated text transcripts, which contain keywords and informative phrases displayed on slides in academic presentations. To fuse ASR and OCR generated texts, we introduce a novel **guided attention based fusion mechanism** which attends the complementary features in both the sources and filters out repetitive and redundant words. After the incorporation of OCR transcript, the baselines and `FLORAL` yield $[0.7 - 3.6]$ ROUGE-L points performance improvement on AVIATE.

5. `FLORAL` reports benchmark results in terms of both automatic and manual evaluation metrics on How2 for short videos. It beats the best baseline by 1.39 ROUGE-L points. On AVIATE, `FLORAL` also turns out to be highly effective – it beats the best baseline by 2.74 ROUGE-L points.

6. Finally, we report the transferability of `FLORAL` between How2 and AVIATE. When trained on AVIATE and tested on How2, `FLORAL` yields 49.9 ROUGE-L score, which is only 6.9 points less than ROUGE-L obtained when both trained and tested on How2. The diverse-length videos of AVIATE make the model transferable, which is tremendously effective for practical applications.

5.2 Dataset

To enable the exploration of abstractive text summarization using multimodal signals and to generalize the task for videos of different lengths, we introduce AVIATE, the first large-scale multimodal text summarization dataset with videos of diverse duration, compiled from academic paper presentations. Currently, the only existing benchmark dataset relevant to our task is the How2 dataset [21], which includes short instructional videos on different topics like cooking, sports, indoor/outdoor activities, music, etc. Our study reveals that deep neural models trained on such short videos fail to produce satisfactory results on longer videos. Moreover, the facial expression of the speakers in academic talks and presentations often plays an important role to preserve the most informative frames, which is not always the case for the How2 dataset.

5.2.1 How2 Dataset

The How2 dataset consists of $2,000$ hours of short instructional videos, where the training, validation, and test set contain $73,993$, $2,965$, and $2,156$ videos respectively, with an average length of 90 seconds. Each video in this dataset is accompanied by a human-generated transcript and a $2-3$ sentence ground-truth summary. The average length of transcripts and summaries is 291 and 33 words respectively.

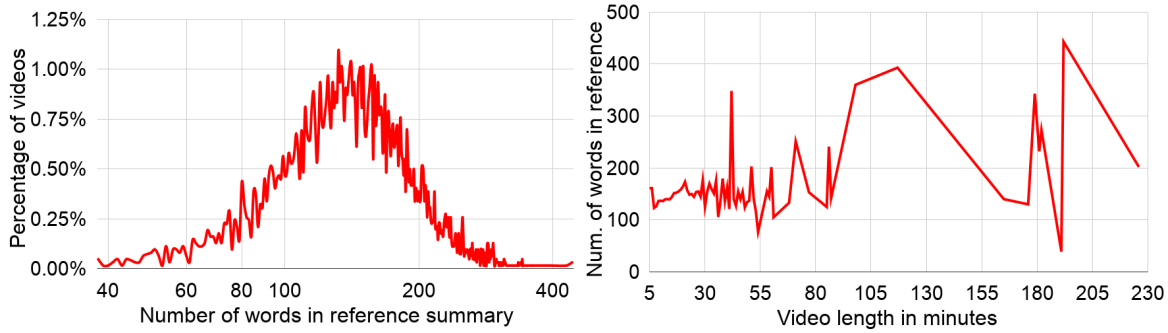
5.2.2 AVIATE Dataset

To collect conference presentation videos, we identified 28 academic conferences in computer science and social science, spanning over various domains such as Machine Learning, Natural Language Processing, Data Mining, Computer Vision, Computational Linguistics, Semantic Web, and Complex Systems. Most of the videos of our dataset come from conferences like NDSS, ICML, NeurIPS, ACL, NAACL, CVPR, EMNLP, ISWC, KDD, etc. To collect the oral and spotlight presentations of these conferences, we scrapped six different academic online video repositories, namely Videolectures.NET[1], ACL Anthology[2], CVF Open Access[3], ICML[4], NeurIPS[5], and NDSS Symposium[6] websites. All the paper presentation videos are accompanied by an abstract, which we use as the ground-truth summary. Thus, unlike the How2 dataset, we did not annotate the summaries ourselves, which significantly improved the quality of ground truth summaries, and hence of the entire dataset.

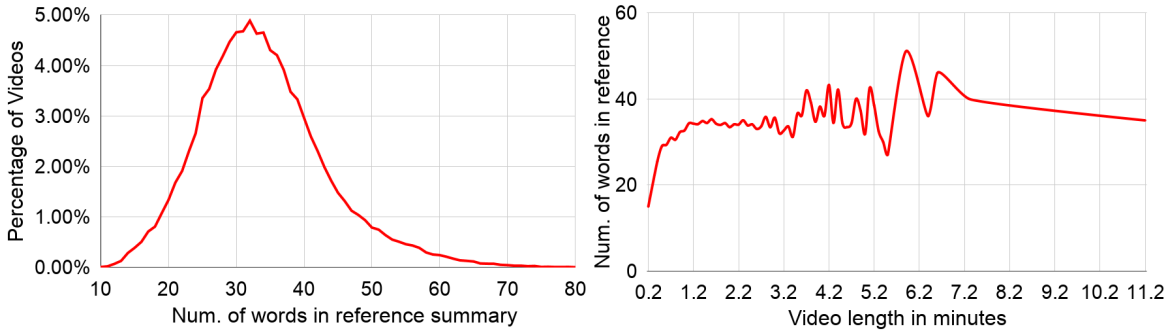
AVIATE consists of a total of $8,201$ videos, which spans over almost $2,300$ hours. Among them, we use $6,680$ videos for training, 662 for validation, and 859 for testing. The length of summaries is mostly between $100-300$ words, with an average of 168 words. A brief source and duration statistics of AVIATE is presented in Figure 5.1.

Transcription: Since we collected all the videos from six different sources, not all of them had subtitles or transcripts readily available. This is particularly the case for videos from ACL Anthology and Videolectures, which contribute the majority of the AVIATE dataset. In the case of videos from NDSS, ICML, NeurIPS, and CVF Open Access corpus, subtitles are available for those videos which are present on Youtube. To maintain uniformity in the quality of transcripts, we apply Deep Speech [1], a pre-trained end-to-end speech recognition algorithm, to extract transcripts for all the videos. To ensure the quality of Deep Speech generated transcripts, we manually transcribe 300 randomly selected videos from our dataset. A low word error rate (24.19%) of the Deep Speech model for those videos indicates the satisfactory standard of the transcripts. An additional normalization step, which includes formatting[7] entities like numbers, dates, times, and addresses, helps us to further reduce the error rate to 20.12%.

[1] `http://videolectures.net/`
[2] `https://www.aclweb.org/anthology/`
[3] `https://openaccess.thecvf.com/`
[4] `https://icml.cc/`
[5] `https://nips.cc/`
[6] `https://www.ndss-symposium.org/`
[7] For example, labeling 'September 16, 2017' as 'september sixteenth twenty seventeen'.



(a) Video length and ground-truth summary length distribution for AVIATE.



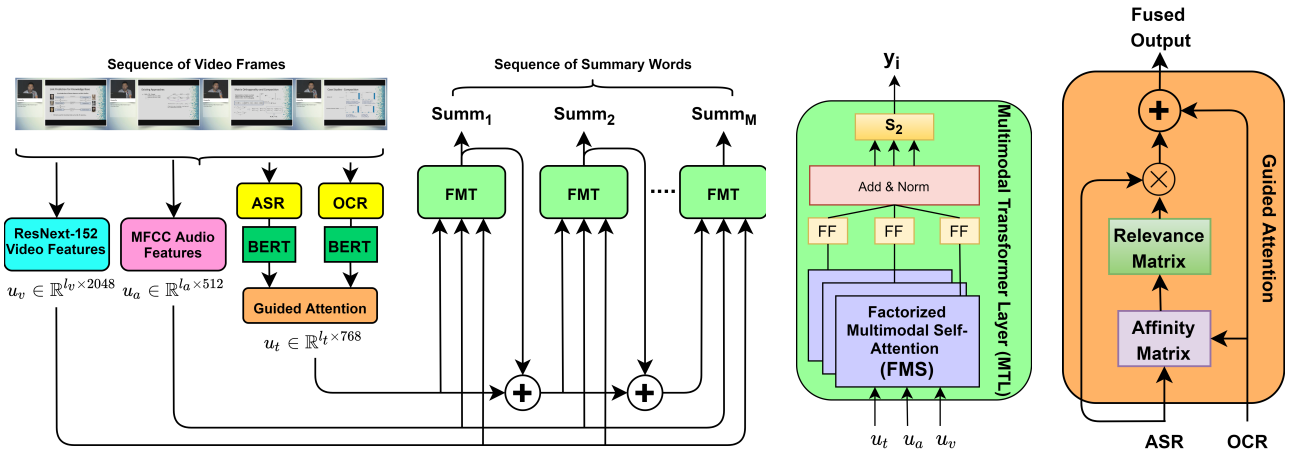
(b) Video length and ground-truth summary length distribution for How2.

Figure 5.2: Correlation between duration of videos and ground-truth summary length for AVIATE and How2 datasets.

Figure 5.2 shows a comparison of video length and ground-truth summary length distribution for AVIATE and How2. For both datasets, longer videos generally have a longer ground-truth summary, which leads to an overall positive correlation between video duration and ground-truth summary length. The average length of AVIATE videos is almost 12 times more than that of How2 videos. The longer videos and lengthier summaries in AVIATE make it harder than How2 to train on, which is explained in Section 5.5.

5.3 FLORAL: Our Proposed System

In this section, we describe our proposed system, **Factorized Multimodal Transformer [3] based Language Model (FLORAL)** for abstractive text summarization using multimodal signals. Figure 5.3 shows the overall architecture of FLORAL. It takes a video, its corresponding audio and text transcript as input and generates an abstractive textual summary. A video generally has three distinct modalities – visual, textual, and acoustic, which supplement each other by providing complementary information, and thus when fused, separately contribute to generating richer and more fluent summaries. The first part of FLORAL extracts unimodal features using respective unimodal feature extraction networks. This phase does not consider the contextual relationship between the three different modalities. In the next part, unimodal features are processed using the



(a) Overview of our proposed model, FLORAL. (b) Architecture of MTL. (c) Architecture of guided attention.

Figure 5.3: The complete architecture of FLORAL, our proposed Factorized Multimodal Transformer based decoder-only Language Model. The Factorized Multimodal Transformer [3] consists of a stack of Multimodal Transformer Layers (MTL), which is shown in Figure 5.3(b). Figure 5.3(c) shows the architecture for guided attention layer used for the fusion of ASR and OCR generated text transcripts.

Factorized Multimodal Transformer (FMT) based decoder-only network over multiple steps, which in turn generates one summary word in each step. After every step, the generated word is appended to the source text with a delimiter. Therefore, FLORAL considers the entire summarization problem as a language modeling task, simplifying traditional encoder-decoder architecture. The remaining part of this section discusses individual modules of FLORAL in detail.

5.3.1 Video Feature Extraction

The visual features are extracted using a pre-trained action recognition model, $ResNeXt-152\ 3D$ Convolutional Neural Network [174] trained on the Kinetics dataset [175] to recognize 400 different human actions. All the frames, computed at a rate of 5 FPS, are first preprocessed by resizing, center-cropping, and normalization to have a resolution of 112×112. For every 16 non-overlapping frames in a video, $ResNeXt-152$ extracts a 2048 dimensional feature vector. Therefore, the result is a sequence of feature vectors per video rather than a global one. The sequential feature vector, $u_v = \{u_i^v\}_{i=1}^{l_v}$, is then used as the visual embedding input to the FMT.

5.3.2 Speech Feature Extraction

The acoustic modality is expected to contribute information related to tonal-specific details of the speaker [176]. To achieve this, we obtain low-level features from the audio stream for each video. Similar to [177], we use the popular speech processing library, Librosa [178] and perform the steps mentioned next. First, the audio sample for a video is stacked as a time-series signal with a sampling rate of 16000 Hz. Next, we remove the echos and background noise from the audio signals

by integrating it with Audacity instance[8], which is a free and open-source audio editor. Then, we segment the audio signals into d_w non-overlapping windows with a window size of 25 ms and successive window shift of 10 ms to extract low-level features that include Mel Frequency Cepstral Coefficients (MFCCs) with hamming window and the related temporal derivatives. Padding and segmentation are performed to achieve a fixed-length representation of the audio sources which are otherwise variable in length. At last, we concatenate all the extracted features to compose a $d_a = 512$ dimensional joint representation for each window. Final MFCC features are obtained by applying a log Mel frequency filter bank over 0 to 8000 Hz and applying discrete cosine transformation (DCT). Similar to the visual features, the audio features, $u_a = \{u_i^a\}_{i=1}^{l_a}$, are also sequential for every video sample and are then used as the acoustic embedding input of FMT.

5.3.3 Textual Feature Extraction

Both How2 and AVIATE datasets contain textual transcripts corresponding to video samples. For How2, the transcripts are manually annotated, while for AVIATE, a pre-trained automatic speech recognition (ASR) algorithm, Deep Speech [1], is used to extract the transcripts for all the videos (as discussed in Section 5.2.2). Since the AVIATE dataset consists of conference presentation videos, we observe that in the majority of video samples of AVIATE, the speaker uses presentation slides that contain the most informative key-phrases. Thus, we extract the text shown in the slides using Google OCR Vision API[9] and fuse the OCR-generated text with ASR-generated text using a novel *guided attention mechanism* to attend complementary and non-redundant words of both sources.

Guided Attention: At first, we represent the text in both ASR and OCR-generated transcripts using pre-trained BERT [179], which provides dynamic embedding for every word. In particular, we use the sequence of 768-dimensional hidden states at the output of the last layer of the BERT model. Let $F \in \mathbb{R}^{n\times 768}$ and $H \in \mathbb{R}^{m\times 768}$ be the BERT representations for ASR and OCR texts respectively, where n and m are the respective token counts. The guided attention mechanism begins with defining an affinity matrix $C \in \mathbb{R}^{n\times m}$, whose element c_{ij} denotes the similarity between the feature vector pairs, $h_i \in \mathbb{R}^{768}$ and $f_i \in \mathbb{R}^{768}$:

$$C = tanh(H\mathbf{W}^b F^\top) \tag{5.1}$$

where $\mathbf{W}^b \in \mathbb{R}^{768\times 768}$ is a correlation matrix to be learned during training.

Subsequently, we compute a normalized weight α_{ij}^h to denote the relevance of the i^{th} ASR-generated word to j^{th} OCR-generated word. Therefore, the weighted summation of the ASR transcript, a_j^h, can be represented as,

$$a_j^h = \sum_{i=1}^{n} \alpha_{ij}^h h_i \tag{5.2}$$

$$\text{where, } \alpha_{ij}^h = exp(c_{ij}) / \sum_{i=1}^{n} exp(c_{ij}) \tag{5.3}$$

[8] https://github.com/officeonlinesystems/audacityonline_audioeditor/

[9] https://cloud.google.com/vision/docs/ocr\#vision_text_detection-python

Since our goal is to emphasize the dissimilar features between the ASR and OCR transcripts, we define the relevance matrix $R(f_i, a_j^h)$ as cosine distance between the attended ASR sentence vector a_j^h and OCR word embedding f_i –

$$R(f_i, a_j^h) = 1 - \frac{f_i^\top \cdot a_j^h}{\|f_i\| \|a_j^h\|} \tag{5.4}$$

Now, the weighted summation of all word embeddings produces the modified ASR representation U computed as,

$$U = \sum_{j=1}^{m} R(f_i, a_j^h) \cdot f_i \tag{5.5}$$

where $R(f_i, a_j^h)$ acts as a filter for the ASR encoding f_i.

Finally, we concatenate the attended ASR word representations with OCR word embeddings to get the sequential textual features $u_t = \{u_i^t\}_{i=1}^{l_t}$, which is used as the textual embedding input of FMT.

5.3.4 Language Model Pre-training

The pre-trained Language Model (LM) has recently been shown to have superior performance in abstractive summarization, particularly to enhance sample efficiency [180]. This decoder-only network, known as Transformer LM, takes a pre-trained transformer [4] as its base module and treats summarization as a language modeling task where each generated summary word in every step is appended to its source article. We extend the concept of Transformer LM to a multimodal setting, where we use **Factorized Multimodal Transformer [3] based Language Model (`FLORAL`)** for multimodal sequential learning. After each step of summary generation, we append the generated summary word to its source text transcript, along with a delimiter, and train the transformer on this reformulated data. `FLORAL` has three crucial advantages over traditional encoder-decoder based summarization networks:

1. In contrast to encoder-decoder architecture, `FLORAL` uses a single network to encode the source and generate the target, and thus, avoids the loading of same pre-trained weights into separate encoder and decoder.

2. Compared to the encoder-decoder network, `FLORAL` has fewer number of parameters.

3. Most critically, all the parameters of `FLORAL` can be pre-trained.

Since there is no available large-scale multimodal corpus, we pre-train `FLORAL` on the text-only 2-billion word corpus[10] based on Wikipedia, called WikiLM [180], and fine-tune on the AVIATE and How2 datasets.

[10] `https://github.com/tensorflow/tensor2tensor`

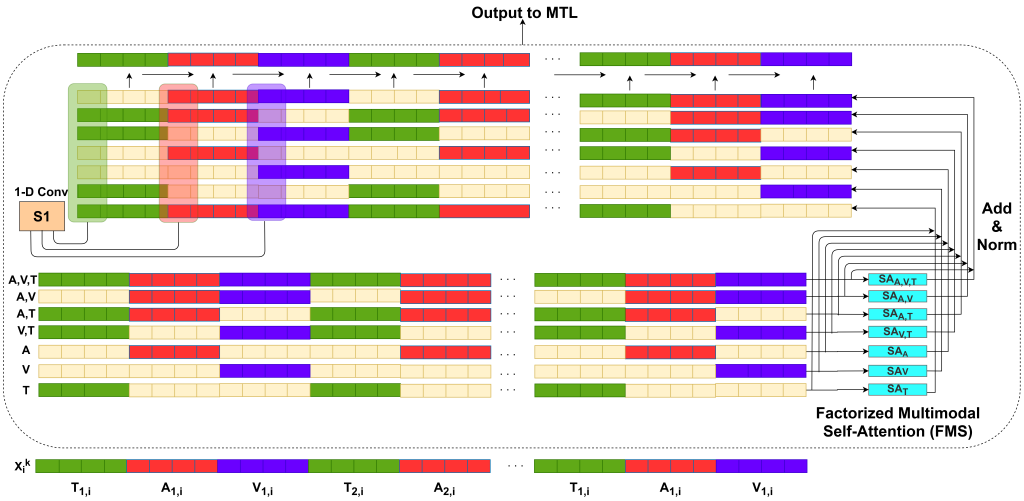


Figure 5.4: Overview of a single Factorized Multimodal Self-attention (FMS) in MTL. Each FMS consists of 7 distinct self-attention [4] layers, which inherently capture inter-modal and intra-modal dynamics within the asynchronous multimodal input sequence. Blue, red and green colors are used to illustrate the propagation of visual, acoustic and textual features within FMS.

5.3.5 Factorized Multimodal Transformer LM

Factorized Multimodal Transformer (FMT) [3], which is the current state-of-the-art model for multimodal emotion recognition and multimodal speaker traits recognition on well-studied IEMOCAP [181] and POM [182] datasets, applies seven distinct self-attention mechanisms to simultaneously capture all possible uni-modal, bi-modal, and tri-modal interactions across its multimodal input. We use the FMT architecture as the backbone of our decoder-only Transformer LM. Before feeding into FMT, the unimodal embeddings are resampled using a reference clock so that modalities can follow the same frequency [18]. Additionally, zero paddings are used to unify the length of all samples of the entire dataset to a desired fixed length L. Hence, the i^{th} data point consists of three distinct sequences of embeddings corresponding to three modalities – visual, acoustic, and language:

$$D = \{x_i = [x_{(t,i)} = \langle u_v^{(t,i)}, u_v^{(t,i)}, u_v^{(t,i)} \rangle]_{t=1}^{L},\ [tar_i^t]_{t=1}^{M}\}_{i=1}^{N} \tag{5.6}$$

where $x_i \in \mathbb{R}^{L \times d_x}$ and $tar_i \in \mathbb{R}^{M \times d_y}$ are the inputs and target summaries respectively, M is the length of the summary; d_x, d_y denote the input and output dimensionality at each time step respectively; N is the total number of samples within the dataset. Positional embeddings are also added to the input.

The FMT consists of a stack of Multimodal Transformer Layers (MTL), which captures factorized dynamics within multimodal data and aligns the time asynchronous information both within and across modalities using multiple Factorized Multimodal Self-attentions (FMS), each of which has 7 distinct self-attention layers. Each attention has a unique receptive field with respect to modalities $f \in F = \{\mathtt{L}, \mathtt{V}, \mathtt{A}, \mathtt{LV}, \mathtt{LA}, \mathtt{VA}, \mathtt{LVA}\}$. The high dimensional output of FMS is controlled by a summarization network S_1 to have a reduced dimension $\mathbb{R}^{L \times d_x}$ which goes through feedforward and normalization layers. If there are a total of P number of FMS units inside MTL,

the dimensionality of the normalization layer is $\mathbb{R}^{P \times L \times d_x}$ which is again mapped to $\mathbb{R}^{L \times d_x}$ using a secondary summarization network S_2. The output of the last MTL of FMT, thus computed, is fed into a Gated Recurrent Unit (GRU) to have a d_y dimensional predicted summary word embedding, $Summ_i$. An overview of a single Factorized Multimodal Self-attention (FMS) block in MTL is presented in the 5.4.

The summary word predicted by the FMT in the first step, $Summ_1$, is appended to the text transcript and fed into the same FMT in the next step to predict the second summary word, $Summ_2$. This process is continued until the model generates a stop-word or a predetermined summary length is reached. We only compute loss over the target sequence, as suggested by [180].

5.4 Experiments

To explore the role of multimodality in abstractive text summarization, we conduct multiple experiments evaluating textual and visual modalities separately and jointly on both How2 and AVIATE datasets. Additionally, we investigate the role of OCR-generated text for the academic presentation videos in AVIATE for improving summary generation.

5.4.1 Training

We train `FLORAL` using Pytorch framework on NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, with CUDA-10 and cuDNN-7 installed. We pre-train all the parameters of `FLORAL` using WikiLM [180], and fine-tune on the summarization datasets (How2 and AVIATE). Similar to encoder-decoder models, we only compute loss over the target sequence. In Table 5.1, we present the details of hyper-parameters used in the baselines and in `FLORAL`.

Table 5.1: Hyperparameters of different abstractive baseline models compared to `FLORAL`.

Modality	**Models**	**Hyperparameters**					
		Batch-size	**#Steps**	**Peak LR**	**Optimizer**	**Dropout**	**#Parameters**
Unimodal (`Text` Only)	PG	16	230K	0.01	Adagrad	0.2	42m
	PG-MMR	16	230K	0.01	Adagrad	0.2	42m
	Hi-MAP	32	200K	0.01	Adagrad	0.1	36m
	CopyTransformer	16	200K	0.05	Adam	0.2	105m
Multimodal (`Text` + `Audio` + `Video`)	Multimodal HA	64	300k	0.05	Adam	0.3	8m
	MulT En-De	24	300k	0.01	Adam	0.2	477m
	FMT En-De	32	300k	0.01	Adam	0.2	495m
	MulT LM	32	500k	0.01	Adam	0.1	242m
	`FLORAL`	16	500k	0.01	Adam	0.1	260m

5.4.2 Baselines

We compare the performance of the following extractive and abstractive unimodal and multimodel text summarization models both on How2 and AVIATE datasets.

Extractive Summarizers (Text Only)

- **Lead3** is the most common baseline which simply selects the leading three sentences of the document as its summary.
- **KLSumm** [120] is a greedy algorithm that minimizes the Kullback-Lieber (KL) divergence between the original document and the ground-truth summary.
- **TextRank** [118] runs a modified version of PageRank on a weighted graph, consisting of nodes as sentences and edges as similarities between sentences.
- **LexRank** [117] is a graph-based algorithm that represents sentences as vertices, and edges represent the similarity.

Abstractive Summarizers (Text Only)

- **Pointer Generator (PG)** [38] network is one of the most popular sequence to sequence (seq2seq) summarization architectures. PG allows both generating words from the vocabulary or copying from the source document.
- **Pointer Generator-MMR** [40] uses MMR along with PG for better coverage and redundancy mitigation. Here MMR computes a similarity score of sentences with the source text and modifies the attention weights for a better summary generation.
- **Hi-MAP** [12] is a hierarchical MMR-attention based PG model, which extends the work of PG and MMR. Here, MMR scores are calculated at word level and incorporated in the attention weights for a better summary generation.
- **CopyTransformer** (Bottom-up Abstractive Summarization) [15] uses the transformer parameters proposed by [45]. It uses a content selection module that over-determine phrases in the source document.

Abstractive Summarizers (Video + Audio + Text)

- **Multimodel Hierarchical Attention** [17] extends the work of [183], which was originally proposed for multimodal machine translation. This model fuses visual and textual modalities and captures the context of visual and textual features along with hierarchical attention to generate summaries.
- **MulT Encoder-Decoder** is an encoder-decoder based summarization architecture, which uses MulT (Multimodal Transformer for Unaligned Multimodal Language Sequences) model [19] as its encoder and decoder unit.
- **FMT Encoder-Decoder** is an encoder-decoder network, similar to MulT encoder-decoder. This baseline uses Factorized Multimodal Transformer [3] as the encoder and decoder units.

Table 5.2: Ablation results after incorporating OCR generated text into the ASR generated text transcript using guided-attention for different extractive and abstractive unimodal and multimodal text summarization systems on AVIATE.

	Model	Modality	AVIATE Dataset		
			R-1	R-2	R-L
Extractive (Text only)	KLSumm	ASR	22.19	2.05	15.59
		ASR+OCR	24.27	2.31	16.92
	TextRank	ASR	22.15	2.71	16.42
		ASR+OCR	24.55	2.72	22.1
	LexRank	ASR	22.63	2.49	15.68
		ASR+OCR	24.55	2.72	22.1
Abstractive (Text only)	PG	ASR	26.27	2.01	22.96
		ASR+OCR	27.77	2.05	23.81
	PG-MMR	ASR	27.34	2.72	22.63
		ASR+OCR	27.82	3.97	23.93
	Hi-MAP	ASR	27.62	3.16	22.1
		ASR+OCR	28.13	3.87	22.5
	CopyTransformer	ASR	29.93	3.73	25.13
		ASR+OCR	30.27	3.94	27.06
Multimodal (Text + Audio + Video)	Multimodal HA	ASR+A+V	27.51	4.83	25.32
		(ASR+OCR)+A+V	28.14	4.91	26.12
	MulT Encoder-Decoder	ASR+A+V	29.65	4.12	26.47
		(ASR+OCR)+A+V	30.89	4.34	27.2
	FMT Encoder-Decoder	ASR+A+V	31.8	4.49	26.1
		(ASR+OCR)+A+V	32.85	4.6	27.65
	MulT LM	ASR+A+V	31.71	4.07	27.58
		(ASR+OCR)+A+V	33.47	4.12	28.73
	FLORAL	ASR+A+V	33.26	6.38	28.52
		(ASR+OCR)+A+V	**37.13**	**11.04**	**31.47**

- **MulT LM** is MulT-based architecture. This is most akin to our proposed FLORAL model; only the FMT module of FLORAL is replaced by MulT [19] to have this multimodal summarization baseline.

5.5 Experimental Results

We present a quantitative analysis of the summaries using the standard metrics for abstractive summarization – ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) [9, 184] that measure the unigrams, bigrams, and longest common sequence between the ground-truth and the generated summaries, respectively. Additionally, we perform extensive qualitative analysis using human experts to primarily understand the fluency and informativeness of the summaries. We also analyze the word distributions in the transcriptions and summaries.

5.5.1 Quantitative Analysis

At first, we evaluate the performance of commonly used extractive and abstractive text summarization models both on the How2 and AVIATE datasets. Note that the average length of text transcripts in How2 is much less than that of AVIATE. Following our intuition, PG-based text summarization networks perform relatively well on How2 as shown in Table 5.3; but their performance drastically drops on AVIATE. This result can be attributed to the fact that attention-based encoder-decoder networks often fail to capture long-term dependencies when the source text is long and noisy. Hence, we decide to use transformer-based pre-trained BERT [179] as the text-embedding layer in our model.

In addition to text-only models, we train two video-only models – the first one uses a single convolutional and pooling layer for feature extraction from the entire video, while the second one applies a single layer RNN over these vectors in time. We observe in Table 5.3 that even using only action features in the videos leads to almost competitive R-1, R-2, and R-L scores compared to text-only models, in some cases often better than extractive text-only systems. This result demonstrates the importance of both modalities for summarization.

Table 5.3: `FLORAL` achieves highest performance in ROUGE-1, ROUGE-2 and ROUGE-L over text based extractive system (Lead3, KLSumm, TextRank and LexRank) and abstractive systems (Seq2Seq, PG, PG-MMR, Hi-MAP and CopyTransformer) and multimodel baselines (Multimodel HA, MulT based encoder decoder, FMT based encoder decoder, MulT based language model and `FLORAL`) in How2 and AVIATE datasets.

Modality	**Model**	**How2**			**AVIATE**		
		R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Extractive Systems (`Text` only)	Lead3	43.97	10.31	35.76	19.23	1.82	14.69
	KLSumm	28.64	12.3	16.2	24.27	2.31	16.92
	TextRank	27.48	12.41	16.55	22.15	2.71	16.42
	LexRank	27.95	12.48	16.98	22.63	2.49	15.68
Abstractive Systems (`Text` only)	Seq2Seq	55.32	23.06	53.9	26.62	2.88	23.32
	PG	51.67	22.65	50.21	27.77	2.05	23.81
	PG-MMR	52.9	23.21	50.24	27.34	2.72	22.63
	Hi-MAP	49.2	21.36	47.36	28.13	3.87	22.5
	CopyTransformer	53.7	23.87	48.04	30.27	3.94	27.06
`Video` only	Action Ft. only	45.21	12.74	38.5	22.87	2.78	17.13
	Action Ft. + RNN	48.23	19.1	46.3	23.54	2.92	18.21
Multimodel Systems (`Video+Audio+Text`)	Multimodal HA	55.87	26.32	54.9	28.14	4.91	26.12
	MulT En-De	55.89	26.79	55.1	30.89	4.34	27.2
	FMT En-De	55.98	26.83	55.4	32.85	4.6	27.65
	MulT LM	56.13	26.89	55.41	33.47	4.12	28.73
	`FLORAL`	**56.89**	**26.93**	**56.80**	**37.13**	**11.04**	**31.47**

Incorporation of OCR

Since the AVIATE dataset is composed of conference presentation videos, we observe that in almost 94.8% videos in the entire dataset, the speaker shows slides during the presentation. The text in

these slides is succinct and contains the most important key-phrases which are crucial for summary generation. Table 5.2 shows the performance improvement for every summarization model when the OCR is fused with ASR transcript using our guided attention mechanism. We also consider direct concatenation of OCR transcript with the ASR-transcript; however, it resulted in lower performance as compared to guided attention fusion. The guided attention ensures the filtering of redundant and repetitive words in OCR and ASR transcripts. For every summarization model, we only use the first 500 tokens of the OCR transcript. We did not consider incorporating OCR for How2 as this dataset only contains instructional videos, and there is no text shown in the frames of instructional videos.

Table 5.2 shows that unimodal extractive text summarization models, namely KLSumm, TextRank and LexRank, yield an improvement of $[2.1 - 2.3]$ R-1 points, $[0.1 - 0.3]$ R-2 points and $[0.6 - 6.5]$ R-L points after incorporating OCR-generated text. Similarly for abstractive summarization, the very popular PG-MMR network produces $0.5, 1.25$, and 2.3 points performance improvement in terms of R-1, R-2, and R-L scores, respectively. The other abstractive summarization networks, namely PG, Hi-MAP, CopyTransformer, also support our hypothesis and show an improvement of $[0.5 - 4]$ points in terms of all three evaluation metrics.

Influenced by the performance of unimodal summarization models, we incorporate the OCR transcripts into all of our multimodal baselines. Supporting our intuition, the multimodal systems obtain significant performance enhancement with OCR transcripts as shown in Table 5.2. The multimodal hierarchical attention model, MulT, and FMT-based encoder-decoder models show $[0.8 - 1.5]$ points improvement in the R-L score. Our proposed `FLORAL` model yields the highest performance boost with OCR among all the multimodal systems, showing 3.87, 4.66, and 2.95 point enhancement in R-1, R-2, and R-L scores respectively. The performance boost can be easily attributed to the keywords in the OCR-generated transcript, which guides the text-embeddings to attend the most important portions in a very long ASR transcript. Hence, in the rest of our discussion, we always report results with `(ASR + OCR)` transcript, fused with guided attention, as the textual modality.

Complementarity of Multiple Modalities

Table 5.3 shows the ROUGE scores for different unimodal and multimodal text summarization systems on the How2 and AVIATE datasets. Among the unimodal variants, the abstractive text summarization systems generally perform much better than the extractive systems, especially on AVIATE. Note that despite being a very strong extractive baseline, Lead3 does not perform well on AVIATE, as the text transcripts of academic presentation videos do not tend to be structured with the most important information at the beginning. The two video-only models, simple conv-pool action features and action features with RNN perform very close to the abstractive text-only baselines, which clearly indicates the necessity of visual modality in addition to the textual modality.[11] As presented in Table 5.3, the MulT, and FMT multimodal baselines and the proposed `FLORAL` model beat most of the unimodal systems by a large margin, on both the

[11]We do not evaluate the performance of acoustic modality separately, as the MFCC audio features are typically incorporated to capture the pitch, intonation, and other tonal-specific details of the speaker, which do not contribute individually to the summarization task.

datasets. This result is expected because of the inherent ability of MulT and FMT to capture the intra-model and inter-modal dynamics within asynchronous multimodal sequences and incorporate diverse information in a single network. Overall, the combination of visual, acoustic, and textual signals significantly improves over the unimodal variants, with an improvement of 1.57, 3.04, and 3 R-1, R-2, and R-L points on How2 and 6.86, 7.1 and 4.41 on AVIATE.

We manually investigate some video samples of AVIATE where the multimodal system generates a better summary than the unimodal system. In most of these samples, the textual transcript is very noisy and contains many irrelevant words that are not much required for the summary generation. Figure 5.5 shows an example training instance of the AVIATE dataset with three different modalities. A closer look into the ASR and OCR transcripts reveals the presence of irrelevant and noisy words. For example, the very first sentence of the ASR transcript "hi i'm lisa ann hendricks and today" does not contribute to the summary generation. As a result, these samples require additional cues for performance improvement, which are availed from the multimodal signals. The variation of outputs from various unimodal and multimodal summarization networks for a single video sample is shown in Table 5.10.

Comparative Study on How2

Table 5.3 shows that the performance of unimodal and multimodal summarization systems on How2 as compared to AVIATE. In contrast to prior work on news-domain summarization [185], the seq2seq model performs the best among all unimodal systems on How2, achieving 55.32, 23.06, and 53.9 R-1, R-2, and R-L scores, respectively. As indicated by [17], the PG model performs lower than seq2seq on How2 due to the lack of overlaps between input and output, which is the important feature of PG networks. Among the multimodal systems, our proposed `FLORAL` model yields the best results; however, the other multimodal baselines reach almost competitive ROUGE scores compared to `FLORAL` on this dataset. Noticeably, despite having a simple structure, the multimodal hierarchical attention model performs very well on How2. On this dataset, `FLORAL` achieves 56.89, 26.93, and 56.80 R-1, R-2, and R-L scores, respectively, which are $[0.1 - 2]$ points higher than the scores achieved by other multimodal baselines.

Comparative Study on AVIATE

AVIATE contains longer videos than How2, resulting in longer transcripts and ground-truth summaries. As shown in Table 5.3, the best performing unimodal summarization model on AVIATE is CopyTransformer. As the ASR and OCR generated summaries are very long, the extractive systems do not perform well on this dataset. While PG and seq2seq yield 23.32 and 23.81 R-L scores respectively, CopyTransformer produces a 27.06 R-L score, outperforming all other unimodal systems. The superior performance of CopyTransformer over PG and seq2seq can be attributed to the self-attention mechanism of transformers which helps to capture long-term dependencies. The incorporation of visual and acoustic modalities significantly improves the ROUGE scores on this dataset. `FLORAL` beats all the transformer-based encoder-decoder networks and language models. `FLORAL` produces 37.13, 11.04 and 31.47 R-1, R-2 and R-L scores respectively, where the second-ranked model on AVIATE, `MulT-LM` obtains 33.47 R-1, 4.12 R-2, 28.73 R-L scores, which are almost $[2.7 - 7]$ points lower than that of `FLORAL`. Since AVIATE contains $6,680$ training

samples, which may not be enough for today's deep neural models, the factorization mechanism on FMT, which allows an increasing number of self-attention to better model the multimodal phenomena, results in its superior performance, without encountering difficulties even on the relatively low-resource setup of AVIATE. Pre-training of all the parameters of FLORAL also has an immense impact, which helps in beating all other baselines by a significant margin.

Table 5.3 also shows that all the unimodal and multimodal summarization models obtain almost $[18 - 25]$ points higher R-1 and R-L scores and $[3 - 6]$ points higher R-2 score on How2 over AVIATE. For example, FLORAL yields 56.89, 26.93, and 56.80 R-1, R-2, and R-L scores on How2 and 37.13, 11.04 and 31.47 R-1, R-2 and R-L scores on AVIATE. We can observe from Table 5.3 that all the baseline models as well as FLORAL yield higher R-1, R-2 and R-L scores on How2 than AVIATE. The overall better performance of every system on How2 than AVIATE can be attributed to two factors – firstly, the text transcripts of How2 are manually annotated. In contrast, we use ASR and OCR outputs as the transcripts for AVIATE. The large margin of ASR and OCR errors in some of the train and test samples significantly affect the model performance. Secondly, since the video length, transcript length, and reference summary length are much longer in AVIATE than How2, the summarization task becomes more challenging in AVIATE. Furthermore, since AVIATE comprises many scientific presentation videos, the audio transcript contains complex academic words, leading to a larger dictionary for the language generation task. Overall, the results in Table 5.3 conclude that the AVIATE dataset is more exacting than How2, indicating room for further research with fine-grained and sophisticated multimodal models for long videos.

In our next experiment, we demonstrate how the summarization task becomes more challenging with longer videos. We divide the AVIATE dataset into three portions - short videos (duration less than 10 minutes), medium videos (duration between 10 minutes and 30 minutes) and long videos (duration more than 30 minutes). We split each portion in $4:1$ ratio and train and test all the multimodal systems on each segment. Table 5.4 shows the performance reduction of each model with the increase in video length. In general, we observe that all four multimodal baselines yield R-L score in the range of $[27.03 - 34.09]$ on the short videos. However, the score reduces to $[23.19 - 28.69]$ for the long videos. The performance of FLORAL also decreases from short to medium and long videos; however, the span of reduction of R-L score is only $[0.39 - 0.64]$, which is relatively less than all other baselines. We also notice that the LM-based systems generally capture

Table 5.4: Performance of multimodal baseline models and FLORAL on short (< 10 min), medium (> 10 min & < 30 min) and long (> 30 min) videos of AVIATE. As the video length and the corresponding reference summary length increase, the performance of all baseline models decreases heavily. However, FLORAL performs well across all video lengths.

Model	**AVIATE Dataset**											
	Short Videos			**Medium Videos**			**Long Videos**			**Whole Dataset**		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Multimodal HA	30.11	5.12	27.03	26.54	4.97	25.31	24.18	4.65	23.19	28.14	4.91	26.12
MulT En-De	31.44	5.32	27.38	27.73	4.93	26.89	25.62	4.67	24.12	30.89	4.34	27.2
FMT En-De	33.62	5.81	31.06	32.1	5.91	28.31	31.48	4.96	27.02	32.85	4.6	27.65
MulT LM	33.13	5.65	30.58	34.09	5.95	29.27	33.37	5.31	28.69	33.47	4.12	28.73
FLORAL	36.13	11.62	31.44	35.59	11.64	31.05	34.31	10.39	30.8	37.13	11.04	31.47

Table 5.5: Significance of multimodal cues in FLORAL. The combination of visual, textual, and acoustic signals significantly improves over the unimodal variants, with a relative improvement of R-1, R-2 and R-L scores of 9.99%, 8.11% and 11.80% respectively over the best unimodal variant.

Model	Modality	AVIATE		
		R-1	R-2	R-L
FLORAL	T	**27.14**	**2.93**	**19.67**
	A	21.64	1.97	16.83
	V	22.31	2.29	17.04
	T+A+V	**37.13**	**11.04**	**31.47**
$\Delta_{multi-unimodal}$		↑ 9.99	↑ 8.11	↑ 11.80

long-term dependencies better than traditional encoder-decoder based systems.

The complementarity of multiple modalities in the performance of FLORAL is shown in Table 5.5. To understand the importance of visual modality, we feed zero input in other two modality channels of FLORAL and continue the process for all three modalities. We observe that the textual modality provides the best performance among unimodal variants. The addition of visual and acoustic features improves significantly over the unimodal baselines and achieves the best performance - with an increase in R-1, R-2 and R-L score of 9.99, 8.11 and 11.80 respectively over the best unimodal variant.

Table 5.6: Transferability of the proposed FLORAL model on the two available multimodal abstractive text summarization datasets - How2 and AVIATE. The network is trained on the dataset in each row, and is tested on the dataset shown in each column. The second row indicates the performance of FLORAL on the How2 videos whose transcripts are generated from ASR.

	How2			AVIATE		
	R-1	R-2	R-L	R-1	R-2	R-L
How2	56.89	26.93	56.80	21.35	4.88	23.11
How2 with ASR	54.19	24.07	52.11	22.77	5.98	24.07
AVIATE	52.68	21.33	49.9	37.13	11.04	31.47

Table 5.7: Transferability of the proposed FLORAL model when instead of the annotated transcript, Automated method for Transcription generation [1] is used on the two available multimodal abstractive text summarization datasets - How2 and AVIATE. The network is trained on the dataset in each row and is tested on the dataset shown in each column.

	How2			AVIATE		
	R-1	R-2	R-L	R-1	R-2	R-L
How2	54.19	24.07	52.11	19.67	3.54	21.87
AVIATE	52.68	21.33	49.9	37.13	11.04	31.47

Table 5.8: Transferability of the proposed FLORAL model on videos of different length in the AVIATE dataset. The network is trained on the videos in each row, and tested on the videos shown in each column.

		AVIATE								
		Short Videos			Medium Videos			Long Videos		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
AVIATE	Short Videos	31.05	8.49	28.25	27.22	7.36	23.74	24.71	6.12	21.53
	Medium Videos	31.43	8.74	28.34	30.21	8.48	26.90	28.66	8.29	26.54
	Long Videos	32.84	9.43	28.56	31.08	9.64	26.98	30.23	8.89	27.25

Transferability of FLORAL

Table 5.7 shows the transferability property of FLORAL between How2 and AVIATE. When trained and tested on the same dataset, FLORAL produces the best ROUGE scores, which is expected. However, when trained on AVIATE and tested on How2, FLORAL yields an R-L score of 49.90, which is just 6.9 decrease in R-L score (11.83% reduction in performance) than the one when trained and tested on How2. The vice-versa is not true, i.e., when trained on How2 and tested on AVIATE, the performance drop is drastic (26.56% reduction in performance). As the videos in How2 have human-annotated transcripts and those in AVIATE have ASR-generated transcripts, for fair comparison of transferability, we extract the ASR transcripts of the How2 videos and train FLORAL. The results of this experiment are shown in the second row of Table 5.7. We observe that the ASR transcript reduces the test performance on How2, which is expected due to the noise in the ASR output. The transferability score on AVIATE improves a bit, but the performance drop is still heavy (25.51% reduction in performance). From all these experiments, we can conclude that since the videos of How2 are very short, the learned weights do not perform well for longer videos. However, AVIATE consists of diverse-length videos, and thus, the trained model on AVIATE yields good results on How2 as well.

Table 5.8 shows the transferability of FLORAL across short, medium and long videos of AVIATE. When trained on the long videos, FLORAL performs the best across all three portions. However, when trained on short videos, the model can not learn long-term dependency for lengthier videos. The same property supports the results on Table 5.7. Since the How2 dataset contains only short videos, the model does not perform well when trained on How2 and tested on AVIATE. The longer videos in the training set helps the model to generalize well across videos of various lengths.

5.5.2 Qualitative Analysis

In addition to ROUGE scores, we conduct a qualitative analysis by performing a human evaluation to understand the standard of the summary outputs. Following the abstractive summarization human annotation work of [186], the summaries were evaluated by five annotators[12] to rate the generated summaries on a scale of $[1-5]$ on four parameters - informativeness, relevance, coherence, and fluency. For the evaluation, we randomly sampled 300 videos from the test sets of How2 and AVIATE. Table 5.9 shows the average human evaluation scores for 4 text-only, 1 video-only and 3

[12]We employed five annotators who are experts in NLP, and their age ranges between 24-35 years.

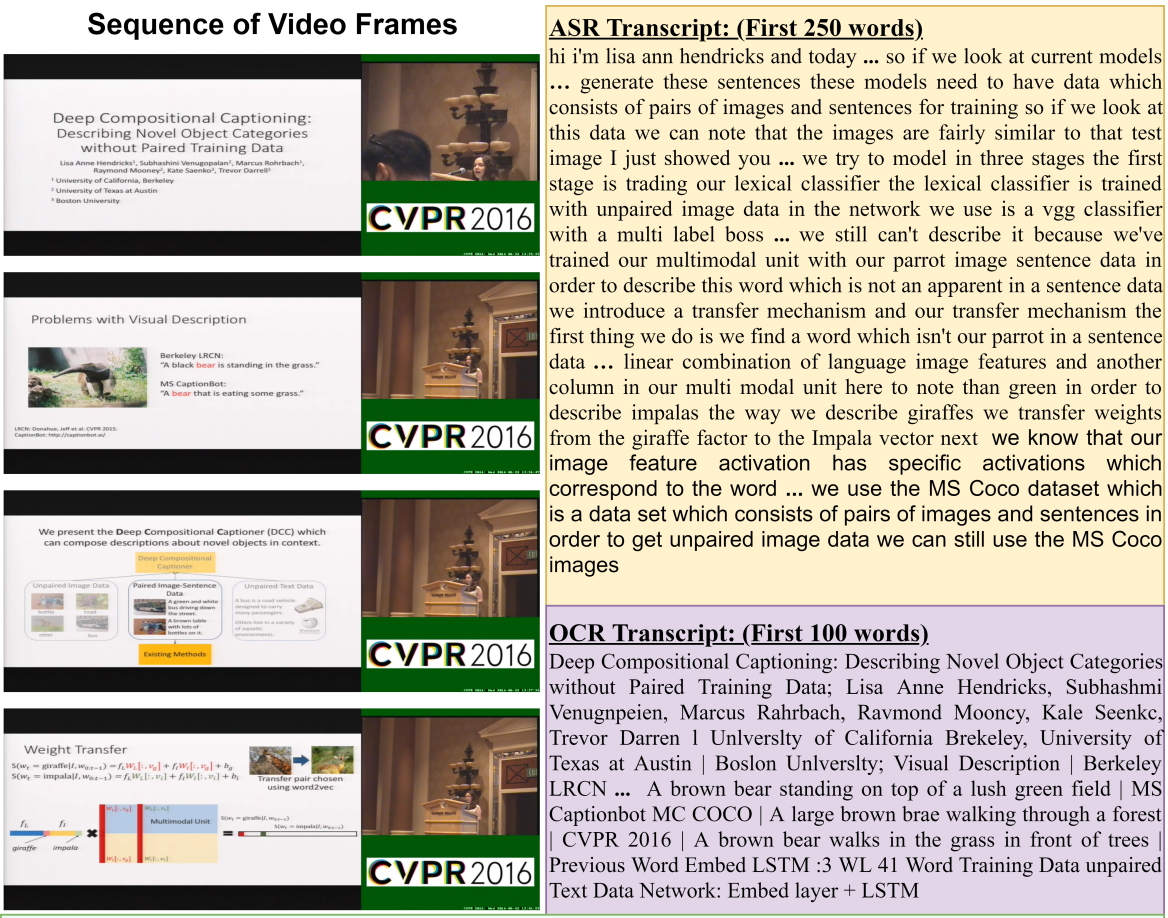
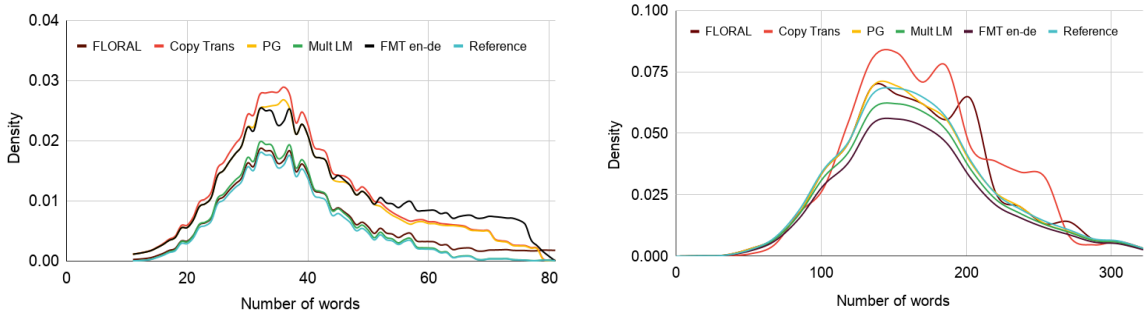


Figure 5.5: Example of AVIATE dataset with three different modalities. To obtain the text transcripts from the acoustic modality, we apply Deep Speech [1], a pre-trained end-to-end automatic speech recognition (ASR) system. We extract the text shown in the slides in the presentation videos using Google OCR Vision API. We use the abstracts of corresponding research papers as the ground-truth summaries.

multimodal models. In general, we observe that PG has difficulty in summarizing articles with repetitive information and tends to assign a lower priority to less occurring important keywords.

Table 5.9: Scores for human evaluated metrics - Informativeness (INF), Relevance (REL), Coherence (COH), Fluency (FLU) over text based extractive systems (KLSumm and TextRank), abstractive systems (PG and CopyTransformer), video based abstractive systems (Action features with RNN) and multimodel systems (FMT Encoder Decoder, MulT Language Model and `FLORAL`) on How2 and AVIATE datasets.

Modality	**Model**	**How2**				**AVIATE**			
		INF	**REL**	**COH**	**FLU**	**INF**	**REL**	**COH**	**FLU**
Extractive Systems (`Text only`)	KLSumm	2.82	2.54	2.98	3.14	1.91	1.56	2.14	2.13
	TextRank	2.92	2.73	2.82	3.12	2.1	1.83	2.13	2.12
Abstractive Systems (Text only)	PG	3.45	3.17	3.12	3.32	3.49	3.38	3.41	**3.62**
	CopyTrans.	3.46	3.18	3.18	3.36	3.54	3.39	3.48	3.56
Abstractive Systems (Video only)	Action Ft.+RNN	3.54	3.20	3.21	3.40	3.52	3.27	3.31	3.41
Multimodel Systems (Video+Audio+Text)	FMT En-De	3.61	**3.39**	3.37	3.67	3.64	3.32	3.37	3.49
	MulT LM	3.57	3.38	3.34	3.68	3.67	3.39	3.38	3.45
	`FLORAL`	**3.62**	3.38	**3.41**	**3.71**	**3.89**	**3.41**	**3.41**	3.41



(a) Density curve on How2 dataset.

(b) Density curve on AVIATE dataset.

Figure 5.6: Word distribution of machine-generated summaries in comparison with the ground-truth summaries for different unimodal and multimodal systems on How2 and AVIATE datasets.

The extractive summarization systems sometimes pick sentences extraneous to the summary. For example, we notice some summaries generated by KLSumm starting with *"Good afternoon everyone, I am ·"*, which is the very first line of the transcript. In contrast, the multimodal summarization models generate summaries with greater relevance and informativeness. Our proposed `FLORAL` model obtains high scores on informativeness, relevance, and coherence on AVIATE, but sometimes seems to generate less fluent summaries. This fluency problem mostly stems from errors in ASR and OCR generated text. Some of these phenomena are illustrated with instances from AVIATE in Table 5.10.

We also analyze the word distributions of the ground-truth summaries and different system-generated summaries. The density curves in Figure 5.6 shows that for both How2 and AVIATE, the abstractive unimodal and multimodal summarization models generate summaries shorter than the ground-truth summary. The average length of summaries is highest for CopyTransformer. Interest-

Table 5.10: Comparison of ground-truth summary and outputs of 7 different unimodal and multimodel abstractive text summarization systems - FLORAL, MultT LM, MultT Encoder-Decoder, CopyTransformer, multimodal hierarchical attention (HA), Pointer Generator (PG) and Pointer Generator with MMR (PG-MMR) - arranged in the order of best to worst ROUGE-L scores in this table. Red highlighted text indicates a positive correlation of context w.r.t. ground-truth summary while blue color represents a negative correlation with ground-truth summary.

Model	R-L	Output
Ground-truth	-	We study the problem of semi-supervised question answering—utilizing unlabeled text to boost the performance of question answering models. We propose a novel training framework, the Generative Domain-Adaptive Nets. In this framework, we train a generative model to generate questions based on the unlabeled text, and combine model-generated questions with human-generated questions for training question answering models. We develop novel domain adaptation algorithms, based on to learn richer context-aware model on the insight and the human-generated data distribution. Experiments show that our proposed framework obtains substantial improvement from unlabeled text.
FLORAL (ours)	31.47	We study the problem of semi-supervised concepts unlabeled text to boost the performance of without without models. We propose a novel training framework, the Generative Domain-Adaptive Nets. In this framework, we train a database model and generate concepts based on procedure on how partitioning in the levels between t, and combine model-generated attention with human-generated types for training associations directed-generation models. We develop novel algorithms, based on model, to understand the discrepancy between the generated data results and the human-generated data distribution. Experiments show model obtains improvement from text.
MultT LM	28.73	In learn paper are neural paper, we propose a new model and the paper, we study demonstrated the character score We investigate the problem answering utilizing unlabeled text to understand trained models. we train a understanding model to generate understanding based on the text, and combine trained understanding with human generated understanding to training data trained we assume like, on a public unsupervised Subsequently, prediction trained to learn richer context-aware model on the insight In trained facilitates smaller annotations previously reported understanding understanding on approach and questions along five perform two then performance The the semantic best dataset and the proposed
MultT En-De	27.2	We present the problem of supervised new utilizing text of new new models. We propose a problem model review a new model In this paper, we propose a simple model Experimental resultsprove the standard model achieves method to then the proposed among review model to also performance based on the proposed CoNLL-2012 dataset. In this we train a new model to generate a model and combine model. of machine morphology into more prediction on the drawback of the system classes. using error-correcting codes collection. We evaluate investigation of NMT and induced Experiments errors. shed light affect vs. scores on a word its dataset with the proposed based
CopyTrans.	27.06	We propose a novel method for novel method for state-of-the-art question answering we show that the method achieves state-of-the-art performance on the state-of-the-art performance of the proposed method on the systems over sequential text tasks or as independent, parallel tasks. In this framework model generate text trained benefits of the new semantic high a image and at them as
HA	26.12	We propose the novel model has the paper, we propose a generative model In both learning can can the benchmark model To capable to also performance of existing the pipelined of the model-generated model: distribution for the human-generated data distribution. Experiments show that our proposed framework to the consistency of the art data distribution and event feedforward network. The an unsupervised dependency framework performance and the human data distruction .
PG	23.81	In this paper we consider the problem of learning a deep neural network, we propose a novel neural network architecture based on novel neural network architecture that is trained end-to-end trainable convolutional neural network (CNN) architecture is able to train a convolutional neural network architecture that is capable of achieving state-of-the-art performance compared to state-of-the-art performance on three benchmark datasets.
PG-MMR	22.63	We propose a novel method for object detection based on a novel object detection method that uses a novel model based on the posterior distribution of the posterior distribution over the parameters of the number of claims and We show that models can be used as well as compared to In this paper, we study the a algorithm that the proposed method can outperforms the state-of-the-art methods on a large number of

ingly, FLORAL and PG generated summaries are similar in length. However, FLORAL outperforms PG by a large margin, which illustrates that for improvements in ROUGE scores, an informative summary is more crucial than a lengthier summary.

5.6 Limitations

While our method demonstrates the capability to integrate guidance from multiple modalities such as audio, video, and text to produce high-quality summaries, there are several limitations to consider.

First, the model employs a regressive approach for generation, meaning it generates one token at a time in an iterative manner. This approach requires the model to make 250-350 passes to produce a single summary, making it computationally intensive. Each pass involves processing the entire model, leading to significant compute resource consumption.

Second, the iterative nature of our model, which is akin to an entailment style, poses challenges for generating short summaries. The model requires generating at least 30-50 tokens to properly understand the context and produce coherent and relevant summaries. As a result, very short summaries may not capture the intended meaning or key information effectively. This limitation suggests that while our model excels in producing detailed and contextually rich summaries, it may struggle with tasks that demand extremely concise outputs.

Third, the model performs optimally when all three modalities (audio, video, and text) are available. Its performance significantly drops when only one or two modalities are used. In contrast, the method proposed in the later chapter shows better performance even with a combination of two modalities, highlighting its robustness and versatility in scenarios where not all three modalities are available.

These limitations underscore the need for continued refinement and adaptation of our approach to ensure its practicality and efficiency across various summarization tasks and resource constraints.

5.7 Conclusion

In this chapter, we explore the role of multimodality in abstractive text summarization. All the previous studies in this direction have used either images or short videos as the visual modality and generate one or two lines long summary, and thus, fail to perform on longer videos. Moreover, there exists no benchmark dataset for abstractive text summarization of medium and long videos. In this work, we introduce AVIATE, the first large-scale dataset for abstractive text summarization with videos of diverse duration, compiled from paper presentation videos in renowned academic conferences. We then propose FLORAL, a Factorized Multimodal Transformer based decoder-only Language Model, which uses an increasing number of self-attentions to inherently capture inter-modal and intra-modal dynamics within the asynchronous multimodal sequences, without encountering difficulties during training even on relatively low-resource setups. To evaluate FLORAL, we perform extensive experiments on How2 and AVIATE datasets and compare them against several unimodal and multimodal baselines. Overall, FLORAL achieves superior performance over previously proposed models across two datasets.

Chapter 6

Abstractive Extreme Text Summarization using Multimodal Signals

The realm of scientific text summarization has experienced remarkable progress due to the availability of annotated brief summaries and ample data. However, the utilization of multiple input modalities, such as videos and audio, has yet to be thoroughly explored. At present, scientific multimodal-input-based text summarization systems tend to employ longer target summaries like abstracts, leading to an underwhelming performance in the task of text summarization.

In this chapter, we address the novel task of *extreme abstractive text summarization* (*aka TL;DR generation*) by leveraging multiple input modalities. To this end, we introduce `mTLDR`, a first-of-its-kind dataset for the aforementioned task, comprising videos, audio, and text, along with both author-composed summaries and expert-annotated summaries. The `mTLDR` dataset accompanies a total of $4,182$ instances collected from various academic conference proceedings, such as ICLR, ACL, and CVPR.

Subsequently, we present `mTLDRgen`, an encoder-decoder-based model that employs a novel dual-fused hyper-complex Transformer combined with a Wasserstein Riemannian Encoder Transformer, to dexterously capture the intricacies between different modalities in a hyper-complex latent geometric space. The hyper-complex Transformer captures the intrinsic properties between the modalities, while the Wasserstein Riemannian Encoder Transformer captures the latent structure of the modalities in the latent space geometry, thereby enabling the model to produce diverse sentences. `mTLDRgen` outperforms 20 baselines on `mTLDR` as well as another non-scientific dataset (How2) across three Rouge-based evaluation measures.

However, while `FLORAL`, a Factorized Multimodal Transformer based decoder-only Language Model, shows superior performance for general summaries and excels in handling long videos by effectively capturing inter-modal and intra-modal dynamics, it falls short in the task of extreme abstractive text summarization. Our experiments demonstrate that `FLORAL` struggles with generating concise TL;DR summaries, highlighting the necessity for models specifically tailored for extreme summarization tasks.

Furthermore, based on qualitative metrics such as BERTScore and FEQA, and human evaluations, we demonstrate that the summaries generated by `mTLDRgen` are fluent and congruent to the original source material.

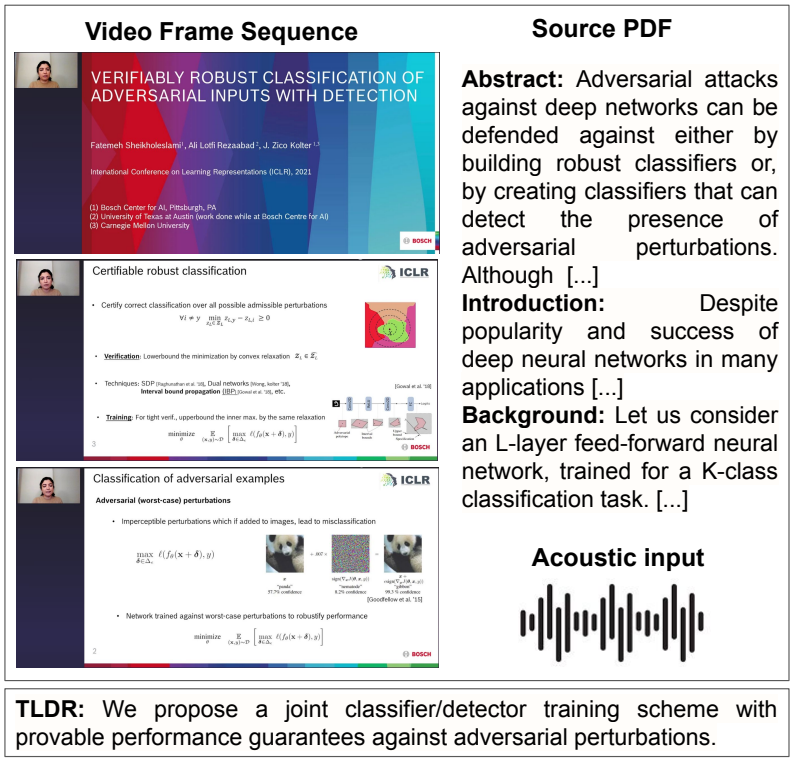


Figure 6.1: A sample of `mTLDR` dataset with video, text and audio modalities along with the target TLDR. The feature representations for video frames are obtained by ResNext, audio features are extracted using Kaldi, and the text is extracted from the pdf of the article.

6.1 Introduction

With the emergence of deep learning architectures like LSTM, Attention, and Transformer, the literature in the scientific community has skyrocketed. It is extremely hard to keep up with the current literature by going through every piece of text in a research article. The abstract of a paper often serves as a bird's eye view of the paper, highlighting the problem statement, datasets, proposed methodology, analysis, etc. Recent studies [23] re-purpose abstracts to generate summaries of scientific articles. However, it is cumbersome to go through the abstract of each paper. The abstracts are nearly 300 tokens long, and reading the complete abstract of every paper to figure out the mutual alignment is tedious. The task of TL;DR (*aka*, tl;dr, too long; didn't read) [67, 5] was introduced to generate an extremely concise summary from the text-only article highlighting just the high-level contributions of the work. Later, [68] introduced the CiteSum dataset for generating text-only extreme summaries. However, the text alone can not comprehend the entire gist of the research article. The multimodal information, including the video of the presentation and audio, often provide crucial signals for extreme text summary generation.

Problem statement: In this work, we propose a new task of multimodal-input-based TL;DR generation for scientific contents which aims to generate an extremely-concise and informative text

summary. We incorporate the visual modality to capture the visual elements, the audio modality to capture the tonal-specific details of the presenter, and the text modality to help the model align all three modalities. We also show the generalizability of the proposed model on another non-academic dataset (How2).

State-of-the-art and limitations: The pursuit of multimodal-input-based abstractive text summarization can be related to various other fields, such as image and video captioning [166, 167, 168, 169, 170], video story generation [171], video title generation [172], and multimodal sentence summarization [173]. However, these works generally produce summaries based on either images or short videos, and the target summaries are easier to predict due to the limited vocabulary diversity. On the other hand, scientific documents have a complex and structured vocabulary, which the existing methods [17] of generating short summaries are not equipped to handle. Recently, [23] proposed as a novel dataset for the multimodal text summarization of scientific presentations; however, it uses the abstract as the target summary, which falls short in producing coherent summaries for the extreme multimodal summarization (TL;DR) task.

In summary, this chapter offers the following contributions:

- **Novel problem:** We propose the task of extreme abstractive text summarization for scientific contents, by utilizing videos, audio and research articles as inputs.

- **Novel dataset:** The development and curation of the first large-scale dataset `mTLDR` for extreme multimodal-input-based text summarization of scientific contents. Figure 6.1 shows an excerpt from the `mTLDR` dataset. This dataset has been meticulously compiled from five distinct public websites and comprises articles and videos obtained from renowned international conferences in Computer Science. The target summaries are a fusion of manually-annotated summaries and summaries written by the authors/presenters of the papers.

- **Novel model:** We propose `mTLDRgen`, a novel encoder-decoder-based model designed to effectively capture the dynamic interplay between various modalities. The model is implemented with a dual-fused hyper-complex Transformer and a Wasserstein Riemannian Encoder Transformer. The hyper-complex Transformer projects the modalities into a four-dimensional space consisting of one real component and three imaginary components, thereby capturing the intrinsic properties of individual modalities and their relationships with one another. Additionally, the Wasserstein Riemannian Encoder Transformer is employed to apprehend the latent structure of the modalities in the geometry of the latent space.

- **Evaluation:** We benchmark `mTLDR` over six extractive (text-only), eight abstractive (text-only), two video-based and four multimodal summarization baselines, demonstrating the effectiveness of incorporating multimodal signals in providing more context and generating more fluent and informative summaries. We evaluate the benchmark results over the quantitative (Rouge-1/2/L) and qualitative (BERTScore and FEQA) metrics. Our proposed modal, `mTLDRgen`, beats the best-performing baseline by +5.24 Rouge-1 and +3.35 Rouge-L points. We also show the generalizability of `mTLDRgen` on another non-scientific dataset (How2).

Table 6.1: Statistics of the used datasets (mTLDR and How2) – the number of samples (#source), average token length of source documents (avg source len), average tokens in the target summaries (avg target len), and abstractness percentage (Abs) of datasets.

Dataset	#source	avg source len	avg target len	%Abs
How2	73993	291	33	14.2
mTLDR	4182	5K	18	15.9

6.2 Related Work

6.3 Proposed Dataset

To explore the efficacy of multimodal signals and enable enriched abstractive summaries aided by various modalities, we introduce mTLDR, the first large-scale **m**ultimodal-input based abstractive summarization (**TL;DR**) dataset with diverse lengths of videos. mTLDR is collected from various well-known academic conferences like ACL, ICLR, CVPR, etc. The only comparable dataset to mTLDR is the How2 dataset, which comprises short instructional videos from various topics like gardening, yoga, sports, etc. Compared to How2, mTLDR contains structured and complex vocabulary, which requires attention to diverse information while generating summaries.

Our compilation encompasses video recordings from openreview.net and videolecture.net, in addition to the accompanying source pdf and metadata information, including the details of the authors, title, and keywords. The collected dataset comprises a total of $4,182$ video recordings, spanning a duration of over $1,300$ hours. Of these, we designated $2,927$ instances as the training set, 418 for validation, and 837 for testing. The average length of the videos is 14 minutes, and the TLDR summary has an average of 19 tokens. The target summaries for the data are a combination of human-annotated and author-generated summaries. In terms of abstractness, mTLDR contains 15.9% novel n-grams. Each data instance includes a video, audio extracted from the video, an article pdf, and a target summary. We opted not to annotate or retain multiple summaries for a single instance to ensure efficient training and testing processes. We assert that a single extreme summary is sufficient to convey the essence of the paper. The target summaries for papers obtained from the ACL anthology were annotated as they lacked any author-generated summaries. Of the $4,182$ videos, a total of $1,128$ summaries were manually annotated by 25 annotators. During the annotation process, the annotators were instructed to thoroughly read the abstract, introduction, and conclusion and to have a general understanding of the remaining content. Each summary was then verified by another to confirm that it accurately represents the paper's major contributions.

In contrast, the How2 dataset [21] consists of $73,993$ training, $2,965$ validation, and $2,156$ test instances. The average token length for the source documents is 291, while for the target summary, it is 33. Compared to the source document, the target summaries contains 14.2% novel n-grams. The transcripts for videos and the target summary are human-annotated. Table 6.1 shows brief statistics of the How2 and mTLDR datasets.

6.4 Proposed Methodology

This section presents our proposed system, `mTLDRgen`, a **m**ultimodal-input-based extreme abstractive text summary (**TL;DR**) **gen**erator. Figure 6.2 shows a schematic diagram. During an academic conference presentation, there are typically three major modalities present – visual, audio, and text, each of which complements the others, and when combined, contributes to a rich and expressive feature space, leading to the generation of coherent and fluent summaries. `mTLDRgen` initially extracts features from the independent modalities and then feeds them to the dual-fused hyper-complex Transformer (DFHC) and the Wasserstein Riemannian Encoder Transformer (WRET) blocks. Cross-modal attention is used to fuse the visual and audio features with the text representations. Finally, the fused representation is fed to a pre-trained BART [161] decoder block to produce the final summary. The rest of this section delves into the individual components of `mTLDRgen`.

6.4.1 Video Feature Extraction

The video modality in an academic presentation often comprises variations in frames and kinesthetic signals, highlighting key phrases or concepts during a presentation. To capture visual and kinesthetic aspects, we utilise the ResNeXt-152-3D [175] model as it is pre-trained on the Kinetics dataset for recognition of 400 human actions. Four frames per second are extracted from the video, cropped to 112×112 pixels and normalized, and a 2048-dimensional feature vector is extracted from the ResNeXt-152-3D model for every 12 non-overlapping frames. The 2048-dimensional vector is then fed to the mean pooling layer to obtain a global representation of the video modality. Later, a feed-forward layer is applied to map the 2048-dimensional vector to a 512-dimensional vector.

6.4.2 Speech Feature Extraction

To capture the variations in the speaker's voice amplitudes, which are considered to signify the importance of specific topics or phrases [176], we extract audio features from the conference video. This is accomplished by extracting audio from the video using the FFMPEG package[1], resampling it to a mono channel, processing it to a $16K$ Hz audio sample, and dividing it into overlapping windows of 30 milliseconds. The extracted audio is then processed to obtain 512-dimensional Mel Frequency Cepstral Coefficients (MFCC) features. The final representation is obtained by applying a log Mel frequency filter bank and discrete cosine transformation, and the feature sequence is padded or clipped to a fixed length.

6.4.3 Textual Feature Extraction

In order to extract the feature representations for the article text, the pdf content is obtained through the Semantic Scholar Open Research pipeline (SSORP) [187]. SSORP uses the SCIENCEPARSE[2] and GROBID[3] APIs for text extraction from pdf. For the How2 dataset, the video transcriptions

[1] https://ffmpeg.org/
[2] https://github.com/allenai/science-parse
[3] https://github.com/kermitt2/grobid

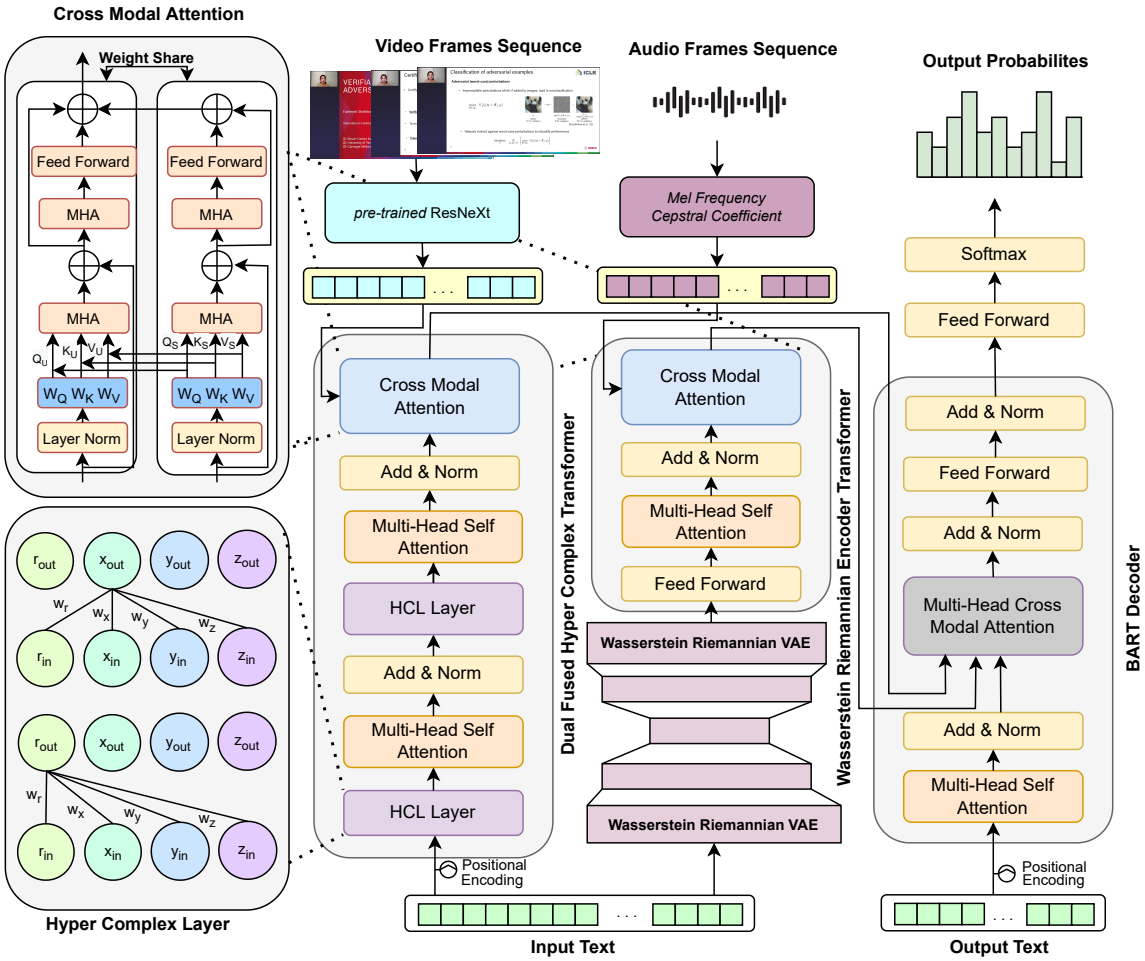


Figure 6.2: An overview of the proposed model – `mTLDRgen`. It houses two parallel encoders, one with a hyper-complex layer fused with the video embeddings using cross-model attention and the other with Wasserstein Riemannian Encoder Transformer with audio embeddings fused with cross-model attention. The individual encoder representations are later fused with the multi-head attention of the pre-trained BART decoder to generate the final summary.

are manually annotated and transformed into a text feature set for training. In contrast, the acoustic features for `mTLDRgen` are not transformed into the text as they are characterized by a variety of non-native English accents and a high error rate for speech-to-text models. Both the textual representations are tokenized using the vanilla BART tokenizer and transformed to word vectors using standard Transformer positional encoding embeddings.

6.4.4 Dual-fused Hyper-complex Transformer

We propose a dual-fused hyper-complex Transformer (DFHC) for the task of multimodal text summarization. Compared to multi-head attention, the hyper-complex layer allows `mTLDRgen` to

efficiently capture the intricacies between different modalities and learn better representations [188] in the hyper-complex space. For the block DFHC, we represent the text as X and pass it through the hyper-complex layer to extract the (Q) Query, (K) Key and (V) Value transformations as follows: $Q, K, V = \Phi(\text{HCL}(X))$, where $HCL(X) = Hx + b$. Here $H \in \mathbb{R}^{m \times n}$ is constructed by a sum of Kronecker products and is given by $H = \sum_{i=1}^{n} P_i \otimes Q_i$. The P_i and Q_i are the parameter matrices, and $\otimes$ represents the Kronecker product.

The final attention score A is obtained as:

$$A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

Here Q represents the query value, K represents the key value, and d_k represents the key dimension.

The HCL layers share attention weights among the multi-head attention heads. The multi-head attention weights are concatenated and represented as

$$X = \text{HCL}([H_1 + ... + H_{Num_h}])$$

Here Num_h represents the attention head. The final output obtained from the HCL layer is represented as:

$$Y = \text{HCL}(\text{ReLU}(\text{HCL}(X))),$$

The transformation Y is passed through a multi-head attention block and is fused with the visual embeddings using the cross-model attention as discussed in Section 6.4.6.

6.4.5 Wasserstein Riemannian Encoder Transformer

We base our idea from [189] to repurpose the Wasserstein Riemannian Autoencoder to Wasserstein Riemannian Encoder Transformer (WRET) in the summarization setting.

For an input X and a manifold M, the Riemannian manifold is represented as (M, G), where G represents the Riemannian tensor unit. For two vectors u and v in the tangent space $T_z M$, the inner product is computed using $\langle u, v \rangle_G = u^T G(z) v$. The Wasserstein block acts as a Variational Autoencoder. However, we extract the feature dimension from the last layer and feed it to the attention block.

The Wasserstein Autoencoder optimizes the cost between the target data distribution $A_x(x)$ and the predicted data distribution $B_x(x)$ using:

$$Dist(A_X, B_G) = \inf_{Q(Z|X) \in Q} E_{P_X} E_{Q(Z|X)}[c(X, G(Z))] + \lambda MMD(Q_Z, P_Z)$$

where G is the generator function, λ is a learnable metric, c is the optimal cost, and D_z is approximated between B_G and $Q_Z(z) = \int q(z|x)p(x)dx$ using the Maximum Mean Discrepancy (MMD) [190]. The MMD is computed using

$$MMD_k(P_Z, Q_Z) = || \int_{\mathcal{Z}} k(z, \cdot) dP_Z - \int_{\mathcal{Z}} k(z, \cdot) dQ_Z ||$$

We formulate a RNF function $F = f_K \ldots f_1$, and optimize the following RNF-Wasserstein function,

$$\begin{aligned} Dist(A_X, B_G) = \inf_{Q(Z|X) \in \mathcal{Q}} & P_X Q(Z|X)[c(X, G(Z'))] \\ & + \lambda MMD(Q_{Z'}, P_{Z'}) \\ & + \alpha(KLD(q(z|x)||p(z)) - \sum \log |det \frac{\partial f'}{\partial z}|) \end{aligned}$$

where $Z' = F(Z)$, KLD is KL [191] divergence, and $p(z)$ represents the posterior probability distribution. The MMD term is approximated using the Gaussian kernel $k(z, z') = e^{-||z-z'||^2}$. The term $G(Z')$ represents the reconstructed feature set, which is then passed to a feed-forward layer. The attention weights are computed for $G(Z')$ and fused with the audio feature using cross-modal attention as discussed in Section 6.4.6.

6.4.6 Cross-model Attention

We fuse the text-video and text-audio features using cross-modal attention to align the attention distribution obtained from the last layer. The text feature set projects the Query (Q) value, while the video and audio features project the key (K) and value (V), respectively. The obtained Q, K, and V representations are passed through cross-modal attention, and the final encoder representation E_s is obtained.

$$\begin{aligned} Q = Z_t W_q;\ K = Z_v W_k;\ V = Z_v W_v \\ E_s = \text{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta} \end{aligned}$$

The amalgamated representation, represented as E_s, is subsequently integrated with the multi-head attention mechanism of the BART decoder to yield the final summary.

To the best of our knowledge, the application of hyper-complex Transformer and Wasserstein Riemannian flow for abstractive text summarization has not been explored yet.

6.5 Experiments

We perform extensive ablations to evaluate the efficacy of the proposed modules of `mTLDRgen` and individual modalities. We explore how text, visual and acoustic features perform separately and jointly over `mTLDR` and How2.

Table 6.2: Performance benchmark over six text-only Extractive (Extr) baselines (Lead-2, Lexrank, TextRank, MMR, ICSISumm, and BERTExtractive), eight text-only Abstrative (Abst) baselines (Seq2Seq, Pointer Generator (PG), CopyTransformer, Longformer, BERT, BART, T5, and Pegasus), two video-only baselines (Action feature, and Action feature with RNN), and four Multimodal baselines (HA, FLORAL, MFFG, and ICAF) over the datasets – `mTLDRgen` and How2. The benchmarks are evaluated over the Quantitative metric – Rouge (Rouge-1 (R1), Rouge-2 (R2), and Rouge-L (RL)), and Qualitative metric – BERTScore (BERTSc.) and FEQA.

Type	System	mTLDR					How2				
		R1	R2	RL	BERTSc.	FEQA	R1	R2	RL	BERTSc.	FEQA
Extr-text	Lead-2	22.82	4.61	15.47	61.27	32.45	43.96	13.31	39.28	71.56	32.28
	LexRank	27.18	6.82	17.22	63.23	34.21	27.93	12.88	16.93	64.52	31.89
	TextRank	27.43	6.86	17.41	63.34	34.29	27.49	12.61	16.71	64.55	31.92
	MMR	29.54	8.19	18.84	64.59	35.67	28.24	13.12	17.86	64.87	31.98
	ICSISumm	31.57	9.52	19.42	65.84	36.14	28.53	13.44	17.93	65.14	32.16
	BERTExtractive	31.52	9.49	19.31	65.83	36.13	27.18	12.47	15.38	63.47	31.67
Abst-text	Seq2Seq	23.54	5.61	15.48	62.47	31.57	55.37	23.08	53.86	76.15	36.48
	PG	23.59	5.78	16.21	62.71	31.84	51.68	22.63	50.29	73.47	35.37
	CopyTransformer	25.63	7.82	18.54	63.11	37.86	52.94	23.25	50.26	73.58	35.43
	Longformer	21.37	6.47	15.12	61.05	32.14	49.24	21.39	47.41	72.39	35.28
	BERT	24.87	8.85	18.33	62.91	31.89	53.74	23.86	48.06	73.45	35.62
	BART	26.13	9.69	19.62	64.24	38.64	53.81	23.89	48.15	73.51	35.68
	T5	25.87	9.24	18.63	64.13	38.45	53.21	22.51	47.48	73.42	35.65
	Pegasus	26.66	9.83	19.26	64.85	36.98	53.87	23.91	48.17	73.61	35.70
Video only	Act. feat.	26.38	6.47	15.37	62.48	30.41	45.24	24.42	38.47	69.74	31.28
	RNN (Act. feat.)	26.73	6.51	15.75	63.14	31.35	48.27	27.74	46.37	72.32	35.11
Multimodal	HA	29.32	11.84	26.18	67.24	39.37	55.82	38.31	54.96	77.15	38.55
	FLORAL	31.69	13.54	31.55	69.56	41.19	56.84	39.86	56.93	79.84	39.14
	MFFG	33.19	18.88	33.28	71.54	43.13	61.49	44.61	57.21	80.16	41.59
	ICAF	*36.38*	*20.54*	*34.52*	*73.94*	*45.63*	*63.84*	*44.78*	*58.24*	*82.39*	*42.86*
	`mTLDRgen`	**41.62**	**22.69**	**37.87**	**78.39**	**48.46**	**67.33**	**48.71**	**61.83**	**84.11**	**44.82**
$\Delta_{\texttt{mTLDRgen}}$	BEST	↑ 5.24	↑ 2.15	↑ 3.35	↑ 4.45	↑ 2.83	↑ 3.49	↑ 3.93	↑ 3.59	↑ 1.72	↑ 1.96

Table 6.3: Ablation study to show the efficacy of each module of `mTLDRgen`.

System	mTLDR				How2			
	Rouge-1	Rouge-2	Rouge-L	BERTScore	Rouge-1	Rouge-2	Rouge-L	BERTScore
Transformer	25.63	7.82	18.54	63.11	52.94	23.25	50.26	73.58
+ DFHC	29.37	11.78	23.19	67.81	57.34	28.71	56.02	77.31
+ WRET	34.52	14.82	26.54	72.06	61.12	36.89	58.1	81.44
+ DFHC & WRET	37.34	18.32	32.49	74.58	64.23	42.61	59.02	82.45
`mTLDRgen`	41.62	22.69	37.87	78.39	67.33	48.71	61.83	84.11

Evaluation measures: We evaluate the performance of `mTLDRgen` using both quantitative metrics - Rouge-1, Rouge-2, and Rouge-L and qualitative metrics – BERTScore and FEQA. Rouge measures the recall of unigrams (Rouge-1), bigrams (Rouge-2), and n-grams (Rouge-L) between the generated and target summaries. BERTScore assesses the similarity between the generated and target summaries in the embedding space through the cosine similarity of the BERT embeddings. FEQA, a question-answer generation metric, evaluates the quality of the generated summaries by determining the number of answers mapped to questions generated from the target summaries.

We further perform human evaluations[4] In the evaluations, we benchmark the summaries over

[4]The human evaluations were performed by 15 individuals with sufficient background in NLP, machine learning

the following parameters — Informativeness, Fluency, Coherence and Relevance. We randomly sample 40 instances from the test set and evaluate them against the target summaries. We perform human evaluations for BART (text-only), T5 (text-only), MFFG (multimodal), FLORAL (multimodal) and `mTLDRgen` (multimodal).

6.5.1 Training

The implementation of the `mTLDRgen` model was carried out by utilizing the Pytorch 1.8.1 framework on an NVIDIA A6000 GPU equipped with 46 GB of dedicated memory and CUDA-11 and cuDNN-7 libraries. The model was initialized with pre-trained BART LM weights for the encoder and decoder and fine-tuned on the summarization dataset. In the implementation, the loss computation was only performed over the target sequence in adherence to the encoder-decoder paradigm.

Table 6.4: Comparison of target summary with six models – Extractive (ICSISumm), Abstractive (Pointer Generator (PG), BART, Pegasus) and multimodal (ICAF and `mTLDRgen`) models.

Model	Output
Target	In this paper, we propose an adversarial multi-task learning framework, where the shared and private latent feature spaces donot interfere with each other. This task is achieved by introducing orthogonality constraints.
ICSISumm	To prevent the shared and private latent feature spaces from interfering with each other, we introduce two strategies: adversarial training and orthogonality constraints.
PG	propose multi-task learning for the generative , propose latent feature for multi task learning where the shared knowledge regarded as off the self knowledge and trasferred to new task.
BART	In this paper, we conduct experiment on 16 tasks demonstrate the benefits and propose multi-task learning framework, The dataset are shared and latent feature spaces. the dataset is prone.
Pegasus	In this paper, we propose an multitask learning framework, where we conduct experiments on 16 text classification tasks. our model is off the shelf and donot interfere with each other.
ICAF	we propose an adversarial multi-task framework, where we conduct experiments demostrating private feature space do not interefere with eachother. The model is regarded as off the shelf and transferred to new task.
`mTLDRgen`	In this paper, we propose an multi-task framework, the shared and private latent feature spaces not interfere with each other. We conduct experiments on 16 text classification tasks.

6.5.2 Quantitative Analysis

Table 6.2 compares the performance of `mTLDRgen` with that of its baselines across the How2 and `mTLDR` datasets. Our results demonstrate that `mTLDRgen` outperforms the best baseline, ICAF, with a Rouge-1 score of 41.62 and a Rouge-L score of 37.87, an improvement of +5.24 and +3.35, respectively. When benchmarked against the How2 dataset, `mTLDRgen` exhibits superior results, obtaining Rouge-1 of 67.33 and Rouge-L of 61.83, outperforming the best baseline (ICAF) by +3.49 and +3.59 points, respectively. With respect to the best text-only abstractive baseline, Pegasus (Rouge-1 and Rouge-2) and BART (Rouge-L), `mTLDRgen` shows an improvement of +14.96 Rouge-1, +12.86 Rouge-2, and +18.25 Rouge-L. Similarly, `mTLDRgen` surpasses ICSISumm, the best extractive

and deep learning. The participants were aged between 22-28 years.

baseline, with an improvement of +10.05, +13.17, and +18.45 on Rouge-1, Rouge-2 and Rouge-L, respectively.

We also perform ablations to study the efficacy of individual modalities and modules of `mTLDRgen`. Table 6.5 demonstrates the performance improvements obtained when all three modalities are fused, while Table 6.3 showcases the contribution of individual modules of `mTLDRgen`. These results serve to affirm our hypothesis that models specifically designed for longer summarization sequences are inadequate in extreme summarization tasks and that the integration of multiple modalities with text modality enhances the quality of the generated summary.

Congruency of multi-modalities The performance of various unimodal and multimodal text summarization systems is shown in Table 6.2. The results demonstrate that for unimodal variants, the lead2, which was reported to be a strong baseline for datasets like CNN/Dailymail [20] and MultiNews [12], fails to perform effectively, indicating that the latent structure of the scientific text is distinct, and the information has a heterogeneous distribution throughout the document. In a similar vein, the text-only abstractive baselines perform inadequately over both the How2 and `mTLDR` datasets. On the other hand, the extractive baselines, which are able to identify the prominent sentences that start with "we propose" or "we introduce", perform better than the text-only abstractive baselines yet still provide only a limited context of the whole article. Meanwhile, the two video-only baselines display performance that is comparable to the best abstractive baselines, signifying that multimodal features do indeed contribute to generating more informative and coherent summaries. No baselines using audio-only features were run as audio captures only the amplitude shift and intonations, which do not constitute a sufficient feature set in the vector space. As indicated in Table 6.2, the multimodal baselines outperform the text-only and video-only baselines. The HA model outperforms the best abstractive baseline by +2.66 Rouge-1 and +6.92 Rouge-L, demonstrating the significance of combining multimodal signals with text-only modalities. The fusion of video with text helps the model attain better context in the vector space, even the audio features aid in the mutual alignment of modalities leading to more diverse and coherent summaries. Evidently, all the remaining multimodal baselines show a remarkable improvement in performance over all the text-only extractive, text-only abstractive, and video-based baselines.

Table 6.5: Performance benchmark for each modality of `mTLDRgen`.

Modality	Rouge-1	Rouge-2	Rouge-L
Text +Audio	27.46	7.47	19.62
Audio +Video	27.62	7.53	20.11
Text +Video	28.05	7.83	24.49
Text +Audio +Video	41.62	22.69	37.87

6.5.3 Qualitative Analysis

We also conduct a qualitative evaluation of the generated summaries utilizing BERTScore and FEQA (c.f. Table 6.2). Both metrics use the text modality from the source and target to assess the quality. On the `mTLDR` data, `mTLDRgen` achieves 78.39 BERTScore and 48.46 FEQA, surpassing the

Table 6.6: Human evaluation scores over the metrics – Informativeness (Infor.), Fluency, Coherence, and Relevance for the text-based baselines (BART and T5), multimodal baselines (MFFG, FLORAL, and mTLDRgen) on the mTLDRgen and How datasets.

Modality	System	mTLDR				How2			
		Infor.	Fluency	Coherence	Relevance	Infor.	Fluency	Coherence	Relevance
Abstractive-text	BART	2.81	2.51	2.94	2.85	2.34	2.37	2.46	2.54
Abstractive-text	T5	2.78	2.49	2.81	2.74	2.33	2.28	2.43	2.54
Multimodal	FLORAL	3.2	3.03	3.02	3.11	3.13	3.14	3.08	3.13
Multimodal	MFFG	3.21	3.05	3.09	3.11	3.17	3.21	3.04	3.11
Multimodal	mTLDRgen	**3.46**	**3.32**	**3.27**	**3.29**	**3.34**	**3.27**	**3.21**	**3.18**

best baseline (ICAF) by +4.45 and +2.83 points, respectively. For the How2 dataset, mTLDRgen obtains 84.11 BERTScore and 44.82 FEQA, outperforming the best baseline (ICAF) by +1.72 and +1.96 points, respectively. Similar to the quantitative benchmarks, the multimodal baselines outperform the text-only extractive, abstractive, and video-only baselines by a substantial margin.

6.5.4 Human Evaluation

The quantitative enhancements are further reinforced by human assessments. As shown in Table 6.6, mTLDRgen scores highest over the datasets – mTLDR and How2, demonstrating that the generated summaries are highly faithful, pertinent, and coherent in comparison to the other baselines. Although mTLDRgen demonstrates some deficiencies in the coherence criterion in the human evaluations, it still performs significantly better than the other baselines. A manual examination of the generated summaries and an analysis of the findings are presented in Section 8.7.

6.5.5 Error Analysis

The limitations of extractive and abstractive baselines in generating extreme summaries are evident. Extractive systems rely on direct copying of phrases from the source document, often resulting in a single-line summary containing limited information diversity. This is reflected in their performance compared to abstractive text-only and a few multimodal (HA and FLORAL) baselines, as seen in Table 6.4 and Table 6.6. The text-only abstractive baselines like Seq2Seq and PG fail to extract the paper's main contributions, while Transformer based methods like Longformer, BERT, etc., struggle to summarize the contributions in very few lines.

However, mTLDRgen stands out as it is able to condense the three key contributions of the source article into a single sentence, demonstrating superiority over the other baselines. A manual inspection of instances where mTLDRgen failed to generate a good summary reveals that the cause was often due to noisy text modality extracted from the article pdf, leading to non-coherent connections between phrases. Further, the data noise in the video and audio modalities arising due to different aspect ratios of presentation and speaker and non-native English accent speakers adds to the perplexity of modality alignment.

6.6 Conclusion

We introduced a novel task of extreme abstractive text summarization using multimodal inputs. we curated `mTLDR`, a unique large-scale dataset for extreme abstractive text summarization that encompasses videos, audio, and text, as well as both author-written and expert-annotated summaries. Subsequently, we introduced `mTLDRgen`, a novel model that employs a dual fused hyper-complex Transformer and a Wasserstein Riemannian Encoder Transformer to efficiently capture the relationships between different modalities in a hyper-complex and latent geometric space. The hyper-complex Transformer captures the intrinsic properties between the modalities, while the Wasserstein Riemannian Encoder Transformer captures the latent structures of the modalities in the latent space geometry, enabling the model to generate diverse sentences. To assess the `mTLDRgen`, we conducted thorough experiments on `mTLDR` and How2 datasets and compared their performance with 20 baselines. Overall, `mTLDRgen` demonstrated superior performance both qualitatively and quantitatively.

Chapter 7

Fingerprinting Corpus Bias and its Effects on Systems

Multi-document summarization (MDS) aims to condense key points from diverse documents into a concise text paragraph, facilitating aggregation of news, tweets, and product reviews from various sources. However, the absence of a standard definition for MDS leads to datasets with varying overlap and conflict among documents, complicating evaluation. Moreover, the lack of standard criteria for defining summary information further adds to the challenge. New systems often report results on specific datasets, potentially lacking correlation with their performance on other datasets. In this work, we address these challenges by studying the heterogeneous nature of MDS using widely-used datasets and state-of-the-art models. We quantify the quality of summarization corpora and provide guidelines for creating new MDS datasets. Additionally, we analyze reasons for the absence of an MDS system achieving superior performance across all datasets. We also assess the quality of training samples and evaluate the effectiveness of data sampling by benchmarking various language models.

7.1 Introduction

Multi-document summarization (MDS) deals with compressing more than one document into a textual summary. It has a wide range of applications – gaining insights from tweets related to similar hashtags, understanding product features amongst e-commerce reviews, summarizing live blogs related to an ongoing match, etc. Most studies on MDS were performed during the DUC[1] and TAC[2] challenges starting in the early 2000s. Each version of the challenges released a new dataset. Most of the MDS systems submitted to these challenges were unsupervised and extractive in nature. Gradually, the data released in these challenges became the *de facto* for MDS. These datasets were manually curated and had less than a hundred instances each. The recent development of deep neural architecture has led to a significant increase in the number of supervised document summarization systems. Large labeled datasets which are mostly crowd-sourced have

[1]https://duc.nist.gov/
[2]http://tac.nist.gov

been introduced to meet the training requirements of the supervised systems. However, the crowd-sourced datasets widely differ in quality based on factors like genre, size of the community, presence of moderation in the community, etc. This is further aggravated by the complexity of the task, the hardness of accumulating labeled data, or more so in the definition of what constitutes a multi-document summary.

Recently, several large datasets for multi-document summarization (MDS) have been introduced [12, 192]. However, there has been a lack of studies measuring the relative complexity or quality of samples within these datasets. We observe notable variations in the behavior of existing MDS systems across different datasets, with systems achieving state-of-the-art performance on one corpus often failing to perform adequately on another. Although ROUGE scores of MDS systems are on the rise, manual inspection reveals an increasing presence of bias in generated summaries. Moreover, new systems are introduced and evaluated on a limited number of datasets, making it challenging to discern whether bias originates from the system itself or is inherent in the training corpus. Additionally, limited research exists on identifying the importance of individual training samples within the corpus. The target summary plays a crucial role in effectively training deep learning systems to accurately extract and map essential sections. Unlike tasks such as machine translation or dialogue generation, where systems often focus on local context windows, summarization systems require a broader context, both within and across documents. Therefore, to facilitate diverse information learning, summarization systems must portray information spread in the vector space, showcasing deep learning systems' ability to learn various representations without fixating on local context windows or disregarding the input document's central theme.

We divide our investigation into two main components:

1. Evaluating the quality of datasets and their impact on systems.
2. Examining representation bias and analyzing the implications of training dynamics on dataset samples.

Initially, we focus on addressing the first task by thoroughly investigating the quality of datasets and their implications on system performance in Section 7.2. Subsequently, in Section 7.4, we delve into the second task, which involves an in-depth discussion on representation bias and an analysis of the effects of training dynamics on dataset samples.

7.2 Evaluating Dataset Quality and their Impact on Systems.

To evaluate the datasets and systems, we examine five MDS datasets – DUC [193], TAC [194], Opinosis [195], Multinews [12], and CQASumm [13]. We consider eight popular summarization systems – LexRank [117], TextRank [118], MMR [142], ICSISumm [143], PG [38], PG-MMR [40], Hi-Map [12], and CopyTransformer [15].

Our major contributions for this section are four-fold:

• We propose a suite of metrics to model the quality of an MDS corpus in terms of – Abstractness, Inter Document Similarity (IDS), Redundancy, Pyramid Score, Layout Bias and Inverse-Pyramid Score.
• We develop an interactive web portal for imminent datasets to be uploaded and evaluated based on our proposed metrics.
• We explore different biases that the MDS systems exhibit over different datasets and provide insight into properties that a universal MDS system should display to achieve reasonable performance on all types of dataset.
• We look into metrics to capture bias shown by MDS systems and explore the extent to which corpus properties influence them.

7.3 Related Work

7.3.1 Background and Proposed Metrics

Throughout this chapter, we use the term **candidate documents** for the documents participating in summarization, and the term **reference** to indicate the ground-truth summary.
Oracle summary is the extractive set of sentences selected from the candidate documents, exhibiting maximum ROUGE-N score w.r.t. the reference summary. It is an NP-hard problem [196], and approximate solutions can be found greedily or using ILP solvers.

Here, we briefly introduce a suite of corpus and system metrics proposed by us to better understand the MDS task.

Corpus Metrics

• **Abstractness**: It is defined as the percentage of non-overlapping *higher order n-grams* between the reference summary and candidate documents. A high score highlights the presence of more distinctive phrases in reference summary. The intuition behind quantifying the number of new words is to sync with the basic human nature of paraphrasing while summarizing.
• **Inter Document Similarity (IDS)**: It is an indicator of the degree of overlap between candidate documents. Inspired by the theoretical model of relevance [6], we calculate IDS of a set of documents as follows:

$$IDS(D_i) = \frac{\sum_{D_j \in S} Relevance(D_j, D_i)}{|S|} \tag{7.1}$$

where D_i is the i^{th} candidate document, and S is the set of all documents other than D_i. Here, $Relevance(.,.)$ is defined as:

$$Relevance(A, B) = \sum_{\omega_i} P_A(\omega_i).\log(P_B(\omega_i)) \tag{7.2}$$

where $P_A(\omega_i)$ represents the probability distribution of the i^{th} semantic unit[3] in document A. The further this score is from 0, the lesser inter document overlap there is in terms of semantic unit

[3]An atomic piece of information

distribution. As shown in Equation 7.1, the numerator calculates relevance which can be interpreted as the average surprise of observing one distribution while expecting another. This score is small if the distributions are similar i.e., $P_A \approx P_B$ from Equation 7.2.

• **Pyramid Score**: We propose the metric Corpus Pyramid score to measure how well important information across documents is represented in the ground truth. As introduced by [197], Pyramid score is a metric to evaluate system summaries w.r.t. the pool of ground-truth summaries. We instead use this metric to quantitatively analyze the ground-truth summary w.r.t. candidate documents. The entire information set is split into *Summarization Content Units* (SCUs[4]), and each SCU is assigned a weight based on the number of times it occurs in the text. A pyramid of SCUs is constructed with an SCU's weight, denoting its level, and a score is assigned to a text, based on the number of SCUs it contains. Pyramid score is defined as the ratio of a reference summary score and an optimal summary score. Higher values indicate that the reference summary covers the SCUs at the top of the pyramid better. SCUs present at the top are the ones occurring in most articles and thus can be deemed as important.
• **Redundancy**: The *amount of information* in a text can be measured as the negative of Shannon's entropy (H) [6].

$$H(D) = -\sum_{\omega_i} P_D(\omega_i).\log(P_D(\omega_i)) \tag{7.3}$$

where P_D represents the probability distribution of documents D, and ω_i represents the i^{th} semantic unit[3] in the distribution. $H(D)$ would be maximized for a uniform probability distribution when each semantic unit is present only once. The farther this score is from 0, the better a document is distributed over its semantic units in the distribution, hence lesser the redundancy. As evident from Equation 7.5, redundancy is maximized if all semantics units have equal distribution i.e., $P(\omega_i) = P(\omega_j)$. The idea behind using redundancy is to quantify how well individual documents cover sub-topics, which might not be the core content but important nonetheless. Thus

$$Redundancy(D) = H_{max} - H(D) \tag{7.4}$$

Since H_{max} is constant, we obtain

$$Redundancy(D) = \sum_{\omega_i} P_D(\omega_i).\log(P_D(\omega_i)) \tag{7.5}$$

• **Layout Bias**: We define *Layout Bias* across a document as the degree of change in importance w.r.t. the ground-truth over the course of candidate documents. We divide the document into k segments, calculate the importance of each segment w.r.t. the ground-truth by a similarity score, and average over the sentences in the segment. Positional importance of D_j, the j^{th} sentence in the document is denoted by:

$$PositionalImportance(D_j) = \max_{1 \le i \le n} sim(\overrightarrow{D_j}, \overrightarrow{R_i}) \tag{7.6}$$

where, $\overrightarrow{R_i}$ is the vector representation of the i^{th} sentence in the reference, sim is a similarity metric between two sentences, and n is the total number of sentences in the reference summary.

[4]They are subsentential units based on semantic meaning

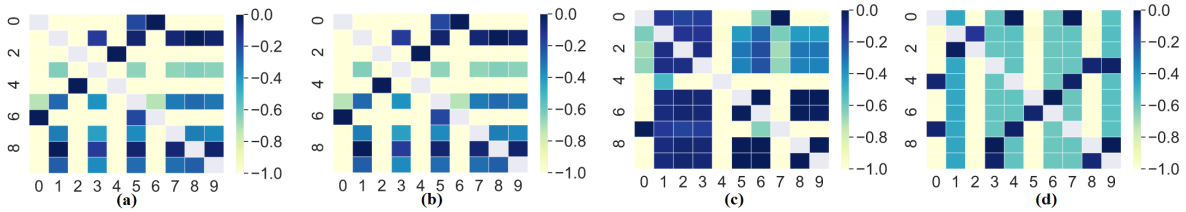


Figure 7.1: Heatmap depicting the corpus metric: Inter document similarity. We explain with a single instance of (a) DUC-2004, (b) DUC-2003, (c) TAC-2008, and (d) CQASumm, and highlight inter-document overlap.

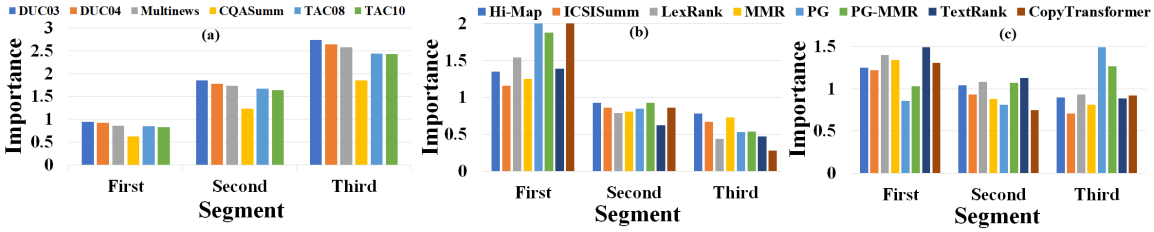


Figure 7.2: (a) Layout Bias across datasets, highlighting cumulative cosine similarity (importance) values (y-axis) between segments (first, second and third) of candidate documents and the reference summary. (b) Change in layout importance across systems over source segments when divided in three uniform segments. (c) Change in layout importance across systems when candidate documents are internally shuffled and divided into three uniform segments.

A lower shift indicates that while generating reference summaries, all segments have been given similar importance within any 3-fold segmented article.

• **Inverse-Pyramid Score (Inv Pyr)**: We propose Inverse-Pyramid score to quantify the bias that a reference summary exhibits w.r.t. its set of candidate documents. It measures the importance given to each document in the candidate set by the reference summary as:

$$InvPyr(D, S) = Var_j\big(D_j \cap S_u\big) \tag{7.7}$$

Here, D and S are the set of candidate documents for MDS and their summary respectively, Var is the variance, D_j and S_u are the sets of SCUs[4] in the j^{th} document of the candidate set and the reference summary respectively.

Higher Inv Pyr scores suggest the difference in importance given to each document while generating the summary is higher. As evident from Equation 7.7, Variance across the similarities is high if the similarity scores across the document-summary pairs are uneven.

System Metrics

• **ROUGE:** is a metric which computes the n-gram overlap recall value for the generated summary w.r.t. the reference summary.

• **F1 Score with Oracle Summaries**: Oracle summaries reflect the extractive selection of sentences that achieve the highest ROUGE score over the candidate documents given a reference

summary. Similar to ROUGE-1, this metric also combines both precision and recall between the oracle and system summaries to calculate F1 Score. It is a better indicator of the presence of non-essential n-grams than ROUGE as it also takes precision into account.

• **System Abstractness**: Analogous to corpus abstractness, we compute the percentage of novel higher order n-grams in the generated summary w.r.t. the candidate documents. System abstractness is calculated using

$$Coverage(D, S) = \frac{\sum_{i \in 1..n}(D \cap S_i)}{C_n(S)}$$

where D represents the set of n-grams in the candidate document, and S represents the set of n-grams in the i^{th} system summary.
The denominator denotes the total count of n-grams in a system summary. Finally, the values of all articles is normalized to get the score for the system

• **Layout Bias**: We propose this metric to capture which sections of the candidate documents comprise a majority of the information in the generated summary. For neural abstractive systems, we concatenate candidate documents to form one large document and feed it to the neural model. We study two variations of this metric – The first variation involves segmenting this large document into k parts and then computing the similarity of n-gram tokens of system summaries w.r.t. the candidate document segment. The second variation includes shuffling the candidate documents before concatenating and then computing the n-gram similarity with the generated summary.

• **Inter Document Distribution (IDD)**: We propose this metric to quantify the extent of contribution of each candidate document to form the generated summary. The relevance for system summaries is calculated by,

$$Relevance(A, B) = \sum_{\omega_i} P_A(\omega_i).\log(P_B(\omega_i))$$

where P_A represents the probability distribution of system summary S, and ω_i represents the i^{th} semantic unit in the distribution.

$$IDD(D_i) = \frac{\sum_{D_j \in S} Relevance(D_j, D_i)}{Cardinality(S)}$$

• **Redundancy**: It measures the degree to which system summaries can cover the distribution across semantic units generated from the candidate documents. Redundancy for candidate documents is given by Eq.,

$$Redundancy(D) = \sum_{\omega_i} S_D(\omega_i).\log(S_D(\omega_i))$$

where S_D represents the probability distribution of a system summary D. ω_i represents the i^{th} semantic unit in the distribution.

7.3.2 Experimental Setup

MDS Datasets

• **DUC** [193] is a news dataset built using newswire/paper documents. The 2003 (DUC-2003) and 2004 (DUC-2004) versions comprise 30 and 50 topics respectively with each topic having 4 manually curated reference summaries.
• **TAC** [194] is built from the AQUIANT-2 collection of newswire articles where NIST assessors select 48 and 44 topics for the 2008 and 2010 versions, respectively. Each topic consists of 4 summaries.
• **Opinosis** [195] is an accumulation of user reviews collected from various sources like TripAdvisor, Edmunds.com and Amazon. There are 51 topics, with each topic having approximately 4 human-written summaries.
• **CQASumm** [13] is a community question answering dataset, consisting of $100,000$ threads from the Yahoo! Answers L6 dataset. It treats each answer under a thread as a separate document and the best answer as a reference summary.
• **Multinews** [12] is a news dataset comprising news articles and human-written summaries from newser.com. It has $56,216$ topics, with summaries of 260 words on average written by professional editors.

MDS Systems

To identify bias in system-generated summaries, we study a few non-neural extractive and neural abstractive summarization systems, which are extensively used for multi-document summarization.
• **LexRank** [117] is a graph based algorithm that computes the importance of a sentence using the concept of eigen vector centrality in a graphical representation of text.
• **TextRank** [118] runs a modified version of PageRank [145] on a weighted graph, consisting of nodes as sentences and edges as similarities between sentences.
• **Maximal Marginal Relevance (MMR)** [142] is an extractive summarization system that ranks sentences based on higher relevance while considering the novelty of the sentence to reduce redundancy.
• **ICSISumm** [143] optimizes the summary coverage by adopting a linear optimization framework. It finds a globally optimal summary using the most important concepts covered in the document.
• **Pointer Generator (PG)** network [38] is a sequence-to-sequence summarization model which allows both copying words from the source by pointing or generating words from a fixed vocabulary.
• **Pointer Generator-MMR:** PG-MMR [40] uses MMR along with PG for better coverage and redundancy mitigation.
• **Hi-Map:** Hierarchical MMR-Attention PG model [12] extends the work of PG and MMR. MMR scores are calculated at word level and incorporated in the attention weights for a better summary generation.
• **Bottom-up Abstractive Summarization (CopyTransformer)** [15] uses transformer parameters proposed by [45]; but one of the attention heads chosen randomly acts as a copy distribution.

Dataset	Metric						
	Abstractness			Red	IDS	Pyr	Inv
	1-gram	2-gram	3-gram				
DUC	11.5	54.66	79.29	-0.21	-6.6	0.35	2.64
Opinosis	11.5	50.36	76.31	-0.02	-5.53	0.26	2.8
Multinews	32.28	67.53	80.45	-0.8	-1.03	0.4	3.8
CQASumm	41.41	80.72	88.79	-0.22	-9.16	0.05	5.2
TAC	9.91	50.26	76.17	-0.19	-4.43	0.32	2.9

Table 7.1: Values of corpus metrics: Abstractness, Redundancy (Red), Inter Document Similarity (IDS), Pyramid Score (Pyr) and Inverse-Pyramid Score (Inv).

7.3.3 Inferences from Corpus Metrics

• News derived datasets show a strong layout bias where significant reference information is contained in the introductory sentences of the candidate documents (Fig. 7.2).
• Different MDS datasets vary in compression factors with DUC at 56.55, TAC at 54.68, Multinews at 8.18 and CQASumm at 5.65. A high compression score indicates an attempt to pack candidate documents to a shorter reference summary.
• There has been a shift in the size and abstractness of reference summaries in MDS datasets over time – while DUC and TAC were small in size and mostly extractive (11% novel unigrams); crowd-sourced datasets like CQASumm are large enough to train neural models and highly abstractive (41.4% novel unigrams).
• Candidate documents in Opinosis, TAC and DUC feature a high degree of redundant information as compared to Opinosis and CQASumm, with instances of the former revolving around a single key entity while that of the latter tending to show more topical versatility.
• MDS datasets present a variation in inter-document content overlap as well: while Multinews shows the highest degree of overlap, CQASumm shows the least and the rest of the datasets show moderate overlap (see Fig. 7.1).
• Pyramid Score, the metric which evaluates if the important and redundant SCUs[4] from the candidate documents have been elected to be part of the reference summary, shows considerably positive values for DUC, TAC and Multinews as compared to crowdsourced datasets like CQASumm (Fig. 7.3.b).
• Inverse-Pyramid Score, the metric which evaluates how well SCUs[4] of the reference summary are distributed amongst candidate documents, also shows better performance on human-annotated datasets compared to crowd-sourced ones (Fig. 7.3(b)).
• A comparison amongst corpus metrics presents a strong positive correlation between IDS and Pyramid Score (Pearson's $\rho = 0.8296$) and a strong negative correlation between the metrics of Redundancy and IDS (Pearson's $\rho = -0.8454$).

7.3.4 Inferences from System Metrics

• MDS systems under consideration are ranked differently in terms of ROUGE on different datasets; leading to a dilemma whether to declare a system superior to others without testing on all types of

System	Metric	Dataset				
		DUC	TAC	Opinosis	Multinews	CQASumm
LexRank	R1	35.56	33.1	33.41	38.27	32.22
	R2	7.87	7.5	9.61	12.7	5.84
	F1	31.34	31.51	31.05	41.01	49.71
	Red.	-0.136	-0.104	-0.278	-0.29	-0.364
	IDD	-3.377	-1.87	-3.526	-2.53	-2.17
	IDDV	0.239	1.62	0.221	0.242	1.232
TextRank	R1	33.16	44.98	26.97	38.44	28.94
	R2	6.13	9.28	6.99	13.1	5.65
	F1	40.8	29.69	31	38.44	46.3
	Red.	-0.25	-1.553	-0.342	-0.208	-0.247
	IDD	-0.196	-5.97	-2.745	-1.879	-2.137
	IDDV	0.799	1.48	0.025	0.146	0.744
MMR	R1	30.14	30.54	30.24	38.77	29.33
	R2	4.55	4.04	7.67	11.98	4.99
	F1	30.57	28.3	31.8	42.07	45.48
	Red.	-0.266	-0.068	-0.255	-0.17	-0.288
	IDD	-2.689	-2.135	-3.213	-1.83	-2.059
	IDDV	1.873	0.231	0.222	0.157	0.126
ICSI -Summ	R1	37.31	28.09	27.63	37.2	28.99
	R2	9.36	3.78	5.32	13.5	4.24
	F1	24.27	27.82	29.83	44.71	50.98
	Red.	-0.327	-0.283	-0.328	-0.31	-0.269
	IDD	-3.357	-1.903	-3.244	-3.14	-2.466
	IDDV	0.694	0.403	1.134	0.239	0.242
PG	R1	31.43	31.44	19.65	41.85	31.09
	R2	6.03	6.4	1.29	12.91	5.52
	F1	23.08	26.32	16.08	43.89	21.85
	Abs.	0.017	0.01	0.04	0.28	0.065
	Red.	-0.16	-0.2542	-0.188	-0.28	-0.12
	IDD	-2.1	-1.93	-2.1	-2.103	-0.5
	IDDV	0.248	0.398	0.168	0.391	0.391
PG-MMR	R1	36.42	40.44	19.8	40.55	36.54
	R2	9.36	14.93	1.34	12.36	6.67
	F1	24.3	26.9	16.39	43.93	21.72
	Abs.	0.019	0.02	0.04	0.275	0.069
	Red.	-0.17	-0.26	-0.172	-0.29	-0.142
	IDD	-2.4	-1.87	-1.9	-1.98	-0.72
	IDDV	0.441	0.274	0.192	0.249	0.318
Transformer	R1	28.54	31.54	20.46	43.57	30.12
	R2	6.38	5.9	1.41	14.03	4.36
	F1	15.72	17.82	16.38	44.54	21.35
	Red.	-0.1771	-0.17	-0.189	-0.18	-0.273
	Abs.	0.09	0.09	0.049	0.319	0.092
	IDD	-1.9148	-1.8677	-1.589	-1.89	-2.239
	IDDV	0.138	0.172	0.249	0.126	1.184
Hi-Map	R1	35.78	29.31	18.02	43.47	31.41
	R2	8.9	4.61	1.46	14.89	4.69
	F1	25.89	24.3	20.36	42.55	19.84
	Abs.	0.14	0.147	0.08	0.267	0.07
	Red.	-0.1722	-0.2002	-0.16	-0.23	-0.26
	IDD	-1.6201	-1.652	-1.8	-1.788	-2.223
	IDDV	0.185	0.155	0.209	0.209	0.448
Highest	R1	94.01	94.07	44.53	79.94	64.45
	R2	49.85	50.17	5.73	42.41	18.38

Table 7.2: Various metrics (Met) showing ROUGE Scores (ROUGE-1, ROUGE-2), F1 Score (F1) between candidate documents and oracle summaries, Abstractness (Abs.) of abstractive systems, Redundancy (Red.) in system generated summaries, Inter Document Distribution (IDD) and Inter Document Distribution Variance (IDDV) of system summaries in dataset DUC, TAC, Opinosis, Multinews and CQASumm.

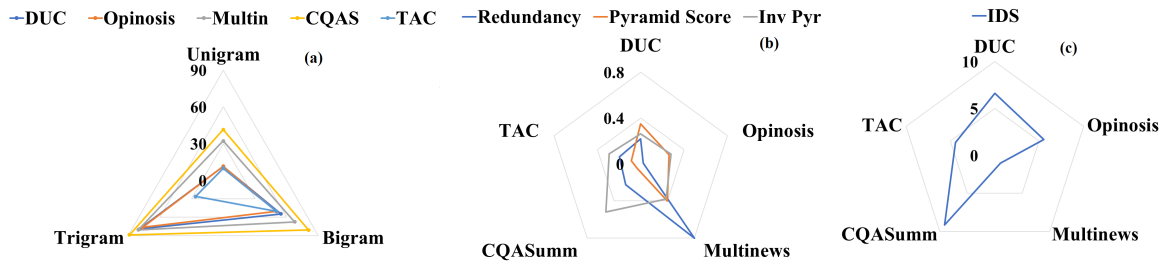


Figure 7.3: (a) Abstractness across datasets. (b) Redundancy, Pyramid Score and Inverse-Pyramid Score (Inv Pyr scaled down by a factor of 10 for better visualization with other metrics) across datasets. (c) Inter Document Similarity (IDS) across datasets.

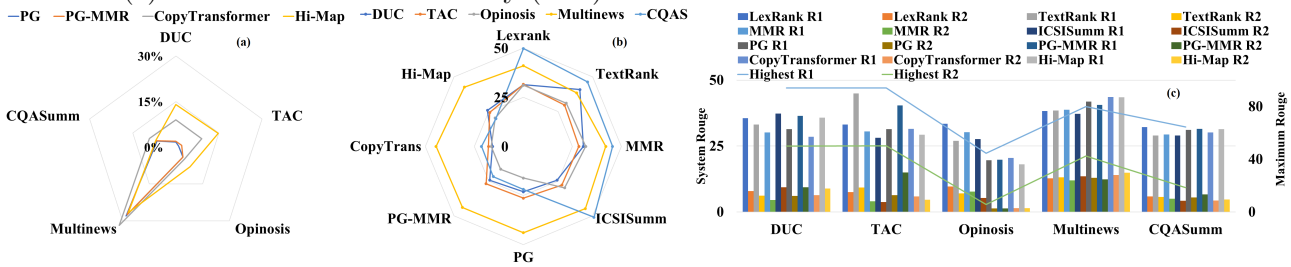


Figure 7.4: (a) Level of abstractness of systems w.r.t. candidate documents and the system generated summaries. (b) F1 Score of various systems between oracle summaries and system-generated summaries. (c) ROUGE scores of various system summaries on the left axis and maximum ROUGE score over a dataset on the right axis.

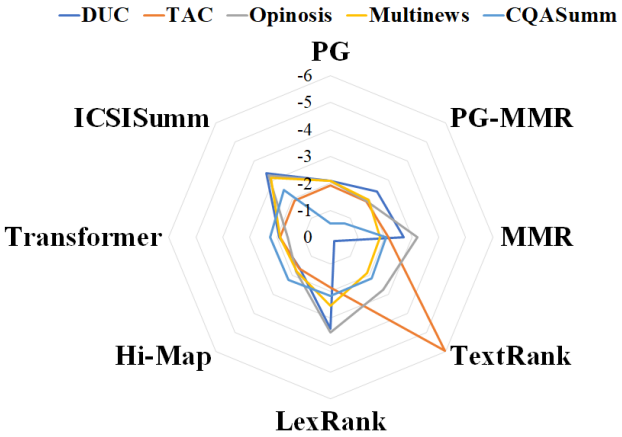


Figure 7.5: Redundancy of various systems across DUC, TAC, Opinosis, Multinews and CQASumm.

datasets (Fig. 7.4(c)) and Table 7.2).

• MDS systems under consideration outperform abstractive summarization systems by up to 10% on ROUGE-1 and up to 30% on F1 Scores, showing contradictory behavior in comparison to single-document summarization systems where state-of-the-art abstractive systems are known to outperform the former (Figs. 7.4(b)-(c)).

• The best summarization system on each corpus obtains a score 39.6%, 47.8%, 75.02%, 54.5%, 49.9% of the oracle upper bound on DUC, TAC, Opinosis, Multinews and CQASumm respectively, indicating that summarization on Opinosis and Multinews is a partially solved problem, while DUC, TAC and CQASumm exhibit considerable scope for improvement (Fig. 7.4(c)).
• Hi-Map and CopyTransformer generate more abstract summaries (17.5% and 16% novel unigrams respectively) in comparison to PG and PG-MMR (Fig. 7.4(a)).
• Averaging over systems and comparing datasets, we notice that Multinews and CQASumm achieve the highest abstractness (27% and 7% respectively), which might be a result of these two datasets having the most abstract reference summaries (Fig. 7.4(a) and (Table 7.2)).
• Abstractive systems exhibit a 55% shift in importance between the first and the second segments of generated summaries, whereas extractive systems show an average shift of only 40%, implying that abstractive systems have a stronger tendency to display layout bias (Fig. 7.2(b) and Fig. 7.2(c)).
• While DUC, TAC and Opinosis summaries generated from PG trained models exhibit lower novel unigrams formation, the same for CopyTransformer and Hi-Map on DUC, TAC and Opinosis shows a higher unigram formation on average (Fig. 7.4(a)).
• In terms of Inter Document Distribution, LexRank summary for TAC and CQASumm shows more variance across documents compared to DUC, Opinosis and Multinews. TextRank summary on DUC, TAC and CQASumm, MMR summary on DUC, and Hi-Map summary on CQASumm show higher variances as well. Systems such as PG, PG-MMR and CopyTransformer show minimal deviation in the document participation across datasets (Table 7.2).
• In terms of Topic Coverage, extractive systems show better coverage than abstractive systems (Table 7.2), which might be a result of extractive systems being based on sentence similarity algorithms which find important sentences, reduce redundancy and increase the spread of information from different segments of the candidate document. (Fig. 7.5).

7.3.5 Discussion

Q1. How should one model the quality of an MDS corpus as a function of its intrinsic metrics? What guidelines should be followed to propose MDS corpora for enabling a fair comparison with existing datasets? The quality of an MDS corpus is a function of two independent variables: the *quality of the candidate documents* and the *quality of the reference summary*. Our findings suggest that a framework for future MDS datasets should provide scores measuring their standing w.r.t. both the above factors. The former is usually crowd-source dependent, while the latter is usually annotator dependent. While Inter Document Similarity, Redundancy, Layout Bias and Inverse-Pyramid Score are indicators of the properties of the candidate document, metrics such as Abstractness of the reference summary and Pyramid Score are ground-truth properties. We divide the above metrics into two categories: *objective* and *subjective*. While all these metrics should be reported by imminent corpora proposers to enable comparisons with existing datasets and systems, we feel that the objective metrics average *Pyramid Score* and *Inverse-Pyramid Score* must be reported as they are strong indicators of generic corpus quality. Other subjective metrics such as IDS, Redundancy, Abstractness etc. can be modeled to optimize task-based requirements.

System	Metric				
	Abs. corr	R-1 corr	Layout correlation		
			First	Second	Third
LexRank	-	0.08	0.88	0.06	0.96
TextRank	-	-0.24	0.91	0.76	0.97
MMR	-	0.32	0.86	0.09	0.97
ICSISumm	-	0.11	0.39	0.53	0.72
PG	0.57	0.65	0.80	-0.80	-0.98
PG-MMR	0.57	0.33	0.84	-0.69	-0.91
CopyTrans.	0.47	0.50	0.84	-0.31	-0.79
Hi - Map	0.11	0.45	0.74	-0.11	-0.46

Table 7.3: Pearson correlation between corpus and system with column 4 (**First**) between Abstractness of corpora and system, column 5 (**Second**) between Abstractness of corpora and ROUGE-1 score of systems across datasets and column 6 (**Third**) showing Layout Bias correlation between system and corpora.

Q2. Why do the ROUGE-based ranks of different MDS systems differ across datasets? How should an MDS system which is to achieve reasonably good ROUGE score on* all *corpora look like? From Table 7.2 within studied systems, in terms of ROUGE-1, ICSISumm achieves the best score on DUC, TextRank on TAC, LexRank on Opinosis, CopyTransformer on Multinews and LexRank on CQASumm. Hence as of today, no summarization system strictly outperforms others on every corpus. We also see that CopyTransformer which achieves state-of-the-art performance on Multinews achieves 10 points less than the best system on DUC. Similarly, LexRank, the state-of-the-art performer on CQASumm, achieves almost 12 points less than the best system on TAC. Therefore, a system that performs reasonably well across all datasets, is also missing. This is because *different corpora are high on various bias metrics, and summarization systems designed for a particular corpus take advantage and even aggravate these biases.* For example, summarization systems proposed on news based corpora are known to feed only the first few hundred tokens to neural models, thus taking advantage of the layout bias. Feeding entire documents to these networks have shown relatively lower performance. Systems such as LexRank are known to perform well on candidate documents with high inter-document similarity (e.g., Opinosis). Solving the summarization problem for an unbiased corpus is a harder problem, and for a system to be able to perform reasonably well on any test set, it should be optimized to work on such corpora.

Q3. Why do systems show bias on different metrics, and which other system and corpus attributes are the reason behind it? We begin by studying how *abstractness of generated summaries is related to the abstractness of corpora the system is trained on.* For this, we calculate the Pearson correlation coefficient between the abstractness of generated summaries and references across different datasets. From Table 7.3, we infer that PG, PG-MMR and CopyTransformer show a positive correlation which implies that they are *likely to generate more abstract summaries if the datasets on which they are trained have more abstract references.* Lastly, we infer *how Layout Bias in system-generated summaries is dependent on the layout bias of reference summaries.* The last three highlighted columns of Table 7.3 infer that the abstractive systems such as PG, PG-MMR, Hi-Map and CopyTransformer show a high negative correlation for the end

segments while maintaining a strongly positive one with the starting segment. On the other hand, extractive systems such as LexRank, TextRank, MMR and ICSISumm maintain a strongly positive correlation throughout the segments. On shuffling the source segments internally, we observe that extractive systems tend to retain their correlation with corpora while abstractive systems show no correlation at all (Fig. 7.2), proving that in supervised systems, *the layout bias in system summaries propagates from the layout bias present in corpora.*

Q4. Is the task of MDS almost solved, or there is still plenty of scope remaining for improvement? In the previous sections, we computed the oracle extractive upper bound summary using greedy approaches to find the summary that obtains the highest ROUGE score given the candidate documents and references. We observe that the best summarization system on each corpus today obtains a score which is 39.6% of the extractive oracle upper bound on DUC, 47.8% on TAC, 75.02% on Opinosis, 54.5% on Multinews and 49.9% on CQASumm. This shows that there is enough scope for MDS systems to achieve double the ROUGE scores obtained by the best system to date on each corpus except Opinosis. Therefore, we believe that the task of MDS is only partially solved and considerable efforts need to be devoted to improving the systems.

7.4 Representation Bias in Summarization systems

In this section, we discuss how the representation bias is characterized, quantified and exploited for quantitative and qualitative performance.

7.4.1 What is Representation Bias?

A set of samples in a dataset represents the variation of the population. Any form of data which lacks the consideration of outliers, uneven data points, diversity and anomaly factors give rise to representation bias. For example, the paucity of varying geographical presence in the Imagenet [198] dataset formulates a representation bias towards the white population. Similarly, in case of abstractive summarization, the scarcity of data points portraying as outliers, uneven themes and non-templatic target summaries lead to representation bias towards a specific set of data samples leading to degraded performance. This also abstains deep learning based systems from generalizing across all genres of datasets. We study this representation bias for the task of abstractive summarization and exploit it to gain improvements with qualitative data points.

7.4.2 Characterizing and Modeling Bias in Training Dynamics

Inspired by the study of biases [199, 200, 7] portraying that the current deep learning based summarization systems show qualitative issues like repetitions, unfaithfulness and non-coherence, we dive deep into the reasoning behind the same.

Given a source document and the corresponding target summary, we train an end-to-end deep abstractive summarization system keeping the batch size as 1. The corresponding embeddings and the encoder outputs learned at each time step are saved. We save weights at every epoch to have the representation of the trained embeddings and the encoder representation learned. After the

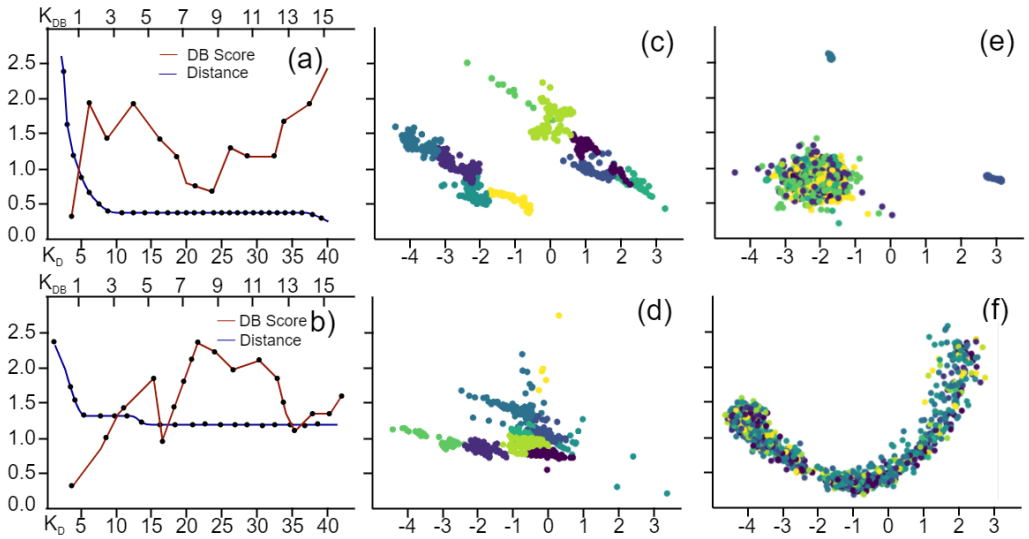


Figure 7.6: Visualizations of (a) Elbow and Davies Bouldin score for LSTM models, (b) Elbow and Davies Bouldin score for Transformer models, (c) LSTM embedding vector visualization using PCA, (d) LSTM encoder vector visualization using PCA, (e) Transformer embedding vector visualization using PCA, (f) Transformer encoder vector visualization using PCA.

convergence of the system, we cluster the embedding representations and the encoder weights. The intuition behind clustering is to understand the distance or the diversity of the data points in the vector space. The idea is that the data points should be spread far apart to represent the diversity of varied examples in the training sample.

For clustering of embedding and encoder weights, we use the KNN+ algorithm. We consider the Elbow [201] method and the Davies Bouldin [202] method to find the optimum number of the clusters for both the embedding and the encoder weights. For the LSTM-based systems, the optimal number of clusters for the embeddings by elbow method as well as the Davies Bouldin method came out to be 9, as shown in Fig. 7.6(a). We visualize the formulated clusters for embeddings by reducing the dimensions using Principal Component Analysis (PCA) in Fig. 7.6(c).

We also follow the same steps for the encoder weights. The optimal number of clusters for encoder weights is also 9 using both the Elbow and the Davies Bouldin methods. Fig. 7.6(d) shows the encoder representation by reducing the dimension size using PCA. The visualization of encoder weights shows the reduced data points when the number of clusters is set to 9. From the visualizations, we infer that the vector space representation of both the embedding and the encoder outputs is not diverse in the space and is heavily saturated. Comparatively, encoder weights show a more diverse spread on the embedding space than the embeddings; yet the data points look heavily saturated for both the representations.

Similarly, for the Transformer-based systems, we again use the KNN+ algorithm to perform clustering of embeddings and the encoder weights. To find the most optimal number of clusters, we use the Elbow and the Davies Bouldin methods. From both the methods, the most optimum number of clusters came out to be 5. Keeping 5 as the cluster parameter, we visualize the embedding data

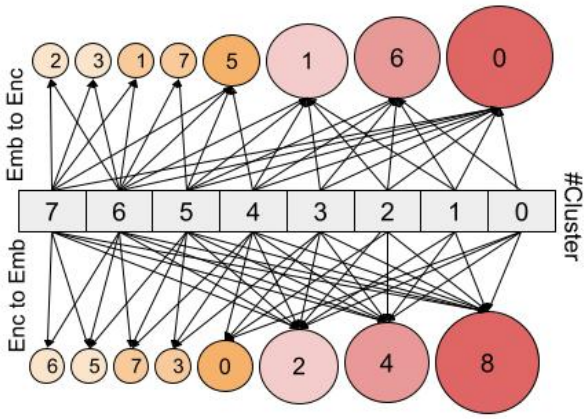


Figure 7.7: Visualizations showing the mapping of clusters for LSTM-based systems from embedding-to-encoder (array to upper cluster cloud) and encoder-to-embedding (array to lower cluster cloud).

points. However, the visualizations obtained came out to be heavily clustered and saturated at a single point. The points that are at the farthest position are three times the average Euclidean distance, making them outliers. We take the second best optimal cluster count as 14 and visualize the corresponding embedding points.

We visualize the corresponding data points using PCA in Figs. 7.6(e) and 7.6(f), inferring that the Transformer-based systems show a more diverse representation of data points at the encoder side than any LSTM-based systems.

To better understand the delineation between the embeddings and the encoder representation, we map the data points originating from the embedding clusters and falling in the encoder clusters for the LSTM-based models shown in Fig. 7.7. For embedding-to-encoder map, we see that most of the data points fall into three major encoder clusters, namely #0, #6 and #1. For cluster #0 in the encoder, data points map to clusters #0 at 71%, and cluster #7 at 21%, while the rest of the embedding majority is the same. For the encoder-to-embedding mapping, most data points from the encoder cluster #3 fall in embedding clusters #4, #2, and #8 except for the encoder cluster #0, where #2 is the majority cluster. Similarly, for the Transformer-based systems, a similar pattern is observed (Fig. 7.8) in the embedding-to-encoder and the encoder-to-embedding mappings. For embedding-to-encoder, the majority of the data points fall in clusters #3, #4, and #6 with minor variations. Similarly, for the encoder-to-embedding mapping, the majority falls to clusters #13, #7, and #5 with minor variability across classes. This taxonomy of cluster data point mapping between the embedding weights and the encoder weights shows that the deep learning based systems learn a similar representation irrespective of the training style or the architecture. This saturation of weights forming limited clusters thwarts the performance and makes the model learn a fixed-type of output template that is repeated in the generated summary irrespective of the input document.

7.4.3 Exploiting Bias for Performance

As the number of clusters and the representations are dense and saturated for both the embedding and encoder weights, we sample the data points from each cluster to obtain diverse representations

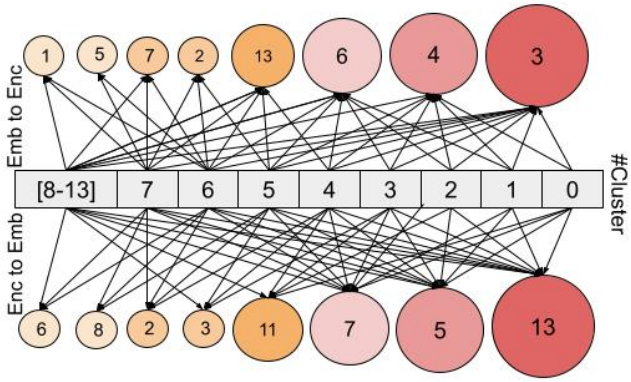


Figure 7.8: Visualizations showing the mapping of clusters for Transformer-based systems from embedding-to-encoder (array to upper cluster cloud) and encoder-to-Embedding (array to lower cluster cloud).

to make the systems more generalizable, easier to train, and perform better across the dataset. The subsampling helps in the data distillation as to use only the essential data points during training.

For subsampling of the data points, we extend the study of [203] to filter the data points. For every data point in the training set, we use the cosine similarity based KNN+ algorithm. The cluster C_i is formulated using

$$C_i = \kappa(f_i, F, k) \tag{7.8}$$

where $F = feat(X)$ is the feature matrix, and k is the nearest neighbour. Now, for the cluster C_i, we define another sub-cluster S_i,

$$S_i = \gamma(C_i) \tag{7.9}$$

with f_i as the centre of cluster C_i and f_j forming as the neighbor of f_i. The similarity of all the points is computed with respect to the centre as a vector $\{S_{i,j}\}_{j=1}^{k}$. The data points in S_i are sorted concerning the similarity function as shown in Fig. 7.9. The subsamples based on the upper limit are extracted uniformly from each S_i.

The training data points obtained from the sorted clustered samples are benchmarked by finetuning them over the BART [204] language model. We finetune BART over the complete CNN/Dailymail, Multinews and CQASumm dataset, and obtain the overall benchmark results. We then perform the subsampling to obtain $10K$ subsamples for the CNN/Dailymail data, Multinews and CQASumm. We finetune BART on the $10K$ subsampled data and compare the performances with the complete data (Table 7.4).

We observe that for CNN/Dailymail data, the $10K$ subsampled instances attain a 39.49 R1 and 36.38 RL, an improvement of +4.92 R1 and +5.04 RL over the Random $10K$ and only 0.64 R1 and 0.46 RL lower than the performance on the complete dataset. Similarly, FEQA and the Pyramid score gain an improvement of +3.73 and +0.06, respectively against the random $10K$ samples showing that the qualitative performance can be achieved with just 10% of data without compromising on the quantitative performance. Similarly, when $40k$ subsampled data is finetuned, the performance gain is only 0.17 R1 and 0.26 RL lower than the complete dataset. For Multinews,

Dataset	No. of samples	Rouge-1	Rouge-2	Rouge-L F1	Rouge-L	BERTScore	FEQA	Pyramid
CNN/Dailymail	Random 10K	34.57	14.83	31.34	30.29	22.63	12.79	0.28
	10K	39.49	17.84	36.38	36.31	26.47	16.52	0.34
	40K	39.96	17.66	36.58	36.59	26.44	16.48	0.34
	100K	40.18	17.74	36.67	36.73	26.49	16.49	0.35
	Complete	40.13	17.76	36.84	36.79	26.57	16.57	0.36
Multinews	Random 10K	34.1	10.24	13.64	13.4	21.54	9.33	0.23
	10K	42.76	13.86	17.07	17.02	33.88	13.24	0.27
	20K	42.89	13.94	17.09	16.95	33.89	13.24	0.26
	40K	43.41	13.91	17.13	17.08	33.83	13.25	0.27
	Complete	43.49	14.02	17.21	17.16	34.21	13.27	0.28
CQASumm	Random 10K	22.41	3.12	14.62	14.52	18.41	7.02	0.13
	10K	29.67	4.87	19.64	19.44	22.12	9.64	0.21
	30K	29.41	4.86	19.77	19.67	22.04	9.65	0.19
	50K	29.38	4.91	20.17	19.92	22.06	9.68	0.22
	Complete	31.47	5.02	20.28	20.24	22.16	9.71	0.24

Table 7.4: Benchmarking scores over various subsampled data on the metrics – Rouge-1, Rouge-2, Rouge-L, Rouge-L F1, BERTScore, FEQA, and Pyramid score for the datasets – CNN/Dailymail, Multinews and CQASumm.

the subsampled $10K$ data produces 42.76 R1 and 43.49 RL, a reduction of just 0.73 R1 and 0.14 RL while giving an improvement of +0.4 FEQA and +0.01 Pyramid score. We notice similar trend for the CQASumm dataset on which the gain is +0.04 FEQA and +0.02 Pyramid score (Table 7.4).

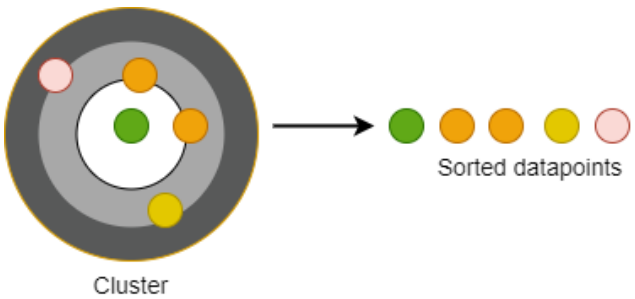


Figure 7.9: A cluster showing the samples with respect to the distance from the center. The datapoints are rearranged according to the distance from minimum to maximum.

7.5 Analysis

In the previous section, we assess the diversity of the embedding weights and the encoder weights in the vector space learned during the model training. Our intuition is that the deep learning based models should learn a diverse representation so as to handle variable source input. We analyze these characteristics for both types of the summarization systems – LSTM-based and Transformer-based. Our study indicates that both types of models often learn similar weight matrices across the

embedding as well as in the encoder vector space shunning the variable input representation. Even though the embeddings can represent the variation in the vector space but the summarization system, during its training, maps these data points to a saturated point at the encoder side. From Fig. 7.7(c), we observe that the embeddings show large variability; however, the encoder space shows the points getting saturated. In case of Transformers, the embedding weights are saturated; however, Fig. 7.8(f) shows that the embedding points are again mapped to a saturated continuous sequence of clusters. This indicates that the current systems, irrespective of the deviation in the input source, try to learn similar weights at the encoder side, giving a templatic and repetitive output in the generated summaries. After the analysis, we filter out the data points using the hierarchical clustering algorithm discussed in the previous Section and finetune them over the BART [204] language model. When compared with the performance of the systems trained on the complete data, the subsampled $10K$ finetuned systems stand comparable while the performance is at par against those finetuned on the randomly-sampled data. This shows that an uncontrolled urge to extract more data for the models is not the right way; limited qualitatively labelled data can provide the model with sufficient information to make it learn diverse representations, generalize across datasets and easier to train.

7.6 Discussion

We discuss some of the insights obtained from the analysis performed over the biases characterized.

Does deep learning model favour generalization? Deep learning models, in general, exploit the biases to learn the representations between the source data and the target data. The relations obtained act as weights to map the contextual relationship for providing inference on the unseen data. However, the exploitation does not always lead to better benchmark scores. For a given problem, the datasets are often collected from a crowdsourced platform that may affect the quality of the dataset due to various reasons such as absence of moderation, limited themes, and more engagement over the trending topics. These limitations make the data less robust and limited to be usable over a particular genre or a problem. Often the proposed models are benchmarked on limited datasets as they fail to perpetuate the performance across similar problems.

Does diversity in training samples help in generalizing the systems? Deep learning based systems require a massive amount of data for training. The available training data is an example of the representation that the system should accept during inference for generating output. The system learns the deviations in the individual train instance to adjust its parameters in order to handle variability and diversity during inference. The diversity in the training data helps the system learn these variations. However, the degree of these variations poses a question of whether to treat the sample as a variation or an outlier. If treated as a variation, the outlier sample may inflate the error metric, causing the system diverge from its intended goal. The threshold of what a system should treat as an outlier and a regular sample acts as a foundation of diversity in the training dataset. Since most of the training examples lie very close to each other due to similar topics or themes, the threshold of a system to consider a data instance as an outlier decreases, making the model treat the diverse training examples as outliers and learn a limited representation as to the actual mapping.

Can the biases be exploited for performance? Biases in summarization systems have been exploited to favour performance. [199] exploited the lead bias in the news datasets to attain high quantitative benchmarks. News datasets contain a majority of the vital information in the first few paragraphs, and the systems can attain this really well, leading to high performance. [54] used data upsampling and transfer of weights from a language model to make models perform with a few shot transfer learning. For a dataset, each model exploits characteristics which may or may not be known to the researchers to visualize and tweak. Our work focuses heavily on the aspect that the biases obtained during the training phase can be exploited to attain similar performance as obtained when a model is trained on the complete dataset.

Can the biases be rectified for datasets? What actually constitutes a bias? The definition of bias is subjective and can not be defined as a function of either a dataset or a model. For some, a bias can be towards gender-imbalanced inferences, and for others, a bias can be a generation of templatic output for different source inputs. Each type of bias requires a specific solution and a prescription either in the data or the model. As discussed by [54], a single system can not rectify all errors. Hence, a study is needed to find the forms of biases that either help in terms of required improvements or limit the performance of the model qualitatively or quantitatively. In addition, biases can also be a result of internal factors like data shuffling, overfitting or external factors like a biased opinion of a crowdsourced platform. However, in this study, we mainly focus on the biases that thwart the performance of the model qualitatively. We study the system's training dynamics and the representations learned by the system at the embedding and encoder levels. Later, we exploit these biases to gain improvements over the current summarization datasets in terms of faithfulness and qualitative inference.

7.7 Biases in Multimodal Datasets and Systems

It is crucial to note that, to date, there have been limited datasets available [21] for multimodal summarization, hindering comprehensive studies on biases in multimodal data. Recognizing this gap, we have taken a proactive step by introducing two new multimodal datasets [22, 23] as part of our research initiatives. As we continue to advocate for fairness and accuracy in summarization, the introduction of these datasets stands as a pivotal contribution towards fostering unbiased representation in multimodal data and systems.

7.8 Conclusion

In this chapter, we have undertaken a comprehensive exploration of the challenges inherent in multi-document summarization (MDS) and deep learning-based models. Our aim was to study the heterogeneous task of MDS, analyzing interactions between widely used corpora and state-of-the-art systems to derive key conclusions. We defined MDS as a mapping from a set of non-independent candidate documents to a synopsis covering important and redundant content in the source. Proposing intrinsic metrics to evaluate MDS corpus quality and a framework for future corpus creation, we laid the groundwork for further research in the field.

Furthermore, we delved into the characterization of representation bias propagating from the dataset into the model. We discussed the limitations hindering the performance of deep learning models across varied datasets, highlighting the skewed representation of data in embeddings and models in the encoder space. Utilizing hierarchical clustering, we demonstrated the efficacy of subsampling data to achieve comparable performance against complete datasets. Leveraging representation bias, we showcased performance improvements on subsampled datasets across various metrics. Additionally, we provided insights into why current datasets and summarization systems struggle to generalize across data.

Chapter 8

Bringing Fairness in Summarization

Multi-document Summarization (MDS) characterizes compressing information from multiple source documents to its succinct summary. An ideal summary should encompass all topics and accurately model cross-document relations expounded upon in the source documents. However, existing systems either impose constraints on the length of tokens during the encoding or falter in capturing the intricate cross-document relationships. These limitations impel the systems to produce summaries that are non-factual and unfaithful, thereby imparting an unfair comprehension of the topic to the readers. To counter these limitations and promote the information equivalence between the source document and generated summary, we propose `FABRIC`, a novel encoder-decoder model that uses pre-trained BART to comprehensively analyze linguistic nuances, simplicial complex layer to apprehend inherent properties that transcend pairwise associations and sheaf graph attention to effectively capture the heterophilic properties. We benchmark `FABRIC` with eleven baselines over four widely-used MDS datasets – Multinews, CQASumm, DUC and Opinosis, and show that `FABRIC` achieves consistent performance improvement across all the evaluation metrics (syntactical, semantical and faithfulness). We corroborate these improvements further through qualitative human evaluation.

8.1 Introduction

Multi-document summarization (MDS) aims to formulate a summary that captures the essence and main points of multiple documents on a specific topic. In contrast to single document summarization (SDS) that focuses on generating summaries from a single source, MDS faces additional challenges such as dealing with a larger search space, redundant documents, and conflicting opinions. These challenges pose difficulties for deep learning models, often resulting in the generation of summaries that are the results of hallucination and they often lack faithfulness [205]. The development of Large Language Models (LLMs) such as BERT [51], BART [204], and T5 [53] has significantly advanced the field of text summarization. However, generating factually accurate and faithful summaries remains a persistent challenge [206].

Recent studies have focused on enhancing the faithfulness and factuality of summarization models, which is broadly categorized into three types – (i) post editing models for correcting generated summaries [103, 100], (ii) using multi-task problems like question-answering [66, 147],

entailment [115, 63] etc., and (iii) using external knowledge [113, 207, 208, 207, 112] to support the model during summary generation. In contrast to these methods, we propose a novel approach that promotes topic coherence and inter-document connections, all without the need for additional parameters, post-editing, external knowledge, or auxiliary tasks.

Additionally, capturing semantic connections and multi-level representation can further aid the model in discerning the redundant and pivotal information in multiple documents. Graph neural networks [209] are constrained in terms of subpar performance in heterophilic settings[1] and oversmoothing [210] when utilized with multi-layer neural networks. To mitigate these concerns, we propose `FABRIC`[2] in which we introduce simplicial complex [211] layer and sheaf graph [212] attention for multi-document summarization. Simplicial complexes are employed to apprehend the interconnections among diverse elements of the text, encompassing words, phrases, or sentences. By representing these associations as simplices (geometric shapes formed by combining vertices, edges, and higher-dimensional counterparts), a simplicial complex furnishes a structure to examine the connectivity and coherence within the text. On the other hand, sheaf graphs facilitate the assimilation of diverse node types and attributes within the graph structure. Each node can symbolize a specific element, such as a word or a phrase, and convey its own attributes or features. By considering these heterogeneous characteristics, sheaf graphs can apprehend and model the relationships among distinct types of nodes. This empowers `FABRIC` to comprehend various elements and their relationships, enabling the generation of faithful and accurate summaries.

Moreover, when dealing with MDS, a critical challenge that neural networks often encounter is processing large documents. Many recent studies have attempted to concatenate multiple documents into a flat sequence and train SDS models on them [12]. However, this approach fails to consider the inter-document relationship and the presence of redundant long input vectors. Even very large language models like Long-T5 [136] and BART-4096 [204], which are capable of handling input vectors beyond 2000 tokens, face practical limitations due to the quadratic growth of input memory space. Furthermore, such extensive global attention may exhibit lower performance compared to alternative methods. To address this issue, we propose a novel approach, called *topic-assisted document segmentation*. It involves using a segment of the document as a context vector, compressing it into an oracle summary while covering major topics, and appending it to the next segment to generate the subsequent context vector. By employing this method, summarization models can learn the representation of the document from a compressed form, overcoming the challenges associated with processing large documents.

In short, our contributions are as follows.

1. We propose a novel topic assisted document segmentation method, which allows any language model to generate contextual vectors for any input length input.
2. We propose `FABRIC`, a novel encoder-decoder model that uses pre-trained BART to comprehensively analyze linguistic nuances, Simplicial Complex layer to apprehend inherent properties that transcend pairwise associations and sheaf graph attention to more effectively apprehend the heterophilic properties.
3. We evaluate `FABRIC` using four standard metrics – ROUGE [133], BARTScore [213], FactCC [214],

[1] In a heterophilic setting, diverse node and edge types can complicate the message-passing process of GNN, therefore affecting its performance.

[2] FABRIC: FAirness using BaRt, sImplicial Complex and sheaf

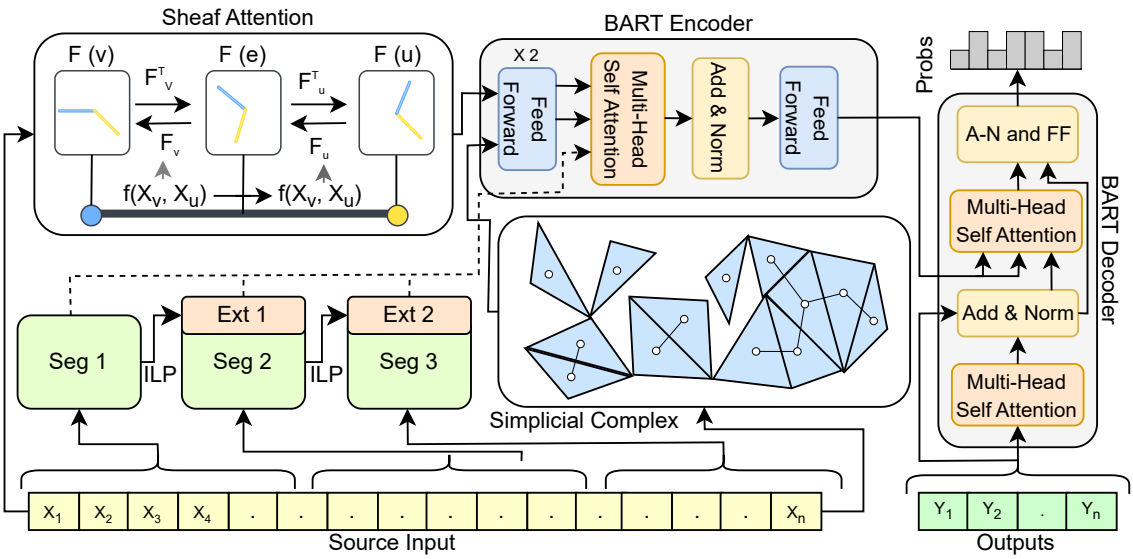


Figure 8.1: A schematic architecture of FABRIC. We adept the BART encoder and introduce simplicial complex layer and sheaf graph attention and fuse them with multi-head attention of BART. The fused representation is passed to the BART decoder to generate candidate summaries.

and SummaC [106], to assess the quantitative, and qualitative performances and faithfulness. We showcase that FABRIC fares well against eleven widely-popular abstractive baselines. FABRIC beats the best baseline, PRIMER [215] by +0.34 ROUGE-L on Multinews, +4.90 Rouge-L on CQASumm, +2.94 on DUC and +2.28 Rouge-L on the Opinosis dataset. Our qualitative analyses on semantic similarity by BARTScore and faithfulness by FactCC and SummaC also show considerable gains compared to the baseline-generated summaries. We further perform an exhaustive human evaluations and comparison with ChatGPT[3] to compare the quality of the system-generated summaries.

8.2 Proposed Methodology

We present FABRIC, a novel multi-encoder-decoder model for multi-document abstractive text summarization. Our approach harnesses the power of topic assisted document segments and leverages higher order topological spaces to adeptly extract essential information. Figure 8.1 shows a schematic diagram of FABRIC. This section explains each component of FABRIC in detail.

8.2.1 Topic Assisted Document Segments

Transforming the entire document to a contextual vector is very expensive pertaining to the quadratic growth of memory and computations in Transformer-based models. LMs like Longformer [49] are able to process documents in a linear complexity. However, they fail to efficiently capture

[3]https://openai.com/blog/chatgpt/

the context of the whole document. To address this limitation, we propose a new data modelling strategy, called *Topic Assisted Document Segmentation with Oracle proxies*. At first, the source document D is split into segments $D_1, D_2, D_3, \ldots, D_n$ based on major topics identified. We use BERTopic [216] for topical identification. Next, we formulate oracle summaries for segment D_1 using an Integer Linear Programming (ILP) solver [217] and penalize it with BERTopic [216] to aid the module in generating oracle summaries and to cover all major topics of the document. We append the oracle summary obtained O_1 to the next segment $O_1 + D_1$ and formulate the next oracle summary using only $O_1 + D_1$. Each segment generated is limited to 768 tokens for BART to process the input vector effectively. LM to process. Each segment, therefore, obtains an independent encoder representation of the document, which when fused with the other modules, is fed to the BART decoder to generate the final summary.

8.2.2 Simplician Complex Layer

Summarizing documents does not always entail information in pairwise connections. Consequently, methods based on single-dimensional graphs fall short in capturing the comprehensive context and its intricate relationships. Conversely, the utilization of a simplicial complex layer enables the incorporation of higher-order graphs, where the inclusion of simplices (vertices within a single-order graph) empowers the modeling of information that extends beyond pairwise associations. This approach formulates an exceedingly diverse and extensive feature set, facilitating the model's ability to comprehend inter-document relations in a more comprehensive and nuanced manner.

We use the simplicial complex (SC) layer in the self-attention setting similar to [211]. Mathematically, we define the input to SC layer as SC_i and learnable weights as W_{sc}. The transformations are applied to the lower irrotational filter H^i, and upper solenoidal filter H^s. The transformations are represented by $H^i = [SC_i]_i W_{sc}^{(i)} \in R^{F_{l+1}}$, and $H^s = [SC_i]_i W_{sc}^{(i)} \in R^{F_{l+1}}$. The representation H^i and H^s are combined for each iteration, and self-attention is computed using the following equations,

$$\begin{aligned} e_{l,i,j}^{(u)} &= a_l^{(i)}\Big(h_{l,p}^{(i)}, h_{l,q}^{(i)}\Big) \qquad \text{for } j \in \mathcal{N}_i^{(u)}, \\ e_{l,i,j}^{(d)} &= a_l^{(s)}\Big(h_{l,p}^{(s)}, h_{l,q}^{(s)}\Big) \qquad \text{for } j \in \mathcal{N}_i^{(d)} \end{aligned}$$

These representations are normalized using the softmax function as $\alpha_l^{(u)} = \text{softmax}_j(e_i^{(u)})$, and $\alpha_l^{(s)} = \text{softmax}_j(e_l^{(s)})$. The attention weights of the SC layer are computed using,

$$Z_{l+1} = \sigma_l\Big(\sum_{p=1}^{J_l^{(d)}} \big(L_l^{(d)}\big)^p Z_l W_{l,p}^{(d)} + \sum_{p=1}^{J_l^{(u)}} \big(L_l^{(u)}\big)^p Z_l W_{l,p}^{(u)} + \widehat{P}_l Z_l W^{(h,l)}\Big)$$

where the coefficients of the upper and lower attentional Laplacians, $\mathbf{L}_l^{(u)}$ and $\mathbf{L}_l^{(d)}$, respectively are obtained . The filter weights $\big\{\mathbf{W}_{l,p}^{(d)}\big\}_p$, $\big\{\mathbf{W}_{l,p}^{(u)}\big\}_p$, $\mathbf{W}_l^{(h)}$ and the attention mechanism parameters $\mathbf{a}_l^{(u)}$ and $\mathbf{a}_l^{(d)}$ are learnable parameters, while the order $J_l^{(d)}$ and $J_l^{(u)}$ of the filters, the number F_{l+1} of output signals, and the non-linearity $\sigma_l(\cdot)$ are hyperparameters to be chosen at each layer. The final representation is passed through a feed-forward layer.

8.2.3 Sheaf Graph Attention

Similar to graph neural networks (GNNs) [209] and graph attention networks (GATs) [218], the contextual representation of simplicial complex (SC) layer encounters limitations in performance, particularly in heterophilic settings and over-smoothing. Moreover, relying on a single module to capture the inter-document relationship may introduce biases in representing factual knowledge. To overcome this limitation, we introduce a new module, called sheaf graph attention in the MDS setting, taking inspiration from the prior sheaf attention networks [212]. Mathematically, we define the input vector as X_s and introduce the learnable weight matrices W_a and W_b. The identity matrix is denoted by I, and the sheaf Laplacian is represented as ψ. We perform a Kronecker product between I and W_a. The sheaf Laplacian ψ is defined as a vector in an undirected graph G_u, where vertices are defined by v and edges by e. Mathematically, it can be expressed as follows:

$$Sheaf(X) = relu((I - \psi)(I \otimes W_a)XW_b \tag{8.1}$$

Here, X is the data transformation block, and $\otimes$ is Kronecker product.

8.2.4 Encoder Setting

The first module introduces a document segments via topic guidance, which enables `FABRIC` to generate multiple target sequence vectors with segmented source documents aided with oracle proxies. The sequence vectors are learned using the BART encoder model. These feature vectors helps the model to understand the linguistic properties and retain multiple context vectors. The second module introduces the simplician complex layer, helping the model to understand the inter-document relations and aiding the weights of the vectors to promote the factualiy and faithfulness. As the feature vector from the SC layer are not same as attention input, the output is passed through a feed-forward layer and fused with the multi-head attention module of BART. Finally, the vectors from the sheaf graph attention model also undergo normalization via a feed-forward layer for feature vector normalization, followed by fusion with the heads of the multi-head attention in the BART encoder. The SC layer and sheaf graph contribute to the model's understanding of inter and intra-document relationships and the weighting of contexts, thereby promoting factual and faithful representations for the decoder generator model.

8.2.5 Decoder Setting

The amalgamated representation derived from the encoder module of `FABRIC`, consisting of the BART encoder, SC layer, and sheaf graph attention, is forwarded to the BART decoder module. In this configuration, the BART decoder generates summaries utilizing the encoder representation and the pointing mechanism. Much like a pointer generator, the pointing mechanism enables the model to directly copy text from the source document.

8.3 Dataset

We benchmark our study on four widely popular datasets. (i) **Multinews** [12] comprises news articles as source documents and human-written summaries as target summaries. (ii) **CQASumm** dataset consists of question-answer pairs, where the accepted answer is used as the target summary and the remaining answers form the source document. (iii) **DUC** [193] includes news articles as source documents, with target summaries manually annotated by editors. (iv) **Opinosis** [195] combines user reviews from various platforms, where individual opinions serve as the source document, while the target summaries are human-annotated.

8.4 Abstractive Baselines

(i) The **Pointer Generator (PG)** approach [38] combines attention, and pointing mechanism to apprehend inter-document relationships. (ii) **Himap** [12] amalgamates MMR sentence weights in the PG network to emphasize pivotal information in the summary. (iii) **Bottom-up Transformers** [15] employ the Transformer architecture [45] with one random attention head functioning as the copy pointer. (iv) **BERT** [51] is an encoder-only language model trained using token masking techniques. (v) **BART** [204] employs an encoder-decoder model trained on the text span masking technique. (vi) **T5** [53], akin to BART, is an encoder-decoder-based model that treats all downstream tasks as text-to-text problems during training. (vii) **LongT5** [136] scales the T5 architecture and utilizes sentence masking during pretraining to enrich language generation. (viii) **Pegasus** [36] adopts an innovative sentence masking technique in the language model to capture sentence-level representations. (ix) **Longformer** [49] incorporates a Transformer with sparse attention to handle elongated sequences.(x) **BRIO** [219] employs a stochastic approach, rather than maximum likelihood, to train the network.(xi) **PRIMERA** [215] extends the Longformer architecture [49] by pretraining it on the downstream task.

8.5 Evaluation Setup

We benchmark `FABRIC` on the quantitative metrics – Rouge-1 (R1), Rouge-2 (R2) and Rouge-L (RL) [133], to evaluate the lexical overlap, as well as on qualitative metrics – BARTScore [213] to compute the semantic overlap, and FactCC [214], and SummaC [106] for faithfulness assessment.

8.5.1 Evaluation Metrics

To compare the quality of the summaries, we employ the following evaluation metrics. (i) **Rouge** computes the lexical overlap between the target and the generated summary. (ii) **BARTScore** computes the semantic overlap between target and the generated summary using the pre-trained BART [204] LM. (iii) **SummaC** evaluates the consistency of summary sentences w.r.t the source article, taking into account any inconsistencies that may occur throughout the source text. (iv) **FactCC** leverages a model trained to assess the consistency between a text/summary and its source article using synthetic data generated through various transformations.

System	Multinews			CQASumm			DUC			Opinosis		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PG	41.85	12.91	20.79	31.09	5.52	21.85	31.43	6.03	23.08	19.65	1.29	20.08
HiMAP	44.17	16.05	24.38	27.13	4.48	19.87	31.44	6.11	23.17	18.02	1.46	19.84
Transfor.	44.32	15.11	28.07	27.52	4.53	21.0	32.14	6.24	23.34	20.46	1.41	21.35
BERT	44.27	16.23	31.41	27.32	4.55	21.14	34.64	6.39	24.18	20.41	4.62	21.84
BART	48.47	18.41	33.21	27.84	4.65	21.74	35.41	6.48	24.67	19.8	6.82	26.87
T5	44.08	16.39	37.68	28.11	3.74	21.45	34.13	6.32	24.13	27.41	6.97	27.41
LongT5	48.17	19.43	38.94	28.74	4.84	22.12	34.58	6.37	24.32	29.46	7.04	28.64
Pegasus	41.79	16.58	39.72	28.24	4.87	22.41	34.61	6.41	24.71	31.28	7.15	29.39
Longfor.	46.89	18.50	41.83	28.01	5.05	23.86	36.31	6.76	27.11	33.32	8.63	30.49
BRIO	47.24	19.34	44.57	29.13	5.12	24.17	37.18	7.02	29.37	31.15	8.52	29.64
PRIMER	49.90	21.11	46.23	31.54	5.58	26.57	37.84	7.21	31.18	34.64	9.12	33.20
FABRIC	50.68	21.26	46.57	35.81	6.91	31.47	39.64	8.75	34.12	39.21	11.42	35.48
Δ - bsln	↑ 0.78	↑ 0.11	↑ 0.34	↑ 4.27	↑ 1.33	↑ 4.90	↑ 1.80	↑ 1.54	↑ 2.94	↑ 4.57	↑ 2.30	↑ 2.28
+SD & SC	48.24	20.54	44.87	33.24	6.14	29.14	36.21	8.29	32.71	38.54	10.84	34.29
+SC	45.58	19.12	42.35	31.85	6.07	27.65	35.78	8.11	31.63	37.68	10.51	32.18
+SD	42.18	18.74	41.42	29.18	5.89	26.19	33.19	8.02	30.12	36.84	10.18	31.73

Table 8.1: Comparative analysis on four datasets – Multinews, CQASumm, DUC and Opinosis. We report ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) for eleven baselines. Our model FABRIC beats all the competing baselines.We also perform ablations over FABRIC. Addition of simplicial complex layer (SC) and Sheaf graph attention (SD) further improves the performance.

8.5.2 Human Evaluation Setup

We conducted a human evaluation to assess the qualitative aspects of the summaries generated by FABRIC. The evaluation focused on five parameters – Informativeness (Inf), Relevance (Rel), Coherence (Coh), Fluency (Flu), and Topic coherence (Topic). To ensure a representative evaluation, we randomly selected 50 test samples from the system-generated summaries and engaged 30 annotators[4] with expertise in the evaluation process. To minimize bias, each sample was annotated by at least two annotators, and annotations with a divergence of more than two degrees were excluded to maintain consistency. The assigned scores by the annotators were then averaged for each sample.

8.6 Experimental Results

We perform quantitative and qualitative evaluation on the four datasets. We benchmark the datasets over eleven baselines and present quantitative and qualitative results.

[4]The annotators were linguistics/subject experts and their age ranged between 25-35.

System	Multinews			CQASumm			DUC			Opinosis		
	BS (↓)	FC (↑)	Sa (↑)	BS (↓)	FC (↑)	Sa (↑)	BS (↓)	FC (↑)	Sa (↑)	BS (↓)	FC (↑)	Sa (↑)
PG	-1.71	62.4	70.4	-2.61	48.3	47.6	-3.57	45.1	70.5	-3.77	55.4	56.1
HiMAP	-1.73	65.8	71.4	-2.62	48.5	47.7	-3.58	45.2	71.4	-3.86	56.1	57.6
Transf.	-1.85	67.8	73.8	-2.68	49.8	48.1	-3.60	47.1	72.1	-3.97	58.4	59.6
BERT	-1.89	73.1	79.8	-2.73	51.7	49.7	-3.71	49.6	74.2	-4.02	60.1	62.8
BART	-2.14	76.8	81.2	-2.82	53.9	51.8	-3.95	51.8	75.6	-4.25	62.7	64.1
T5	-1.92	74.2	81.4	-2.86	51.8	50.1	-3.87	50.1	74.2	-4.13	61.2	62.8
LongT5	-2.34	75.4	83.4	-3.02	54.2	50.4	-4.31	50.8	73.5	-4.18	62.1	63.4
Pegasus	-2.51	76.2	84.2	-3.08	55.1	51.1	-4.34	51.4	76.2	-4.21	62.4	64.7
Longf.	-2.66	76.8	86.7	-3.13	56.2	52.9	-4.37	52.9	77.8	-4.28	62.7	65.8
BRIO	-2.43	75.4	85.2	-3.11	55.4	51.2	-4.21	52.5	76.4	-4.02	61.2	62.8
PRIMER	-2.73	77.3	87.1	-3.17	57.1	53.8	-4.68	53.6	78.7	-4.33	63.4	66.5
FABRIC	-2.81	78.1	88.2	-3.21	57.6	54.2	-4.74	54.9	79.1	-4.37	64.8	68.3

Table 8.2: Performance of the competing method in terms of BARTScore (BS), FactCC (FC), and SummaC (Sa). Lower BS indicates better performance.

System	Multinews						CQASumm					
	Inf	Coh	Rel	Flu	Topic	BERTopic	Inf	Coh	Rel	Flu	Topic	BERTopic
BART	3.13	2.84	2.17	2.37	2.31	2.31	2.28	1.92	2.35	2.14	2.52	1.47
Longformer	3.02	2.92	2.12	2.29	2.43	2.34	2.19	1.93	2.37	2.15	2.51	1.49
PRIMER	3.27	3.12	2.94	2.48	2.47	2.37	2.24	2.33	1.97	2.17	2.53	1.50
FABRIC	3.25	3.16	2.94	2.49	2.49	2.39	2.32	2.35	2.02	2.17	2.59	1.53

Table 8.3: Scores for five human evaluation metrics - Informativeness (Inf), Relevance (Rel), Coherence (Coh), Fluency (Flu) and Topic Coherence (Topic) and one automatic metric - BERTopic over three baselines and our proposed model, FABRIC.

8.6.1 Quantitative Evaluation

Table 8.1 presents the lexical overlap between the various generated summaries and the reference summaries. Our results demonstrate that FABRIC attains 50.68 R1 and 46.57 RL, beating the best baseline (PRIMER) by +0.78 R1 and +0.34 RL points. For CQASumm, FABRIC attains 35.81 R1 and 31.47 RL beating PRIMER by +4.27 R1 and +4.90 RL points. Similarly, for DUC and Opinosis, FABRIC attains 39.64 R1 and 34.12 RL, respectively and 34.12 R1 and 35.48 RL respectively, beating PRIMER by +1.80 R1, +4.57 R1, +2.94 RL, and +2.28 RL respectively.

To demonstrate the effectiveness of each module in FABRIC, we conduct an ablation study. The results of each ablation are presented in Table 8.1. Our base model, augmented with the simplicial complex (SC) layer, enhances the R1 and RL scores compared to the multinews baseline by +2.89 R1 and +9.14 RL points, respectively. Furthermore, the integration of sheaf graph attention contributes to additional improvements, boosting the performance by +2.66 R1 points and +2.52 RL points. These findings corroborate our hypothesis that language models alone are insufficient to capture inter-document relationships. However, when combined with context captured on higher dimension, they significantly enhance the model's ability to achieve high performance.

8.6.2 Qualitative Evaluation

We also conducted a human evaluation to assess the performance of FABRIC. Table 8.2 presents the results, demonstrating that our model generates high-quality summaries that are redundant, faithful, and easily comprehensible to humans. In addition to the human evaluation, we employed the BARTScore [213], FactCC [214], and SummaC [106] metrics to evaluate the semantic overlap and faithfulness between the system-generated summaries and the reference summaries.

Table 8.2 provides an assessment of the semantic overlap, faithfulness, and factuality of the generated summaries. The results demonstrate a significant improvement in all three qualitative metrics. In terms of semantic overlap, FABRIC achieves a score of -2.81 compared to multinews, exhibiting an improvement of $+0.08$ over the best baseline, PRIMER. Similarly, for the FactCC and SummaC metrics, FABRIC shows improvements of $+0.8$ and $+1.1$, respectively. For the CQASumm dataset, FABRIC achieves a semantic overlap of -3.21, a FactCC score of 57.6, and a SummaC score of 54.2, surpassing the best baseline, PRIMER, by $+0.04$, $+0.6$, and $+0.4$ points, respectively. Table 8.2 summarizes the improvements across all baselines on four datasets. These findings confirm our hypothesis that language models alone are insufficient to comprehend the relationships between multiple entities, necessitating the incorporation of additional modules to ensure the accuracy and coherence of the generated summaries.

Model	Output
Target-1	A teenage would-be bank robber in Detroit found himself imprisoned—and presumably regretting one or two life choices—within seconds of being handed money at a bank on Monday, police say. A police spokesman tells the Detroit Free Press that the 15-year-old gave a teller a note claiming he had a bomb and demanding money. [...]
FABRIC	A bank teller in Detroit managed to stop a would-be bank robber before he could make it to the sidewalk. Police say a 16-year-old boy walked into a Chase Bank on the cityś east side Monday afternoon and told the teller he was strapped with a bomb and to give him all the money, reports the Detroit Free Press. The teller did, and the teen set off for the doorway, "likely exhilarated by the scent of escaping with his new wealth," reports.[...]
ChatGPT	Kelly Preston, in an interview with Natalie Morales on 'Today,' discusses the "silent birth" technique commonly practiced by Scientologists. This technique involves refraining from spoken words during childbirth to create a peaceful environment. Preston credits this technique for the calm and peaceful demeanor of her children. [...]
Target-2	A preservationist in New York City thinks heś found the site of a 1750s tavern [...] Adam Woodward says he found what he believes to be hand-hewn and hand-planed joists and foundation walls in the basement.
FABRIC	So he searched the basement, and "found myself in what I am pretty certain is the 1750s historic tavern," he says. Specifically, he found what he thinks are hand-hewn and hand-planed joists and foundation walls from Colonial.
ChatGPT	Adam Woodward believes he has found evidence that a building in Manhattan could be the former site of the Bull's Head Tavern, a famous 18th-century tavern where George Washington is believed to have visited during the American Revolution.

Table 8.4: Comparison of target summary with the summary generated by FABRIC and ChatGPT.

The quantitative improvements achieved by our model are further supported by human assessments. Table 8.3 presents the results of these assessments, revealing FABRIC's consistent performance across all datasets. With the exception of informativeness on the multinews dataset compared to the PRIMER baseline, FABRIC achieves the highest scores in all metrics on multinews and CQAsumm. This indicates that the generated summaries are highly faithful, relevant, and coherent when compared to the other baselines. Although FABRIC exhibits some shortcomings in terms of informativeness according to the human evaluations, it still outperforms other baselines by a significant margin. A detailed examination of the generated summaries and an analysis of the findings can be found in Section 8.7.

8.7 Error Analysis

As indicated in Table 8.3, significant improvements are observed across all four metrics of human evaluation. The generated summaries successfully capture the essence of the source documents and cover the major topics discussed. For instance, in sample #1 in Table 8.4, the generated summary not only comprehensively conveys the main incident but also provides additional relevant information. This additional information is captured by the simplicial complex layer and the sheaf graph attention, which pass it on to the decoder for inclusion in the final summary. However, in the case of sample #2, the source document presents the information about the year 1750 as a hypothesis, while another document presents it as a quotation. This discrepancy leads FABRIC to treat this information as definitive and present it as a quoted fact in the summary. This highlights the challenge of properly handling such cases and emphasizes the notion that target summaries and quantitative metrics alone may not suffice as the true performance measure [164] for the task of abstractive summarization.

8.8 FABRIC vs ChatGPT

We conducted a comparison between the summaries generated by FABRIC and ChatGPT as shown in Table 8.4. We randomly generated 50 summaries from the multinews test set for this analysis. When comparing the two, we found that the summaries generated by ChatGPT are often overly general, lacking relevance and informativeness in relation to the source documents. For instance, in sample #1, ChatGPT discusses a general topic that deviates from the main focus by expanding on its own knowledge graph. Although it showcases linguistic capabilities, it fails to align the generated summary with the specific factual information from the source document. Similarly, in sample #2, ChatGPT provides a summary that captures the main idea of the source but neglects to mention important factual details such as the year or any correlations with specific locations. This comparison highlights that general purpose LLMs like ChatGPT have a tendency to focus on linguistic aspects but struggle to ensure fidelity to the factual information and alignment with the original source. As a result, they can be considered unfaithful as they deviate from the source and expand the generated output based on their own knowledge.

8.9 Fairness in Multimodal Systems

To address the limitations in the availability of diverse and well-annotated datasets, this research has introduced two novel datasets – AVIATE [23], and mTLDR [22], meticulously labeled through a combination of automated and manual processes. AVIATE, the first large-scale dataset for abstractive text summarization with videos of varying durations, is sourced from presentations in renowned academic conferences like NDSS, ICML, NeurIPS, and others. Comprising diverse modalities, including videos, audio, and text, AVIATE is complemented by reference summaries derived from corresponding research paper abstracts, ensuring both quality and uniformity. On the other hand, mTLDR, designed for extreme abstractive text summarization with multimodal inputs, stands as a first-of-its-kind dataset. With instances gathered from academic conference proceedings such as ICLR, ACL, and CVPR, mTLDR includes videos, audio, and text, accompanied by both author-composed and expert-annotated summaries. These datasets serve as invaluable resources for advancing research in abstractive multimodal text summarization, offering diverse challenges and insights that contribute to the ongoing evolution of summarization methodologies. Further details and analyses are presented in Chapter 5.

Simultaneously, the research presents two innovative summarization systems— FLORAL [23] and mTLDRgen [22] — meticulously designed to cater to the challenges of biased content and fidelity issues in generated summaries. In the case of mTLDRgen, designed for extreme abstractive text summarization with multimodal inputs, the model leverages a dual-fused hyper-complex Transformer and Wasserstein Riemannian Encoder Transformer. This architecture is adept at capturing intricate relationships between different modalities in a hyper-complex latent geometric space, allowing mTLDRgen to produce diverse and contextually rich summaries. On the other hand, FLORAL, dedicated to multimodal abstractive text summarization, adopts a factorized multi-modal Transformer-based decoder-only approach. FLORAL employs an increasing number of self-attentions to capture both intra-modal and inter-modal dynamics within diverse input modalities, showcasing its prowess in generating nuanced and accurate textual summaries. Both FLORAL and mTLDRgen thus represent robust approaches that push the boundaries of multimodal summarization, demonstrating superior performance over existing baselines in their respective domains. Further discussion and contributions are discussed in Chapter 5.

8.10 Conclusion

In this study, we presented `FABRIC`, a encoder-decoder model designed to enhance topic coherence and inter-document relations for multi-document abstractive summarization. The key components of `FABRIC` include BART for capturing linguistic aspects and implicial complex Layer and sheaf graph attention for capturing inter-document relationships. We compared `FABRIC` with eleven baselines on four widely popular MDS datasets. The results consistently demonstrated that `FABRIC` surpasses existing systems and achieves significant improvements across both quantitative and qualitative measures. These findings were further backed by human evaluation, validating the effectiveness of `FABRIC` in generating more accurate and faithful summaries that better represent the content of the source documents.

Chapter 9

Conclusion and Future Works

This comprehensive research work has introduced a diverse set of novel models and datasets, each addressing distinct challenges in the evolving landscape of text summarization. First, `REISA` has significantly advanced the field by introducing a reinforced attention calibration-based model. This model constrains the encoder to attend to a fixed number of tokens at each step and recalibrates attention scores, resulting in substantial improvements in both ROUGE and BERTScore metrics on Multinews and CQASumm datasets. Furthermore, its faithfulness to generated summaries is highlighted through the QAEval metric and human evaluations, positioning it as a robust solution for abstractive summarization tasks. Looking ahead, the plan to extend its application to low-resource summarization tasks and diverse domains showcases its potential for real-world applications.

In a parallel effort, `ExGrapf2`, an informative extreme summarization model, leverages Fast Fourier Transform (FFT), fractality, Graph Convolutional Networks (GCN), and contrastive loss. This model has demonstrated superior performance quantitatively and qualitatively, surpassing fifteen state-of-the-art models. Human evaluations underscore its prowess in generating highly faithful, coherent, and fluent extreme summaries, solidifying its contribution to the realm of extreme abstractive text summarization.

Addressing the challenge of multimodal abstractive text summarization, `FLORAL` and AVIATE emerge as pivotal contributions. AVIATE, the first large-scale dataset for abstractive text summarization with videos of diverse durations, enriches the resources available for training and evaluating models in multimodal scenarios. `FLORAL`, a Factorized Multimodal Transformer-based decoder-only Language Model, showcases superior performance over various baselines on How2 and AVIATE datasets. This work provides a robust foundation for the exploration of multimodal summarization, a critical aspect in the evolving landscape of NLP.

Taking a step further into the domain of extreme abstractive text summarization with multimodal inputs, `mTLDRgen` and `mTLDR` make significant strides. The curated `mTLDR` dataset, comprising videos, audio, and text, addresses the lack of benchmark datasets for extreme abstractive text summarization of scientific videos. `mTLDRgen`, utilizing a dual fused hyper-complex Transformer and Wasserstein Riemannian Encoder Transformer, exhibits superior performance qualitatively and quantitatively. These contributions expand the capabilities of summarization models to handle diverse modalities and extreme summarization tasks, contributing to the broader applicability of NLP technologies.

Further, a dedicated investigation into heterogeneous multi-document summarization illuminates the intricate challenge of summarizing non-independent candidate documents. This study proposes intrinsic metrics for evaluating corpus quality and offers insights into system behavior across diverse corpora. Establishing a foundational understanding of biases in system predictions, this research paves the way for future endeavors in comprehending and mitigating biases. The envisioned future work entails a causal analysis of corpus bias and its impact on model predictions, promising measures to de-bias NLP algorithms both with and without de-biasing the corpora.

In parallel, our study introduced `FABRIC`, an encoder-decoder model tailored for multi-document abstractive summarization, with a focus on enhancing topic coherence and inter-document relations. Comprising BART for linguistic aspects and simplicial complex layers with sheaf graph attention for inter-document relationships, `FABRIC` underwent rigorous evaluation against eleven baselines on four widely used MDS datasets. The consistently superior performance of `FABRIC` across quantitative metrics, qualitative assessments, and human evaluations reaffirms its effectiveness in generating more accurate and faithful summaries that authentically represent the content of source documents.

In the area of future work, several promising avenues emerge from the current research landscape, offering opportunities to enhance and extend the capabilities of text summarization models. One potential trajectory involves a dedicated exploration of bias in multimodal summarization. Understanding and mitigating biases in system predictions, particularly when handling diverse modalities, is crucial for ensuring fair and equitable summarization outputs. Future work in this area could involve a comprehensive causal analysis of corpus bias and its impact on model predictions. Additionally, measures to de-bias NLP algorithms, both with and without de-biasing the corpora, present promising directions for advancing the fairness and reliability of multimodal summarization systems. Moreover, incorporating interactive and controlled generation mechanisms into multimodal summarization models could empower users to tailor summaries according to their preferences and specific information needs, opening new avenues for user-centric summarization experiences.

Another noteworthy future direction is the proposal and development of more fair multimodal summarization systems. Addressing issues related to bias and fairness, this line of work could involve the incorporation of fairness-aware algorithms, explicitly designed to mitigate biases and ensure equitable representations in generated summaries. By introducing fairness as a key criterion in the evaluation and design of multimodal summarization models, future research can contribute to more ethically grounded and inclusive NLP technologies. Furthermore, a novel and intriguing prospect lies in the exploration of proposing multimodal summaries. Going beyond unimodal text summaries, the development of methodologies to generate coherent and informative summaries that seamlessly integrate information from diverse modalities, such as text, image, audio, and video, represents an exciting avenue for future investigation. This research direction could open new possibilities for creating more comprehensive and contextually rich summarization outputs, catering to the evolving demands of multimodal content in various real-world applications.

In the quest for explainable AI, future work could focus on making summarization systems more transparent and interpretable. Exploring methodologies to provide clear insights into how models arrive at specific summarization decisions can enhance user trust and understanding. Developing interpretable summarization models will not only contribute to the responsible deployment of NLP technologies but also facilitate effective collaboration between AI systems and human users.

Moreover, a compelling challenge for future research involves the development of a unified model capable of handling both text-only and multimodal inputs seamlessly. Creating a versatile model that can effectively process and generate summaries for diverse types of input data, ranging from traditional text to complex multimodal content, holds immense potential. Such a model would be instrumental in addressing the growing need for versatile summarization systems capable of handling the evolving nature of information across various modalities and sources.

Real-time summarization is another exciting direction, where models capable of performing real-time summarization for live events or streaming content are developed. This requires highly efficient algorithms that can process and summarize content on-the-fly without significant delays, valuable for applications in news broadcasting, live sports events, and virtual conferences. Personalized summarization, focusing on tailored summaries based on user preferences and specific information needs, is also promising. This could involve user profiles, interests, and past behavior to deliver more relevant and engaging summaries, potentially integrating recommendation system techniques. Cross-lingual summarization, where input and output languages differ, is another area worth exploring, beneficial for multilingual societies and global information dissemination.

Enhancing robustness to noisy or imperfect data is crucial for real-world applications. Future work could explore techniques to improve model performance with incomplete, ambiguous, or erroneous inputs, including error detection and correction methods. Developing new evaluation metrics that better capture summary quality, going beyond current metrics like ROUGE and BLEU, is also important. These new metrics should assess aspects such as fluency, factual correctness, coherence, and relevance. Human-in-the-loop systems, incorporating human feedback into the summarization process, can significantly improve model performance and user satisfaction by refining summaries to better meet user expectations.

Domain adaptation and transfer learning techniques to improve summarization models across various domains is another valuable direction. Models trained in one domain often struggle when applied to different domains with distinct characteristics. Developing methods to transfer knowledge and adapt models to new domains with minimal additional training would enhance versatility. Addressing ethical considerations and bias mitigation remains critical, ensuring summarization models do not propagate harmful biases or generate discriminatory content. Future work could involve developing frameworks to identify and mitigate biases, ensuring fair and unbiased summaries across different demographics and content types.

Creating interactive summarization tools that allow users to customize and refine summaries based on their specific needs, such as highlighting important sections, asking for more details on specific topics, or adjusting summary length and focus dynamically, can greatly enhance usability. Integrating summarization models with knowledge graphs to improve the extraction and presentation of key information is another area of interest. Knowledge graphs provide contextual information and relationships between entities, enhancing the accuracy and coherence of summaries, opening new possibilities for creating more comprehensive and contextually rich summarization outputs, catering to the evolving demands of multimodal content in various real-world applications.

Bibliography

[1] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014.

[2] Sam Shleifer and Alexander M. Rush. Pre-trained summarization distillation, 2020.

[3] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. *Elsevier Information Fusion Journal*, 2020.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[5] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online, November 2020. Association for Computational Linguistics.

[6] Maxime Peyrard. A simple theoretical model of importance for summarization. In *ACL*, pages 1059–1073, 2019.

[7] Alvin Dey, Tanya Chowdhury, Yash Kumar, and Tanmoy Chakraborty. Corpora evaluation and system bias detection in multi-document summarization. In *EMNLP Findings*, pages 2830–2840, Online, November 2020. ACL.

[8] Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors, 2023.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023.

[11] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models, 2023.

[12] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *ACL*, pages 1074–1084. ACL, July 2019.

[13] Tanya Chowdhury and Tanmoy Chakraborty. Cqasumm: Building references for community question answering summarization corpora. In *CODS-COMAD*, page 18–26. ACM, 2019.

[14] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*, 2022.

[15] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. Bottom-up abstractive summarization. In *EMNLP*, pages 4098–4109. ACL, October-November 2018.

[16] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France, May 2020. ELRA.

[17] Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 6587–6596, Florence, Italy, July 2019. ACL.

[18] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171, 2017.

[19] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. ACL. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[20] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Neur*, NIPS'15, page 1693–1701. MIT Press, 2015.

[21] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*, pages 26.1–26.12. NeurIPS, 2018.

[22] Yash Kumar Atri, Vikram Goyal, and Tanmoy Chakraborty. Fusing multimodal signals on hyper-complex space for extreme abstractive text summarization (tl; dr) of scientific contents. *arXiv preprint arXiv:2306.13968*, 2023.

[23] Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*, 227:107152, 2021.

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Neur*, NIPS'14, page 3104–3112, 2014.

[25] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[26] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421. ACL, September 2015.

[27] Junyang Lin, Xu Sun, Xuancheng Ren, Muyu Li, and Qi Su. Learning when to concentrate or divert attention: Self-adaptive attention temperature for neural machine translation. In *EMNLP*, pages 2985–2990. ACL, October-November 2018.

[28] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In *EMNLP*, pages 866–874. ACL, September 2017.

[29] Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *ACL, pp. 208–224.*

[30] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, ICML'15, page 2048–2057. JMLR.org, 2015.

[31] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. Entity-aware image caption generation. In *EMNLP*, pages 4013–4023, October-November 2018.

[32] Chenyang Huang, Osmar Zaïane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *NACL:HLT, pp. 49–54.*

[33] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *EMNLP*, pages 1192–1202. ACL, November 2016.

[34] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. Abstractive document summarization with a graph-based attentional neural model. In *ACL, pp. 1171–1181.*

[35] Hardy Hardy and Andreas Vlachos. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *EMNLP*, pages 768–773. ACL, October-November 2018.

[36] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, pages 11328–11339, Jul 2020.

[37] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *SIGNLL-CNLL*, pages 280–290. ACL, August 2016.

[38] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. pages 1073–1083. ACL, July 2017.

[39] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. pages 1693–1701, 2015.

[40] Logan Lebanoff, Kaiqiang Song, and Fei Liu. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *EMNLP*, pages 4131–4141. ACL, October-November 2018.

[41] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *NACL-HLT, pp. 615–621.*

[42] Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. Efficiently summarizing text and graph encodings of multi-document clusters. In *NACL: HLT*, pages 4768–4779, Online, June 2021. ACL.

[43] Tanya Chowdhury, Sachin Kumar, and Tanmoy Chakraborty. Neural abstractive summarization with structural attention. In *IJCAI, np. 7*, IJCAI'20.

[44] Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Ziqiang Cao, Sujian Li, Hua Wu, and Haifeng Wang. BASS: Boosting abstractive summarization with unified semantic graph. In *ACL)*, pages 6052–6067, Online, August 2021. ACL.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, volume 30, 2017.

[46] Shusheng Xu, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. Unsupervised extractive summarization by pre-training hierarchical transformers. In *EMNLP Findings*, pages 1784–1795, Online, November 2020. ACL.

[47] Shashi Narayan, Joshua Maynez, Jakub Adamek, Daniele Pighin, Blaz Bratanic, and Ryan McDonald. Stepwise extractive summarization and planning with structured transformers. In *EMNLP, pp. 4143–4159.*

[48] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *NeurIPS*, volume 33, pages 17283–17297, 2020.

[49] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv e-prints*, page arXiv:2004.05150, April 2020.

[50] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3730–3740. ACL, November 2019.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, June 2019.

[52] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[53] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

[54] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *NACL: HLT*, pages 704–717, Online, June 2021. ACL.

[55] Yi-Syuan Chen and Hong-Han Shuai. Meta-transfer learning for low-resource abstractive summarization. *AAAI*, 35(14):12692–12700, May 2021.

[56] Travis Goodwin, Max Savery, and Dina Demner-Fushman. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. In *EMNLP Findings*, pages 3215–3226, Online, November 2020. ACL.

[57] Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K. Reddy. Deep reinforcement learning for sequence-to-sequence models. *IEEE Trans. Neural Netw. Learn. Syst.*, (7):2469–2489, 2020.

[58] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Ranking sentences for extractive summarization with reinforcement learning. In *NACL: HLT*, pages 1747–1759. ACL, June 2018.

[59] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. BanditSum: Extractive summarization as a contextual bandit. In *EMNLP*, pages 3739–3748. ACL, October-November 2018.

[60] Gyoung Ho Lee and Kong Joo Lee. Automatic text summarization using reinforcement learning with embedding features. In *IJCNLP*, pages 193–197, November 2017.

[61] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304, 2017.

[62] Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. The summary loop: Learning to write abstractive summaries without examples. In *ACL*, pages 5135–5150, Online, July 2020. ACL.

[63] Ramakanth Pasunuru and Mohit Bansal. Multi-reward reinforced summarization with saliency and entailment. In *NACL: HLT*, pages 646–653, June 2018.

[64] Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. In *EMNLP-IJCNLP*, pages 3110–3120. ACL, November 2019.

[65] Alexander R. Fabbri2021MultiPerspectiveAA, Xiaojian Wu, Srini Iyer, and Mona T. Diab. Multi-perspective abstractive answer summarization. *CoRR*, abs/2104.08536, 2021.

[66] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *ACL*, pages 5055–5070, Online, July 2020. ACL.

[67] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to Learn Automatic Summarization. In Giuseppe Carenini, Jackie Chi Kit Cheung, Fei Liu, and Lu Wang, editors, *Workshop on New Frontiers in Summarization at EMNLP 2017*, pages 59–63. Association for Computational Linguistics, September 2017.

[68] Yuning Mao, Ming Zhong, and Jiawei Han. Citesum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision, 2022.

[69] Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. In *ACL*, pages 5070–5081. ACL, July 2019.

[70] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019.

[71] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[72] Yash Atri, Arun Iyer, Tanmoy Chakraborty, and Vikram Goyal. Promoting topic coherence and inter-document consorts in multi-document summarization via simplicial complex and sheaf graph. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2154–2166, Singapore, December 2023. Association for Computational Linguistics.

[73] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[74] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*, 2019.

[75] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[76] Athar Sefid and C. Lee Giles. Scibertsum: Extractive summarization for scientific documents. In Seiichi Uchida, Elisa Barney, and Véronique Eglin, editors, *Document Analysis Systems*, pages 688–701, Cham, 2022. Springer International Publishing.

[77] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[78] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

[79] Peng Cui, Le Hu, and Yuanchao Liu. Enhancing extractive text summarization with topic-aware graph neural networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5360–5371, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[80] Chen An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Enhanccing scientific papers summarization with citation graph. In *AAAI Conference on Artificial Intelligence*, 2021.

[81] Jingqiang Chen, Chaoxiang Cai, Xiaorui Jiang, and Kejia Chen. Comparative graph-based summarization of scientific papers guided by comparative citations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5978–5988, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[82] Sajad Sotudeh and Nazli Goharian. GUIR @ MuP 2022: Towards generating topic-aware multi-perspective summaries for scientific documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 273–278, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics.

[83] Arman Cohan, Guy Feigenblat, Tirthankar Ghosal, and Michal Shmueli-Scheuer. Overview of the first shared task on multi perspective scientific document summarization (mup). In *SDP*, 2022.

[84] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on EMNLP*, pages 1092–1102, 2017.

[85] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):996–1009, 2018.

[86] Marco Baroni. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13, 2016.

[87] Douwe Kiela. Deep embodiment: grounding semantics in perceptual modalities. Technical report, University of Cambridge, Computer Laboratory, 2017.

[88] Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. Exercise? i thought you said'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 513–524, 2021.

[89] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online, August 2021. Association for Computational Linguistics.

[90] Jingqiang Chen and Hai Zhuge. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 Conference on EMNLP*, pages 4046–4056, 2018.

[91] Zechao Li, Jinhui Tang, Xueming Wang, Jing Liu, and Hanqing Lu. Multimedia news summarization in search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):1–20, 2016.

[92] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018.

[93] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on EMNLP*, pages 4154–4164, Brussels, Belgium, October-November 2018. ACL.

[94] Tsutomu Hirao, Takahiro Fukusima, Manabu Okumura, Chikashi Nobata, and Hidetsugu Nanba. Corpus and evaluation measures for multiple document summarization with multiple sources. In *ICCL*, COLING '04, page 535–es. ACL, 2004.

[95] Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *COLING*, pages 1039–1050, December 2016.

[96] Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *EMNLP-IJCNLP*, pages 3324–3335, November 2019.

[97] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics.

[98] Maxime Peyrard. Studying summarization evaluation metrics in the appropriate scoring range. In *ACL*, pages 5093–5100, 2019.

[99] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *NAACL*, pages 708–719, June 2018.

[100] Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[101] Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[102] Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Ziqiang Cao, Sujian Li, and Hua Wu. FRSUM: Towards faithful abstractive summarization via enhancing factual robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3640–3654, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[103] Alex Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. Improving factual consistency in summarization with compression-based post-editing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9149–9156, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[104] Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[105] Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim, and Kyomin Jung. Factual error correction for abstractive summaries using entity

retrieval. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 439–444, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[106] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Revisiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

[107] Meng Cao, Yue Dong, Jingyi He, and Jackie Chi Kit Cheung. Learning with rejection for abstractive text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9768–9780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[108] Asish Ghoshal, Arash Einolghozati, Ankit Arun, Haoran Li, Lili Yu, Yashar Mehdad, Scott Wen tau Yih, and Asli Celikyilmaz. Improving faithfulness of abstractive summarization by controlling confounding effect of irrelevant sentences, 2022.

[109] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online, June 2021. Association for Computational Linguistics.

[110] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online, June 2021. Association for Computational Linguistics.

[111] Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535, Seattle, United States, July 2022. Association for Computational Linguistics.

[112] Yuanjie Lyu, Chen Zhu, Tong Xu, Zikai Yin, and Enhong Chen. Faithful abstractive summarization via fact-aware consistency-constrained transformer. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1410–1419, New York, NY, USA, 2022. Association for Computing Machinery.

[113] Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online, July 2020. Association for Computational Linguistics.

[114] Lulu Zhao, Zeyuan Yang, Weiran Xu, Sheng Gao, and Jun Guo. Improving abstractive dialogue summarization with conversational structure and factual knowledge, 2021.

[115] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. Factually consistent summarization via reinforcement learning with textual entailment feedback, 2023.

[116] Mengli Zhang, Gang Zhou, Wanting Yu, and Wenfen Liu. Far-ass: Fact-aware reinforced abstractive sentence summarization. *Information Processing Management*, 58(3):102478, 2021.

[117] Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, page 457–479, dec 2004.

[118] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *EMNLP*, pages 404–411. ACL, July 2004.

[119] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, page 1606–1618, nov 2007.

[120] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *NACL-HLT*, pages 362–370. ACL, June 2009.

[121] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *ACL:HLT, pp. 93–98.*

[122] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL, pp. 1631–1640.*

[123] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *ACL, pp. 76–85*, pages 76–85. ACL, August 2016.

[124] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *NeurIPS*, volume 28. Curran Associates, Inc., 2015.

[125] Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In *ACL: HLT*, pages 1419–1436, Online, June 2021. ACL.

[126] Yichen Jiang, Asli Celikyilmaz, Paul Smolensky, Paul Soulos, Sudha Rao, Hamid Palangi, Roland Fernandez, Caitlin Smith, Mohit Bansal, and Jianfeng Gao. Enriching transformers with structured tensor-product representations for abstractive summarization. In *ACL: HLT*, pages 4780–4793, Online, June 2021. ACL.

[127] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. Discourse-aware neural extractive text summarization. In *ACL, pp. 5021–5031.*

[128] Potsawee Manakul and Mark Gales. Long-span summarization via local attention and content selection. In *ACL*, pages 6026–6041, Online, August 2021. ACL.

[129] Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Walter Chang, and Fei Liu. A cascade approach to neural abstractive summarization with content selection and fusion. In *AACL*, pages 529–535. ACL, December 2020.

[130] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *ACL*, pages 2214–2220. ACL, July 2019.

[131] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *EMNLP-IJCNLP*, pages 540–551, Hong Kong, China, November 2019. ACL.

[132] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *ACL*, pages 1906–1919, Online, July 2020. ACL.

[133] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. ACL, July 2004.

[134] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020.

[135] Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. Deep reinforcement learning with distributional semantic rewards for abstractive summarization. In *EMNLP-IJCNLP*, pages 6038–6044. ACL, November 2019.

[136] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences, 2021.

[137] Peng Cui and Le Hu. Topic-guided abstractive multi-document summarization. In *EMNLP Findings*, pages 1463–1472. ACL, November 2021.

[138] Hangyu Mao, Zhengchao Zhang, Zhen Xiao, and Zhibo Gong. Modelling the dynamic joint policy of teammates with attention multi-agent ddpg. In *AAMAS*, AAMAS '19, page 1108–1116, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[139] P. Parnika, Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, and Shalabh Bhatnagar. *Attention Actor-Critic Algorithm for Multi-Agent Constrained Co-Operative Reinforcement Learning.* 2021.

[140] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992.

[141] Tsutomu Hirao, Masaaki Nishino, Jun Suzuki, and Masaaki Nagata. Enumeration of extractive oracle summaries. In *EACL, pp. 386–396.*

[142] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[143] Daniel Gillick, Benoit Favre, and Dilek Hakkani-Tür. The icsi summarization system at tac 2008. In *Tac*, 2008.

[144] Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. page 4453–4460. AAAI, 2018.

[145] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, WWW7, page 107–117, NLD, 1998. Elsevier Science Publishers B. V.

[146] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[147] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *TACL*, 9:774–789, 2021.

[148] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642. ACL, September 2015.

[149] John Ioannidis, Richard Klavans, and Kevin W Boyack. Thousands of scientists publish a paper every five days, 2018.

[150] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313, Seattle, United States, July 2022. Association for Computational Linguistics.

[151] Benoit Mandelbrot. *Fractals*. Freeman San Francisco, 1977.

[152] Elham Najafi and Amir H Darooneh. The fractal patterns of words in a text: a method for automatic keyword extraction. *PloS one*, 10(6):e0130617, 2015.

[153] M. Dolores Ruiz and Antonio B. Bailón. Summarizing structured documents through a fractal technique. In Joaquim Filipe, José Cordeiro, and Jorge Cardoso, editors, *Enterprise Information Systems*, pages 328–340, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[154] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[155] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[156] B. CAPRILE, A.C. LEVI, and L. LIGGIERI. Random rain simulations of dendritic growth. In Luciano PIETRONERO and Erio TOSATTI, editors, *Fractals in Physics*, pages 279–282. Elsevier, Amsterdam, 1986.

[157] Hao Zheng and Mirella Lapata. Sentence centrality revisited for unsupervised summarization. *arXiv preprint arXiv:1906.03508*, 2019.

[158] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[159] Koray Tuğberk GÜBÜR. Extractive summarization with bert extractive summarizer, Sep 2022.

[160] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*, 2020.

[161] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[162] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[163] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[164] Lei Li, Wei Liu, Marina Litvak, Natalia Vanetik, Jiacheng Pei, Yinan Liu, and Siya Qi. Subjective bias in abstractive summarization, 2021.

[165] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14(4):975–985, 2012.

[166] Jonghwan Mun, Minsu Cho, and Bohyung Han. Text-guided attention model for image captioning. In *AAAI*, pages 4233–4239, 2017.

[167] Sheng Liu, Zhou Ren, and Junsong Yuan. Sibnet: Sibling convolutional encoder for video captioning. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1425–1434, New York, NY, USA, 2018. ACM.

[168] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on CVPR Workshops*, pages 958–959, 2020.

[169] Xiangxi Shi, Jianfei Cai, Jiuxiang Gu, and Shafiq Joty. Video captioning with boundary-aware hierarchical language decoding and joint video prediction. *Neurocomputing*, 417:347–356, 2020.

[170] Xiangqing Shen, Bing Liu, Yong Zhou, Jiaqi Zhao, and Mingming Liu. Remote sensing image captioning via variational autoencoder and reinforcement learning. *Knowledge-Based Systems*, page 105920, 2020.

[171] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on EMNLP*, pages 968–974, 2018.

[172] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Title generation for user generated videos. In *European conference on computer vision*, pages 609–625. Springer, 2016.

[173] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh IJCAI-18*, pages 4152–4158. IJCAI, 7 2018.

[174] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[175] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *CoRR*, 2017.

[176] Joseph Tepperman, David Traum, and Shrikanth Narayanan. Yeah right: Sarcasm recognition for spoken dialogue systems. pages 1838–1841, 01 2006.

[177] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). pages 4619–4629, 01 2019.

[178] Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Vincent Lostanlen, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, et al. Wzy. *Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stter, Matt Vollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis" Fjord" Hawthorne, CJ Carr, Joo Felipe Santos, JackieWu, Erik, and Adrian Holovaty,"librosa/librosa: 0.6*, 2, 2018.

[179] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[180] Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. Sample Efficient Text Summarization Using a Single Pre-Trained Transformer. *arXiv preprint arXiv:1905.08836*, pages 1,7, 2019.

[181] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[182] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57, 2014.

[183] Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the ACL (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada, July 2017. ACL.

[184] Yvette Graham. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on EMNLP*, pages 128–137, 2015.

[185] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. ACL.

[186] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. ACL.

[187] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[188] Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters, 2021.

[189] Prince Zizhuang Wang and William Yang Wang. Riemannian normalizing flow on variational Wasserstein autoencoder for text modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 284–294, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[190] Ilya O Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[191] James M. Joyce. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[192] Tanya Chowdhury and Tanmoy Chakraborty. Cqasumm: Building references for community question answering summarization corpora. Association for Computing Machinery, 2019.

[193] DUC. Document Understanding Conferences. [online] Available at: https://duc.nist.gov/.

[194] TAC. Text Analysis Conferences. [online] Available at: https://tac.nist.gov/.

[195] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Coling*, pages 340–348. Coling, August 2010.

[196] Tsutomu Hirao, Masaaki Nishino, Jun Suzuki, and Masaaki Nagata. Enumeration of extractive oracle summaries. In *EACL*, pages 386–396, April 2017.

[197] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[198] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[199] Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1462–1471, New York, NY, USA, 2021. Association for Computing Machinery.

[200] Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 80–89, Hong Kong, China, November 2019. Association for Computational Linguistics.

[201] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538, 2018.

[202] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[203] Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D. Bui, and Khoa Luu. Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10842–10851, 2021.

[204] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[205] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *ACL*, pages 1906–1919, Online, July 2020. ACL.

[206] OpenAI. Gpt-4 technical report, 2023.

[207] Qianren Mao, Jianxin Li, Hao Peng, Shizhu He, Lihong Wang, Philip S. Yu, and Zheng Wang. Fact-driven abstractive summarization by utilizing multi-granular multi-relational knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1665–1678, 2022.

[208] Jiaao Chen and Diyi Yang. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online, June 2021. Association for Computational Linguistics.

[209] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[210] Yunchong Song, Chenghu Zhou, Xinbing Wang, and Zhouhan Lin. Ordered GNN: Ordering message passing to deal with heterophily and over-smoothing. In *The Eleventh International Conference on Learning Representations*, 2023.

[211] L. Giusti, C. Battiloro, P. Di Lorenzo, S. Sardellitti, and S. Barbarossa. Simplicial attention neural networks, 2022.

[212] Jakob Hansen and Thomas Gebhart. Sheaf neural networks, 2020.

[213] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021.

[214] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics.

[215] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[216] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[217] Daniel Gillick, Benoit Favre, and Dilek Z. Hakkani-Tür. The icsi summarization system at tac 2008. *Theory and Applications of Categories*, 2008.

[218] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[219] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland, May 2022. Association for Computational Linguistics.