Multimodal Sarcasm Explanation

by

Poorav DESAI
MT19010

Under the Supervision of
Dr. Md. Shad Akhtar,
Dr. Tanmoy Chakraborty

Indraprastha Institute of Information Technology Delhi
August, 2021

Multimodal Sarcasm Explanation

by

Poorav DESAI
MT19010

Submitted
in partial fulfillment of the requirements for the degree of
Master of Technology in Computer Science & Engineering
(Specialization in AI)

to

Indraprastha Institute of Information Technology Delhi
August, 2021

Certificate

This is to certify that the thesis titled "**Multimodal Sarcasm Explanation**" being submitted by **Poorav DESAI** to the Indraprastha Institute of Information Technology Delhi, for the award of the *Master of Technology in Computer Science & Engineering (Specialization in AI)*, an original research work carried out by him under our supervision. In our opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2021

Dr. Md. Shad Akhtar
Department of Computer Science & Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Dr. Tanmoy Chakraborty
Department of Computer Science & Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgements

I would like to express my sincere gratitude and indebtedness to Dr. Md Shad Akhtar for his exemplary guidance, supervision and constant encouragement throughout.

I am grateful and remain indebted to Dr. Tanmoy Chakraborty for his constant support and his insightful comments and suggestions at every stage of the research project.

I would also like to thank esteemed committee members Dr. Mukesh Mohania and Dr. Koteswar Rao Jerripothula for evaluating my thesis work.

I would also like to thank members of LCS2 lab for being a constant source of motivation.

Finally, I would like to thank my supportive family and friends who encouraged me and kept me motivated throughout the Thesis.

Abstract

Sarcasm is a pervading linguistic phenomenon and highly challenging to explain due to its subjectivity, lack of context and deeply-felt opinion. In the multimodal setup, sarcasm is conveyed through the incongruity between the text and visual entities. Although recent approaches consider it as a classification problem, it is unclear why an online post is identified as sarcastic. Without proper explanation, end users may not be able to perceive the underlying use of irony. In this paper, we propose a novel problem – **Multimodal Sarcasm Explanation** (MSE) – given a multimodal sarcastic post containing an image and a caption, we aim to generate a natural language explanation to reveal the intended sarcasm. To this end, we develop a novel dataset, MORE, with explanation for 3510 sarcastic multimodal posts. Each explanation is a natural language (English) sentence that describes the hidden irony. We then propose EXMORE, a multimodal transformer-based architecture to address MSE. It incorporates a cross-modal attention in transformer's encoder which attends the distinguishing features between two modalities. Subsequently, a BART-based auto-regressive decoder is used as the generator. Empirical results demonstrate the efficacy of EXMORE over six baselines (adopted for MSE) and shows $>$ 10% improvement compared to the best baseline across five evaluation metrics.

Contents

List of Figures

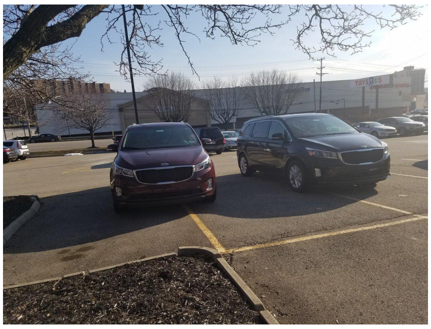
List of Tables

Chapter 1

Introduction

According to merriam webster[1], Sarcasm refers to the use of words that mean the opposite of what you really want to say, especially in order to insult someone, or to show irritation, or just to be funny. For example, saying "they're really on top of things" to describe a group of people who are very disorganized is using sarcasm. Understanding sarcasm is challenging due to its subjective nature, importance of context and deeply-felt opinions. The advance in technology and rise of social media platforms have led to generation of a huge amount of multimodal content in various forms such as social media posts, messages, product reviews, forums and more on daily basis. This means sarcasm can be found in abundance in such a variety of forms. In such situation, it becomes essential to not only detect the sarcasm but also understand the reason "why something is sarcastic?". For various applications ranging from product reviews, sentiment analysis tools, sensitive social media posts and more, understanding the real intended meaning of the speaker is crucial. Hence, this gives a rise to need for explaining the sarcasm in the detected sarcastic content. As mentioned earlier, context plays an important role in, first, deciding whether the content is sarcastic and then understanding why so, if it is. Given the increasing generation of multimodal content consisting of audiovisual aspects along with the text, it becomes necessary to considered other modes in addition to text to get the context in its entirety. In addition, sarcasm can be arising from different modalities and this makes it insufficient to rely on just one modality. Sarcasm detection is a well known problem and it has been actively studied by the research community. However, the task of explaining the sarcasm by revealing the intended meaning of the speaker has not been explored in the past. In this work, we introduce a novel task of Multimodal Sarcasm Explanation (MSE) which aims at generating a natural language explanation for a given multimodal sarcastic post explaining why the given post is sarcastic. We consider image and text modality to better understand the context of the post. Figure 1.1 shows an instance where, in the absence of the image, it would not be possible to decide if the post is sarcastic, since the caption alone can be a genuine complement as well. This shows the importance of context and the role of image modality to capture it. To address this novel task we curate a new dataset, MORE, and to benchmark it we also propose a transformer based encoder-decoder model, EXMORE, which considers image and text modality to generate an explanation for a given multimodal sarcastic post.

[1] https://www.merriam-webster.com/dictionary/sarcasm



Caption: nice parking jobs, guys.
Explanation: these are bad parking jobs, the cars are out of the slots.

FIGURE 1.1: An instance of MSE task showing importance of the image modality.

1.1 Problem Definition

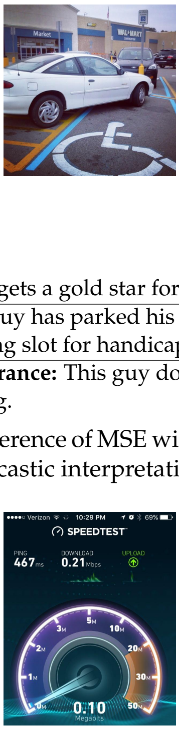
In this work, we introduce a novel task of multimodal sarcasm explanation with the objective of revealing the intended irony in a sarcastic multimodal post. Unlike other explainable systems that uses attention heatmaps or similar mechanisms (e.g., SHAP(Lundberg and Lee, 2017), LIME(Ribeiro, Singh, and Guestrin, 2016), etc.) to explain the model behaviour, we deal with the explanation as the natural language generation task. Thus, MSE's output should be a cohesive and coherent English sentence. Furthermore, we draw the difference between MSE and the non-sarcastic interpretation task proposed by Dubey, Joshi, and Bhattacharyya, 2019 in Figure 1.2a. The first difference is the incorporation of multimodality in MSE compared to the text-based non-sarcastic interpretation. The second and the prime difference is that the non-sarcastic interpretations (Dubey, Joshi, and Bhattacharyya, 2019) are primarily the negation of the sarcastic text. In comparison, MSE is primarily defined to explain the incongruity not necessarily with the use of negation; however, we have a few examples for which the explanation can be termed as the non-sarcastic interpretation (c.f. Figure 1.2b).

We formulate MSE as follows: For a given multimodal sarcastic post $P = < I,\ T[t_1, t_2, ..., t_N] >$, we aim to reveal the intended irony by generating a natural language explanation $E[e_1, e_2, ..., e_D]$, where, $\forall t_i, e_j \in Vocab^{English}$.

1.2 An Overview of the Research

To address the novel task of MSE, we develop a new dataset, called MORE. We, then, propose EXMORE, a Transformer-based neural architecture that considers the image and text modality of a sarcastic post for generating an explanation. We compare the performance of EXMORE against six existing systems adapted for MSE, and observe that it performs significantly better than all of them.

Our main contributions are as follows:



Caption: This guy gets a gold star for excellent parking.

Explanation: this guy has parked his car partially covering the parking slot for handicapped.

Non-Sarcastic utterance: This guy does not get a gold star for bad parking.

(A) Difference of MSE with non-sarcastic interpretation

Caption: Internet is awesome. i love my dialup internet! show me the 70mbps! #wtf.

Explanation: This internet is terrible, I hate dialup internet.

Non-Sarcastic utterance: This internet is not awesome, I hate dial up internet.

(B) Similarity of MSE and non-sarcastic interpretation.

FIGURE 1.2: Example scenarios showing the multimodal sarcasm explanation task compared to the (textual) non-sarcastic interpretation task (**joshi:non-sarcastic:2019**). Non-sarcastic interpretation is primarily negation of caption.

- We introduce MSE, a new task that aims at generating a natural language explanation for a given sarcastic post explaining why the given post is sarcastic.

- We introduce the MORE dataset consisting of 3510 triples (image, caption and explanation) for the MSE task.

- We propose EXMORE, a transformer based model. The experimental results show that it outperforms all the 6 baseline models, chosen from related tasks, across 5 evaluation metrics on the MORE dataset.

- We perform analysis which includes Linguistic analysis and Human evaluation which further shows the superiority of EXMORE over the best baseline model.

Chapter 2

Background and Related Work

Sarcasm analysis is an important component of sentiment analysis, where several studies explore different facets of the problem. Most prior work focuses on detection of sarcasm using one or more modalities, while a few explore the problem of converting the instances into their non-sarcastic utterances. The task of explaining sarcasm by highlighting the underlying meaning of the speaker has not been explored in the past. Sarcasm detection helps in semantic understanding of utterances that are sarcastic and what differentiates these with non sarcastic utterances, while on the other hand conversion of sarcastic to non sarcastic utterances gives an insight into what the utterance would be if quoted in a non sarcastic tone. Related generative tasks such as Summarization and Machine Translation are closely related with the generative nature of the task undertaken.

2.1 Sarcasm Detection

Various methods have been proposed for the problem of Sarcasm detection (Bedi et al., 2021), earlier methods including (Bouazizi and Otsuki Ohtsuki, 2016) and (Felbo et al., 2017) use hand engineered feature representations such as punctuation marks, Part-Of-Speech tags, use of emojis, emotion lexicons etc. Recently studies have pushed more for more complex approaches, utilizing several different inputs like video, speech and images. Castro et al., 2019 introduced a new dataset named MUStARD for the task of multimodal sarcasm detection, their experiments highlighted the contributions of the additional modalities in improving the performance for sarcasm detection. Schifanella et al., 2016 use images along with the corresponding captions from three platforms- Instagram, Tumblr and Twitter, for sarcasm detection, and through their study highlight the role of images in sarcastic posts. Y. Cai, H. Cai, and Wan, 2019 use a hierarchical fusion for the text and image modalities, using two stages of fusion of image and text representations, for extracting a fused vector for the task of classification.

2.2 Sarcastic Utterances into their Non-Sarcastic Interpretation

The task of conversion of Sarcastic Utterances into their Non-Sarcastic utterance hasn't been explored extensively. A few key studies like (Peled and Reichart, 2017) and (Dubey, Joshi, and Bhattacharyya, 2019) use a parallel corpus consisting of tweets containing a #sarcasm for sarcastic utterances, and then use human annotated non sarcastic utterances for each. Both employ systems based on machine translation like MOSES (Koehn et al., 2007) and Neural machine translation systems made using RNNs and Attention networks. They also employ systems for summarization (Pointer generator networks See, P. J. Liu, and Manning, 2017) for the task considering summarization as monolingual machine translation.

2.3 Related generative tasks

In the field of natural language processing, Machine translation and Summarization are popular and well studied problems that fall under the category of generative tasks. Recent approaches have also focused on harnessing multimodality for these tasks. Given the generative nature of the proposed task which also deals with multimodal input, it can be related to the machine translation and summarization tasks.

Machine translation The task undertaken can be modelled as machine translation wherein the domain for the input is the set of all sarcastic utterances and for the output is the corresponding explanation. Bahdanau, Cho, and Bengio, 2016 use RNNs with attention mechanism to jointly learn to align and translate. Yao and Wan, 2020 propose multimodal transformer model with cross-modal attention mechanism over image and text for the task of multimodal machine translation.

Summarization Summarization can also be interpreted as monolingual machine translation where the long text input is translated into a summarized text and thus directly follows from the above related task of machine translation. Alternatively, the task undertaken can also be interpreted as producing the explanation as the summary of the given sarcastic input. See, P. J. Liu, and Manning, 2017 propose pointer generator networks with copy mechanism which facilitates for generation of unknown tokens in the input text for summarization. Exploring other modalities, Li et al., 2018 introduce a multi-modal sentence summarization task that produces a short summary from a pair of sentence and image, and propose a modality-based attention mechanism to pay different attention to image patches and text units, and design image filters to selectively use visual information to enhance the semantics of the input sentence. N. Liu et al., 2020 propose a system for summarizing open domain videos.

Chapter 3

Dataset

3.1 Dataset details

In this chapter , we describe our effort in developing the **M**ultim**O**dal sa**R**casm **E**xplanation (MORE) dataset. Since, MSE demands a sarcastic post, we explore two existing multimodal sarcasm detection datasets – (Schifanella et al., 2016), and (Sangwan et al., 2020), to extract the sarcastic posts. Schifanella et al., 2016 used hashtag-based approach (#sarcasm or #sarcastic) to collect 10000 sarcastic examples from Twitter, Instagram, and Tumblr. On the other hand, Sangwan et al., 2020 manually annotated 1600 sarcastic posts. Additionally, we explore another multimodal sarcasm detection dataset[1] to collect 10560 sarcastic posts. In total, we collect 22, 160 sarcastic posts.

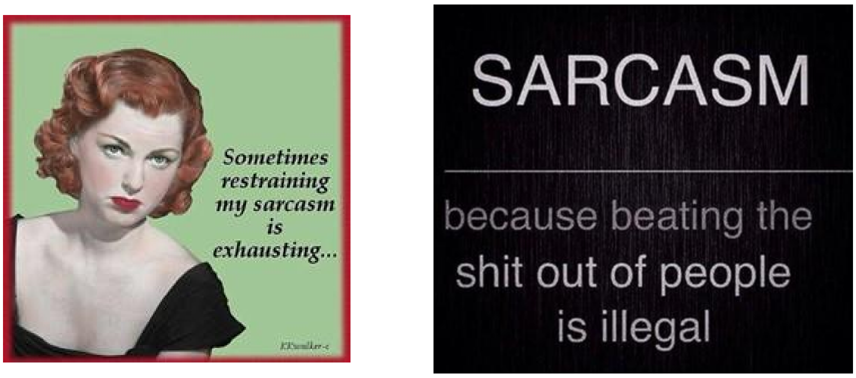
3.1.1 Annotation

We adopt the following annotation guidelines to generate an explanation for each post.

- Non-sarcastic posts are discarded.
- Sarcastic posts with explicit mention of sarcasm are discarded.
- Post describing the intra-incongruity (within text, or within image) or inter-incongruity (between image and text) are considered.
- All entities including image, caption, hashtags, emojis, etc. are considered for interpreting the irony and generating the appropriate explanation.
- In case the underlying sarcasm can be explained in multiple ways, the shorter and simpler explanation is preferred.
- Any unrelated topic in explanation is avoided.

We obtain services of two annotators – who carefully examined each post in our collection. Following the guidelines, annotators generate explanation for 3510 sarcastic posts. The remaining posts are discarded due to one of the reasons mentioned above. Two such examples are shown in Figure 3.1.

[1] `https://github.com/headacheboy/data-of-multimodal-sarcasm-detection`


Sometimes restraining my sarcasm is exhausting...
SARCASM
because beating the shit out of people is illegal

Caption: #sarcasm #sarcastic

Caption: Some idiots just don 't get it .. thoughts #idiot #idiotlist.

FIGURE 3.1: Posts that are discarded due to explicit sarcasm and do not suffice for an explanation.

3.1.2 Dataset Statistics

We use 85:5:10 split as the train (2983), validation(175) and test (352) sets. A brief statistics of the dataset is presented in Table 3.1. We can observe that we do not need lengthy sentences for explanation (15.43 vs. 19.68 avg tokens in caption) and the objective can be achieved by highlighting the incongruency.

Split	# of Posts	Caption		Explanation	
		Avg. Len	$\|V\|$	Avg. Len	$\|V\|$
Train	2983	19.75	9677	15.47	5972
Val	175	18.85	1230	15.39	922
Test	352	19.43	2172	15.08	1527
Total	3510	19.68	10865	15.43	6669

TABLE 3.1: Statistics of the MORE dataset.

3.1.3 OCR Extraction

Depending on the nature of the sarcastic post, the image may or may not contain text. A sarcastic post is referred to as an OCR instance if the image contains some text and this text contributes to or causes the sarcasm along with the caption. For analysis purpose, we extract the text from the images using the OCR of google vision API. The OCR output is used to separate the OCR instances from the Non-OCR instances in order to perform separate analysis over these two subsets of data.

Chapter 4

EXMORE: Proposed Methodology

In this chapter, we discuss the proposed methodology and explain the EXMORE architecture in detail. EXMORE is a multimodal transformer-based encode-decoder approach. To generate explanation, we take both the inputs i.e image and the textual caption and pass them through the image encoder and text encoder respectively. Next, we feed the image representation and caption representation obtained from the image and text encoders respectively into the cross-modal encoder for cross-modal learning. Finally, the representation produced by the cross-modal encoder is concatenated with the caption representation to obtain the final encoder representation and passed to the explanation decoder which generates the explanation. Figure 4.1 shows the complete architecture of EXMORE. Each module of the architecture is discussed in detail in the following sections.

4.1 Image Encoder

We use VGG19 (Simonyan and Zisserman, 2014) pre-trained on object detection task as the image encoder. The input image is passed through the VGG19 and the last convolutional layer output is considered as the image representation ($x_I \in \mathbb{R}^{q \times d^I}$). Here, q is the number of regions (7×7 in VGG19) and d^I is the dimension of each region representation (512 in VGG19).

4.2 Text Encoder

We use the pre-trained BART (Lewis et al., 2019) encoder (BART-base version) as the text encoder. The input caption is passed though the BART encoder to obtain the caption representation ($x_T \in \mathbb{R}^{r \times d^T}$). Here, r is the number of tokens in the caption and d^T is the dimension of each token representation (768 in BART-base).

4.3 Cross-modal Encoder

The cross-modal encoder consists of Multi-head Cross-modal Attention layer and Fully-connected layer. Each of these layers are followed by layer normalization and residual connections. The image x_I and caption x_T representations are passed through the cross-modal encoder to get the cross-modal representation ($c \in \mathbb{R}^{r \times d^T}$). This module is intended to facilitate a mechanism to capture the incongruence between image and caption modality.

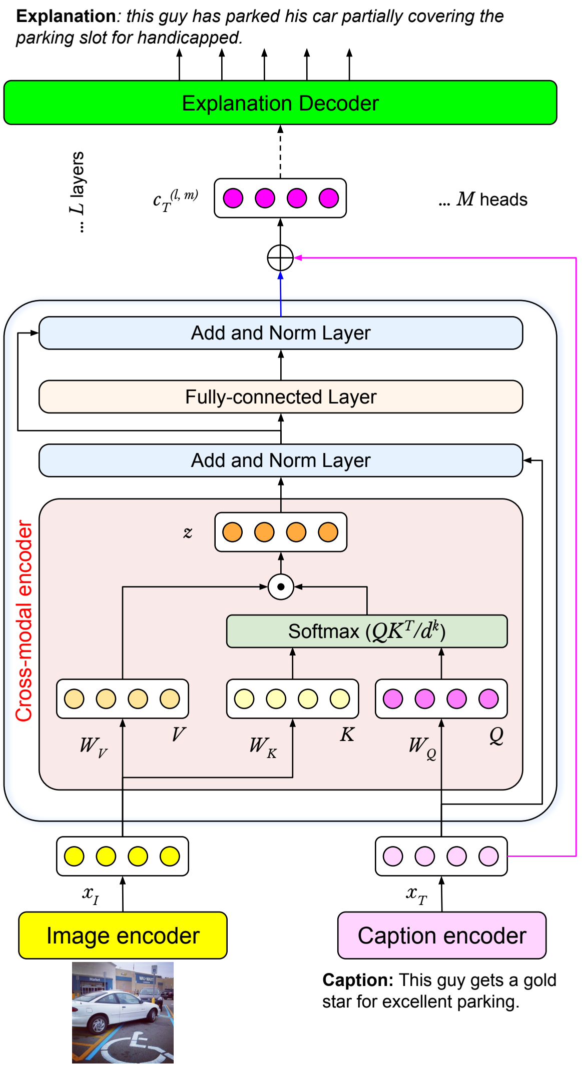


FIGURE 4.1: Architecture of EXMORE.

4.4 Multi-head Cross-modal Attention

The multi-head cross-modal attention takes image x_I and caption representation x_T as input. Unlike traditional transformer architecture Vaswani et al., 2017, where the same input is projected as *'query'*, *'key'* and *'value'*, in this cross-modal variant, we project the caption representation as *'query'* ($Q \in \mathbb{R}^{r \times d^k}$) and image representation as *'key'* ($K \in \mathbb{R}^{q \times d^k}$) and *'value'* ($V \in \mathbb{R}^{q \times d^k}$). Subsequently, we apply the conventional self-attention mechanism to compute the cross-modal attentive representation $z \in \mathbb{R}^{r \times d^k}$. Taking $d^k = \frac{d^T}{M}$, we incorporate $M = 4$ heads for the computation.

4.5 Fully-connected layer

This is a simple 2-layered fully connected feed forward network with a ReLU layer in between. Both the layers are of size 2048.

4.6 Explanation Decoder

The pre-trained BART decoder (BART-base) is used as the explanation decoder. The cross-modal representation c and the caption representation x_T are then concatenated to produce the final encoder representation $C_T \in \mathbb{R}^{2r \times d^T}$. This representation C_T is fed to the explanation decoder along with the decoder input. The decoder, auto-regressively, produces the explanation.

Chapter 5

Experiments, Results and Analysis

In this chapter, we dwell deep into the experiments carried out as a part of the research which include baselines as well as our proposed model. We show experimental results on the overall test set and in addition, we also show results for Non-OCR instances and OCR instances in test set separately. We perform analysis in terms of Linguistic analysis and Human evaluation to further support the experimental results and show that EXMORE qualitatively outperforms the best baseline.

5.1 Baselines

Since the problem undertaken is novel, we adapt various related existing systems for comparison.

5.1.1 Text Based

For text-based baselines we employ transformer (Vaswani et al., 2017) and pointer generator network (See, P. J. Liu, and Manning, 2017) for generating explanations. Both of them work on the related task of summarization and thus give estimates for performance expectations considering only the textual modality.

5.1.2 Text and Image Based

For a simple baseline model we consider LSTM+InceptionV3(+Image OCR). It consists of LSTM's for text representations from the caption and the text extracted from the image using OCR. It uses InceptionV3 (Szegedy et al., 2016) for extracting image representation which is added to the OCR text representation to form the final image representation. Finally, the caption representation and image representation are used for generating the explanations. We also consider the system proposed by N. Liu et al., 2020 for summarizing open domain videos as a viable baseline. Here, the input for the video is a set of frames captured at intervals of 16 frames, however for the purpose of our experiment we use a single frame composed of the accompanying image for a particular instance. Lastly, we consider the multimodal transformer proposed by Yao and Wan, 2020 for multimodal machine translation, for analysing performance of our system.

5.2 Evaluation Metrics

The experiments were performed on the mentioned train, validation and test splits. As the evaluation metrics, we employ BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BertScore (Zhang* et al., 2020) and cosine similarity between sentence level embeddings of the target and predicted explanations. The cosine similarity-based metric is obtained by computing the sentence level embedding for each target and predicted explanation using Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019) and then computing the cosine similarity between those embeddings. The cosine similarity between these gives an estimate of how close the target and predicted explanations are in terms of semantics.

5.3 Experimental settings

The EXMORE encoder is first trained and evaluated on Multimodal sarcasm detection (MSD) data. This enables the encoder to detect the sarcasm from the input which in turn can help encode useful features that are passed to the EXMORE decoder to generate the explanation. The github data[1] and the gold set from Sangwan et al., 2020 are combined to form the MSD data. The MSD data is split using 70:20:10 split into train(17929), validation(2562) and test(5123) set. This pretrained EXMORE encoder is then used to train the EXMORE model as a whole on the MORE dataset. The output of EXMORE encoder has dimension (sequence length, d_text), the mean across sequence length dimension gives a vector of dimension d_text. This vector is passed to the classifier for classification. The classifier consists of a single linear layer of size 768 (bart-base d_model) followed by a tanh activation. For the input text, BART tokenizer is used with maximum length set to 256 and padding and truncation enabled. We use 1 Cross-model encoder layer and dropout with probability of 0.1. During the training over MSD data, we use AdamW (Loshchilov and Hutter, 2017) optimizer with learning rate of 1e-4, batch size of 32 and run it for 70 epochs. F1 score is monitored over the validation set during the training. Finally, the best performing model checkpoint is considered. During the training over MORE dataset, we use AdamW (Loshchilov and Hutter, 2017) optimizer with learning rate of 1e-5 for the Cross-modal encoder and 3e-4 for the LM head of decoder, batch size of 16 and run it for 125 epochs. Cross entropy loss is monitored over the validation set during the training. For both the trainings, the image encoder is freezed.

5.4 Experimental Results

Table 5.1 shows the results of all the baseline models and the EXMORE model. It can be observed that EXMORE outperforms all the baselines across all the evaluation metrics shown. Among the baselines, M-transf outperforms the rest of the baselines, except Pointer Generator Network, across all the evaluation metrics shown. M-transf outperforms Pointer Generator Network in 9 out of 12 evaluation metrics and, hence, it is considered the best baseline.

[1] https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

Modality	**Model**	**BLEU**				**Rouge**			**METEOR**	**BERT-Score**			**Sent-BERT**
		B1	B2	B3	B4	R1	R2	R*L*		Pre	Rec	F1	(Cosine)
Text	Transformer	11.44	4.79	1.68	0.73	17.78	5.83	15.90	9.74	83.4	84.9	84.1	52.55
	Pointer Generator Network	17.54	6.31	2.33	1.67	17.35	6.90	16.00	15.06	84.8	85.1	84.9	49.42
Text + Image	LSTM+InceptionV3 (+ Image OCR)	9.96	3.76	0.92	0.18	12.84	3.59	12.18	9.06	83.8	83.7	83.8	51.04
	MFFG-RNN	14.16	6.10	2.31	1.12	17.47	5.53	16.21	12.31	81.5	84	82.7	44.65
	MFFG-Transf	13.55	4.95	2.00	0.76	16.84	4.30	15.14	10.97	81.1	83.8	82.4	41.58
	M-Transf	14.37	6.48	2.94	1.57	20.99	6.98	18.77	12.84	86.3	86.2	86.2	53.85
	EXMORE	19.26	11.21	6.56	4.26	27.55	12.49	25.23	19.16	88.3	87.5	87.9	59.12

TABLE 5.1: Performance on Overall data

Modality	**Model**	**BLEU**				**Rouge**			**METEOR**	**BERT-Score**			**Sent-BERT**
		B1	B2	B3	B4	R1	R2	R*L*		Pre	Rec	F1	(Cosine)
Text + Image	M-Transf	14.91	6.90	2.66	0.83	21.05	7.08	19.34	13.91	86.5	86.3	86.4	51.77
	EXMORE	19.47	11.69	6.82	4.27	27.12	12.12	24.92	19.20	88.3	87.6	88	56.95

TABLE 5.2: Performance on Non-OCR instances

Table 5.2 and 5.3 shows the performance of M-transf and EXMORE on the Non-OCR samples and OCR samples of the test data. Non-OCR samples are those in which the image does not have text content contributing towards the sarcasm and OCR samples are those in which the image has text content contributing towards the sarcasm. To separate the Non-OCR samples from the OCR samples for performance analysis, the OCR text from the image is generated and if the number of tokens in the OCR text is less than or equal to 8 then it is considered as a Non-OCR sample and if the number of tokens is greater than or equal to 6 then it is considered as an OCR sample. It can be observed that EXMORE outperforms M-transf across all the evaluation metrics in case of both Non-OCR samples and OCR samples.

Modality	**Model**	**BLEU**				**Rouge**			**METEOR**	**BERT-Score**			**Sent-BERT**
		B1	B2	B3	B4	R1	R2	R*L*		Pre	Rec	F1	(Cosine)
Text + Image	M-Transf	14.06	6.25	3.22	2.28	21.04	7.01	18.42	12.06	86.2	86.1	86.1	55.66
	EXMORE	19.40	11.31	6.83	4.76	28.02	13.10	25.66	19.55	88.2	87.5	87.9	60.82

TABLE 5.3: Performance on OCR instances

5.5 Analysis

To further investigate the results, we perform qualitative analysis of the explanations generated from EXMORE and the best baseline M-transf. Linguistic analysis reveals the quality of the explanations in linguistic feature space i.e the similarity in the distribution of various POS tags and overlap of tokens between ground truth and predicted explanations. Human evaluation is focused on the semantic quality and the syntactic quality of explanations. Semantic quality refers to the extent to which the explanation justifies and reveals the sarcasm in the given instance, whereas syntactic quality refers to the quality of the explanation in terms of a proper English sentence.

5.5.1 Linguistic Analysis

Adjective, adverb, verb and noun POS tags are extracted from the ground truth (GT) explanation and the prediction (Pred) explanation for each instance in the test set. "GT count"

Model	GT count	Pred count	Difference	Overlap	Synonym count
M-Transf	1.32	0.86	1.04	0.04	0
EXMORE	1.31	1.04	0.91	0.16	0.02
M-Transf (Non-OCR)	1.19	0.80	0.91	0.05	0
EXMORE (Non-OCR)	1.19	0.91	0.85	0.15	0.01
M-Transf (OCR)	1.40	0.91	1.13	0.04	0
EXMORE (OCR)	1.40	1.14	0.95	0.17	0.03

TABLE 5.4: Linguistic Analysis: Adjective

Model	GT count	Pred count	Difference	Overlap	Synonym count
M-Transf	1.03	0.78	0.79	0.27	0
EXMORE	1.03	0.69	0.72	0.24	0
M-Transf (Non-OCR)	0.96	0.73	0.84	0.22	0
EXMORE (Non-OCR)	0.96	0.61	0.71	0.18	0
M-Transf (OCR)	1.09	0.80	0.76	0.31	0
EXMORE (OCR)	1.09	0.76	0.74	0.28	0

TABLE 5.5: Linguistic Analysis: Adverb

and “Pred count” are the average counts of a given POS tag in the GT explanations and Pred explanations, respectively, in the test set. For each instance, the difference in count of a given POS tag between GT and Pred explanation is computed, the average difference across the test set gives the “Difference”. For each instance, for a given POS tag, the count of tokens common in the GT and Pred explanation gives the overlap for that POS tag, the average overlap across the test set gives the “Overlap”. For each instance, for a given POS tag, the count of tokens in Pred explanation that are synonyms of tokens in GT explanation gives the synonym count, the average synonym count across the test set gives the “Synonym count”.

Tables 5.4, 5.5, 5.6 and 5.7 show the results of these 5 measures for M-transf and EXMORE

Model	GT count	Pred count	Difference	Overlap	Synonym count
M-Transf	2.78	2.67	1.24	0.45	0.14
EXMORE	2.78	2.41	1.18	0.60	0.16
M-Transf (Non-OCR)	2.71	2.51	1.16	0.38	0.13
EXMORE (Non-OCR)	2.71	2.38	1.15	0.51	0.21
M-Transf (OCR)	2.80	2.81	1.32	0.52	0.14
EXMORE (OCR)	2.80	2.46	1.17	0.69	0.12

TABLE 5.6: Linguistic Analysis: Verb

Model	GT count	Pred count	Difference	Overlap	Synonym count
M-Transf	3.76	3.68	1.80	0.68	0.01
EXMORE	3.75	3.57	1.81	1.03	0.02
M-Transf (Non-OCR)	3.71	3.62	1.80	0.63	0
EXMORE (Non-OCR)	3.70	3.53	1.88	1.05	0.02
M-Transf (OCR)	3.81	3.73	1.80	0.73	0.015
EXMORE (OCR)	3.79	3.68	1.76	1.06	0.02

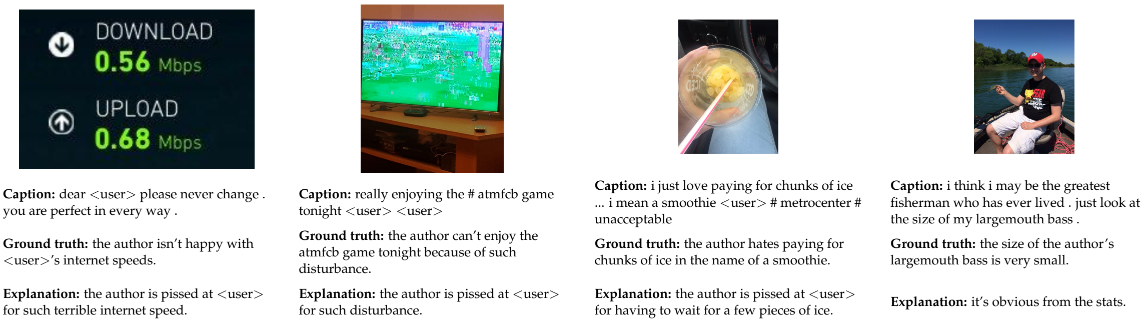
TABLE 5.7: Linguistic Analysis: Noun

explanations. The results are computed over the entire test set (1st and 2nd rows), the Non-OCR instances (3rd and 4th rows) and the OCR instances (5th and 6th rows) separately as well. Difference is less in case of EXMORE as compared to M-transf for adjective, adverb and verb across all sets of instances. In case of noun, the Difference is similar for M-transf and EXMORE when computed over the entire test set, it is slightly less for M-transf for Non-OCR instances and slightly more for OCR instances. This shows that in the case of EXMORE the linguistic features in Pred explanations are closer to that of GT explanations as compared to M-transf. The Overlap is more in case of EXMORE as compared to M-transf for adjective, verb and noun across all sets of instances. In case of adverb, the Overlap of EXMORE is slightly less than that of M-transf. This shows that EXMORE Pred explanations are more likely to be Related to Input as compared to M-transf Pred explanations since higher Overlap means more number of tokens common between GT explanation and Pred explanation. Synonym count of EXMORE is slightly higher than or close in certain cases to that of M-transf which again suggests Pred explanations to be likely to be Related to Input.

5.5.2 Human Evaluation

The human evaluators are asked to rate each explanation (i.e prediction) in terms of its Adequacy and Fluency. Adequacy represents how well a prediction explains the given sarcastic post i.e how well does it highlight the intended meaning of the author (one who created the post) and hence, explain the sarcasm. Fluency represents the quality of the prediction in terms of a coherent English statement, irrespective of whether it is related to the given sarcastic post or not. i.e a prediction can have poor adequacy and still have good fluency. Inspired by Kayser et al., 2021, for Adequacy, the human evaluators are provided with 4 rating options (Justify, Weakly Justify, Related to Input, Not Related to Input). These options are explained with an example as follows:
Example:
Image: a courier parcel which is delivered in a poor condition (torn or partially open).
Caption: wow I love how well the parcel is delivered.
Now,

- "Justify" represents a prediction which highlights the "specific reason" why the input is sarcastic. i.e 'The author is disappointed with the poor condition of the parcel.'
- "Weakly Justify" represents a prediction which highlights the semantic incongruence but not the specific reason why the input is sarcastic. i.e 'it is not a good delivery.'
- "Related to Input" represents a prediction which does not explain or convey the sarcasm but talks about or refers to some entity related to the input (either in image or caption). i.e 'the author loves the parcel'. (it is related to input since it refers to parcel or delivery)
- "Not related to Input" represents a prediction which is totally unrelated to the input i.e 'the author hates the weather.' (It is unrelated since it has nothing to do with the input).



Caption: dear <user> please never change . you are perfect in every way .

Ground truth: the author isn't happy with <user>'s internet speeds.

Explanation: the author is pissed at <user> for such terrible internet speed.

Caption: really enjoying the # atmfcb game tonight <user> <user>

Ground truth: the author can't enjoy the atmfcb game tonight because of such disturbance.

Explanation: the author is pissed at <user> for such disturbance.

Caption: i just love paying for chunks of ice ... i mean a smoothie <user> # metrocenter # unacceptable

Ground truth: the author hates paying for chunks of ice in the name of a smoothie.

Explanation: the author is pissed at <user> for having to wait for a few pieces of ice.

Caption: i think i may be the greatest fisherman who has ever lived . just look at the size of my largemouth bass .

Ground truth: the size of the author's largemouth bass is very small.

Explanation: it's obvious from the stats.

FIGURE 5.1: Examples of Adequacy Ratings (Justify, Weakly Justify, Related to Input, Not Related to Input) from left to right respectively.

Model	Adequacy	Fluency
M-Transf	0.37	3.71
EXMORE	0.69	4.48

TABLE 5.8: Adequacy and Fluency of M-transf and EXMORE.

These Adequacy ratings (Justify, Weakly Justify, Related to Input, Not Related to Input) are then mapped to numeric values (1, $\frac{2}{3}$, $\frac{1}{3}$, 0) respectively. For Fluency, the annotators are asked to rate predictions on the scale of 1 (Low) to 5 (High). To compute the final Adequacy score and Fluency score of a model, first, for each explanation the scores of the annotators are averaged, and then the sample average is taken. To compute the Adequacy Rating distribution of a model, for each explanation the most frequent rating is considered and then the percentage of explanations with a rating 'r' gives the distribution of rating 'r'.

Model	Adequacy Rating			
	Justify	Weakly Justify	Related to Input	Not Related to Input
M-Transf	15%	15%	35%	35%
EXMORE	65%	5%	20%	10%

TABLE 5.9: Adequacy Rating distribution of M-transf and EXMORE.

Table 5.8 shows the Adequacy and Fluency scores of M-transf and EXMORE. It can be observed that EXMORE significantly outperforms M-transf in terms of Adequacy which indicates better semantic quality of EXMORE compared to M-transf. In an ideal case the Adequacy score would be 1 when all the explanations are rated as "Justify". EXMORE has significantly better Fluency score compared to M-transf which indicates better generation quality of EXMORE compared to M-transf. In ideal case the Fluency score would be 5 when all the explanations are rated 5. Table 5.9 shows the Adequacy Rating distribution of M-transf and EXMORE. It can be observed that in M-transf more explanations lie in "Related to Input" and "Not Related to Input" space as compared to those in "Justify" and "Weakly Justify" space. On the other hand, in EXMORE more explanations are concentrated in "Justify" space as compared to the rest of the ratings. Also, explanations lying in "Not Relatd to Input" space are significantly more in case of M-transf as compared to EXMORE. This shows the superiority of EXMORE in terms of Adequacy as compared to M-transf. Figure

5.1 shows the examples for each Adequacy category with respect to EXMORE.

Chapter 6

Conclusion

In this work, we introduced a novel task MSE which aims at generating a natural language explanation for a given sarcastic post. The purpose of the explanation is to reveal the underlying irony and in turn explain why the given sarcastic post is sarcastic. To address MSE, we curated a new dataset, MORE which consists of 3510 (image, caption, explanation) triples. To benchmark the dataset, we proposed EXMORE, a transformer based encoder-decoder model incorporating a cross-modal encoding mechanism for the task. We performed experiments to compare the performance of EXMORE against the six baseline systems and showed significant improvements across 5 evaluation metrics. We also performed qualitative analysis in terms of Linguistic analysis and Human evaluation showing that EXMORE generated explanations that are superior in semantic and syntactic quality as compared to the best baseline. This work opens a new area in the study of sarcasm aiming to provide explanation of sarcasm.

Chapter 7

Publication

- Poorav Desai, Tanmoy Chakraborty, Md Shad Akhtar. Nice perfume. How long did you marinate in it? Multimodal Sarcasm Explanation, AAAI, 2022. (Desai, Chakraborty, and Akhtar, 2021)

Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv: 1409.0473 [cs.CL].

Bedi, Manjot et al. (2021). "Multi-modal sarcasm detection and humor classification in code-mixed conversations". In: *IEEE Transactions on Affective Computing*.

Bouazizi, M. and T. Otsuki Ohtsuki (2016). "A Pattern-Based Approach for Sarcasm Detection on Twitter". In: *IEEE Access* 4, pp. 5477–5488. DOI: 10.1109/ACCESS.2016.2594194.

Cai, Yitao, Huiyu Cai, and Xiaojun Wan (July 2019). "Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2506–2515. DOI: 10.18653/v1/P19-1239. URL: https://www.aclweb.org/anthology/P19-1239.

Castro, Santiago et al. (July 2019). "Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4619–4629. DOI: 10.18653/v1/P19-1455. URL: https://www.aclweb.org/anthology/P19-1455.

Desai, Poorav, Tanmoy Chakraborty, and Md Shad Akhtar (2021). "Nice perfume. How long did you marinate in it? Multimodal Sarcasm Explanation". In: *arXiv preprint arXiv:2112.04873*.

Dubey, Abhijeet, Aditya Joshi, and Pushpak Bhattacharyya (2019). "Deep models for converting sarcastic utterances into their non sarcastic interpretation". In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pp. 289–292.

Felbo, Bjarke et al. (Sept. 2017). "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1615–1625. DOI: 10.18653/v1/D17-1169. URL: https://www.aclweb.org/anthology/D17-1169.

Kayser, Maxime et al. (2021). "e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks". In: *arXiv preprint arXiv:2105.03761*.

Koehn, Philipp et al. (June 2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180. URL: https://www.aclweb.org/anthology/P07-2045.

Lewis, Mike et al. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv: 1910.13461 [cs.CL].

Li, Haoran et al. (2018). "Multi-modal Sentence Summarization with Modality Attention and Image Filtering." In: *IJCAI*, pp. 4152–4158.

Liu, Nayu et al. (2020). "Multistage Fusion with Forget Gate for Multimodal Summarization in Open-Domain Videos". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1834–1845.

Loshchilov, Ilya and Frank Hutter (2017). "Fixing Weight Decay Regularization in Adam". In: *CoRR* abs/1711.05101. arXiv: 1711.05101. URL: http://arxiv.org/abs/1711.05101.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777.

Peled, Lotem and Roi Reichart (July 2017). "Sarcasm SIGN: Interpreting Sarcasm with Sentiment Based Monolingual Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1690–1700. DOI: 10.18653/v1/P17-1155. URL: https://www.aclweb.org/anthology/P17-1155.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://www.aclweb.org/anthology/D19-1410.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv: 1602.04938 [cs.LG].

Sangwan, S. et al. (2020). "I didn't mean what I wrote! Exploring Multimodality for Sarcasm Detection". In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206905.

Schifanella, Rossano et al. (2016). "Detecting sarcasm in multimodal social platforms". In: *Proceedings of the 24th ACM international conference on Multimedia*. ACM, pp. 1136–1145.

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 2017). "Get To The Point: Summarization with Pointer-Generator Networks". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://www.aclweb.org/anthology/P17-1099.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Szegedy, Christian et al. (2016). "Rethinking the Inception Architecture for Computer Vision". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* URL: http://arxiv.org/abs/1512.00567.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., pp. 5998–

6008. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa
Paper.pdf.

Yao, Shaowei and Xiaojun Wan (July 2020). "Multimodal Transformer for Multimodal Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4346–4350. DOI: 10.18653/v1/2020.acl-main.400. URL: https://www.aclweb.org/anthology/2020.acl-main.400.

Zhang*, Tianyi et al. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkeHuCVFDr.