VENUE DISCOVERY USING SUPERVISED GENERATIVE ADVERSARIAL CROSS-MODAL HASHING

By

Himanshu Aggarwal

Under the supervision of
Dr. Rajiv Ratn Shah, MIDAS Lab, IIIT Delhi
Co-Supervisor
Dr. Yi Yu, NII Japan

Indraprastha Institute of Information Technology, Delhi
July, 2019

VENUE DISCOVERY USING SUPERVISED GENERATIVE ADVERSARIAL CROSS-MODAL HASHING

By

Himanshu Aggarwal

Submitted
in partial fulfilment of the requirements for the degree of
Master of Technology in Computer Science
to

Indraprastha Institute of Information Technology, Delhi
July, 2019

Certificate

This is to certify that the thesis titled "**Venue Discovery using Supervised Generative Adversarial Cross-modal Hashing**" being submitted by **Himanshu Aggarwal** to Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my 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.

July, 2019

Dr. Rajiv Ratn Shah

Department of Computer Science
Indraprastha Institute of Information Technology Delhi
New Delhi, India

Dr. Yi Yu

National Institute of Informatics Japan
Tokyo, Japan

Abstract

With a massive amount of online multimedia data (e.g., images, videos, text articles, etc.) and increasing needs of the people, venue discovery using multimedia data has become an underlined research topic. We are referring to business and travel locations as venues in this study and aim to improve efficiency of venue discovery by hashing. Previously, a lot of work has been done in the field of cross-modal retrieval for reducing the heterogeneous gap between multiple modalities, so that samples from those modalities can be compared directly. Such techniques have also shown their application in venue discovery. However, improved technology has increased the size of the multimedia data and thus has made the retrieval more difficult and slower. Therefore, hashing techniques are being developed to project features from different modalities into a common hamming space. Hash features take very less storage space, and they can be compared faster than the real-valued features, using hamming distance. In this thesis, we propose an adversarial learning-based approach of generating hash code for venue-related heterogeneous multimedia data to ease the task of venue discovery without any location information.

Previous works have shown the great ability of Generative Adversarial Networks (GANs) to model the distribution of the data and learn discriminative representations. We show how GANs can be used to learn to generate hash codes with category and pairwise information that occur naturally in the data. Most existing supervised cross-modal hashing methods map data in different modalities to Hamming space, where the semantic information is exploited to supervise data in different modalities during the training stage. However, previous works neglect pairwise similarity between data in different modalities, which lead to degraded performance of the model for finding exact matches for the queries. To address this issue, we propose a supervised Generative Adversarial Cross-modal Hashing method by Transferring Pairwise Similarities (SGACH-TPS). This work has three significant contributions: i) we propose a model for making efficient venue discovery on a new dataset, WikiVenue, of real-world images produced by the people, ii) the supervised generative adversarial network to construct a hash function that can map multimodal data of image-text pairs to a common hamming space, and, iii) a simple transfer training strategy for the adversarial network is suggested to supervise different modalities of samples where we transfer the pairwise similarity to the fine-tuning stage of training. To generalize our work in the field of cross-modal retrieval, we showed experiments with the benchmark datasets, Wiki, and NUS-WIDE.

Acknowledgement

I sincerely thank my supervisor Dr. Rajiv Ratn Shah, IIITD for giving me the opportunity to work on this project at NII, Japan and for his valuable guidance over the course of the thesis. I would also like to thank my co-supervisor Dr. Yi Yu, NII for her constant supervision and valuable insights.

A special thanks to all my colleagues at Dr.Yu's lab and Dr.Rajiv's MIDAS lab for making the journey of this thesis memorable and helping me putting pieces together.

I also thank my family for their support throughout the entire process by keeping me cheery and confident.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Research Motivation

Multimedia is an integral part of our lives. With the ever-increasing development of the Internet and other technologies, vast amounts of multimedia data—including images, videos, text, audio, and more—is created every second. Information is now present as a mixture of different media and expresses comprehensive knowledge. Since, this data is available and accessible to people from a variety of platforms, like social media websites (such as Facebook, Twitter, Youtube, etc.), the responsibility of a researcher is to design techniques to better access this data [16, 29, 33, 39, 43]. Some automatic mechanisms are needed to setup similarity links between the different modality data to achieve maximum advantage from this rich multimedia information. When a user wants to search for something from this multimedia data, then she should be able to do that with any modality query that is available to her. For this purpose, better cross-modal retrieval techniques are necessary.

Cross-modality refers to integrating information acquired from different modalities like images and texts, or audio and video, etc. This term is mostly associated with retrieval where it means that given a query data in one modality, we want to retrieve results from different modality data.

One primary application of such retrieval systems is discovering or finding a venue using its photographs or description. For instance, assume a person is visiting some site for the first time. He is not aware of the place. So, he takes a photo of this venue and searches for the location. Venue discovery will be useful at this time. It will help to find the place where the user took the photograph. Also, it can help find the similar venues whose visual and textual features match that location in the photograph [39]. This is the exact case that needs venue recommendation and discovery. We represent this idea in Figure 1.1.

In the earlier time, accurate venue discovery was not possible since the data available was mostly just images. Moreover, there were no reliable sources for the venue data. On the other hand, today, with extensive mobile user engagements, multimedia innovations, and business-related data on social media have led to an ever-expanding source of multimedia information. Most prominent examples of these are travel and business-related venue photos on Foursquare, Wikipedia articles, and

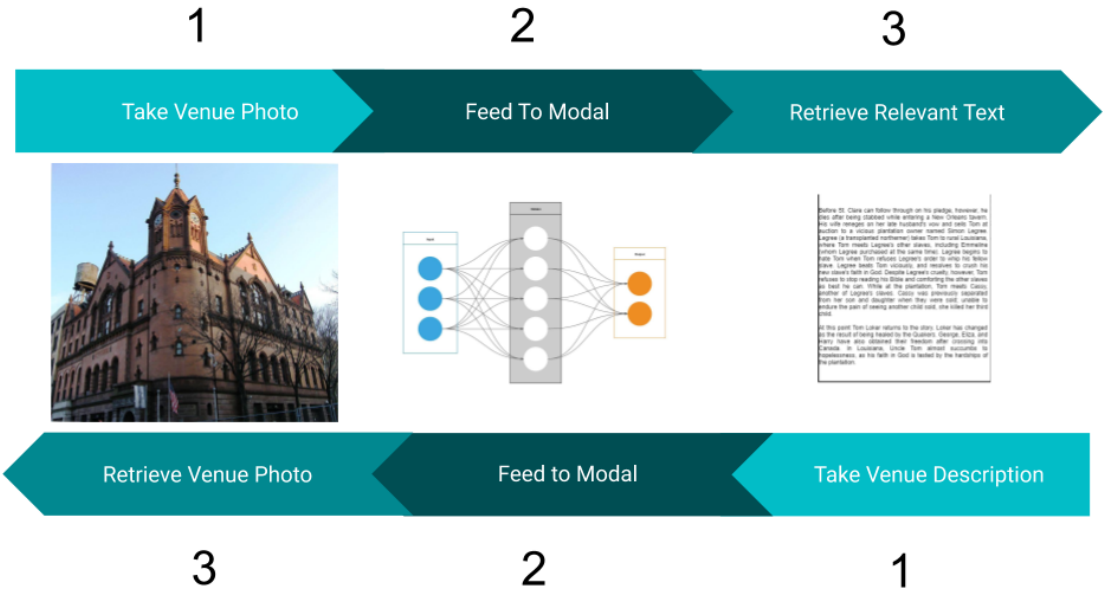


Figure 1.1: Diagram explaining the idea behind our work.

advertisements on YouTube. Moreover, the growth of such venue data with visual business-related content has made several businesses more visible over the Internet and lead to an increase in sales in the physical world. Venue discovery brings new opportunities with new multimedia data available today.

There are, however, several issues that lie with cross-modal retrieval. First of all, the features of different modalities are very different in terms of their statistical properties. This prevents a direct comparison of the features of different modalities. So the current research focuses on resolving this in two ways: correlation maximization and feature selection. Correlation maximization includes techniques to increase the correlation between modalities by Subspace learning. Subspace learning is trendy research area where features from different modalities are projected to a common space where they have similar statistical characteristics and same dimension. Several works like [34, 39, 43, 44, 45] have shown significant progress using this technique. Nevertheless, this has its disadvantages like large storage space due to real values, and slow retrieval since distance calculation between samples can be time-consuming.

Today, to improve the issues as mentioned earlier, hashing techniques have become very popular for this task. Hashing means mapping values to keys using hash functions. But in multimedia, it refers to mapping the feature vectors of some media from real-valued space to binary-valued vectors of shorter length in hamming space, again using a hash function. It has two major advantages when applied to multimedia retrieval: (i) It can do fast retrieval because using Hamming distance can be quickly computed using the binary codes based on bit operations. (ii) Binary codes take much less storage space compared to original high dimensional features.

Now data in different modalities lie in separate space and have a different distribution. Hence, they cannot be compared directly to perform any retrieval task.

We need a way to represent the data in different modalities in a common space. This is done by hashing. Thus, cross-modal hashing refers to mapping rea; valued features from different modalities to a hamming space with same feature-length and a preserved correlation between the data of different modalities. Previous works [41, 19, 35, 38, 40] have shown the ability of hashing in cross-modal retrieval, but more work is still needed in this area to generate better features and reduce the loss of information.

Most of the past works in cross-modal hashing exploits only semantic category information in the training phase. This does not let the model learn about the specific characteristics of the sample of paired data in different modalities, which is essential for getting better retrieval results. Therefore, in this thesis, we use pairwise information that is naturally present in the data, to train our model.

In this thesis, we have developed a Venue Discovery model using the SGACH-TPS model proposed in Chapter 3. We have used the WikiVenue Dataset [39], which is a new dataset for Venue related multimedia data. Further details are discussed in later chapters.

1.2 Literature Review

Previously, several works have been proposed for performing the task of venue discovery and cross-modal retrieval using hashing. Following subsections discuss some of the important works in this area.

1.2.1 Venue Discovery

There is some early literature on venue discovery, like prediction of geographic category of the photo using data-driven scene matching method, [3], coarse prediction of the location of the photo [2], using visual concept of the photos investigating street-view images of the city [7], and many others like [26, 4, 8, 22]. Friedland et al. [7] have exploited multiple information sources to estimate the location of the video. This location discovery was of region-scale granularity. Work in [4, 8] is closely related to location prediction. The model in this work recognizes the objects and locations in the given images and uses images as SIFT features.

Most recent works in venue discovery [22, 39] are deep learning-based methods. These models detect visual words from images and create word vectors from text articles, and use them in the model. Except Category based Deep CCA (C-DCCA) [39], no work uses real-word images generated by users for training and retrieval.

1.2.2 Cross-Modal Hashing

Retrieval of single modality data using hashing techniques has been studied at length during the past decade. In recent years, cross-modal hashing has gained momentum. In simple terms, the hash functions project the higher dimensional data from different modalities to a common hamming space. These projected features then

can be used to perform faster retrieval. The hashing techniques are mainly divided into traditional methods and deep learning methods. The conventional methods are further divided into supervised and unsupervised learning methods based on whether the model learns semantic information. Following are brief reviews of some hashing methods for cross-modal retrieval.

Unsupervised Cross-Modal Hashing Methods: These methods map the heterogeneous data into a common hamming space to maximize the correlation. These are similar to the Canonical Correlation Analysis (CCA) methods [1]. Kumar et al. [9] proposed Cross-view Hashing (CVH), which extends Spectral Hashing [5]. CVH consider both intra-view and inter-view similarities with a generalized eigenvalue formulation. Similarly, to preserve intra-media and inter-media consistency, Inter-Media Hashing (IMH) [13] was proposed to learn a common Hamming space. Predictable Dual-view Hashing (PDH) [12] proposed an objective function to preserve the predictability of pre-generated binary codes and optimize it using an iterative method. Ding et al. [27] proposed Collective Matric Factorization Hashing (CMFH), which uses a latent factor model to learn unified hash codes from different modalities of one instance. Likewise, Latent Semantic Sparse Hashing (LSSH) was proposed by Ding et al. [19], which is again a matrix factorization technique to learn semantic features for image and text. Unsupervised techniques achieve limited accuracy on retrieving semantically similar data since they try to learn cross-modal hash functions using the data distributions.

Supervised Cross-Modal Hashing Methods: These methods take advantage of the semantic information present with the data in the form of labels. This leads to better accuracy for the retrieval task as compared to the unsupervised methods. Bronstein et al. [6] proposed Cross-modality Similarity Sensitive Hashing (CMSSH) as a way to model the hashing as a classification problem. Zen et al. [10] in Co-Regularized Hashing (CRH), proposed a method to learn the hash function of each bit sequentially so that the bias introduced by each hash function is minimized. Heterogeneous Translated Hashing (HTH) [17] was proposed to learn different hamming spaces for different modalities. It then aligns these spaces using learned translators for cross-modal retrieval. In Iterative Multi-view Hashing (IMVH), Hu et al. [15] propose to learn the hash functions using within-view similarities and between-view correlations. Wu et al. [24] recommended Quantized Correlation Hashing (QCH) to optimize quantization error along with correlation. Supervised methods for cross-modal hashing produce better results by applying the semantic information available with the data.

Deep Learning based Methods Advancement in neural networks has led inspiration to apply deep learning to cross-modal retrieval task and then to learn hash functions for the same. Cross-Media Neural Network Hashing (CMNNH) [21] proposed to preserve intra-modal discriminative capability and inter-modal pairwise correspondence. In Deep Multimodal Hashing with Orthogonal Regularization (DMHOR), Wang et al. [23] proposed to preserve inter-modal and intra-modal correlation to learn hash functions, and also reducing the redundant information between hash bits. Cao et al. [25] proposed Cross Autoencoder Hashing (CAH) which is based on deep autoencoder structure. It maximizes the feature correlation

between bimodal data and also maximizes the semantic relationship that is given by label information. In Deep Semantic Correlation Learning based Hashing for Multimedia Cross-Modal Retrieval, Gong et al. [36] constructed a semantic similarity matrix based on labels and try to generate a semantic correlation between modalities automatically.

Quantization Techniques Recently, researchers have also started exploring another way to create hashes, which tries to prevent information loss due to hashing. This technique is called Quantization. It has been mostly used for signal processing and data compression. This works by dividing a large number of points or vectors in space into approximately equal-sized groups, where its centroid point represents each point. It is similar to the clustering method. Quantization [11, 18] has shown stronger representation ability than the hashing techniques for single-modal retrieval. For cross-modal retrieval, it has been relatively unexplored. There are only three significant approaches [18, 30, 31] that use quantization for cross-modal search. Among them, only Cao et al.'s approach [31] learns deep representations of the data.

Now with the introduction of the new types of learning techniques like Adversarial learning, unique research opportunities have opened up. Recently few attempts have been made to generate hash functions using generative adversarial networks. The related review is as follows.

1.2.3 Generative Adversarial Networks

Goodfellow et al. [14] first proposed the Generative Adversarial Network (GAN) to estimate the generative model by an adversarial process. After evaluating its capability to learn the data distributions, several attempts have been made to use GAN for many computer vision problems. In simple terms, GAN consists of two models: a generative module G, which learns the data distribution, and a discriminative module D to estimate the probability that the input sample is from real data rather than G. Recently, few attempts have been made to use GAN to generate embedding for cross-modal data to perform the retrieval task. In Adversarial Cross-Modal Retrieval (ACMR), Wang et al. [34] use triplet learning constraint to learn the data distribution using GAN. More recently, Zang et al. [41] in Unsupervised Generative Adversarial Cross-Modal Hashing (UGACH), propose an unsupervised method to learn hash function for cross-modal retrieval. This model uses a generative module to learn the data distribution and automatically select positive and negative samples for given anchor forming a triplet. The discriminative module then uses this triplet to determine the hash function.

1.3 Research Contribution

The problem we tried to solve in this thesis is to 'retrieve relevant venue texts based on venue images'. Here, by 'venue text', we mean text articles related to the venue, and by 'venue image', we mean images of some venue or place. By the term 'relevant'we mean that the venue text and venue image must belong to the same category, eg. Schools and Colleges, Nightlife Spots, etc. This thesis has the following significant contributions:

1. we propose a model for making efficient venue discovery on a new dataset, WikiVenue, of real-world images produced by the people. We have built and tuned our model over heterogeneous data of real-world images from Wikipedia and Foursquare, and text articles from Wikipedia related to the venues, to decrease the distance between theory and application. In reality, normal photographs do not well represent visual information. Models that are trained on sophisticated images will fail on real-world images. Therefore, it is important to build models on such images.

2. a supervised generative adversarial network to construct a hash function that can map multimodal data of image-text pairs to a common hamming space.

3. a simple transfer training strategy for the adversarial network is suggested to supervise data of different modalities where we transfer the pairwise similarity to the fine-tuning stage of training.

4. In addition to above work, we have done experiments on publicly available benchmark image-text paired datasets, UCSD and NUS-WIDE, to generalize our work.

1.4 Thesis Outline

The remaining of this thesis has been structured as follows. Datasets and benchmarks are introduced in Chapter 2. Then, Chapter 3 presents our approach and explains how to use pairwise similarities in the adversarial model to improve the quality of hash embedding. It also talks about the implementation of the proposed model for Venue Discovery task. Experimental evaluation results are shown in Chapter 4. Finally, conclusions are pointed out in Chapter 5.

1.5 Summary

In this chapter, we introduced the problem we are solving in this thesis, i.e., extracting venue articles related to the venue photographs. We highlighted the recent related works for performing venue discovery and cross-modal retrieval. We also mentioned works related to hashing based cross-modal retrieval. Finally, a thesis outline is given to provide an overview of further chapters of the thesis.

Chapter 2

Datasets and Benchmarks

2.1 Datasets

2.1.1 WikiVenue Dataset

This dataset is created by Digital Content and Media Science Research Division lab at the National Institute of Informatics, Japan. Key information of this dataset can be found in Tables 2.1, 2.2 and 2.3. Some samples images are also present in Figure 2.1. It contains text articles and images along with other information, from featured venue articles on Wikipedia website[1]. Wikipedia consists of one featured article and one photograph per venue. However, a single photograph is not enough to completely represent a venue. Therefore, to better represent the visual aspect of the venue, the dataset is added with user-generated images from the Foursquare website[2]. Now multiple images can be used to illustrate a single text article of Wikipedia. Combining from both the sources, the dataset incorporates 19792 photographs and 1994 article descriptions for 1994 venues.

Wikipedia data is collected for five cities, namely New York, Los Angeles, London, Sydney, and Orlando. The photographs from Foursquare, however, just belonging to Los Angeles and London, have been collected. All the images belong to exactly one of the ten category labels. For training and validation, we use all the venue images from Wikipedia and the varying ratio of Foursquare images (10% - 50% images, in different experiments). We use remaining images of Los Angeles and London from Foursquare as the test set.

[1] www.wikipedia.com
[2] www.foursquare.com

19792	Venue Images
1994	Venue Articles
1994	Venues
5	Cities
10	Categories

Table 2.1: Key Information of WikiVenue Dataset.

S.No.	Categories
1.	Arts & Entertainment
2.	College & University
3.	Event
4.	Food
5.	Nightlife Spot
6.	Outdoor & Recreation
7.	Professional & Other Places
8.	Residence
9.	Shop & Services
10.	Travel and Transport

Table 2.2: Categories in WikiVenue Dataset

S.No.	Cities
1.	New York
2.	Los Angeles
3.	Sydney
4.	Orlando
5.	London

Table 2.3: List of cities from where venues are included in WikiVenue Dataset.

Figure 2.1: Sample Images from WikiVenue Dataset

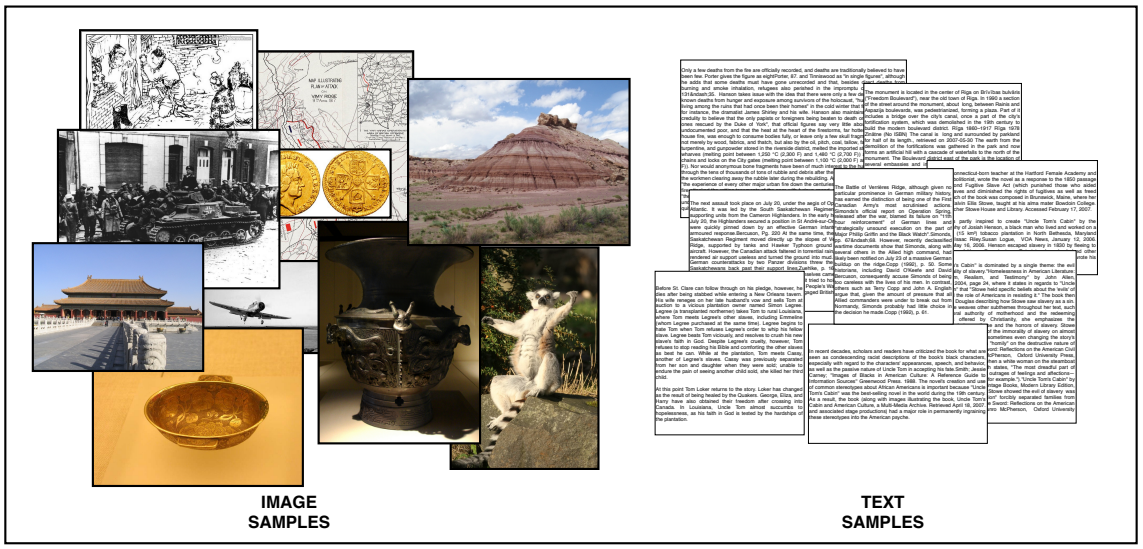


Figure 2.2: Wikipedia Dataset

Dataset	Wiki	NUS-WIDE
No. of Items	2866 image-text pairs	186644 image-text pairs
Image and Text Features	VGG16 & LDA	BoW & Concept List
Categories	10	10 out of 81
Lable Type	Single	Multiple

Table 2.4: Key Information about Wiki and NUS-WIDE Datasets.

2.1.2 Wiki Dataset

This is a public dataset provided by the University of California, San Diego (UCSD) group. It contains 2866 image-text pairs collected from featured articles on Wikipedia. No other image resource is used for compiling this dataset. The related key information is present in Table 2.4 and sample images are present in Figure 2.2. It is a widely used dataset for cross-modal retrieval. The image features are represented as 4096-dimensional deep features extracted from 16-layer VGGNet. Whereas a 10-dimensional topic vector represents the textual features. All the paired data instances are annotated with one of the ten semantic categories. The definition of a category depends on the Wikipedia tags. From this dataset, we randomly selected 75% of the data as the training set and the rest 25% as the query set.

2.1.3 NUS-WIDE Dataset

It is created by NUS laboratory for media search. It contains 269648 images with associated tags which are treated as a textual modality. The key information about this dataset is present in Table 2.4 and some sample images are also present in Figure 2.3. Each of the images is categorized into at least one of the 81 semantic labels. We considered the top 10 most common labels the extracted the corresponding images. It included 186644 images which have an annotation of at least one of the top 10 most common labels. Out of this dataset, we selected 10000 random images as the training set, 2000 random images as query set, and around 100000 images as retrieval set.

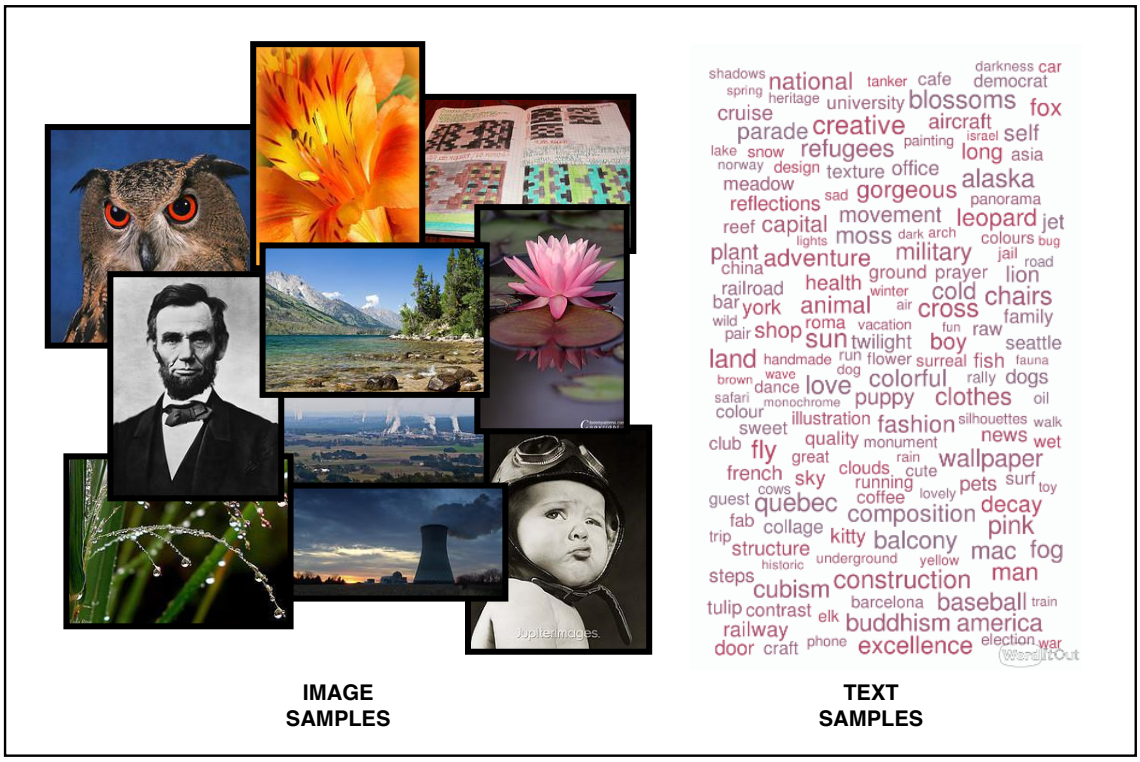


Figure 2.3: NUS-WIDE Dataset

2.2 Benchmarks

There is no existing work that uses WikiVenue Dataset for hashing. Therefore, we can not compare our work with anything. However, we have added a few experiments with the state-of-the-art cross-modal hashing technique to prove the worth of our model.

Apart from that, we are comparing our results on the benchmark datasets with current state-of-the-art methods for cross-modal hashing techniques, LSSH [20], SMFH [28], FSH [32], UGACH [41], and SCRATCH [37].Among these, LSSH, FSH, and UGACH are unsupervised methods, whereas SMFH is a supervised method.

2.3 Summary

In this chapter, we discussed all the datasets we have used in this thesis, i.e., WikiVenue Dataset, Wiki Dataset, and NUS-WIDE Dataset. WikiVenue Dataset contains venue related images and text articles. The other two datasets have been used to generalize our work. Later, some benchmark works are mentioned which are used for comparison of our model's results.

Chapter 3

Supervised Generative Adversarial Cross-Modal Hashing

3.1 Problem Formulation

The focus of our thesis is on the cross-modal retrieval of the image-text data. Let there be a collection of n samples of paired image-text data, which is denoted as, $S = \{s_i\}_{(i=1)}^{n}, s_i = \{u_i, t_i\}$ where u_i is the image feature vector and t_i is the text feature vector. u_i is d_u-dimensional, whereas t_i is d_t-dimensional and generally $d_u \neq d_t$.

Each sample of s_i is also associated with category information, denoted by l_i, $l_i \epsilon [1, c]$, c denotes the total number of semantic categories/labels.

In this thesis, the image feature vector is represented as U, and the text feature vector is represented as T. Since U and T are statistically different with different distributions, they cannot be compared directly for performing cross-modal retrieval. For this purpose, there is a need to find mapping functions to project the image U and text T feature vectors to a common feature space F where they can be compared directly. The projected feature vectors are denoted as F_u and F_t for image and text, respectively.

The goal of this thesis is to generate more effective transformed features as hash codes which can use preserve semantic as well as pairwise information and can produce better results for ANN search. For this purpose, the model learns two mapping functions, $H_1 : U \rightarrow F_U, H_2 : T \rightarrow F_T$, where $F_U, F_T \epsilon R^K$, and K is the length of the hash codes produced by the model. This will be discussed in more detail in later sections.

3.2 Our Model: SGACH-TPS

The architectural diagram in Figure 3.1 presents the model that we are proposing in this thesis. This model is inspired by GAN model and the work of Zang et al.[41] in UGACH. The Feature Extraction part of the diagram consists of the image feature extractor which uses the standard deep convolution network named VGG16 to extract image features. The text feature extractor uses the Doc2Vec technique

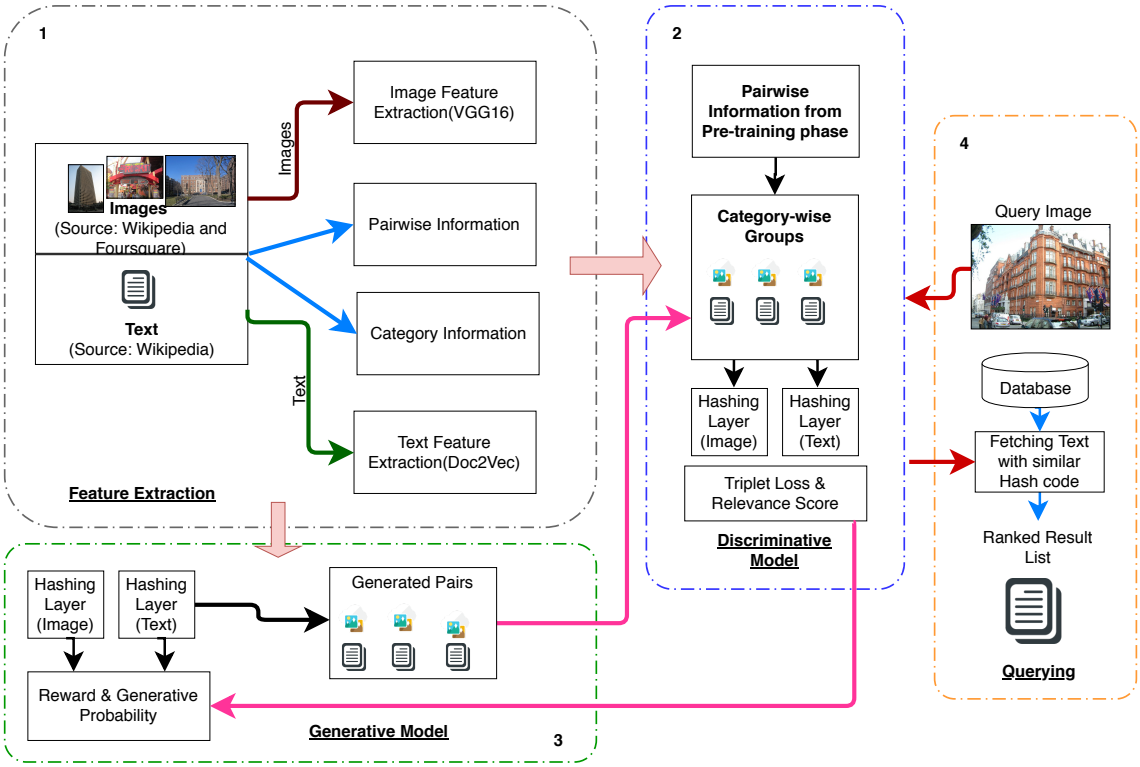


Figure 3.1: Architecture of SGACH-TPS: Proposed Model for learning hash functions for cross-modal data using adversarial training.

to generate bag-of-words features of the text data. A GAN model then uses these extracted features. The architecture consists of a generative and a discriminative model which help optimize each other while training in the form of minimax game. After the training, the discriminator can be independently used for generating hash codes. Lastly, the Querying module sends a query to the discriminator model and receives the corresponding hash code. That hash is then compared with the database to find the nearest hash codes of the other modality data. The retrieved results then can be provided as a ranked list based on the hamming distance.

The discriminative model is optimized by using the triplet loss. Here the model is fed with an anchor data of one modality, along with a negative and positive sample chosen by the generative model and the semantic label information, respectively.

3.3 Generative Module

The generative model has two paths in the architecture and can take as input both image and text features. Each pathway consists of two fully connected layers. The first layer, Common representation layer, maps the features of each modality into a common space. The second fully-connected layer, Hashing layer, works as a hashing function which maps the real-valued features from a common space to a hamming space, with binary hash codes. The equations below represent the process:

$$\varphi(x) = \tanh(W_1 x + b_1) \tag{3.1}$$

$$f(x) = sigmoid(W_2\varphi(x) + b_2) \tag{3.2}$$

Equations 3.1 and 3.2 defines the the generative module's process, where x represents features of image or text sample, $W1$ and $W2$ denote the weights in common representation and hash layer, respectively, and $b1$ and $b2$ are the respective bias parameters. Equation 3.2 outputs the real-valued hash codes representing the input image/text. The below equation is used to convert these features to binary hash codes:

$$h(x) = sgn(f(x) - 0.5), p = 1, 2, ..., k \tag{3.3}$$

The threshold function in Equation 3.3 gives a binary representation to the common space features of the image/text. Here k is the length of the hash code. Now when the features from different modalities are mapped to a common hamming space, the similarity between these different modalities can be measured easily with fast hamming distance. Since it is hard to optimize with binary values, real-valued hash codes are used for the training purpose.

The goal of the generative model G is to fit the data distribution of one modality over its semantic information to select the data of other modality and challenge the discriminative model D. The model calculates the generative probability, $p_\theta(x^U|q_t)$ and $p_\theta(x^T|q_u)$, for both image and text samples in the database, given the query as text or image, respectively. The generative probability is calculated using a softmax function:

$$p_\theta(x|q) = \frac{exp(-\left\|f(q) - f(x)\right\|^2)}{\sum_x exp(-\left\|f(q) - f(x)\right\|^2)} \tag{3.4}$$

For an input query of one modality, we can use the above Equation 3.4 to calculate the relevance score of each database sample of the other modality.

3.4 Discriminative Module

Similar to the generative model, the discriminative model also has a two-way architecture. All other architectural details are also the same as the generative model. The only difference is that the discriminative model has a goal of distinguishing between the pair generated by the generative model, and the pair sampled from the labeled data. This model is input with both the pairs, and it then produces binary hash codes for query and both the samples.

The aim here is to produce hash codes such that the hamming distance between the actual pair, i.e., query and corresponding positive sample from other modality, should be small. Whereas, the distance between the hash codes of query and the negative sample selected by the generative model should be significant.

At this point, we want to introduce the novelty we are proposing in this thesis that is applied in the discriminative module. We present the novel technique of transferring information between different training sessions that we used in our model.

In the previous works, only the label information has been exploited as semantic information. Other information available with the paired data, like the pairwise constraint, has not been used. Here we are proposing the use of this information to learn the distribution of the data better and produce hash codes accordingly. We suggest pre-training of the discriminator with only pairwise details we have from the data. The parameters from this pre-training stage can then be used in the fine-tuning stage of the GAN training as the initial weights of the discriminative model.

Here, learning pairwise information denotes that the data is available as paired samples, and the model should be able to retain this information of the semantic pairs to project that corresponding paired data of other modality as the best-retrieved result for the query of other modality. To accomplish this, the model uses the pairwise constraint information as the ground truth and maps the corresponding image and text closer to each other in the common modality space. For a given query, the relevant positive sample is the paired sample of the other modality only. The network architecture of the discriminative model remains the same. But the information it is learning is different in the pre-training stage. Later, when this knowledge is used in the fine-tuning stage, i.e., the parameters learned in the pre-training stage serves as the initial parameters of the fine-tuning stage, it helps the model to create better hashing functions. We denote this process as transferring pairwise information.

This model tries to produce the relevance score as $g_\phi(x, q)$, for the candidate and query data pair by using triplet ranking loss. The formula for the relevance score can be observed in Equations 3.5 and 3.6.

Pre-training relevance score formula is,

$$g_\phi(x^R, q) = max(0, \beta + \left\|f(q) - f(x^P)\right\|^2 - \left\|f(q) - f(x^R)\right\|^2) \tag{3.5}$$

Fine-tuning relevance score formula is,

$$g_\phi(x^G, q) = max(0, \beta + \left\|f(q) - f(x^L)\right\|^2 - \left\|f(q) - f(x^G)\right\|^2) \tag{3.6}$$

q is the input query, x^R is a randomly paired instance, x^P is pairwise information constrained paired instance, x^L is semantic label information constrained paired instance, and x^G is the selected instance by Generative model G. β here acts as a margin parameter. This equation is used to ensures that the distance between information constrained pair in the fine-tuning stage is smaller than the generated pair by the margin β so that the discriminative model D can clearly distinguish between the generated and semantic pair. Similarly, the loss is used in the pre-training stage also, to make discriminative model D aware of the pairwise information.

This relevance score is used by the discriminative model D to calculate the predictive probability of an instance x by a sigmoid function as defined in Equation 3.7.

$$D(x|q) = sigmoid(g_\phi(x, q)) \tag{3.7}$$

The work of the generative model is to select informative data to challenge the discriminative model, due to which its ability to perform cross-modal retrieval is limited. However, the discriminative model is appropriate for cross-modal retrieval task, since it has been supported by the generative model to learn the better distinction between relevant and not-relevant instances. Therefore, the discriminative model is used to perform cross-modal retrieval by generating hash codes.

3.5 Adversarial Learning

The generative and discriminative models act like players in a minimax game: For a given query of one modality, the generative model tries to select a marginal data sample of other modality that can be classified wrongly by the discriminative model. The discriminative model then takes a query as an input with corresponding true relevant sample of other modality sampled from the ground truth (semantic information) and the sample data selected by the generative model. This model tries to distinguish between these samples. Below Equation 3.8 represent the process for an image/text query where p_{true} represents set of true pairs and p_θ represents set of pairs generated by G:

$$\begin{aligned} V(G, D) = \min_\theta \max_\phi \sum_{j=1}^{n} (&E_{x \sim p_{true}(x^L|q^j)}[log(D(x^L \mid q^j))] \\ &+ E_{x \sim p_\theta(x^G|q^j)}[1 - log(D(x^G \mid q^j))]) \end{aligned} \tag{3.8}$$

The generative and discriminative models can be trained iteratively by minimizing and maximizing the above equation. The genertive model tries to minimize the Equation 3.8 to fit the distribution over semantic information, and discriminative model maximizes the Equation 3.8.

Now, the entire process of training this model is divided into two stages:

1. Pre-training of Discriminative model
2. Fine-tuning or Training of GAN model

Using these two stages, we propose a new way to learn multiple information by the model without making the loss function complex, as well as, shifting from the main aim of the learning process. Here our goal is not to learn the pairwise information but to use the semantic category information to provide relevant results corresponding to a query. But the pairwise information that we are preserving here helps to produce better results by projecting the paired data of different modalities closer together.

3.6 Optimization

Using the definitions of the generative and the discriminative models, the proposed model can be optimized to produce high-quality hash embedding for image and text data. Using the objective function in Equation 3.8, a minimax game can be conducted to train the proposed model. The training parameters of the generative model are fixed while training discriminator and vice versa. We discuss the optimization of both the models separately.

3.6.1 Optimization of Discriminative Module

Discriminative model is trained as a two-stage training process - pre-training and fine-tuning stage. In the pre-training stage, only pairwise information is used by the model to get features retraining pairwise similarity. Later the parameters of the pre-training stage are used as the initial parameters of the fine-tuning stage of GAN training. Optimization of both the stages is discussed separately.

1. Optimization in Pre-training Stage: There is no generative model in this stage. The discriminative model trains on its own. It takes as input the true image-text pairs using the pairwise information of the paired image-text dataset. It also takes some randomly selected fake image-text pairs to fool the discriminator network. The model is then trained using the Equations 3.7 and 3.8. The Equation 3.9 defines the process.

$$\phi^* = arg\max_{\phi} \sum_{j=1}^{n} (E_{x \sim p_{true}(x^R|q^j)}[log(sigmoid(g_\phi(x^P, q^j)))]) \tag{3.9}$$

2. Optimization in Fine-tuning Stage: After the pre-training stage, the preserved information is transferred to the discriminator of the fine-tuning stage. This helps the discriminator to start training with better initial parameters. Now to optimize the parameters of this stage, the generative model's parameters are fixed. We use the generative model of the previous iteration to generate the image-text and text-image fake pairs for image and text as queries, respectively. We further sample true image-text and text-image pairs from the labeled data. The discriminator then tries to maximize the log-likelihood of distinguishing the selected and the generated pairs. The Equation 3.8 can be re-written as following to optimize the model:

$$\begin{aligned} \phi^* = arg\max_{\phi} \sum_{j=1}^{n} (&E_{x \sim p_{true}(x^G|q^j)}[log(sigmoid(g_\phi(x^L, q^j)))] \\ &+ E_{x \sim p_{\theta*}(x^G|q^j)}[log(sigmoid(g_\phi(x^G, q^j)))]) \end{aligned} \tag{3.10}$$

$p_{\theta}*$ is the generative model of the previous iteration. This equation is differentiable with respect to Equations 3.5 and 3.6. Therefore, a stochastic gradient descent algorithm can be used to solve the Equations 3.9 and 3.10.

3.6.2 Optimization of Generative Module

To train this model, now we fix the parameters of the discriminative model. This model tries to minimize equation 3.8 and fits true relevance information. The following Equation 3.11 defines the optimization when the discriminative model is fixed, and the generative model is trained:

$$\begin{aligned}\theta^* = arg\min_{\theta} \sum_{j=1}^{n} (E_{x\sim p_{true}(x^G|q^j)}[log(sigmoid(g_\phi * (x^L, q^j)))] \\ + E_{x\sim p_\theta(x^G|q^j)}[log(sigmoid(g_\phi * (x^G, q^j)))])\end{aligned} \tag{3.11}$$

where $g_\phi*$ refers to the discriminative model in the previous iteration.

The traditional GAN generates new data from continuous noise vector and is optimized using a stochastic gradient descent algorithm. But here we followed a discrete selection policy for selecting triplets from the labeled data, which can not be optimized this way. Therefore, following the work in UGACH, we used reinforcement learning based on the policy gradient to update the parameters of the generative model. The derivation is presented in Equation 3.12.

$$\begin{aligned}&\bigtriangledown_\theta \, E_{x\sim p_\theta(x^G|q^j)}[log(1 + exp(g_\phi(x^G, q^j)))] \\ &= \sum_{k=1}^{m} \bigtriangledown_\theta p_\theta(x_k^G|q^j) log(1 + exp(g_\phi(x_k^G, q^j))) \\ &= \sum_{k=1}^{m} p_\theta(x_k^G) \bigtriangledown_\theta \, log p_\theta(x_k^G|q^j) log(1 + exp(g_\phi(x_k^G, q^j))) \\ &= E_{x\sim p_\theta(x^G|q^j)}[\bigtriangledown_\theta log p_\theta(x^G|q^j) log(1 + exp(g_\phi(x^G, q^j)))] \\ &\simeq \frac{1}{m} \sum_{k=1}^{m} \bigtriangledown_\theta log p_\theta(x_k^G|q^j) log(1 + exp(g_\phi(x_k^G, q^j)))\end{aligned} \tag{3.12}$$

According to the reinforcement learning perspective, action x is taken in the environment q using the policy $\log p(x|q)$. Corresponding to the action, the discriminative model calculates a value $log(1 + e..)$. This acts as a reward value for the generative model in the reinforcement learning environment. For instance, if the discriminative model is performing well, the reward will be a small value. This means the generative model needs to select better samples to confuse the discriminator and get higher rewards.

3.7 Venue Discovery using SGACH-TPS

We train the model on the venue images and venue descriptions (text), and query the trained model with unseen real-world images of the venues, all from WikiVenue dataset. The task within the framework is group venue search. This refers to finding the relevant venues with the same category as the venue in the given an image, without any location information (GPS information). Each venue has an assigned category which is inherited by the venue text and images automatically.

The text and images of the venues are used in the training phase to learn the cross-modal correlation so that the text and image features are highly correlated in the shared space. In the testing phase, the input photo's visual features are compared to the textual features of the venues. This is in contrast to the existing approaches for venue discovery which use only visual features.

To develop an efficient venue discovery model that can visually and textually understand the venue better, we applied two new techniques. First, as already discussed, we used the pairwise information to understand the cross-modal feature space of the dataset better. Further, we used a portion of real-world images of the venues so that the model can visually understand the venues better. We used all the Wikipedia venue images and 10% of the Foursquare images for training and validation of the model and evaluated the model on the remaining Foursquare images corresponding to Los Angeles and London. Similarly, we experimented with 20%, 30%, 40%, and 50% of FS images. Analysis of these experiments is discussed in the next section.

3.8 Summary

In this chapter, we explain our model SGACH-TPS. Firstly, the problem is formulated, and all the variables used in the thesis are defined. The model is then discussed in detail with the help of the diagram of the model architecture 3.1. The adversarial training process of Generative and Discriminative module is discussed. Later optimization of both the modules is performed, and venue discovery using SGACH-TPS is discussed.

Chapter 4

Experimental Results

4.1 Retrieval Tasks and Evaluation Metrics

Two types of retrieval tasks are performed to evaluate our proposed model:
Image-to-Text: Here image is used as query and the as part of retrieval, semantically similar text articles are returned as a ranked list. We denote this task as $I \rightarrow T$.
Text-to-Image: In this task, a text sample is taken as an input query, and it retrieves the semantically similar images as a raked list. We denote this task as $T \rightarrow I$.

Hamming distance is used to generate a ranked list of related samples of one modality, given the query of another modality. The list is expected to contain the semantically similar items to the query. The returned retrieval results are evaluated based on whether the query and the retrieved results share the same semantic labels. We used the recall-precision curve, mean average precision (MAP), and mean reciprocal rank 1 (MRR1) as main metrics.

1. The MAP score is the mean of the average precision (AP) for all queries. AP is calculated as shown in Equation 4.1:

$$AP = \frac{1}{R} \sum_{k=1}^{m} \frac{k}{R_k} \times rel_k \tag{4.1}$$

   where m is the size of database, R is the number of relevant images in the database, R_k is the number of relevant images in the top k retrieved ranking list, and $rel_k = 1$ if the image ranked at k-th position is relevant and 0 otherwise.

2. Precision-Recall Curve (PR-curve) is plotted by taking precision values at a certain level of recall of the retrieved ranking list. It is a widely used measure for information retrieval tasks.

3. Mean Reciprocal Rank 1 (MRR1) is a measure to calculate the mean of the reciprocal of the rank of the first relevant result in the retrieved ranking list, for all the queries. It is calculated as shown in Equation 4.2:

$$MRR1 = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{4.2}$$

Ratio	MAP (16 bits)		MAP (32 bits)		MAP(64 bits)		MAP (128 bits)		MAP (128 bits) [WTT]	
	I $\rightarrow T$	T $\rightarrow I$	I $\rightarrow T$	T $\rightarrow I$	I $\rightarrow T$	T $\rightarrow I$	I $\rightarrow T$	T $\rightarrow I$	I $\rightarrow T$	T $\rightarrow I$
0.1	0.5382	0.6038	0.5631	0.6438	0.6026	0.7047	0.5929	0.7054	0.4575	0.5288
0.2	0.5448	0.6164	0.5827	0.6587	0.6362	0.7142	0.6197	0.7104	0.4634	0.5493
0.3	0.5723	0.6322	0.6196	0.6821	0.6849	0.7279	0.6385	0.7179	0.4802	0.5544
0.4	0.5522	0.6231	0.6035	0.6784	0.6739	0.7277	0.6325	0.7091	0.4953	0.5689
0.5	0.5746	0.6417	0.6234	0.6876	0.6746	0.7232	0.6460	0.7302	0.5066	0.5773

Table 4.1: Mean Average Precision (MAP) for Retrieval Tasks on WikiVenue Dataset using SGACH-TPS with ration of Foursqaure images used in training. MAP[WTT] represents model performance Without using Transfer Training technique.

Ratio	MAP (I $\rightarrow T$)(16$bits$)		MAP (I $\rightarrow T$)(128$bits$)	
	Los-Angeles	London	Los-Angeles	London
0.1	0.5000	0.4903	0.5496	0.5391
0.2	0.5283	0.4417	0.6060	0.5216
0.3	0.4519	0.4703	0.4926	0.5154
0.4	0.5290	0.4377	0.5869	0.5185
0.5	0.4802	0.5080	0.5122	0.5594

Table 4.2: MAP for Los-Angeles and London Queries with 16 bit and 128 bit hash using SGACH-TPS.

$rank_i$ refers to the rank of the first relevant result for a query, and Q is the set of queries.

4.2 Results on WikiVenue Dataset

Evaluation is done by testing the model on queries from the same data distribution as train set, results in Table 4.1, and testing model on real-world images from FourSquare website that are generated by people in day-to-day life, for venues in Los-Angeles and London, in Table 4.2 and Table 4.5. These images were not seen by the model while training. The purpose of these queries is to evaluate the proposed model on real photos, and not just in a conserved environment.

As can be seen, the results on test queries are improving with increasing hash code length. We also present the performance of the model when the transfer training strategy is not used, i.e., no pre-training stage. Results are shown in Table 4.1 as [WTT], Without Transfer Training. It is very clear from these results that our novel training strategy improves the model. The performance on unseen images from Foursquare in Table 4.2 is also comparable to the test results. This proves that the model is almost as good for the real world new images as for a conserved dataset. Precision-recall curves in Figures 4.1 and 4.2, show that with an increase in the ratio of FS images used for training, there is an improvement in the precision values for initial recall values. This shows that better results are appearing in the top ranks of the retrieval.

Ratio	MAP (16 bits)		MAP (128 bits)	
	I $\rightarrow T$	T $\rightarrow I$	I $\rightarrow T$	T $\rightarrow I$
0.1	0.2986	0.3014	0.3034	0.3052
0.2	0.3057	0.3058	0.3142	0.3162
0.3	0.3128	0.3142	0.3198	0.3215
0.4	0.3187	0.3189	0.3253	0.3284
0.5	0.3202	0.3199	0.3343	0.3359

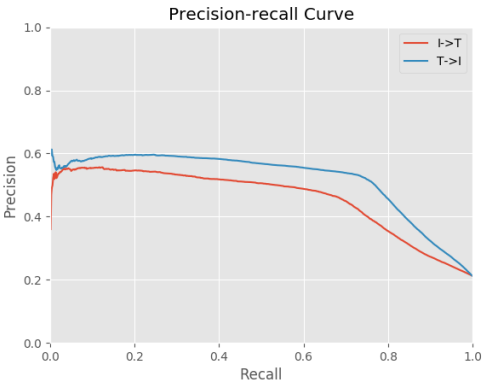
Table 4.3: Mean Average Precision (MAP) using UGACH model for Retrieval Tasks on WikiVenue Dataset with ration of Foursqaure images used in training.

Ratio	MAP (I $\rightarrow T$)(16$bits$)		MAP (I $\rightarrow T$)(128$bits$)	
	Los-Angeles	London	Los-Angeles	London
0.1	0.2823	0.3145	0.2854	0.3176
0.2	0.2773	0.3189	0.2822	0.3205
0.3	0.2745	0.3075	0.2752	0.3189
0.4	0.2786	0.3089	0.2869	0.3145
0.5	0.2753	0.3127	0.2844	0.3153

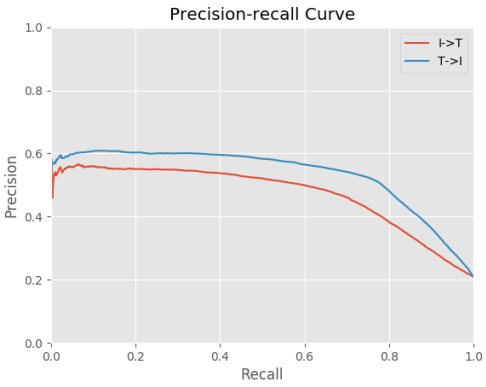
Table 4.4: MAP for Los-Angeles and London Queries using UGACH model with 16 bit and 128 bit hash.

The MRR1 values present in Table 4.5, indicate positive retrieval in the top ranks of the retrieved results. These values also show the loss of information that occurred with a small hash code length of 16 bits. The fluctuation in the MRR1 values with increasing ratio of FS images suggest that the model might not improve further for real images if we keep increasing the images of the same venues in the training data.

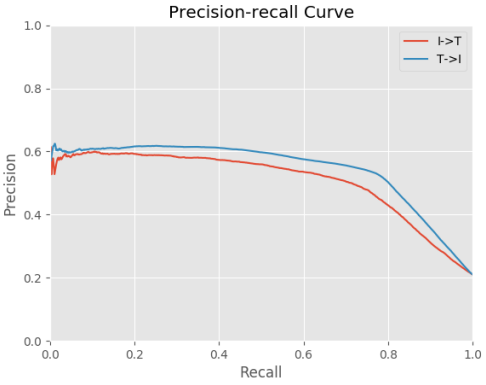
This dataset has never been used before for venue discovery using hashing, and therefore, there are no current state-of-the-art results to compare. However, we present results of experiments with WikiVenue dataset using UGACH [41], in Tables 4.3 and 4.4. We also present a comparison of our model with the latest state-of-the-art cross-modal retrieval method, UGACH, in Table 4.6. We used 0.0 and 0.1 ratios for comparison since they show the performance of both the models with and without using Foursquare images. So, now it can be judged if the earlier models are capable of visually learning about the venue with multiple photographs. We can observe that the results of our supervised method SGACH-TPS are better than the unsupervised method UGACH. The performance of UGACH model decreases when real-world images are added to it in the training phase. This is due to the inability of the unsupervised model to manage the noise from real-world images and learn more visual information about the venue. However, our model does not face such problems. This proves the importance of the semantic information used in our model.



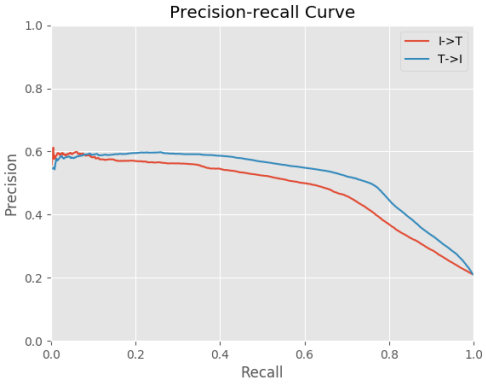
(a) Ratio 0.1, 16 bits



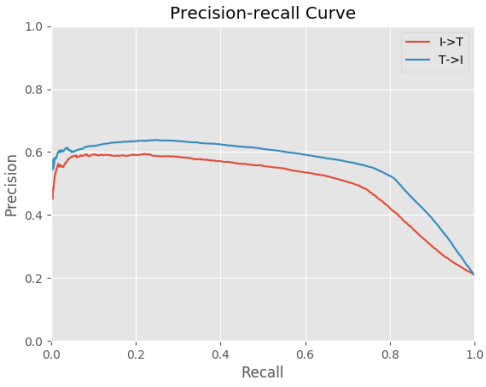
(b) Ratio 0.2, 16 bits



(c) Ratio 0.3, 16 bits

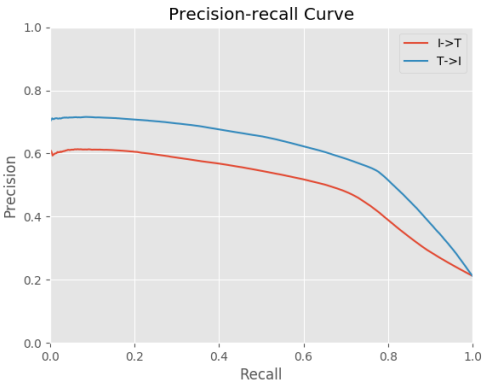


(d) Ratio 0.4, 16 bits

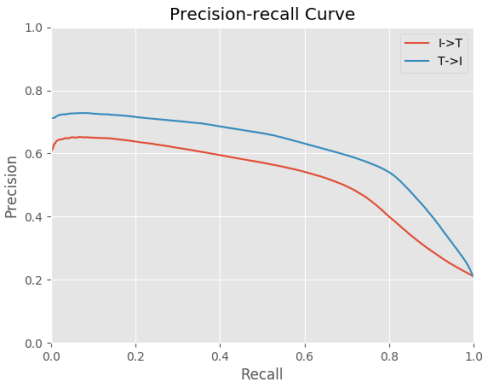


(e) Ratio 0.5, 16 bits

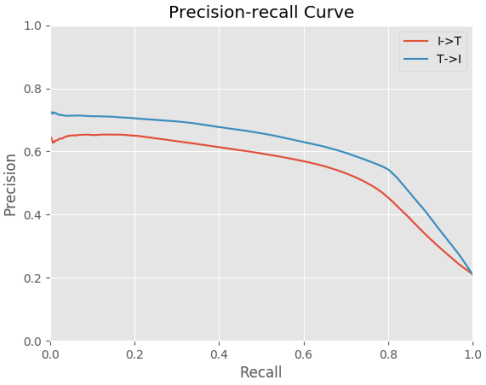
Figure 4.1: Precision-Recall Curve for WikiVenue Dataset Retrieval with varying ratio of Foursquare images used in training. Hash code length used is 16 bits.



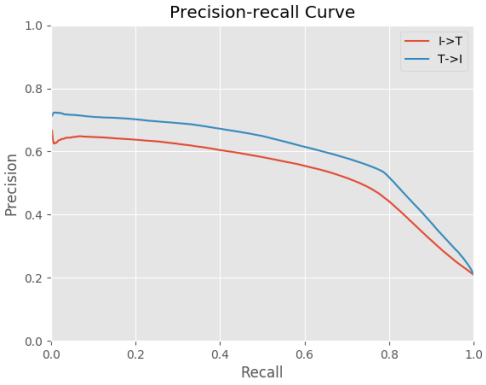
(a) Ratio 0.1, 128 bits



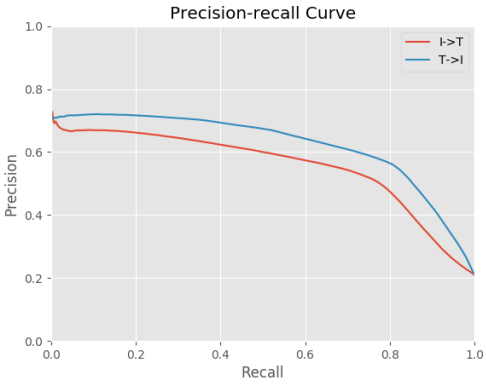
(b) Ratio 0.2, 128 bits



(c) Ratio 0.3, 128 bits



(d) Ratio 0.4, 128 bits



(e) Ratio 0.5, 128 bits

Figure 4.2: Precision-Recall Curve for WikiVenue Dataset Retrieval with varying ratio of Foursquare images used in training. Hash code length used is 128 bits.

Ratio	MRR1 (I $\rightarrow T$)(16$bits$)		MRR1 (I $\rightarrow T$)(128$bits$)	
	Los-Angeles	London	Los-Angeles	London
0.1	0.6412	0.5657	0.6458	0.6212
0.2	0.6401	0.5204	0.6916	0.6321
0.3	0.5652	0.5532	0.5968	0.6075
0.4	0.6147	0.4858	0.6533	0.6168
0.5	0.5523	0.6022	0.6486	0.6557

Table 4.5: MRR1 for Los-Angeles and London Queries with 16 bit and 128 bit hash.

Methods	MAP (128 bits)	
	I $\rightarrow T$	T $\rightarrow I$
SGACH-TPS (Ratio 0.0)	0.4807	0.6521
SGACH-TPS (Ratio 0.1)	0.5929	0.7054
UGACH (Ratio 0.0)	0.4207	0.5164
UGACH (Ratio 0.1)	0.2879	0.2916

Table 4.6: Comparison of SGACH-TPS and UGACH;comparison with no FS images.

4.3 Other Experiments and Results

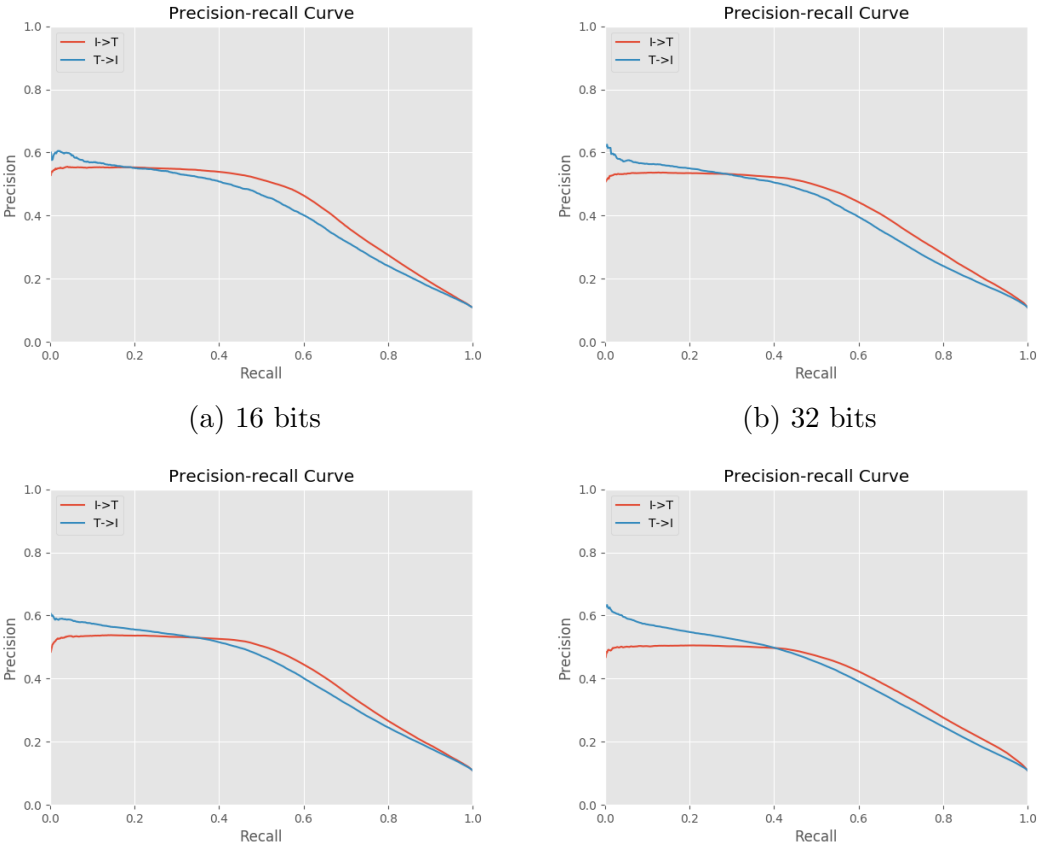
To generalize our model, we conducted experiments with the publicly available benchmark datasets, UCSD, and NUS-WIDE. The results of the experiments are noted in Tables 4.7 and 4.8. We are comparing our results with current state-of-the-art methods for cross-modal hashing techniques, LSSH [20], SMFH [28], FSH [32], and UGACH [41]. Among these, LSSH, FSH, and UGACH are unsupervised methods, whereas SMFH is supervised method.

As explained in Chapter 3, we have produced results by transferring pairwise information to the fine-tuning stage, where only category information is preserved. This experiment is named as SGACH-TPS (CI) in the result tables. However, we also experimented by preserving only pairwise information in the fine-tuning stage as well and ignoring the category information altogether. This experiment is named as SGACH-TPS (PI). Results of both the experiments can be found in the Tables 4.7 and 4.8. Also, the Precision-Recall graphs can be found in Figure 4.3.

Results on Wikipedia

Results of SGACH-TPS for Wikipedia dataset are reported in Table 4.7 in terms of MAP values. Here we present results for two variants of our model, i.e., when both category and pairwise information is used [SGACH-TPS (CI)], and when only pairwise information is used [SGACH-TPS (PI)]. The precision-recall curves for four different lengths are plotted in Figure 4.3. We can observe the following points from these results:

- SGACH-TPS considerably outperforms all the existing state-of-the-art models for Image-to-Text retrieval for all hash lengths. Essential reason can be the use of triplets formed by the generative module in training, which helps to build a better hash function by learning the data distribution.



(a) 16 bits

(b) 32 bits

(c) 64 bits

(d) 128 bits

Figure 4.3: Precision-Recall Curve for Wiki Dataset Retrieval.

Hash Length	16 bits		32 bits		64 bits		128 bits	
Methods	$I \to T$	$T \to I$	$I \to T$	$T \to I$	$I \to T$	$T \to I$	$I \to T$	$T \to I$
LSSH	0.2162	0.4990	0.2164	0.5225	0.2041	0.5287	0.2084	0.5330
SMFH	0.2276	0.5242	0.2470	0.5961	0.2955	0.6608	0.3133	0.6924
FSH	0.2270	0.4864	0.2433	0.5197	0.2366	0.4961	0.2463	0.5247
UGACH	0.3696	0.3771	0.3874	0.3943	0.4051	0.4154	0.442	0.4579
SGACH-TPS (CI)	0.4466	0.4607	0.4660	0.4866	0.4771	0.4932	0.4844	0.5115
SGACH-TPS (PI)	0.5230	0.4889	0.5236	0.4968	0.5165	0.4829	0.5005	0.4815

Table 4.7: Mean Average Precision (MAP) Comparison over Two Retrieval Tasks on Wikipedia Dataset

Hash Length	16 bits		32 bits		64 bits		128 bits	
Methods	$I \to T$	$T \to I$	$I \to T$	$T \to I$	$I \to T$	$T \to I$	$I \to T$	$T \to I$
LSSH	0.5812	0.5917	0.5811	0.5929	0.5805	0.5926	0.5800	0.5918
SMFH	0.5688	0.5586	0.5917	0.5727	0.5953	0.5841	0.5961	0.5828
FSH	0.6061	0.5979	0.6149	0.6114	0.6194	0.6186	0.6242	0.6251
UGACH	0.613	0.603	0.623	0.614	0.628	0.640	0.631	0.641
SGACH-TPS	0.4692	0.4486	0.456	0.4404	0.4498	0.4338	0.4429	0.4303

Table 4.8: Mean Average Precision (MAP) Comparison over Two Retrieval Tasks on NUS-WIDE Dataset

- This model again proves that supervised models outperform the unsupervised models and highlights the importance of preserving semantic information.
- The MAP values for Text-to-Image retrieval are lower than the past works. However, we can see that the performance of both Image-to-Text and Text-to-Image are comparable, which indicated balanced learning of the features of both modalities in the common space.
- The MAP values of the model when only pairwise information is used [SGACH-TPS (PI)], is better than when pairwise and category both information is used. The reason for this can be that the pairwise information is the natural information that is present in the data. However, category information can be vague sometimes. When the model is trained on only pairwise information, the model uses the natural correlation between samples to get a better retrieval for queries. But such results are data-dependent and should not be expected with every kind of data.

Results on NUS-WIDE

The MAP values on NUS-WIDE dataset for SGACH-TPS are reported in Table 4.8. The analysis of the results is as follows:

- Performance of SGACH-TPS on NUS-WIDE dataset is lower than the state-of-the-art results. The reason for this is the difference in the type of dataset. Earlier, both the datasets we used, WikiVenue and UCSD, contained paired samples of different modalities with a single category label associated with each sample. However, with NUS-WIDE, this is not true. It contains paired samples with multiple categories (few of the total 81 categories) associated

with each paired sample. This can be the reason for the model not learning the data distribution correctly.

- With the increase in the hash code length, the model's performance fluctuates. This shows that the model is not properly learning, and the performance is mere coincidence. This again supports the above claim that the model is not capable of performing on the dataset containing multiple labels with a single sample.

4.4 Summary

All the results on WikiVenue, Wiki, and NUS-WIDE datasets using the SGACH-TPS model are discussed in this chapter. For venue discovery task, results on WikiVenue dataset are presented. There are no existing results on this dataset, so we present results on existing state-of-the-art method for cross-modal retrieval for venue discovery task. Our model performs better than the state-of-the-art for this task. For generalizing our work, we also present results on Wiki and NUS-WIDE datasets. Our model outperformed all the existing state-of-the-art methods for general cross-modal retrieval for Image-to-Text retrieval. Detailed analysis of the results is also presented in this chapter.

Chapter 5

Conclusion and Future Work

In this thesis, we have proposed a supervised method for generating hash codes for cross-modal retrieval for venue discovery task, named SGACH-TPS. We use an adversarial model to learn hash functions for the bimodal data. Based on GAN [14], the model has two modules, the generative model and discriminative model. The discriminative model can be independently used as a hash function after the training of the model. We also proposed a novel way of training the GAN model to preserve pairwise information from the dataset by transferring the information from the pre-training to the fine-tuning stage.

For the retrieval task, the hash codes generated using the learned hash functions can be compared in the Hamming space, and a ranked list can be created corresponding to a query. We compared our model with state-of-the-art methods and outperformed them.

We presented several experiments and results to evaluate our model. Most of the findings related to the venue discovery task seem promising for building an application. Our model also performs well on unseen real-world images. However, for the generalized task of cross-modal retrieval, our model is yet to improve. The essential drawback of the model is the inability to learn the data distribution when there are multiple labels associated with each sample of the dataset. Also, Text-to-Image retrieval still lacks behind the state-of-the-art methods for the generalized task.

This work has been accepted in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) [42]. In the future, we wish to overcome all existing drawbacks in our work. Experiments with other sampling strategies for triplet loss function and improvement in generative model architecture can be explored. Exact venue search with and without location information is also an open area of research left to explore. We also hope to experiment with quantization techniques to generate better hash codes.

Bibliography

[1] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. "Canonical Correlation Analysis: An Overview with Application to Learning Methods". In: *Neural Computation* 16.12 (2004), pp. 2639–2664. DOI: 10.1162/0899766042321814.

[2] Grant Schindler, Matthew Brown, and Richard Szeliski. "City-scale location recognition". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer. 2007, pp. 1–7.

[3] James Hays and Alexei A Efros. "IM2GPS: estimating geographic information from a single image". In: *2008 ieee conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[4] Yunpeng Li, David J Crandall, and Daniel P Huttenlocher. "Landmark classification in large-scale image collections". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 1957–1964.

[5] Yair Weiss, Antonio Torralba, and Rob Fergus. "Spectral hashing". In: *Advances in neural information processing systems*. 2009, pp. 1753–1760.

[6] Michael M Bronstein et al. "Data fusion through cross-modality metric learning using similarity-sensitive hashing". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 3594–3601.

[7] Gerald Friedland, Oriol Vinyals, and Trevor Darrell. "Multimodal location estimation". In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010, pp. 1245–1252.

[8] David M Chen et al. "City-scale landmark identification on mobile devices". In: *CVPR 2011*. IEEE. 2011, pp. 737–744.

[9] Shaishav Kumar and Raghavendra Udupa. "Learning Hash Functions for Cross-view Similarity Search". In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*. IJCAI'11. Barcelona, Catalonia, Spain: AAAI Press, 2011, pp. 1360–1365. ISBN: 978-1-57735-514-4. DOI: 10.5591/978-1-57735-516-8/IJCAI11-230. URL: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-230.

[10] Yi Zhen and Dit-Yan Yeung. "Co-regularized hashing for multimodal data". In: *Advances in neural information processing systems*. 2012, pp. 1376–1384.

[11] Tiezheng Ge et al. "Optimized product quantization". In: *IEEE transactions on pattern analysis and machine intelligence* 36.4 (2013), pp. 744–755.

[12] Mohammad Rastegari et al. “Predictable Dual-View Hashing”. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 1328–1336. URL: http://proceedings.mlr.press/v28/rastegari13.html.

[13] Jingkuan Song et al. “Inter-media Hashing for Large-scale Retrieval from Heterogeneous Data Sources”. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’13. New York, New York, USA: ACM, 2013, pp. 785–796. ISBN: 978-1-4503-2037-5. DOI: 10.1145/2463676.2465274. URL: http://doi.acm.org/10.1145/2463676.2465274.

[14] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[15] Yao Hu et al. “Iterative multi-view hashing for cross media indexing”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 527–536.

[16] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. “Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 607–616.

[17] Ying Wei et al. “Scalable heterogeneous translated hashing”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 791–800.

[18] Ting Zhang, Chao Du, and Jingdong Wang. “Composite Quantization for Approximate Nearest Neighbor Search.” In: *ICML*. Vol. 2. 2014, p. 3.

[19] Jile Zhou, Guiguang Ding, and Yuchen Guo. “Latent Semantic Sparse Hashing for Cross-modal Similarity Search”. In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR ’14. Gold Coast, Queensland, Australia: ACM, 2014, pp. 415–424. ISBN: 978-1-4503-2257-7. DOI: 10.1145/2600428.2609610. URL: http://doi.acm.org/10.1145/2600428.2609610.

[20] Jile Zhou, Guiguang Ding, and Yuchen Guo. “Latent semantic sparse hashing for cross-modal similarity search”. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM. 2014, pp. 415–424.

[21] Yueting Zhuang et al. “Cross-media hashing with neural networks”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 901–904.

[22] Tsung-Yi Lin et al. “Learning deep representations for ground-to-aerial geolocalization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5007–5015.

[23] Daixin Wang et al. “Deep multimodal hashing with orthogonal regularization”. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

[24] Botong Wu et al. “Quantized correlation hashing for fast cross-modal search”. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

[25] Yue Cao et al. “Correlation autoencoder hashing for supervised cross-modal search”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM. 2016, pp. 197–204.

[26] Bor-Chun Chen et al. “Business-aware visual concept discovery from social media for multimodal business venue recognition”. In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[27] G. Ding et al. “Large-Scale Cross-Modality Search via Collective Matrix Factorization Hashing”. In: *IEEE Transactions on Image Processing* 25.11 (2016), pp. 5427–5440. ISSN: 1057-7149. DOI: 10.1109/TIP.2016.2607421.

[28] Hong Liu et al. “Supervised matrix factorization for cross-modality hashing”. In: *arXiv preprint arXiv:1603.05572* (2016).

[29] Rajiv Ratn Shah et al. “Leveraging multimodal information for event summarization and concept-level sentiment analysis”. In: *Knowledge-Based Systems* 108 (2016), pp. 102–109.

[30] Ting Zhang and Jingdong Wang. “Collaborative Quantization for Cross-Modal Similarity Search”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[31] Yue Cao et al. “Collective deep quantization for efficient cross-modal retrieval”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[32] Hong Liu et al. “Cross-modality binary code learning via fusion similarity hashing”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7380–7388.

[33] Rajiv Shah and Roger Zimmermann. *Multimodal analysis of user-generated multimedia content*. Springer, 2017.

[34] Bokun Wang et al. “Adversarial cross-modal retrieval”. In: *Proceedings of the 25th ACM international conference on Multimedia*. ACM. 2017, pp. 154–162.

[35] Xiaolong Gong, Linpeng Huang, and Fuwei Wang. “Deep Semantic Correlation Learning Based Hashing for Multimedia Cross-Modal Retrieval”. In: *2018 IEEE International Conference on Data Mining (ICDM)* (2018), pp. 117–126.

[36] Xiaolong Gong, Linpeng Huang, and Fuwei Wang. “Deep Semantic Correlation Learning Based Hashing for Multimedia Cross-Modal Retrieval”. In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, pp. 117–126.

[37] Chuan-Xiang Li et al. “SCRATCH: A Scalable Discrete Matrix Factorization Hashing for Cross-Modal Retrieval”. In: *ACM Multimedia*. 2018.

[38] Cong Li et al. “Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 4242–4251.

[39] Yi Yu et al. “Category-based deep CCA for fine-grained venue discovery from multimodal data”. In: *IEEE transactions on neural networks and learning systems* 99 (2018), pp. 1–9.

[40] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. "Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network". In: *IEEE transactions on cybernetics* (2018).

[41] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. "Unsupervised generative adversarial cross-modal hashing". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[42] Himanshu Aggarwal et al. "Supervised Generative Adversarial Cross-Modal Hashing by Tranferring Pairwise Similarities for Venue Discovery". In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2019.

[43] Yi Yu et al. "Deep cross-modal correlation learning for audio and lyrics in music retrieval". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.1 (2019), p. 20.

[44] Jian Zhang and Yuxin Peng. "Multi-pathway Generative Adversarial Hashing for Unsupervised Cross-modal Retrieval". In: *2019 IEEE Transactions on Multimedia* (2019).

[45] Jian Zhang and Yuxin Peng. "SSDH: Semi-supervised Deep Hashing for Large Scale Image Retrieval". In: *2019 IEEE Transactions on Circuits and System for Video Technology* (2019).