CSE 353 Milestone 1 Report

Topic: Text-Image Similarity

1. Introduction

Cross modal retrieval approach is widely used to learn common representation for the heterogeneous data (e,g., text data and image data) [1]. This report analyzes the research on two categories of cross-modal retrieval--retrieval across different modalities: 1. Binary-valued representation learning, 2. Real-valued representation learning. Each denoted as 2-a and 2-b will be elaborated in the following section. Subsequently, the DSCMR, as one of the state-of-art technologies, will be implemented and analyzed in the experiment section. Lastly, we list the notable points from the related works and suggest the objective of our future works.

2. Related Works

a. Binary-valued representation

Hashing has been widely used for information retrieval as it can perform fast retrieval speed with low stage cost [2]. If we use binary hash codes to represent the original data, the storage cost can be reduced even more. In addition to that, using hash codes to construct an index can achieve a constant or sub-linear time complexity for search [3].

Multi-modal hashing is for the data that has multi-modalities, which can be divided into two main categories. One of them is called cross-modal hashing (CMH), and the other is called multi-source hashing (MSH). There are a number of methods that have been proposed recently in both CMH and MSH. For CMH methods, for example, there are Semantics-Preserving Hashing method (SePH) [4], multi-modal latent binary embedding (MLBE) [5], and deep cross-modal hashing (DCMH) [6]. Among two types of methods, MSH aims to learn hash codes by utilizing all the information from multiple modalities. In general, however, it is difficult to acquire all modalities of all data points [7]. Due to the fact that the application of MSH is limited, we will focus more on CMH than MSH. Figure 1 demonstrates the overview of the hashing method.

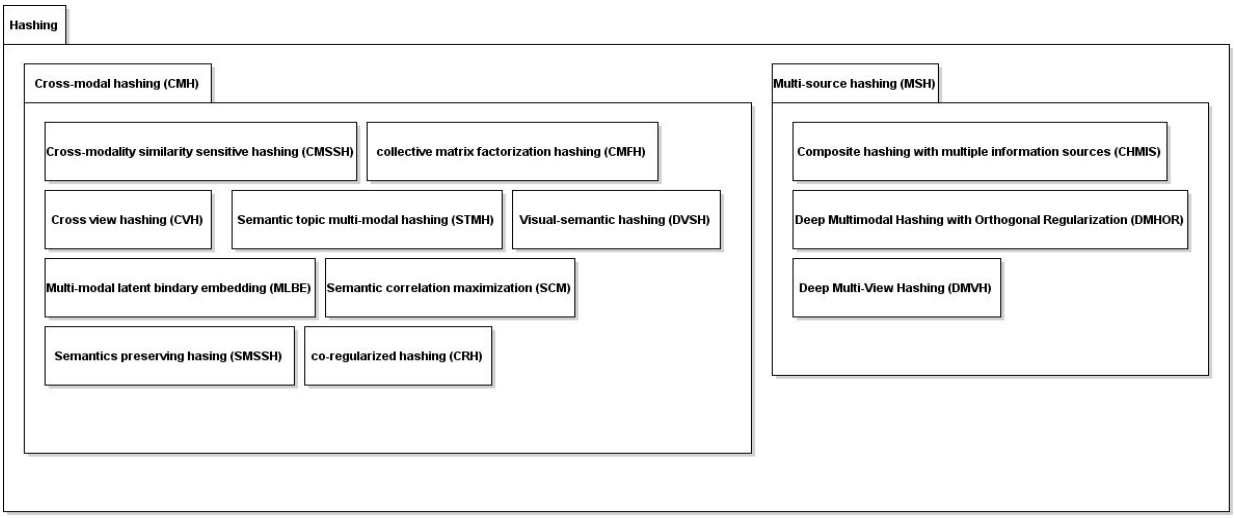

Figure 1: Overview of the Hashing

i. DCMH: Deep Cross-Modal Hashing [7]

This paper proposed a novel cross-modal hashing method called deep cross-modal hashing (DCMH). DCMH is an end to end learning framework with deep neural networks. To solve the hash-code learning problem known as the discrete learning problems, DCMH directly learns the discrete hash codes while most existing cross-modal hashing methods solve the problem by relaxing the discrete into continuous learning problems. DCMH consists of two parts: feature learning part and hash-code learning part. In the feature learning part, there are two deep neural networks as it has two inputs. Figure 2 depicts the architecture of the proposed method.

1. Image modality is a convolution neural network (CNN) with eight layers. The first seven layers are as same as CNN-F of paper, “Return of the Devil in the Details: Delving Deep into Convolutional Nets.” These seven layers use the Rectified Linear Unit (RELU) as their activation function. The last layer is a fully-connected layer with its activation function as identity function.
2. The other deep neural network is for text. Each text is represented as a vector with bag-of-words representation and these vectors are used as the input of a deep neural network with two layers. One layer uses RELU, and the other uses the identity function as their activation function.

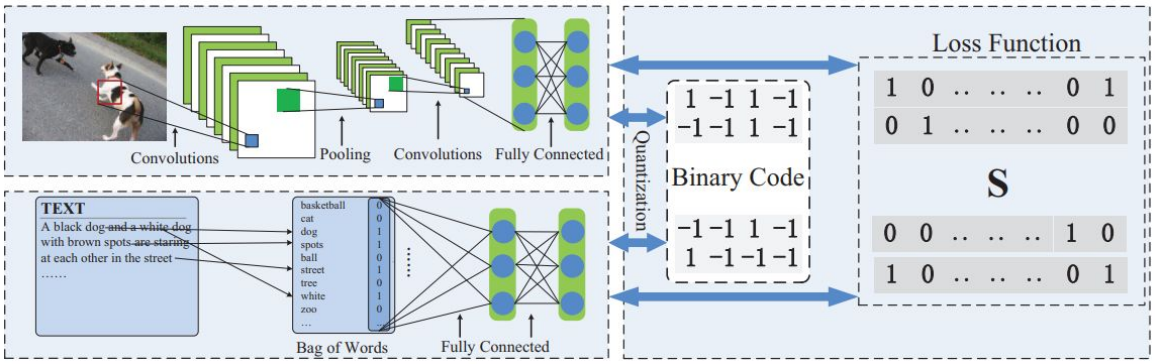

Figure 2: Architecture of DCMH

The objective function of DCMH is defined in the hash-code part as follows:

$$\begin{aligned}\min_{\mathbf{B}^{(x)},\mathbf{B}^{(y)},\theta_x,\theta_y} \quad & \mathcal{J} = -\sum_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}})) \\ & + \gamma(\|\mathbf{B}^{(x)} - \mathbf{F}\|_F^2 + \|\mathbf{B}^{(y)} - \mathbf{G}\|_F^2) \\ & + \eta(\|\mathbf{F1}\|_F^2 + \|\mathbf{G1}\|_F^2) \\ s.t. \quad & \mathbf{B}^{(x)} \in \{-1,+1\}^{c\times n}, \\ & \mathbf{B}^{(y)} \in \{-1,+1\}^{c\times n},\end{aligned}$$

In the paper, the author found that better performance can be achieved if the binary codes from the two modalities are set to be the same for the same training points. Therefore the above formulation can be transformed as follows:

$$\begin{aligned}\min_{\mathbf{B},\theta_x,\theta_y} \quad & \mathcal{J} = -\sum_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}})) \\ & + \gamma(\|\mathbf{B} - \mathbf{F}\|_F^2 + \|\mathbf{B} - \mathbf{G}\|_F^2) \\ & + \eta(\|\mathbf{F1}\|_F^2 + \|\mathbf{G1}\|_F^2) \\ s.t. \quad & \mathbf{B} \in \{-1,+1\}^{c\times n}.\end{aligned}$$

As we did not fully understand the hash-code part of this paper--we will further investigate this hash-code part.

ii. SePH: Semantics-Preserving Hashing for Cross-View Retrieval [8]

The main contribution of this paper is that it proposes an effective Semantics-Preserving Hashing method(SePH) to solve the problem of cross-modal retrieval. To reduce the

storage costs, SepH learns one unified hash code instead of learning different hash codes for each modal.

SePH first learns the semantics-preserving hash codes of training data by minimizing the KL-divergence of the derived probability distribution in Hamming space from that in semantic space. And then, hash functions for each view are learned to project features into the learned hash codes. When there are any unseen instances, predicted hashcode and the corresponding output probabilities from learnt hash functions in all observed views are utilized to determine its unified hash code with a novel probabilistic approach.

b. Real-valued representation learning

Unlike the binary approach, the real-valued approach employs the real data rather than binary information in terms of common representation space. The category includes three subcategories: unsupervised approaches, pairwise approaches, and supervised approaches [9]. The unsupervised approach only utilizes co-occurrence information of different modality to learn common representations [10]. Meanwhile, the pairwise-based method uses more similar pairs for comparing samples from different types of data [11]. Lastly, the supervised approach employs the semantic category labels for the learning of common representation [12]. In this report, we introduce two examples of the real-valued representation learning, Cross-modal Generative Adversarial Networks for Common Representation Learning (CM-GANs) and Deep Supervised Cross-modal Retrieval (DSCMR).

i. CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning [13]

This research utilizes Generative Adversarial Network (GAN) to deal with heterogeneous modalities, compared to most of the existing GANs-based works that mainly focus on unimodal problems such as image synthesis.

The basic model of GANs consists of two components; a generative model G captures the data distribution, and the discriminative model D attempts to discriminate whether the input sample comes from the real data or is generated from G. The general overview of CM-GAN is as follows:

1. Cross-modal convolutional autoencoders for both image and text form the generative model to exploit both cross-modal representation as well as reconstruction information.
2. Two kinds of discriminative models entail intra-modality discrimination and inter-modality, making discrimination on both common representation and reconstruction representation for mutual boosting.

ii. Deep Supervised Cross-modal Retrieval (DSCMR) [9]

The objective of this study is to demonstrate a novel cross-modal retrieval method that minimizes the discrimination loss in label space and common representation space, as well as modality invariance loss. The general architecture of the proposed method is as follows:

1. Image and text are imputed into image CNN (19-layer VGGNet) and text CNN (sentence CNN) respectively and obtain high-level semantic representations.
2. The fully connected layers with ReLU are added on the top of each representation to ensure the two sub-networks map to the common representation, by sharing the weights of the last layers. The number of hidden units for each modality are 2048, and 1024, respectively.
3. A linear classifier is connected to each image modal and the text modal to learn discriminative features by exploiting the label information.

The visual representation of the architecture is as shown in Figure 3.

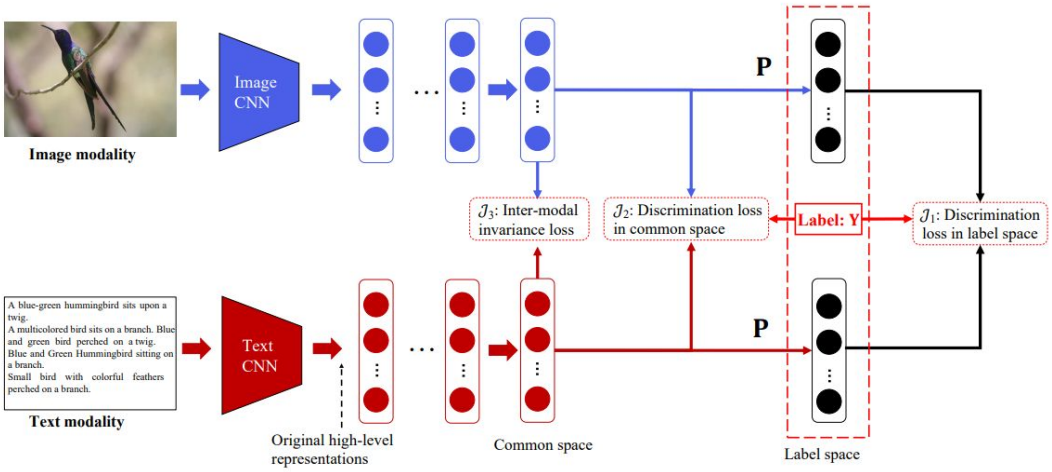

Figure 3: Architecture of DSCMR

$$\mathcal{J} = \mathcal{J}_1 + \lambda\mathcal{J}_2 + \eta\mathcal{J}_3,$$

Figure 4: Objective function of DSCMR

The objective function of DSCMR is as depicted in Figure 4--J1, J2, J3 denotes the discrimination loss in the label space, discrimination loss from two modalities in the common representation space, and the modality invariance loss, respectively.

3. Experiment and Performance Analysis

In this section, we demonstrate the code execution and the analysis of the third approach--DSCMR. This study conducted experiments with the three datasets: Wikipedia dataset, Pascal Sentence dataset, and the NUS-WIDE-10k dataset. To evaluate the method, mean Average Precision (mAP) score is used since it is a widely used statistical metrics for the research regarding cross-modal retrieval. Figure 5 demonstrates the code execution result of the proposed method.

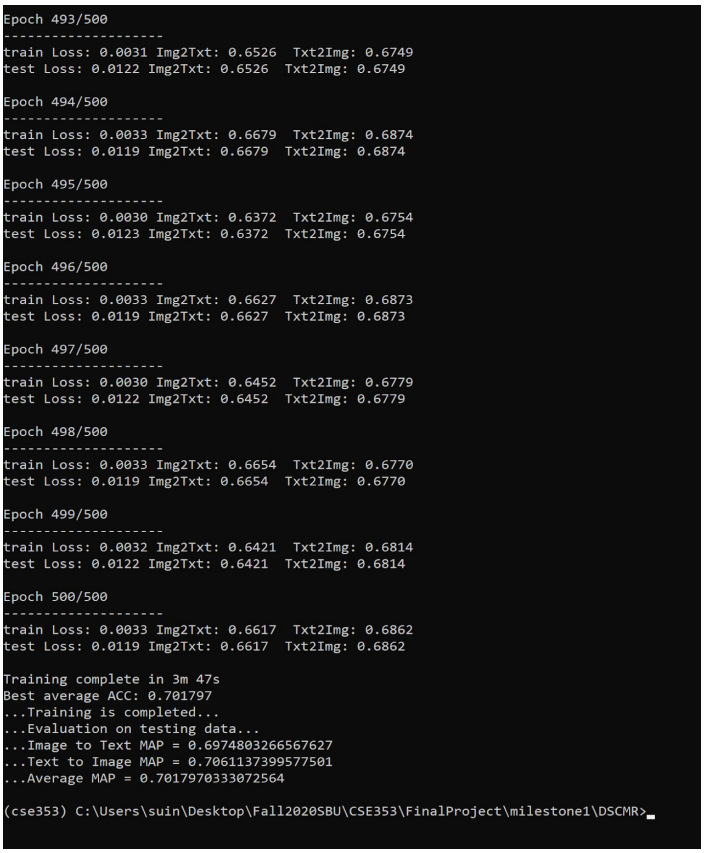

Figure 5: Result of the DSCMR source code execution

As shown above, the study investigated the mAP scores of the two different retrieval tasks: Image to Text and Text to Image.

4. Notable points and our Future Work

The notable conclusions that we could obtain from the related works for our studies are as follows:

1. DCMH, as one of the Binary-valued approaches, ensures computational efficiency when retrieval, by mapping the different modality into common hamming space.
2. DSCMR, as one of the Supervised real-valued approaches, enhanced performance in terms of mAP scores compared to other cross-modal retrieval methods proposed in this report.

Hence, for future works, we will design our architecture, applying and enhancing the distinguishable features from these two approaches.

References

[1] Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017, October). Adversarial cross-modal retrieval. In Proceedings of the 25th ACM international conference on Multimedia (pp. 154-162).

[2] Kulis, B., & Grauman, K. (2009, September). Kernelized locality-sensitive hashing for scalable image search. In 2009 IEEE 12th international conference on computer vision (pp. 2130-2137). IEEE.

[3] Kong, W., & Li, W. J. (2012). Isotropic hashing. In Advances in neural information processing systems (pp. 1646-1654).

[4] Lin, Z., Ding, G., Hu, M., & Wang, J. (2015). Semantics-preserving hashing for cross-view retrieval. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3864-3872).

[5] Zhen, Y., & Yeung, D. Y. (2012, August). A probabilistic model for multimodal hash function learning. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 940-948).

[6] Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P. S. (2016, August). Deep visual-semantic hashing for cross-modal retrieval. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1445-1454).

[7] Jiang, Q. Y., & Li, W. J. (2017). Deep cross-modal hashing. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3232-3240).

[8] Lin, Z., Ding, G., Hu, M., & Wang, J. (2015). Semantics-preserving hashing for cross-view retrieval. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3864-3872).

[9] Zhen, L., Hu, P., Wang, X., & Peng, D. (2019). Deep supervised cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10394-10403).

[10] Wang, K., Yin, Q., Wang, W., Wu, S., & Wang, L. (2016). A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215.

[11] Zhai, D., Chang, H., Shan, S., Chen, X., & Gao, W. (2012). Multiview metric learning with global consistency and local smoothness. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3), 1-22.

[12] Sharma, A., Kumar, A., Daume, H., & Jacobs, D. W. (2012, June). Generalized multiview analysis: A discriminative latent space. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2160-2167). IEEE.

[13] Peng, Y., & Qi, J. (2019). CM-GANs: Cross-modal generative adversarial networks for common representation learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1), 1-24.