Modeling Text with Graph Convolutional Network for Cross-Modal Information Retrieval

Jing Yu\textsuperscript{1}, Yuhang Lu\textsuperscript{1,2}, Zengchang Qin\textsuperscript{3}, Yanbing Liu\textsuperscript{1}, Jianlong Tan\textsuperscript{1}, Li Guo\textsuperscript{1}, Weifeng Zhang\textsuperscript{4}

\textsuperscript{1}Institute of Information Engineering, Chinese Academy of Sciences, China
\textsuperscript{2}School of Cyber Security, University of Chinese Academy of Sciences, China
\textsuperscript{3}Intelligent Computing \& Machine Learning Lab, School of ASEE, Beihang University, China
\textsuperscript{4}Hangzhou Dianzi University, China

\{yujing02, luyuhang, liuyanbing, tanjianlong, guoli\}@iie.ac.cn, zcqin@buaa.edu.cn, zwf.zhang@gmail.com

Abstract

Cross-modal information retrieval aims to find heterogeneous data of various modalities from a given query of one modality. The main challenge is to map different modalities into a common semantic space, in which distance between concepts in different modalities can be well modeled. For cross-modal information retrieval between images and texts, existing work mostly uses off-the-shelf Convolutional Neural Network (CNN) for image feature extraction. For texts, word-level features such as bag-of-words or word2vec are employed to build deep learning models to represent texts. Besides word-level semantics, the semantic relations between words are also informative but less explored. In this paper, we model texts by graphs using similarity measure based on word2vec. A dual-path neural network model is proposed for couple feature learning in cross-modal information retrieval. One path utilizes Graph Convolutional Network (GCN) for text modeling based on graph representations. The other path uses a neural network with layers of nonlinearities for image modeling based on off-the-shelf features. The model is trained by a pairwise similarity loss function to maximize the similarity of relevant text-image pairs and minimize the similarity of irrelevant pairs. Experimental results show that the proposed model outperforms the state-of-the-art methods significantly, with 17% improvement on accuracy for the best case.

1 Introduction

For past a few decades, online multimedia information in different modalities, such as image, text, video and audio, has been increasing and accumulated explosively. Information related to the same content or topic may exist in various modalities and has heterogeneous properties, that makes it difficult for traditional uni-modal information retrieval systems to acquire comprehensive information. There is a growing demand for effective and efficient search in the data across different modalities. Cross-modal information retrieval \cite{rasiwasia2010cross, yu2012cross, wang2016cross} enables users to take a query of one modality to retrieve data in relevant content in other modalities. However, there is no natural correspondence between different modalities. Previous research has made continuous effort on designing appropriate distance measure of similarity and gained great progress. A common solution is to learn a common latent semantic space to compare all modalities of data directly, typically using probabilistic models \cite{blei2003latent}, metric learning \cite{zheng2012metric}, subspace learning \cite{hardoon2004canonical, rasiwasia2010cross, sharma2012metric}, and joint modeling methods \cite{wang2016joint}. A brief survey is available in \cite{wang2016cross}.

Feature representation is the footstone for cross-modal information retrieval. In the case of text-image retrieval, off-the-shelf features learnt by deep models are widely used to represent images. Most methods \cite{zhang2017cross, wang2016joint} use Convolutional Neural Network (CNN) \cite{lecun1998gradient} to learn the visual features obtained from
In this kind of models, bag-of-words (BOW) is commonly used in cross-modal information retrieval [Liu et al., 2016; Wang et al., 2016c]. Intuitively, the text document is represented by a word-frequency vector regardless of the word order. Although some weighting schemes based on word frequency have been proposed to enhance the feature discrimination [Wang et al., 2016b], one common problem is that the relations among words are not considered. Recently, word2vec [Mikolov et al., 2013] becomes one of the best models for semantical modeling of word semantics. It’s pre-trained on GoogleNews to learn the vector representation from the context information. [Wang et al., 2016b] extracts word vectors via word2vec model and adopts Fisher vector encoding to obtain the sentence representation. [Zhang et al., 2017] represents a text by calculating a mean vector of all the word word2vec vectors in a text. Although this kind of word vector is enriched by learning from neighboring words, it still ignores the global structural information inherent in the texts and only treat the word as “flat” features. In light of the common weakness in vector-space models, recent research has found that the relations among words could provide rich semantics of the texts and can effectively promote the text classification performance [Wang et al., 2016a].

In this paper, we represent a text as a structured and featured graph and learn text features by a graph-based deep model, i.e. Graph Convolutional Network (GCN) [Defferrard et al., 2016]. Such a graph can well capture the semantic relations among words. GCN allows convolutions to be dealt as multiplication in the graph spectral domain, rendering the extension of CNN to irregular graphs. (Figure 1 shows the comparison of our model to classical cross-modal retrieval models.) The GCN model has a great ability to learn local and stationary features on graphs, which was successfully used in text categorization [Kipf and Welling, 2017] and brain network matching [Kienzle et al., 2017]. Based on this graph representation for texts, we propose a dual-path neural network, called Graph-In-Network (GIN), for cross-modal information retrieval. The text modeling path contains a GCN on the top of graph representations. The image modeling path contains a neural network with layers of nonlinearities on the top of off-the-shelf image representations. To train the model, we employ a pairwise similarity loss function [Kumar BG et al., 2016], that maximizes the similarity between samples in the same semantic concept and minimizes the similarity between samples in different semantic concepts.

The main contributions can be summarized as follows:

- The model can jointly learn the textual and visual representations as well as text-image similarity metric, providing an end-to-end training mode;
- Experimental results on five benchmark datasets show the superior performance of our model over the state-of-the-art methods, verifying the benefits of using graphs to model the irregular textual data.

2 Methodology

In this paper, we propose a dual-path neural network to simultaneously learn multi-modal representations and similarity metric in an end-to-end mode. In the text modeling path (top in Figure 2), each text is represented by a featured graph and the text GCN is used to learn the feature representation. It has two key steps: graph construction and GCN modeling.

2.1 Text Modeling

Graph Construction: Classical methods semantically model the fundamental features of a text only by word vectors regardless of the structural information. In this work, we represent a text by a featured graph to combine the strengths of structural information with semantic information together. Given a set of text documents, we extract the most common words, denoted as \( W = [w_1, w_2, ..., w_N] \), from all the unique words in this corpus and represent each word by a pre-trained word2vec embedding. For the graph structure, we construct a \( k \)-nearest neighbor graph, denoted as \( G = (V, E) \). Each vertex \( v_i \in V \) is corresponding to a unique word and each edge \( e_{ij} \in E \) is defined by the word2vec similarity between two words:

\[
e_{ij} = \begin{cases} 
1 & \text{if } w_i \in N_k(w_j) \text{ or } w_j \in N_k(w_i) \\
0 & \text{otherwise}
\end{cases}
\]

(1)

where \( N_k(\cdot) \) denotes the set of \( k \)-nearest neighbors by computing the cosine similarity between word word2vec embeddings. \( k \) is the parameter of neighbor numbers (set to 8 in our following experiments). The graph structure is stored by an adjacent matrix \( A \in \mathbb{R}^{N \times N} \). For the graph features, each text document is represented by a bag-of-words vector and the frequency value of word \( w_i \) serves as the 1-dimensional feature on vertex \( v_i \). In this way, we combine structural information of word similarity relations and semantic information of word vector representation in a featured graph. Note that the graph structure is identical for a corpus and we use different graph features to represent each text in a corpus.

GCN Modeling: In modeling text corpora, deep network models have become increasingly popular and achieved breakthroughs in many machine learning areas. However, classical deep network models are defined for grid-structured data and can not be easily extended to graphs. It’s challenging to define the local neighborhood structures and the vertex orders for graph operations. Recently, Graph Convolutional Network (GCN) [Defferrard et al., 2016] is proposed to generalize Convolutional Neural Network (CNN) to irregular-structured graphs. The basic idea is that, based on spectral graph theory, the graph convolutions can be dealt as multiplications in the graph spectral domain. The feature maps can
Figure 2: The structure of the proposed model is a dual-path neural network: i.e., text Graph Convolutional Network (text GCN) (top) and image Neural Network (image NN) (bottom). The text GCN for learning text representation contains two layers of graph convolution on the top of constructed featured graph. The image NN for learning image representation contains layers of non-linearities initialized by off-the-shelf features. They have the same dimension in the last fully connected layers. The objective is a global pairwise similarity loss function.

be obtained by inverse transform from the graph spectral domain to original graph domain. In this paper, the text features are learnt by GCN given the graph representation of a text document.

Given a text, we define its input graph feature vector by $F_{in}$ and we denote the output feature vector after graph convolution by $F_{out}$. Firstly $F_{in}$ is transformed to the spectral domain via graph Fourier transform. This transform is based on the normalized graph Laplacian, defined as $L = I_N - D^{-1/2} A D^{-1/2}$, where $I_N$ and $D$ are respectively the identity matrix and diagonal degree matrix of the graph structure $G$. Then $L$ can be eigendecomposed as $L = U \Lambda U^T$, where $U$ is a set of eigenvectors and $\Lambda$ is a set of real, non-negative eigenvalues. The Fourier transform of $F_{in}$ is a function of $U$ defined as:

$$\hat{F}_{in} = U^T F_{in}$$

While the inverse transform is defined as:

$$F_{in} = U \hat{F}_{in}$$

The convolution of $F_{in}$ with a spectral filter $g_\theta$ is given by:

$$F_{out} = g_\theta \ast F_{in} = U g_\theta U^T F_{in}$$

where parameter $\theta$ is a vector to learn. In order to keep the filter $K$-localized in space and computationally efficient, [Defferrard et al., 2016] proposes a approximated polynomial filter defined as:

$$g_\theta = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L})$$

where $T_k(x) = 2x T_{k-1}(x) - T_{k-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$, $\tilde{L} = \frac{2}{\lambda_{\text{max}}} L - I_N$ and $\lambda_{\text{max}}$ denotes the largest eigenvalue of $L$. The filtering operation can then be written as $F_{out} = g_\theta F_{in}$. In our model, we use the same filter as in [Defferrard et al., 2016]. For the graph representation of a text document, the $i^{th}$ input graph feature $f_{in,i} \in F_{in}$ is the word frequency of vertex $v_i$. Then the $i^{th}$ output feature $f_{out,i} \in F_{out}$ is given by:

$$f_{out,i} = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) f_{in,i}$$

where we set $K=3$ in the experiments to keep each convolution at most 3-steps away from a center vertex.

Our text GCN contains two layers of graph convolutions, each followed by Rectified Linear Unit (ReLU) activation to increase non-linearity. A fully connected layer is successive with the last convolution layer to map the text features to the common latent semantic space. Given a text document $T$, the text representation $f_t$ learnt by the text GCN model $H_t(\cdot)$ is denoted by:

$$f_t = H_t(T)$$

2.2 Image Modeling

For modeling images, we adopt a neural network (NN) containing a set of fully connected layers (bottom in Figure 2). We have three options of initializing inputs by hand-crafted feature descriptors, pre-trained neural networks, or jointly trained end-to-end neural networks. In this paper, the first two kinds of features are used for fair comparison with other models. The input visual features are followed by a set of fully
connected layers for fine-tuning the visual features. Similar to text modeling, the last fully connected layer of image NN maps the visual features to the common latent semantic space with the same dimension as text. In experimental studies, we tune the number of layers and find that only keeping the last semantic mapping layer without feature fine-tuning layers can obtain satisfactory results. Given an image \( I \), the image representation \( f_{\text{img}} \) learnt by the model from image NN \( H_{\text{img}}(\cdot) \) is represented by:

\[
f_{\text{img}} = H_{\text{img}}(I)
\]

(8)

### 2.3 Objective Function

Distance metric learning is applied to estimate the relevance of features learned from the dual-path model. The outputs of the two paths, i.e., \( f_{t} \) and \( f_{\text{img}} \), are in the same dimension and combined by an inner product layer. The successive layer is a fully connected layer with one output score \( u \), denoting the similarity score function between a text-image pair.

The training objective is a pairwise similarity loss function where we sequentially select \( Q \) different semantic concepts. Meanwhile, we also minimise the variance of pairwise similarity score for both matching similarity and non-matching similarity.

\[
\text{Loss} = (\sigma^2 + \sigma^2) + \lambda \max(0, m - (u^+ - u^-))
\]

(9)

where \( \lambda \) is used to balance the weight of the mean and variance, and \( m \) is the margin between the mean distributions of matching similarity and non-matching similarity. \( u^+ = \sum_{i=1}^{Q_1} \frac{\text{score}(T_i, I_t)}{Q_1} \) and \( u^- = \sum_{i=1}^{Q_2} (\text{score}(T_i, I_t) - u^+) \) when text \( T_i \) and image \( I_t \) are in the same class. While \( u^+ = \sum_{j=1}^{Q_1} \frac{\text{score}(T_I, I_j)}{Q_2} \) and \( u^- = \sum_{j=1}^{Q_2} (\text{score}(T_I, I_j) - u^+) \) when \( T_J \) and \( I_J \) are in different classes. We train the model by mini-batch gradient descent with mini-batch size 200. In other words, we sequentially select \( Q_1 + Q_2 = 200 \) text-image pairs from the training set for each mini-batch in the experiments.

### 3 Experimental Studies

To evaluate the performance of our proposed model, we conduct extensive experiments to investigate cross-modal retrieval tasks, i.e., text-query-images and image-query-texts.

#### 3.1 Datasets

Experiments are conducted on four English benchmark datasets, i.e., English Wikipedia, NUS-WIDE, Pascal VOC, and TVGraz. To verify the extensibility of our model, we also conduct experiments on the Chinese Wikipedia dataset. Each dataset contains a set of text-image pairs. Images are represented by off-the-shelf feature vectors while texts are represented by featured graphs.

- **English Wikipedia** dataset (Eng-Wiki for short) \cite{Rasiwasia2010} contains 2,866 image-text pairs divided into 10 classes, where 2,173 pairs are for training and 693 pairs are for testing. Each image is represented by a 4,096-dimensional vector extracted from the last fully connected layer of VGG-19 model \cite{Simonyan2015}. Each text is represented by a graph with 10,055 vertices.
- **NUS-WIDE** dataset consists of 269,648 image-tag pairs, which are pruned from the NUS dataset by keeping the pairs belonging to one or more of the 10 largest classes. We select samples in the 10 largest classes as adopted in \cite{Zhang2017}. For images, we use 500-dimensional bag-of-features. For tags, we construct a graph with 5,018 vertices.
- **Pascal VOC** dataset consists of 9,963 image-tag pairs belonging to 20 classes. The images containing only one object are selected in our experiments as \cite{Sharma2012, Wang2013, Wang2016}. Obtaining 2,808 training and 2,841 testing samples. For the features, 512-dimensional Gist features are adopted for the images and a graph with 598 vertices is used for the tags.
- **Chinese Wikipedia** dataset (Ch-Wiki for short) \cite{Qin2016} is collected from Chinese Wikipedia articles. It contains 3,103 image-text pairs divided into 9 classes, where 2,482 pairs are for training and 621 pairs are for testing. Each image is represented by a 4,096-dimensional output of VGG-19. Each text is represented by a graph with 9,613 vertices.

#### 3.2 Evaluation and Implementation

We compare our proposed GIN with a number of state-of-the-art models, including CCA & SCM \cite{Rasiwasia2010}, TCM & w-TCM & c-TCM \cite{Qin2016}, GMLDA & GMMFA \cite{Sharma2012}, LCFS \cite{Wang2013}, MvDA \cite{Kan2016}, LGCFL \cite{Kang2017}, m-CCA \cite{Ranjan2015}, AUSL \cite{Zhang2017}, JFS-SSL \cite{Wang2016}, PLS \cite{Sharma2011}, BLM \cite{Sharma2012}, CDFE \cite{Lin2006}, CCA-3V \cite{Gong2014}, CM & SM \cite{Pereira2013}, and TTI \cite{Qi2011}. For the same settings with \cite{Zhang2017}, principal component analysis is performed on the original features for CCA, SCM, GMLDA and MvDA.

CCA, PLS and BLM are three popular un-supervised models that adopt pairwise information to maximize the correlation between projected vectors. AUSL and CCA-3V are semi-supervised models that leverage both labelled and unlabelled data to learn the common space. GMLDA, GMMFA, m-CCA, TCM, LCFS, LGCFL, JFSSSL, CDFE, and MvDA are supervised models that use the semantic class information to directly make data from one modality to correlate with data from another modality.

The mean average precision (MAP) is used to evaluate the performance of all the algorithms on the five datasets. Higher MAP indicates better retrieval performance. Meanwhile, the precision-recall (PR) curve \cite{Rasiwasia2010} is also
3.3 Experimental Results

The MAP scores of all the methods on the five benchmark datasets are shown in Table 1. All the other models are well cited in this field. Since not all the papers have tested these five datasets, for fair comparison, we compare our model to methods on their reported datasets with the same preprocessing conditions. From Table 1 we can have the following observations:

First, GIN outperforms all the compared methods over the five datasets for the text-query-image task. On the Eng-Wiki and Pascal datasets, the MAP scores of GIN are 76.72% and 45.15%, which are about 35.70% and 17.14% higher than the second best result from JFSSL. For the NUS-WIDE dataset, the MAP score of GIN is 54.18% and 12.9% higher than the second best result from AUSL. It’s obvious that no matter for the rich text, e.g. Eng-Wiki and TVGraz, or for the sparse tags, e.g. NUS-WIDE and Pascal, our model gains the superior performance for the text-query-image task. The reason is that the proposed model effectively keeps the inter-word semantic relations by representing the texts with graphs, which has been ignored by other methods that represent the texts with only feature vectors, no matter skip-gram vectors or word frequency vectors. Such inter-word relations are enhanced and more semantically relevant words are activated with the successive layers of graph convolutions, resulting in discriminative representations of the text modality.

Second, the MAP score of GIN for the image-query-text task is superior to most of the compared methods. GIN ranks the second best on Eng-Wiki and NUS-WIDE, the third best on Pascal and the best on TVGraz and Chi-Wiki. Table 1 indicates that GIN is only inferior to JFSSL by 1.44% on Eng-Wiki and 4.37% on Pascal. GIN is just 4.54% lower than AUSL on NUS-WIDE. Since GIN uses off-the-shelf feature vectors for image view, it’s normal that the performance is comparable with state-of-the-art results. The retrieval performance can be further improved if the feature extraction network was trained together with the fully connected layers in our model. In this paper, we didn’t focus on the vector feature selection problem.

Third, GIN achieves the best average MAP over all the competitors, especially outperforming the second best method JFSSL by 17.13% on Eng-Wiki. That’s mainly because that our learning framework can jointly seek a common latent semantic space and correlated feature representations of multi-modal data, which can be trained end-to-end. The parameters in the path of graph convolutional networks are learnt referring to the features in the image branch, which enhances the relations between different modal features in their original data domain. Moreover, the learnt distance metric is also improving the separation between matching and non-matching image-text pairs.

Finally, on TVGraz dataset, GIN obtains the best results for both retrieval tasks. The improvement for the image-query-text task is greater than that for the text-query-image task.
task, which is quite different from the observations on other datasets. The reason is that, for the image view, the existing algorithms represent images simply by bag-of-features with SIFT descriptors while we utilize the 4096-dimensional CNN features, which are proved to be much more powerful than the hand-crafted feature descriptors. In addition to English, the representative alphabetic language, we also conduct experiments on Chinese dataset to show the generalization ability of our model. On Ch-Wiki, GIN gains 6.77% and 2.43% improvement for the text query and image query, respectively.

The precision-recall (PR) curves of image-query-text and text-query-image are plotted in Figure 3. Since the competitive models, i.e. w-TCM and c-TCM, haven’t reported PR curves on Ch-Wiki, we compare GIN with random baseline on this dataset. For JFSSL, we show its best MAP after feature selection (see Table 7 in [Wang et al., 2016b]). Since JFSSL hasn’t reported the PR curves corresponding to the best MAP, we use its reported PR curves in [Wang et al., 2016b].

For the text-query-image task, it’s obvious that GIN achieves the highest precision than the compared methods with almost all the recall rate on the five benchmark datasets. For the image-query-text task, GIN outperforms other competitors with almost all the recall rate on Eng-Wiki. For NUS-WIDE dataset, GIN is only inferior to AUSL and LGCFL. For Pascal dataset, GIN is just slightly inferior to JFSSL. On the whole, GIN is comparable with state-of-the-art methods for the image-query-text task.

Discussion. In general, the proposed model shows superior performance for the text-query-image task in all the comparison experiments, especially on the three widely used benchmark datasets (i.e. Eng-Wiki, NUSE-WIDE, and Pascal), achieving about 17%∼35% remarkable improvement on MAP. It’s mainly because that the graph representation for text can well reserve the inherent property of semantic relations between different words, which provides an effective global prior knowledge for the successive GCN to model each text with distinctive feature input. In the graph convolutional procedure, such global prior knowledge guides the convolution to adaptively propagate the distinctive vertex features along semantic paths, which guarantees good generalization ability at conceptual level for the learnt text representation. The remarkable performance on text query proves that, compared with vector-space models, the incorporation of semantic structure is a great benefit and gains better generalization ability for un-seen data.

4 Conclusion

In this paper, we propose a novel cross-modal retrieval model named GIN that takes both irregular graph-structured textual representations and regular vector-structured visual representations into consideration to jointly learn coupled feature and common latent semantic space. A dual path neural network with graph convolutional networks and layers of nonlinearities is trained using a pairwise similarity loss function. Extensive experiments on five benchmark datasets demonstrate that our model considerably outperform the state-of-the-art models. Besides, our model can be widely used in analyzing heterogeneous data lying on irregular or non-Euclidean...
domains.

References

[Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling anotated data. In SIGIR, pages 127–134, 2003.

[Defferrard et al., 2016] Michael Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In NIPS, pages 3837–3845, 2016.

[Gong et al., 2014] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for internet images, tags, and their semantics. TPAMI, 106(2):210–233, 2014.

[Hardoon et al., 2004] David R. Hardoon, Sándor Szédmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 16(12):2639–2664, 2004.

[Kan et al., 2016] Meina Kan, Shiguang Shan, Hairong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. TPAMI, 38(1):188–194, 2016.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[Qi et al., 2011] Guojun Qi, Charu Aggarwal, and Tomas Huang. Towards semantic knowledge propagation from text corpus to web images. In WWW, pages 297–306, 2011.

[Qin et al., 2016] Zengchang Qin, Jing Yu, Yonghui Cong, and Tao Wan. Topic correlation model for cross-modal multimedia information retrieval. Pattern Analysis and Applications, 19(4):1007–1022, 2016.

[Ranjan et al., 2015] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In ACM-MM, pages 251–260, 2010.

[Sharma and Jacobs, 2011] Abhishek Sharma and David W Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution adn sketch. In CVPR, pages 593–600, 2011.

[Sharma et al., 2012] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In CVPR, pages 2160–2167, 2012.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[Wang et al., 2013] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In IJCAI, pages 2088–2095, 2013.

[Wang et al., 2016a] Chenguang Wang, Yangqiu Song, Haoran Li, Ming Zhang, and Jiawei Han. Text classification with heterogeneous information network kernels. In AAAI, pages 2130–2136, 2016.

[Wang et al., 2016b] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. TPAMI, 38(10):2010–2023, 2016.

[Wang et al., 2016c] Liang Wang, Ran He, Zhenan Sun, and Tieniu Tan. Group-invariant cross-modal subspace learning. In IJCAI, pages 1739–1745, 2016.

[Wang et al., 2016d] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In CVPR, pages 5005–5013, 2016.

[Yu et al., 2012] Jing Yu, Yonghui Cong, Zengchang Qin, and Tao Wan. Cross-modal topic correlations for multimedia retrieval. In ICPR, pages 246–249, 2012.

[Zhang et al., 2017] Liang Zhang, Bingpeng Ma, Jianfeng He, Guorong Li, Qingming Huang, and Qi Tian. Adaptively unified semi-supervised learning for cross-modal retrieval. In IJCAI, pages 3406–3412, 2017.

[Zheng and Yeung, 2012] Yi Zheng and Dit-Yan Yeung. Co-regularized hashing for multimodal data. In NIPS, pages 1376–1384, 2012.