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

Jing Yu[1,2], Yuhang Lu[1,2], Zengchang Qin[3(✉)], Weifeng Zhang[4,5], Yanbing Liu[1], Jianlong Tan[1], and Li Guo[1]

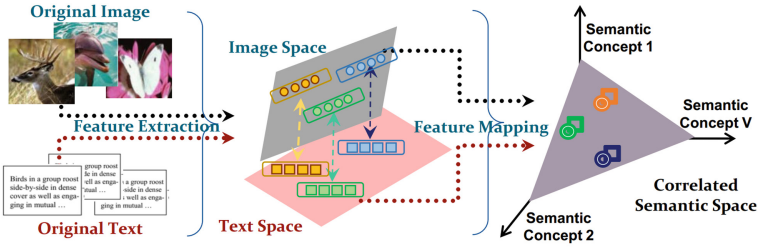
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{yujing02,luyuhang,liuyanbing,tanjianlong,guoli}@iie.ac.cn
[2] School of Cyber Security,
University of Chinese Academy of Sciences, Beijing, China
[3] Intelligent Computing and Machine Learning Lab,
School of ASEE, Beihang University, Beijing, China
zcqin@buaa.edu.cn
[4] Hangzhou Dianzi University, Hangzhou, China
zwf.zhang@gmail.com
[5] Zhejiang Future Technology Institute, Jiaxing, China

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 non-linearities 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.

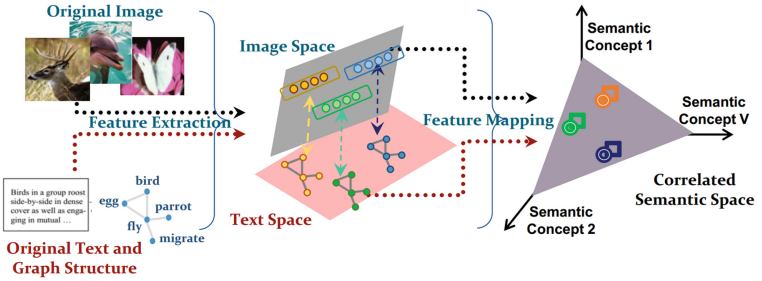
1 Introduction

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

R. Hong et al. (Eds.): PCM 2018, LNCS 11164, pp. 223–234, 2018.
https://doi.org/10.1007/978-3-030-00776-8_21



(a) Overview of classical cross-modal retrieval models.



(b) Overveiw of our proposed cross-modal retrieval model.

Fig. 1. Comparison of classical cross-modal retrieval models to our model. (a) Classical models adopt feature vectors to represent grid-structured multimodal data; (b) Our model can handle both irregular graph-structured data and regular grid-structured data simultaneously.

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 [13,17,20] enables users to take a query of one modality to retrieve data in relevant content in other modalities.

The mainstream solution for cross-modal retrieval is to project the features of different modalities into a common semantic space and measure their similarity directly. Thus, feature representation is the footstone for cross-modal information retrieval. Existing work treats the irregular-structured data (i.e. text, protein network) as "flat" features in a similar way as modeling grid-structured data (i.e. image, audio, video). Take text-image retrieval for example. Recent works [19,22] extract the image features by pre-trained Convolutional Neural Network (CNN) [8], which can leverage the local information in the grid-structured data to represent the visual semantics. For text representation, deep models are also widely applied to extract high-level semantics based on the sequential word embeddings. CNN-based methods yield competitive results in image-sentence retrieval. Meanwhile, Recurrent Neural Networks (RNN) gains remarkable multimodal retrieval accuracy. However, these vector-space models treat the input words as "flat" embeddings for the downstream task. More specifically, they only

consider the context relations in the text modeling regardless of other important relations.

Recent research has found that the global semantic relations among words can provide rich semantics and can effectively promote the text classification performance [16]. Inspired by their work, we aim to combine deep models to explore the global word relations in representing the irregular-structured text data. Such relations are leveraged for enhancing the generalization ability of text in cross-modal retrieval tasks. In this paper, we propose one of the possible solutions, that is, representing a text by a structured and featured graph and learning text features by a graph-based deep model, i.e. Graph Convolutional Network (GCN) [1,5]. Such a graph can well capture the semantic relations among words. The GCN model has a great ability to learn local and stationary features on graphs. Figure 1 shows the comparison of our model to classical cross-modal retrieval models. Based on this graph representation for texts, we propose a dual-path neural network, called **G**raph-**I**n-**N**etwork (**GIN**), for cross-modal information retrieval.

The main contributions can be summarized as follows:(1)We propose to model text by graphs using similarity measure based on word2vec, which realizes cross-modal retrieval between irregular-structured and regular grid-structured data; (2) The model can jointly learn the textual and visual representations as well as similarity metric, providing an end-to-end training mode; (3) Experimental results show the superior performance of our model over the state-of-the-art methods.

2 Related Work

Cross-Modal Information Retrieval. The generic solution for cross-modal retrieval is to learn a common semantic space for different modalities of data and measure their similarity directly. Traditional statistical correlation analysis methods, typically like Canonical Correlation Analysis (CCA) [13], aim to maximize the pairwise correlations between two sets the data of different modalities. In order to leverage the semantic information, semi-supervised methods [17,22] and supervised methods [14,18] are proposed to explore the label information and achieve great progress. With the advances of deep learning in multimedia applications, DNN-based cross-modal methods are in the ascendant. This kind of methods generally construct two subnetworks for modeling data of different modalities and learn their correlations by a joint layer. Wang *et al*. [19] uses two branches of neural networks for learning textual-visual embeddings and realize effective end-to-end fine-tuning. In this work, we also follow the DNN-based routine to model the matched and mismatched text-image pairs.

Graph Convolutional Network (GCN). To render the extension of CNN to irregular graphs, [1] proposes graph convolutional network, which allows convolutions on the graphs to be solved as multiplications in the graph spectral domain. Besides, [5] further simplifies GCN by a first-order approximation of graph spectral convolutions, resulting in more efficient filtering operations. Based on GCN,

recent work [6] proposes a novel method for learning a similarity metric between irregular graphs. A siamese graph convolutional network is introduced for similarity matching. Different from our work, the two branches of the model come from the same image modality and the two branches share weights. Their model can only handle graph-structured modal data, which can been seen as a special case of our framework.

3 Methodology

In this section, we introduce a novel 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 Fig. 2, that the convolution part is referred to the blog of GCNs[1]) contains two key steps: *graph construction* and *GCN modeling*.

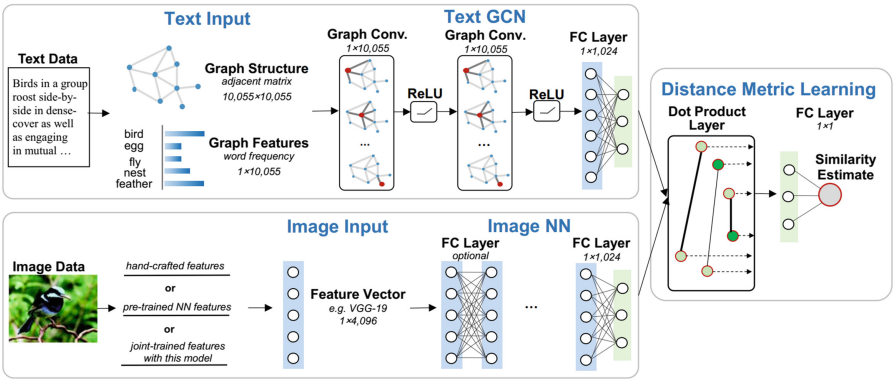


Fig. 2. The structure of the proposed model is a dual-path neural network: i.e., text Graph Convolutional Network (text GCN) (top) and image Neural Network (image NN) (bottom). The text GCN for learning text representation contains two layers of graph convolution on the top of constructed featured graph. 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.

3.1 Text Modeling

Graph Construction: 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 texts, 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 $v_i \in V$

[1] http://tkipf.github.io/graph-convolutional-networks/.

is corresponding to a unique word and each edge $e_{ij} \in E$ is defined by the *word2vec* similarity between two words:

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

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: Deep network models have become increasingly popular and achieved breakthroughs in many text analysis tasks. 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) is proposed to generalize Convolutional Neural Network (CNN) to irregular-structured graphs. 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 denote the output feature vector after graph convolution by F_{out}. In order to keep the filter K-localized in space and computationally efficient, [1] proposes a approximated polynomial filter defined as $g_\theta = \sum_{k=0}^{K-1} \theta_k T_k(\widetilde{L})$, where $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$, $\widetilde{L} = \frac{2}{\lambda_{max}} L - I_N$ and λ_{max} denotes the largest eigenvalue of L. L is the normalized graph Laplacian for the input graph structure. The filtering operation can then be written as $F_{out} = g_\theta F_{in}$. In our model, we use the same filter as in [1]. 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:

$$f_{out,i} = \sum_{k=0}^{K-1} \theta_k T_k(\widetilde{L}) f_{in,i} \tag{2}$$

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 $f_t = H_t(T)$.

3.2 Image Modeling

For modeling images, we adopt a neural network (NN) containing a set of fully connected layers (bottom in Fig. 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_{img} learnt by the model from image NN $H_{img}(\cdot)$ is represented by $f_{img} = H_{img}(I)$.

3.3 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_{img}, are in the same dimension and combined by an inner product layer. The successive layer is a fully connected layer with one output $score(T, I)$, denoting the similarity score function between a text-image pair. The training objective is a pairwise similarity loss function proposed in [7], which outperforms existing works in the problem of learning local image features. In our research, we maximize the mean similarity score u^+ between text-image pairs of the same semantic concept and minimize the mean similarity score u^- between pairs of different semantic concepts. Meanwhile, we also minimises the variance of pairwise similarity score for both matching σ^{2+} and non-matching σ^{2-} pairs. The loss function is formally by:

$$Loss = (\sigma^{2+} + \sigma^{2-}) + \lambda \max(0, m - (u^+ - u^-)) \tag{3}$$

where λ 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. $u^+ = \sum_{i=1}^{Q_1} \frac{score(T_i, I_i)}{Q_1}$ and $\sigma^{2+} = \sum_{i=1}^{Q_1} \frac{(score(T_i, I_i) - u^+)^2}{Q_1}$ when text T_i and image I_i are in the same class. While $u^- = \sum_{j=1}^{Q_2} \frac{score(T_j, I_j)}{Q_2}$ and $\sigma^{2-} = \sum_{j=1}^{Q_2} \frac{(score(T_j, I_j) - u^-)^2}{Q_2}$ when T_j and I_j are in different classes. We sequentially select $Q_1 + Q_2 = 200$ text-image pairs from the training set for each mini-batch in the experiments.

4 Experimental Studies

4.1 Datasets

Experiments are conducted on four widely used benchmark datasets. Each dataset contains a set of text-image pairs. dataset (Eng-Wiki for short) [13]

contains 2,866 image-text pairs divided into 10 classes. Each image is represented by a 4,096-dimensional vector extracted from the last fully connected layer of VGG-19 model [15]. Each text is represented by a graph with 10,055 vertices. **NUS-WIDE** dataset consists of 269,648 image-tag pairs We select samples in the 10 largest classes as adopted in [22]. 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 [14,18] For the features, 512-dimensional Gist features are adopted for the images and a graph with 598 vertices is used for the tags. **TVGraz** dataset contains 2,594 image-text pairs [10]. We choose the texts that have more than 10 words. Each image is represented by a 4,096-dimensional VGG-19 feature and each text is represented by a graph with 8,172 vertices.

4.2 Evaluation and Implementation

To evaluate the performance of our model, we conduct experiments for cross-modal retrieval tasks, i.e. text-query-images and image-query-texts. The mean average precision (MAP) and precision-recall (PR) curves [13] are used to evaluate the performance of all the algorithms on the four datasets. For all the datasets, we randomly select matched and non-matched text-image pairs and form 40,000 positive samples and 40,000 negative samples for training. The ground truth labels are binary denoting whether the pairs are from the same class or not. We train the model for 50 epochs with mini-batch size 200. We adopt the dropout ratio of 0.2, learning rate 0.001 with an Adam optimisation, and regularisation 0.005. m and λ are set to 0.6 and 0.35, respectively. In the semantic mapping layers of both text and image paths, the reduced dimensions are set to 1,024, 500, 256, 1,024, 1,024 for Eng-Wiki, NUS-WIDE, Pascal, and TVGraz, respectively.

4.3 Experimental Results

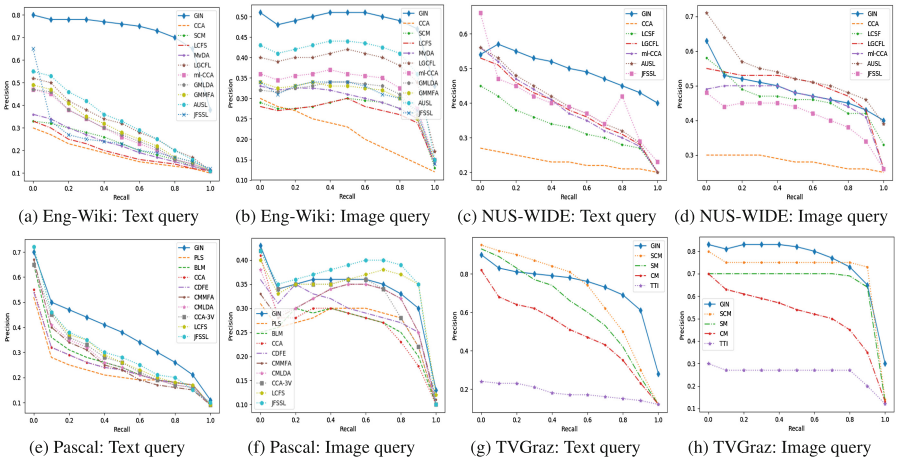
(1) Comparison with State-of-the-Art Methods. We compare our proposed GIN with a number of state-of-the-art models. The MAP scores of all the methods on the five benchmark datasets are shown in Table 1. All the other models are well cited work in this field. Since not all the papers have tested these four 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 four datasets for the text-query-image task. On the Eng-Wiki, Pascal, NUS-WIDE, and TVGRaz datasets, the MAP scores of GIN are about 35.70%, 17.14%, 12.9%, and 1.3% higher than the second best results, respectively. It's obvious that no matter for the rich text or for the sparse tags, our model gains superior performance than other models. The reason is that the proposed model effectively keeps the inter-word semantic relations by representing the texts with graphs, which has

Table 1. MAP score comparison of text-image retrieval on four given benchmark datasets.

Method	Text query	Image query	Average	Dataset
CCA [13]	0.1872	0.2160	0.2016	Eng-Wiki
SCM [13]	0.2336	0.2759	0.2548	
TCM [11]	0.2930	0.2320	0.2660	
LCFS [18]	0.2043	0.2711	0.2377	
LGCFL [3]	0.3160	0.3775	0.3467	
ml-CCA [12]	0.2873	0.3527	0.3120	
GMLDA [14]	0.2885	0.3159	0.3022	
GMMFA [14]	0.2964	0.3155	0.3060	
AUSL [22]	0.3321	0.3965	0.3643	
JFSSL [17]	0.4102	**0.4670**	0.4386	
GIN (ours)	**0.7672**	0.4526	**0.6099**	
CCA [13]	0.2667	0.2869	0.2768	NUS-WIDE
LCFS [18]	0.3363	0.4742	0.4053	
LGFCL [3]	0.3907	0.4972	0.4440	
ml-CCA [12]	0.3908	0.4689	0.4299	
AUSL [22]	0.4128	**0.5690**	0.4909	
JFSSL [17]	0.3747	0.4035	0.3891	
GIN (ours)	**0.5418**	0.5236	**0.5327**	
CCA [13]	0.2215	0.2655	0.2435	Pascal
CDFE [9]	0.2211	0.2928	0.2569	
BLM [14]	0.2408	0.2667	0.2538	
GMLDA [14]	0.2448	0.3094	0.2771	
GMMFA [14]	0.2308	0.3090	0.2699	
CCA3V [2]	0.2562	0.3146	0.2854	
LCFS [18]	0.2674	0.3438	0.3056	
JFSSL [17]	0.2801	**0.3607**	0.3204	
GIN (ours)	**0.4515**	0.3170	**0.3842**	
CM [10]	0.4500	0.4600	0.4550	TVGraz
SM [10]	0.5850	0.6190	0.6020	
SCM [13]	0.6960	0.6930	0.6945	
TCM [11]	0.7060	0.6940	0.6950	
GIN (ours)	**0.7196**	**0.8188**	**0.7692**	

been ignored by other methods that only word frequency or context information. Such inter-word relations are enhanced and more semantically relevant words are activated with the successive layers of graph convolutions, resulting in better generalization ability for un-seen text data.



(a) Eng-Wiki: Text query (b) Eng-Wiki: Image query (c) NUS-WIDE: Text query (d) NUS-WIDE: Image query

(e) Pascal: Text query (f) Pascal: Image query (g) TVGraz: Text query (h) TVGraz: Image query

Fig. 3. Precision-recall curves on the four datasets.

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. 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. Different from the observations on other datasets, the improvement for image-query-text is greater than that for text-query-image. The reason is that, for the image view, the compared algorithms represent images by bag-of-features with SIFT descriptors while we utilize 4096-dimensional CNN features, which are proved to be much more powerful than the hand-crafted feature descriptors. GIN achieves the best average MAP over all the competitors, especially outperforming the second best method JFSSL by 17.13% on Eng-Wiki.

The precision-recall (PR) curves of image-query-text and text-query-image are plotted in Fig. 3. For JFSSL, we show its best MAP after feature selection (see Table 7 in [17]). Since JFSSL hasn't reported the PR curves corresponding to the best MAP, we use its reported PR curves in [17]. 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 four 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.

(2) Comparison with Baseline Models. Besides our proposed model, we implement another four baseline models to evaluate the influence of the variation in text features and image features on the retrieval performance. All the experiments are conducted on the Eng-Wiki dataset. The retrieval performance

Table 2. Comparisons of MAP with baseline methods w.r.t different text and image features.

Text features	Image features	Text query	Image query	Average
LSTM	fixed VGG-19	0.62	0.42	0.52
CNN	fixed VGG-19	0.36	0.30	0.33
GCN	fixed VGG-19	**0.75**	**0.43**	**0.59**
GCN	fixed ResNet-50	0.66	0.39	0.53
GCN	CNN-5	0.28	0.27	0.28

of MAP is given in Table 2. Our proposed model GIN is based on GCN text features and VGG-19 image features. First, we fix the image features of VGG-19 and change the text features by LSTM [21] and CNN [4], respectively. The first three models in Table 2 shows the retrieval performance. It's obvious that GIN outperforms other models especially for the text retrieval task, which indicates the power of GCN in semantic representation of texts. The MAP of LSTM is inferior to GCN while CNN performs the worst. Then we fix the text features of GCN and change the image features by ResNet-50 and CNN with five convolution layers (CNN-5), respectively. Particularly, CNN-5 is trained end-to-end with our proposed model. We obtain the same conclusion that GIN performs the best. The model using ResNet-50 is slightly worse than using VGG-19. CNN-5 performs the worst because that shallow convolutional networks are detrimental to high-level image feature representation. What's more, the training process of GIN is 5 times faster than CNN+VGG-19 and 8 times faster than LSTM+VGG-19.

4.4 Parameters Analysis

We conduct several experiments on the Eng-Wiki datasets to explore how parameters, i.e. m and λ in the loss function, affect the cross-modal retrieval performance. In Table 3, we range the value of m from 0.4 to 0.6 and range λ from 0.25

Table 3. Experiments on the influence of the parameters m and λ.

m	λ	Text Query	Image Query	Average
0.40	0.35	0.553	0.384	0.469
0.50	0.35	0.622	0.463	0.543
0.60	0.35	**0.808**	0.460	**0.634**
0.70	0.35	0.643	**0.473**	0.558
0.80	0.35	0.606	0.448	0.527
0.60	0.25	0.788	0.441	0.615
0.60	0.30	0.795	0.450	0.623
0.60	0.40	0.791	0.452	0.621

to 0.4 and show the model's MAP scores. From the results we can see that the model is not much sensitive to λ in the range of 0.25 to 0.4. On the contrary, the range of m has obvious impact on the final cross-modal retrieval performance. The average MAP scores range from 0.47 to 0.63 when varying the value of λ. In general, 0.35 for λ and 0.6 for m are the relative best settings for our model.

5 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 representaions 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.

Acknowledgments. This work is supported by the National Key Research and Development Program (Grant No. 2017YFC0820700) and the Fundamental Theory and Cutting Edge Technology Research Program of Institute of Information Engineering, CAS (Grant No. Y7Z0351101).

References

1. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NIPS, pp. 3837–3845 (2016)
2. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for internet images, tags, and their semantics. TPAMI **106**(2), 210–233 (2014)
3. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. TMM **17**(3), 276–288 (2017)
4. Kim, Y.: Convolutional neural networks for sentence classification (2014). arXiv preprint arXiv:1408.5882
5. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
6. Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., Rueckert, D.: Distance metric learning using graph convolutional networks: Application to functional brain networks (2017). arXiv. 1703.02161
7. Kumar, B.G.V., Carneiro, G., Reid, I.: Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions. In: CVPR, pp. 5385–5394 (2016)
8. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. IEEE **86**(11), 2278–2324 (1998)
9. Lin, D., Tang, X.: Inter-modality face recognition. In: ECCV, pp. 13–26 (2006)
10. Pereira, J.C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI **36**(3), 521–535 (2014)

11. Qin, Z., Yu, J., Cong, Y., Wan, T.: Topic correlation model for cross-modal multimedia information retrieval. Pattern Anal. Appl. **19**(4), 1007–1022 (2016)
12. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: ICCV, pp. 4094–4102 (2015)
13. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: ACM-MM, pp. 251–260 (2010)
14. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: CVPR, pp. 2160–2167 (2012)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
16. Wang, C., Song, Y., Li, H., Zhang, M., Han, J.: Text classification with heterogeneous information network kernels. In: AAAI, pp. 2130–2136 (2016)
17. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. TPAMI **38**(10), 2010–2023 (2016)
18. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: ICCV, pp. 2088–2095 (2013)
19. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR, pp. 5005–5013 (2016)
20. Yu, J., Cong, Y., Qin, Z., Wan, T.: Cross-modal topic correlations for multimedia retrieval. In: ICPR, pp. 246–249 (2012)
21. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization (2014). arXiv preprint arXiv:1409.2329
22. Zhang, L., Ma, B., He, J., Li, G., Huang, Q., Tian, Q.: Adaptively unified semi-supervised learning for cross-modal retrieval. In: IJCAI, pp. 3406–3412 (2017)