Fine-Grained Correlation Learning with Stacked Co-attention Networks for Cross-Modal Information Retrieval

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

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{luyuhang,yujing02,liuyanbing,tanjianlong,guoli}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
zwf.zhang@gmail.com
[4] Zhejiang Future Technology Institute, Jiaxing, China

Abstract. Cross-modal retrieval provides a flexible way to find semantically relevant information across different modalities given a query of one modality. The main challenge is to measure the similarity between different modalities of data. Generally, different modalities contain unequal amount of information when describing the same semantics. For example, textual descriptions often contain more background information that cannot be conveyed by images and vice versa. Existing works mostly map the global data features from different modalities to a common semantic space to measure their similarity, which ignore their imbalanced and complementary relationships. In this paper, we propose stacked co-attention networks (SCANet) to progressively learn the mutually attended features of different modalities and leverage these fine-grained correlations to enhance cross-modal retrieval performance. SCANet adopts a dual-path end-to-end framework to jointly learn the multimodal representations, stacked co-attention, and similarity metric. Experiment results on three widely-used benchmark datasets verify that SCANet outperforms state-of-the-art methods, with 19% improvements on MAP in average for the best case.

Keywords: Stacked co-attention network · Graph convolution
Fine-grained cross-modal correlation

1 Introduction

With the fast development of Internet and mobile network, multimodal data including image, text, video and audio, has been emerging and accumulated rapidly. Multimedia retrieval becomes a fundamental technique for intelligent

W. Liu et al. (Eds.): KSEM 2018, LNAI 11061, pp. 213–225, 2018.
https://doi.org/10.1007/978-3-319-99365-2_19

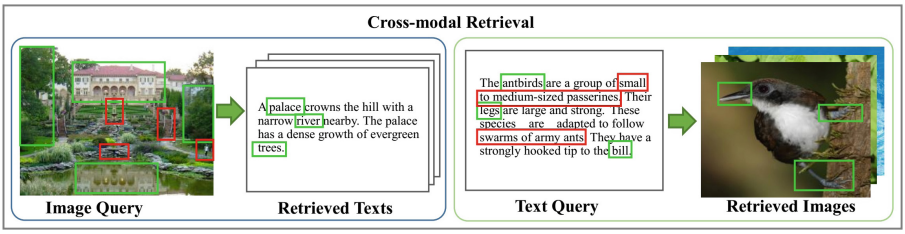

Fig. 1. Examples of image-text retrieval. The samples of different modalities have imbalanced and unequal information. The green boxes indicate related content appeared in both modalities while the red boxes mean extra content existed in only one modality.

search and big data analysis. Traditional information retrieval approaches, such as key word retrieval and content-based image retrieval, belong to single-modal models only obtaining results of the same modality with the query. However, such models limit the flexibility for accessing various modalities of data. Cross-modal retrieval provides semantic relevant information from multimodal data when given a query of one modality. This is much closer to what humans do than matching text or visual features independently.

The main challenge for cross-modal retrieval is to bridge the "heterogeneous gap" between different modalities and learn a appropriate similarity metric between them. A common solution of existing work is to map the features of different modalities into one common latent space and compare their similarity by the learnt common representations. Conventional methods are based on Canonical Correlation Analysis (CCA) [15], probabilistic models [1], metric learning [24], and joint modeling [18]. Recently, Deep Neural Network(DNN) has promoted the advances in cross-modal retrieval, such as deep feature representation with CNN [6] and LSTM, common space learning with multilayer network, and similarity measure with DNN-based metric learning.

The aforementioned approaches map the complete data of different modalities from their feature space to the common semantic space equally to find their feature correlations, which is based on the assumption that semantically relevant data of different modalities has equal amount of information. However, this assumption is not always true in practice. In fact, data conveying the same semantics but from different modalities may express imbalanced information and have complementary relationships. For example, an image is usually accompanied by a text description and vice versa to express the same semantics, but the amount of information between the image and the text is unequal. In Fig. 1, the left image query contains a complex scenery which cannot be completely described by a few objects in the retrieved texts. On the other scenarios, the right query text has more background information beyond the content of only an antbird in the retrieved images. It's not all the fine-grained information between text and image has exact correlations. Therefore, regarding all different modalities equally will weaken some important aligned information while introducing unaligned noise.

Resently, Peng *et al.* [12] has demonstrated the advantages of fine-grained information in modality-specific space for cross-modal correlation modeling. Different from their work, we focus on preserving the mutual fine-grained parts of different modalities in the common semantic space to learn the cross-modal correlations. Inspired by the progress of attention mechanism in image caption and visual question answering, we propose **S**tacked **C**o-**A**ttention **Net**works (SCANet) to explore the mutually attended characteristics of different modalities for strengthening the semantic correlations.

The rest of this paper is organized as follows. We briefly review the related works in Sect. 2. Section 3 introduces our proposed SCANet approach. We report the experimental results in Sect. 4 and conclude our work in Sect. 5.

2 Related Work

DNN-Based Feature Extraction. Feature representation is the footstone for cross-modal retrieval. In the text-image retrieval field, off-the-shelf features leant by deep neural networks are widely utilized to represent images. Most existing works use Convolutional Neural Network (CNN) pre-trained on ImageNet to extract visual features for text-image semantic space mapping. However, such CNN model is pre-trained for object recognition which may ignore some detailed information for other tasks. Therefore, fine-tuning off-the-shelf CNN features for more discriminative embeddings is necessary for cross-modal-specific tasks.

For text representation, the popular vector-space models are usually used to convert a text to a high-level semantic vector based on the sequential word embeddings. Recurrent Neural Networks (RNN) is one of the popular choices in this kind of models. Nam *et al.* [10] applies directional LSTM for text representation and results in remarkable multimodal retrieval accuracy. Peng *et al.* [11] utilizes attention-based LSTM for modality-specific feature learning to refine the cross-modal correlations. Meanwhile, CNN-based text modeling also yields competitive results in image-sentence retrieval. These vector-space models treat the input words as "flat" features and ignore the global semantic structures inherent in the text. Recent research has found that the relations among words could provide rich semantics. Graph Convolutional Network (GCN) [7] is one popular graph-based neural network and has been used to model the semantic relations in a text as a featured graph. It has a great ability to learn local and stationary features and can effectively promote the text classification performance. In this paper, in stead of RNN which is commonly used in text-image retrieval, we explore the usage of GCN for text feature extraction.

Cross-Modal Learning. 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. The traditional statistical correlation analysis methods, typically like Canonical Correlation Analysis (CCA) [15], aim to maximize the pairwise correlations between two sets the data of different modalities. In order to leverage the semantic information, graph-based semi-supervised methods [18] and supervised methods 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. Zheng *et al.* [25] uses two convolution 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.

Attention Mechanism. Recently, attention mechanism has promoted remarkable advances in many multimodal tasks, such as image caption, image question answering, cross-modal retrieval and etc. It allows deep models to focus on the task-driven necessary parts of the features. Yang *et al.* [20] proposes Stacked Attention Networks (SANs), which takes multiple attention steps to progressively focus on the informative parts for image question answering. Attention-based cross-modal retrieval models aim to simultaneously locate the necessary components in both textual and image features to learn more accurate semantic correlations. For example, Zhang *et al.* [23] generates adaptive attention masks and divides features into attended and unattended parts to enhance the robustness of learnt representations. Peng *et al.* [12] designs a recurrent attention network to capture the modality-specific characteristics in textural and image space independently. Different from their work, we use attention mechanism to fully explore the co-attended parts inherent in both of the two modalities for learning better cross-modal correlations.

3 Methodology

In this section, we present **S**tacked **C**o-**a**ttention **Net**owrks (SCANet) using dual-path neural networks. The overall architecture of SCANet is shown in Fig. 2. We describe the major four components of SCANet: the text model, the image model, the stacked co-attention network, and the objective function.

3.1 Text Model

As illustrated in [7], Graph Convolutional Network (GCN) shows strong ability in modeling the semantics of texts and has good performance in text classification [2]. We explore GCN to learn the text features to leverage the prior information of semantic similarities inherent in the text corpus. We first represent each text by a featured graph as the input of GCN. The graph structure is identical for all the texts while the graph features are unique for each text. We extract the most common words from all the unique words in the text corpus and represent each word by a pre-trained *word2vec* [9] embedding. Then each vertex in the graph structure is corresponding to a common word. For each vertex, we compute its k-nearest neighbors of vertices based on the cosine similarity between word *word2vec* embeddings to form the edge set. For the graph features, each text is represented by a *bag-of-words* vector and the word frequency serves as the 1-dimensional feature on the corresponding word vertex.

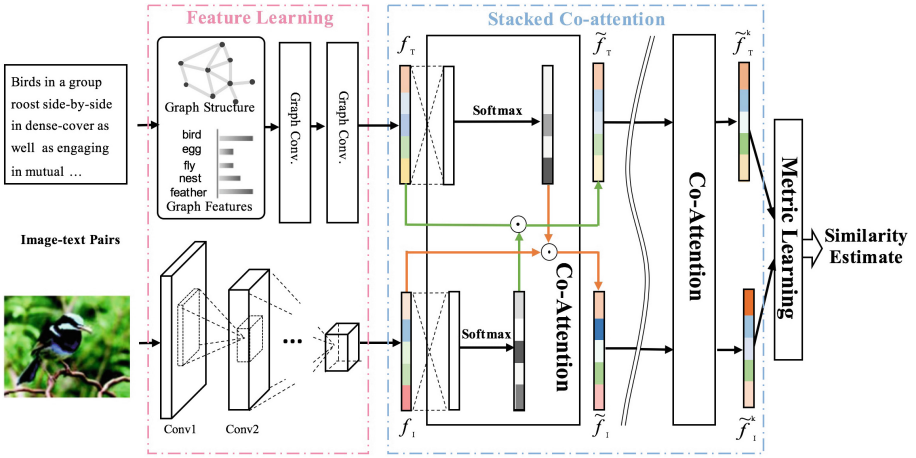

Fig. 2. Overview of SCANet. Above is the text modality path and below is the image modality path. Each path is divided into two parts: feature learning (f_T and f_I) and stacked co-attention (f_T^k and f_I^k). The feature learning maps each modality data into high-level feature representation with the same dimension. The stacked co-attention conducts co-attention process multiple times. For each co-attention process, the attention distribution learnt from image features are used to update the text features and vice versa. Finally, the two paths are joint by metric learning for the similarity measure.

Given the featured graph of each text, we learn the text features using GCN [7]. The input and output features are defined by F_{in} and F_{out}. The i th output feature $f_{out,i} \in F_{out}$ corresponding to the i th input feature $f_{in,i}$ is given by:

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

where θ_k is the parameter to learn and $T_k(\widetilde{L}) = 2\widetilde{L}T_{k-1}(\widetilde{L}) - T_{k-2}(\widetilde{L})$ with $T_0(\widetilde{L}) = 1$ and $T_1(\widetilde{L}) = \widetilde{L}$. $\widetilde{L} = \frac{2}{\lambda_{max}} L - I_N$ and λ_{max} denotes the largest eigenvalue of the normalized graph Laplacian L of the input graph structure. K is set to 3. In our model, GCN contains two layers of graph convolutions, each followed by Rectified Linear Unit (ReLU) activation to increase non-linearity. Then a fully connected layer is used to reduce the feature dimension and map the text to the common latent space with the image. Given a text T, the text representation f_T learnt by the GCN model $H_T(\cdot)$ is denoted by:

$$f_T = H_T(T) \tag{2}$$

3.2 Image Model

For modeling images, we use CNN to get the embeddings of images. Specially, we use the pre-trained VGGNet [17] and extract the features from last fully

connected layers as the image features. After getting the fixed VGG-19 image features, we use a set of fully connected (FC) layers for fine-tuning the feature representations. Similar to the text model, the last FC layer is used to reduce the feature dimension and map the image to the common latent space with the text. In the experimental study, we tune the number of FC layers and find that only keeping the last common space mapping layer without extra feature fine-tuning can obtain the best retrieval performance. Given an image I, the image representation f_I learnt by the image model $H_I(\cdot)$ is demoted by:

$$f_I = H_I(I) \tag{3}$$

3.3 Stacked Co-attention Networks

Given the text features f_T and the image features f_I, the stacked co-attention networks learn the fine-grained semantic correlations between these two modalities via multiple attention layers. In many cases, the amount of information in different modalities is unequal even if they convey the same semantics. Using the global image features and text features for feature alignment could introduce noise from their irrelevant parts. Therefore, the stacked co-attention networks are proposed to progressively attend to the parts that are highly correlated between image and text pairs and gradually filter out the unaligned noise.

As shown in Fig. 2, the inputs of stacked co-attention networks the are features f_T and f_I of two modalities. In the first co-attention layer, the text features f_T is updated by the attention distribution learnt from the image features and we denote the output text features as $\tilde{f_T}$. While the image features f_I is updated by the attention distribution learnt from the text features and we denote the output image features as $\tilde{f_I}$. Specifically, given f_T and f_I, we put them into a fully connected layer and then use a $softmax$ function to generate the attention distribution. Here is the formula for computing the attention distribution:

$$h_T = relu(W_T f_T + b_T) \quad p_T = softmax(h_T) \tag{4}$$

$$h_I = relu(W_I f_I + b_I) \quad p_I = softmax(h_I) \tag{5}$$

where $f_T \in \mathbb{R}^N, f_I \in \mathbb{R}^N, W_T \in \mathbb{R}^{N \times N}$, $W_I \in \mathbb{R}^{N \times N}$, $b_T \in \mathbb{R}^N$ and $b_I \in \mathbb{R}^N$. p_I is the attention distribution learnt from the image feature while p_T is the attention distribution learnt from the text feature. Based on the attention distribution, we calculate the weighted sum of the text features and image features respectively and get the updated features denoted as $\tilde{f_T}$ and $\tilde{f_I}$ according to the following formula:

$$\tilde{f_T} = p_I \circ f_T \qquad \tilde{f_I} = p_T \circ f_I \tag{6}$$

where "$\circ$" is Hadamard product. $\tilde{f_T}$ and $\tilde{f_I}$ are the inputs of the second co-attention layer. The co-attention layer will repeat multiple times to gradually obtain the fine-grained correlations. We denote the outputs of the k-th co-attention layer as $\tilde{f_T^k}$ and $\tilde{f_I^k}$ for the text and image respectively. Formally, the k-th co-attention layer is computed by the following formula:

$$h_T^k = relu(W_T^k f_T^{k-1} + b_T^k) \quad p_T^k = softmax(h_T^k) \tag{7}$$

$$h_I^k = relu(W_I^k f_I^{k-1} + b_I^k) \quad p_I^k = softmax(h_I^k) \tag{8}$$

Based on the attention distribution p_T^k and p_I^k, we can get the updated text features $\tilde{f_T^k}$ and image features $\tilde{f_I^k}$ by the following formula:

$$\tilde{f_T^k} = p_I^k \circ f_T^{k-1} \quad \tilde{f_I^k} = p_T^k \circ f_I^{k-1} \tag{9}$$

The number of co-attention layers depends on the characteristics of the multimodal data. If the semantics of the data are very complex, we need more co-attention layers to achieve better results. However, it should be noted that different pairs of multimodal data have different degree of information imbalance, and the number of co-attention layers needs to adapt to the specific data. Too many or too few layers may reduce the overall performance. In our experiments, using two co-attention layers is often the best choice.

3.4 Objective Function

In our model, we have finally get two modal features from the two path. Distance metric learning is applied to estimate the relevance of the two features. An inner product layer combines the two features and is followed by a single fully connected output layer with a sigmoid activation function and one output, that is the similarity we measure the two features. And the training objective is a pairwise similarity loss function proposed in [8]. The main idea is that we maximize the mean similarity score u^+ between text-image pairs of the same semantic concepts and minimize the mean similarity score u^- between pairs of different semantic concepts. Meanwhile, we also minimize the variance of matching pairs' similarity scores σ^{2+} and non-matching pairs' similarity scores σ^{2-}. We can get the *loss* by:

$$Loss = (\sigma^{2+} + \sigma^{2-}) + \lambda \max(0, m - (u^+ - u^-)) \tag{10}$$

where λ can adjust the proportion of mean, and m controls the upper limit between the mean of matching and non-matching similarity.

4 Experiments

To exhaustively evaluate our model, we conduct experiments on five benchmark datasets to compare our model with 11 state-of-the-art methods. Besides, we implement several baseline models to explore how variation in text features, image features, and the number of co-attention layers affects the performance of the proposed co-attention model. What's more, we also report the sensitive analysis of hyper-parameters including m and λ in the loss function. All the experiments are implemented by Tensorflow-1.6.0 using Python-3.5.0.

Table 1. Descriptions of the three benchmark datasets.

Dataset	#Training	#Testing	#Category	Image features	Text features
Eng-Wiki	2173 (original)	693	10	VGG-4,096	GCN-10,055
NUS-WIDE	40,000 (original)	1,000	10	BOF-500	GCN-5,018
Pascal VOC	2,808 (original)	2,841	20	Gist-512	GCN-598

4.1 Datasets

We use four popular datasets for our experiments, including English Wikipedia (Eng-Wiki for short), NUS-WIDE, Pascal VOC, and TVGraz. The detailed information is listed in Table 1. Based on the original training image-text pairs, we randomly select matched and non-matched pairs and form 40,000 positive samples and 40,000 negative samples for model training. The preprocessing methods for the datasets are the same as [21]. We preprocess the texts of each dataset to construct a featured graph with different vertex number as the input of GCN. We set $k = 8$ in k-nearest neighbors for text graph construction. The learnt text representation by GCN has the same dimension with the input feature. For image features, we adopt 4,096-dimensional VGG-19 [17] features for Eng-Wiki. Since NUS-WIDE and Pascal VOC datasets haven't provided the original images online, we use the off-the-shelf 500-dimensional bag-of-features and 512-dimensional Gist features, respectively.

4.2 Evaluation Measure and Experimental Settings

The mean average precision (MAP) [12] is used for experimental evaluation. Higher MAP indicates better retrieval performance. We train the model for 80 epochs with mini-batch size 256. We adopt the dropout ratio of 0.2 at the input of the last FC layer, learning rate 0.0001 with an Adam optimisation, and regularisation 0.005. m and λ in the loss function are set to 0.6 and 0.35, respectively. In the preceding layer of stacked co-attention, the text and image features are reduced to the same dimensions, which are set to 1,024, 500, 256 for En-Wiki, NUS-WIDE, Pascal, respectively.

4.3 Comparison with State-of-the-Art Methods

We first compare our model with several state-of-the-art methods, including CCA & SCM [15], TCM [13], GMLDA & GMMFA [16], LCFS [19], MvDA [4], LGCFL [5], ml-CCA [14], AUSL [22] and JFSSL [18]. All these methods are well cited work in this field. Since not all the papers have tested on the three datasets, for fair comparison, we compare our model to methods on their reported datasets with the same preprocessing conditions.

From Table 2 we can see that, for the text query task, SCANet outperforms all the other methods on all the datasets. Compared with the second best models, SCANet obtains remarkable improvements in MAP by 40%, 20% and 20% on

Table 2. Comparisons of MAP with state-of-the-art methods on three datasets.

Method	Eng-Wiki			NUS-WIDE			Pascal VOC		
	Text	Image	Average	Text	Image	Average	Text	Image	Average
SCM [15]	0.23	0.28	0.26	-	-	-	-	-	-
TCM [13]	0.29	0.23	0.26	-	-	-	-	-	-
CCA [15]	0.19	0.22	0.20	0.27	0.29	0.28	0.22	0.27	0.24
LCFS [19]	0.20	0.27	0.24	0.34	0.47	0.41	0.27	0.34	0.31
MvDA [4]	0.23	0.30	0.26	-	-	-	-	-	-
LGCFL [5]	0.32	0.38	0.35	0.39	0.50	0.44	-	-	-
ml-CCA [14]	0.29	0.35	0.32	0.39	0.47	0.43	-	-	-
GMLDA [16]	0.29	0.32	0.30	0.24	0.31	0.28	-	-	-
GMMFA [16]	0.30	0.32	0.31	0.23	0.31	0.27	-	-	-
AUSL [22]	0.33	0.40	0.37	0.41	**0.57**	0.49	-	-	-
JFSSL [18]	0.41	**0.47**	0.44	0.38	0.40	0.39	0.28	**0.36**	0.32
SCANet	**0.81**	0.46	**0.63**	**0.58**	0.52	**0.55**	**0.48**	0.34	**0.41**

Eng-Wiki, NUS-WIDE and Pascal VOC respectively. It proves that SCANet can represent the text well no matter for rich texts shch as Eng-Wiki, or for sparse texts such as NUS-WIDE and Pascal VOC. It strongly proves the effectiveness of our proposed stacked co-attention model. The improvement also owes to the text GCN model, which leverage the semantic similarity structure inherent in the text corpus as a prior knowledge to progressively enhance the feature learning. The text embeddings learnt by GCN have good generalization ability.

For the image query task, the MAP of SCANet is superior to most of the compared methods. SCANet ranks second best on all the datasets, which is slightly inferior than the best performance. It is because for the image model of SCANet, we use pre-trained deep model VGGNet [17], which is trained for image classification task, to extract the image features. In future work, we may consider connecting the VGGNet with our model for end-to-end training. The main contribution of this paper doesn't focus on the image feature extraction. For the average performance, SCANet has 19%, 16%, and 9% improvements compared with the second best results on Eng-Wiki, NUS-WIDE and Pascal VOC respectively. The improvements are different on these datasets, which is also related to the image features. Eng-Wiki provides original images and we use VGG-19 to extract images features. While NUS-WIDE and Pascal VOC doesn't provide original images, we just use 500-dimensional bag-of-features and 512-dimensional Gist features respectively as they provided.

4.4 Baseline Comparisons

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

Table 3. Comparisons of MAP with five baseline methods w.r.t different text features, image features, and co-attention layer numbers.

Text features	Image features	Text query				Image query				Average			
		#Attention layers				#Attention layers				#Attention layers			
		0	1	2	3	0	1	2	3	0	1	2	3
LSTM	Fixed VGG-19	0.62	0.64	0.63	0.60	0.42	0.44	0.43	0.42	0.52	0.54	0.53	0.51
CNN	Fixed VGG-19	0.36	0.37	0.41	0.35	0.30	0.31	0.33	0.29	0.33	0.34	0.37	0.32
GCN	Fixed VGG-19	**0.75**	**0.76**	**0.81**	**0.72**	**0.43**	**0.45**	**0.46**	**0.46**	**0.59**	**0.61**	**0.63**	**0.59**
GCN	Fixed ResNet-50	0.66	0.69	0.71	0.70	0.39	0.41	0.43	0.40	0.53	0.56	0.57	0.55
GCN	CNN-5	0.28	0.31	0.33	0.30	0.27	0.28	0.28	0.26	0.28	0.30	0.31	0.28

retrieval performance. Our proposed model SCANet is based on GCN text features and VGG-19 [17] image features. We first fix the image features of VGG-19 and change the text features by LSTM and CNN [6], respectively. Then we fix the text features of GCN and change the image features by ResNet-50 [3] and CNN with five convolution layers (CNN-5), respectively. Particularly, CNN-5 is trained end-to-end with our proposed model. Meanwhile, for each baseline model, we conduct four experiments to variate the number of attention layers in the range of (0,1,2,3) and show its affects on the performance. All the experiments are conducted on the Eng-Wiki dataset. The retrieval performance of MAP is given in Table 3.

The Influence of Co-attention Layer Numbers. When varying the number of co-attention layers, we can see that almost all the models benefit from our stacked co-attention method compared with the original models (i.e. #Attention layers = 0). Generally speaking, when the number of attention layers changes from 1 to 3, the increase of MAP scores of different models is ranging from 2% to 6%. That's because the stacked co-attention layers progressively enhance the mutually attended features of the paired text and image and filter out the unaligned noise. Different models fit for different number of co-attention layers and 2 is a relatively good setting in most cases. SCANet with two co-attention layers obtains the highest MAP compared with other baseline models.

The Influence of Text and Image Features. For the first three models in Table 3, we fix the image VGG-19 features. For each number of attention layers, it's obvious that SCANet 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 than GCN while CNN performs the worst. For the last three models in Table 3, we fix the text GCN features. We also obtain the same conclusion that SCANet performs the best on all the number of attention layers. 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.

The Efficiency of SCANet. The experiments were conducted on a 64-bit Linux machine with 4 Tesla V100-PCIE GPUs each with 16.16 GB memory. The training process on Eng-Wiki costs 1.3 h for GCN+VGG19 model (SCANet), 3.6 h for CNN+VGG19 model and 8.7 h for LSTM+VGG19 model. SCANet costs the least training time since that only 3 parameters need to be learnt for calculating the convolutions within 3 layers of neighborhoods.

Table 4. Experiments on the influence of the parameters m and λ.

m	λ	Text query	Image query	Average
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
0.60	0.30	0.795	0.450	0.623
0.60	0.40	0.791	0.452	0.621

4.5 Parameter Analysis

We conduct several experiments on the Eng-Wiki datasets to explore how parameters affect the cross-modal retrieval performance. We range the value of m from 0.4 to 0.6 and λ from 0.25 to 0.4 and show some of results in Table 4. We can know that the model is not much sensitive to λ and still performs well when λ is in the interval (0.25, 0.40). On the contrary, m has obvious impact on the retrieval performance. The average MAP scores range from 0.47 to 0.63 when varying the value of λ. So, 0.35 for λ and 0.6 for m are the relative best settings.

5 Conclusion

In this paper, we have proposed a stacked co-attention network to progressively align the semantically relevant features of different modalities and strengthen their fine-grained correlations. In the dual-path end-to-end framework, the attention distribution is jointly learnt with both multimodal feature learning and distance metric learning to benefit each other. Experimental results on three benchmark datasets verify that our model outperforms 11 state-of-the-art methods. Meanwhile, the extensive baseline comparisons indicate that the proposed attention approach can promote cross-modal retrieval performance regardless of the feature representations, though GCN+VGG gains the best performance.

Acknowledgement. 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. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 127–134. ACM (2003)
2. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NIPS, pp. 3837–3845 (2016)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
4. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. TPAMI **38**(1), 188–194 (2016)
5. 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)
6. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
7. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
8. 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)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. arXiv preprint arXiv:1611.00471 (2016)
11. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: IJCAI, pp. 3846–3853 (2016)
12. Peng, Y., Qi, J., Yuan, Y.: Modality-specific cross-modal similarity measurement with recurrent attention network. arXiv preprint arXiv:1708.04776 (2017)
13. Qin, Z., Yu, J., Cong, Y., Wan, T.: Topic correlation model for cross-modal multimedia information retrieval. PAA **19**(4), 1007–1022 (2016)
14. Ranjan, V., Rasiwasia, N., Jawahar, C.: Multi-label cross-modal retrieval. In: ICCV, pp. 4094–4102 (2015)
15. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: ACM-MM, pp. 251–260 (2010)
16. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: CVPR, pp. 2160–2167 (2012)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014)
18. 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)
19. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: ICCV, pp. 2088–2095 (2013)
20. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR, pp. 21–29 (2016)
21. Yu, J., et al.: Modeling text with graph convolutional network for cross-modal information retrieval. arXiv preprint arXiv:1802.00985 (2018)
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)