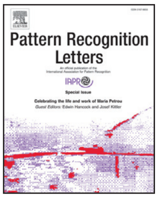
Pattern Recognition Letters 130 (2020) 189–198

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Learning cross-modal correlations by exploring inter-word semantics and stacked co-attention

Jing Yu[a,b], Yuhang Lu[a,b], Weifeng Zhang[c,*], Zengchang Qin[d,**], Yanbing Liu[a], Yue Hu[a]

[a] *Institute of Information Engineering, Chinese Academy of Sciences, China*
[b] *School of Cyber Security, University of Chinese Academy of Sciences, China*
[c] *School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China*
[d] *Intelligent Computing & Machine Learning Lab, School of ASEE, Beihang University, China*

ARTICLE INFO

Article history:
Available online 17 August 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
Cross-modal correlation
Inter-word semantics
Fine-grained correlation
Stacked co-attention
Cross-modal retrieval

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 learn the semantic correlations between different modalities and measure the distance across modalities. For text-image retrieval, 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 explore the inter-word semantics by modelling texts by graphs using similarity measure based on word2vec. Besides feature presentations, we further study the problem of information imbalance between different modalities when describing the same semantics. For example textual descriptions often contain more background information that cannot be conveyed by images and vice versa. We propose a stacked co-attention network to progressively learn the mutually attended features of different modalities and enhance their fine-grained correlations. A dual-path neural network is proposed for cross-modal information retrieval. 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 19% improvement on accuracy for the best case.

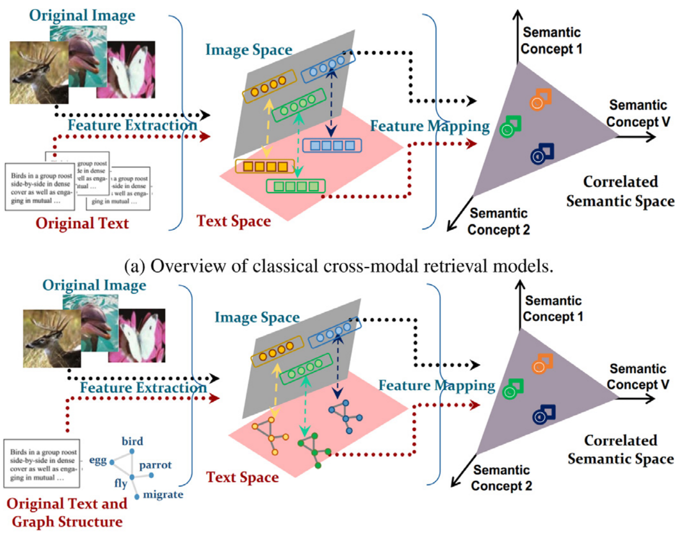
1. Introduction

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. In recent years, much effort has been made to exploit multi-modal data for more comprehensive semantics to improve the performance of uni-modal tasks [42–44]. Besides uni-modal search, there is a growing demand for effective and efficient search in the data across different modalities. Cross-modal information retrieval [25,30,35] enables users to take a query of one modality to retrieve data of other modalities in semantically relevant content. 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 semantic space to compare all modalities of data directly, typically using probabilistic models [2], metric learning [40], subspace learning [6,25,27], and joint modeling methods [30]. A brief survey is available in [30].

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 [3,33,38] use Convolutional Neural Network (CNN) [13] to learn the visual features obtained from the pre-trained model for object recognition on ImageNet. CNN can effectively extract hierarchies of visual feature vectors and the fixed CNN feature can be directly used for text-image semantic space mapping. In this paper, we also employ the same routine to use pre-trained CNN for visual feature learning. For text representation, the popu-

[*] Corresponding author.
[**] Co-Corresponding author.
E-mail addresses: yujing02@iie.ac.cn (J. Yu), Weifeng.zhang@broadsens.com (W. Zhang), zcqin@buaa.edu.cn (Z. Qin).

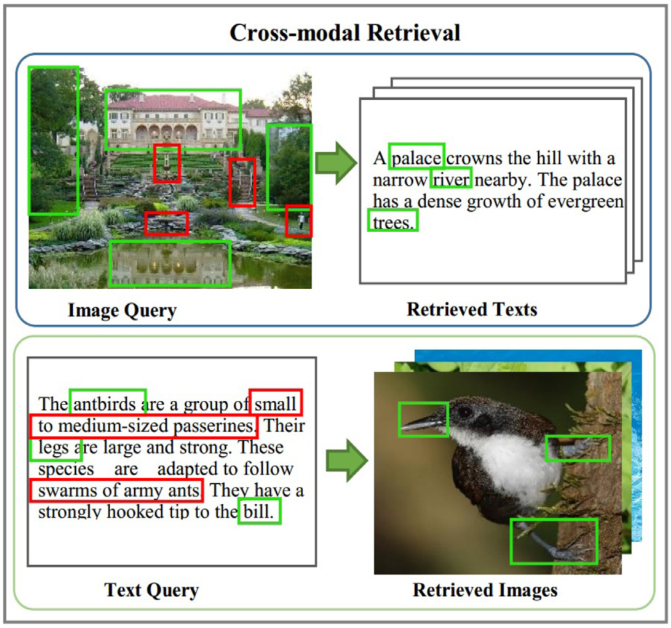
https://doi.org/10.1016/j.patrec.2018.08.017
0167-8655/© 2018 Elsevier B.V. All rights reserved.



(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.



A palace crowns the hill with a narrow river nearby. The palace has a dense growth of evergreen trees.

The antbirds are a group of small to medium-sized passerines. Their legs are large and strong. These species are adapted to follow swarms of army ants. They have a strongly hooked tip to the bill.

Fig. 2. 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

lar *vector* − *space* model is usually adopted to convert a text to a textual vector for learning high-level semantic representations. In this kind of models, *bag* − *of* − *words* (BOW) is commonly used in cross-modal information retrieval [15,32]. 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 [30], one common problem is that the relations among words are not considered. Recently, *word2vec* [18] 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. [33] extract word vectors via *word2vec* model and adopts Fisher vector encoding to obtain the sentence representation. Zhang et al. [38] represent 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 [29].

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) [4,10]. 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. (Fig. 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 [10] and brain network matching [11].

Based on the feature representation of different modalities, the main challenge for cross-modal retrieval is to learn the semantic correlations between different modalities and map them into a common semantic space, in which distance between concepts can be well modeled. Existing 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. 2, the top image query contains a complex scenery which cannot be completely described by a few objects in the retrieved texts. On the other scenarios, the bottom 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, Yuxin et al. [36] have 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 a stacked co-attention mechanism to explore the mutually attended characteristics of different modalities for strengthening their semantic correlations.

A dual-path neural network model, called **S**tacked **C**o-**A**ttention **Net**works (SCANet), jointly exploring inter-word semantics and stacked co-attention for cross-modal information retrieval is proposed. The text modelling path contains layers of GCN on the top of graph representations. The image modelling path contains a neural network with layers of nonlinearities on the top of off-the-shelf image representations. Then the mutually attended parts of the image and text features can be fully explored with the stacked co-attention mechanism. To train the model, we employ a pairwise similarity loss function [12], that maximizes the similarity between samples in the same semantic concept and minimizes the similarity between samples in different semantic concepts. 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 the inter-word semantics and stacked co-attention mechanism.

The rest of this paper is organized as follows. We briefly review the related works in Section 2. Section 3 introduces our proposed SCANet model. We report the experimental results in Section 4 and conclude our work in Section 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. [19] apply directional LSTM for text representation and results in remarkable multimodal retrieval accuracy. Peng et al. [20] utilize 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) [10] 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.

2.1. 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) [25], aim to maximize the pairwise correlations between the projections of data from two modalities; Latent Dirichlet Allocation (LDA) based approaches [1] model the correlations between images and the corresponding annotations by a shared mixture of topic distributions. However, these methods merely consider the pairwise constrains across different modalities, regardless of high-level semantic priority. In addition, these methods separate the shared feature learning from the measure of relevance, thus deteriorating the generalization ability.

In order to enhance the inter-modal correspondence, more semantic information has been exploited. Graph-based semi-supervised methods [30] and supervised methods [14] are proposed to explore the label information and achieve great progress. More recently, ranking-based models [16] have studied the degrees of learning difficulty inherent in the retrieved sentences corresponding to the image query. Based on such degrees, a more optimal common semantic space is gradually learnt from easy to more complex rankings. 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 jointly learn their modal-specific features and semantic correlations. Zheng et al. [41] use 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 based on the information of semantic labels.

2.2. Attention mechanism

Recently, attention mechanism has promoted remarkable advances in many multimodal tasks, such as image caption, image question answering, cross-modal retrieval, etc. It allows deep models to focus on the task-driven necessary parts of the features. Yang et al. [34] propose 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. [39] generate adaptive attention masks and divides features into attended and unattended parts to enhance the robustness of learnt representations. Yuxin et al. [36] design 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 paper, we propose a novel cross-modal information retrieval model as shown in Fig. 3. The model mainly contains four parts: (i) Text model: in the text modelling path (top in Fig. 3, that the convolution part is referred to the blog of GCNs[1]), each text is represented by a featured graph and the text GCN is used to learn the feature representation. (ii) Image model: in the image modelling path (bottom in Fig. 3), we leverage the pre-trained convolutional neural networks to extract the image features and use a set of fully connected layers to increase the non-linearity. (iii) Stacked co-attention: the mutually attended parts of the image and text features are fully explored with the multiple layers of co-attention mechanism. (iv) Distance metric learning: we apply distance metric learning to estimate the relevance of features learned from the dual-path model.

3.1. Text model

3.1.1. 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 $W = [w_1, w_2, \ldots, w_N]$, 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:

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

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

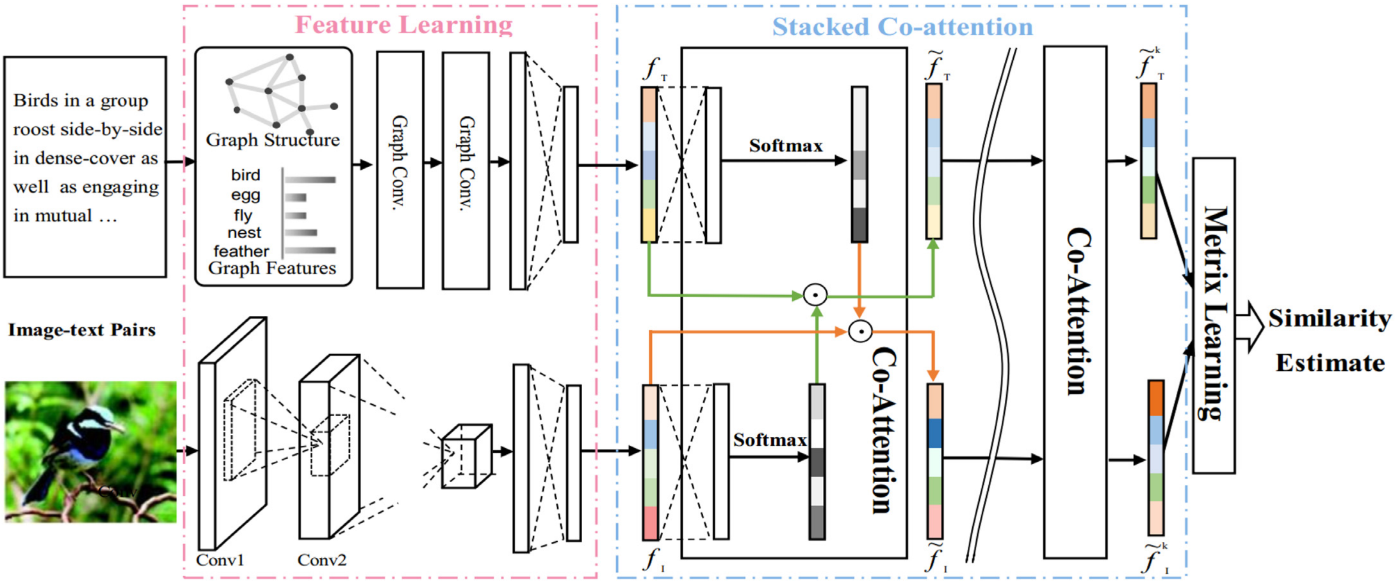


Fig. 3. The structure of the proposed model is a dual-path neural network: i.e., text Graph Convolutional Network (text GCN) (top) and image Neural Network (image NN) (bottom). The text GCN for learning text representation contains two layers of graph convolution on the top of constructed featured graph. 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.

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 inter-word similarity 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.

3.2. 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) 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 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 $L = I_N - D^{-1/2}AD^{-1/2}$, 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 Λ is a set of real, non-negative eigenvalues. The Fourier transform of F_{in} is a function of U defined as:

$$\widehat{F_{in}} = U^T F_{in} \tag{2}$$

While the inverse transform is defined as:

$$F_{in} = U\widehat{F_{in}} \tag{3}$$

The convolution of F_{in} with a spectral filter g_θ is given by:

$$F_{out} = g_\theta * F_{in} = U g_\theta U^T F_{in} \tag{4}$$

where parameter θ is a vector to learn. In order to keep the filter K-localized in space and computationally efficient, Defferrard et al. [4] propose a approximated polynomial filter defined as:

$$g_\theta = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) \tag{5}$$

where $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$, $\tilde{L} = \frac{2}{\lambda_{\max}}L - I_N$ and $\lambda_{\max}$ 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 [4]. 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:

$$f_{out,i} = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}) f_{in,i} \tag{6}$$

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 nonlinearity. A fully connected layer is successive with the last convolution layer to map the text features to the common semantic space. Given a text document T, the text representation f_t learnt by the text GCN model $H_t(\cdot)$ is denoted by:

$$f_t = H_t(T) \tag{7}$$

3.3. Image model

For modeling images, we adopt a neural network (NN) containing a set of fully connected layers (bottom in Fig. 3). 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 our model, the pre-trained CNN, i.e. VGGNet, are used for image feature extraction. 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 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:

$$f_{img} = H_{img}(I) \tag{8}$$

3.4. Stacked co-attention networks

Given the text features f_t and the image features f_{img}, 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 will introduce noise from 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. 3, the inputs of stacked co-attention networks the are features f_t and f_{img} 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_{img} is updated by the attention distribution learnt from the text features and we denote the output image features as $\tilde{f_{img}}$. Specifically, given f_t and f_{img}, we feed them into a fully connected layer and then apply a *softmax* function to generate the attention distribution. Here is the formula for computing the attention distribution:

$$h_t = relu(W_t f_t + b_t) \qquad p_t = softmax(h_t) \tag{9}$$

$$h_{img} = relu(W_{img} f_{img} + b_{img}) \qquad p_{img} = softmax(h_{img}) \tag{10}$$

where $f_t \in \mathbb{R}^N, f_{img} \in \mathbb{R}^N, W_t \in \mathbb{R}^{N\times N}, W_{img} \in \mathbb{R}^{N\times N}, b_t \in \mathbb{R}^N$ and $b_{img} \in \mathbb{R}^N$. p_{img} 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 $\tilde{f_{img}}$ according to the following formula:

$$\tilde{f}_t = p_{img} \circ f_t \qquad \tilde{f_{img}} = p_t \circ f_{img} \tag{11}$$

where "○" is Hadamard product. $\tilde{f}_t$ and $\tilde{f_{img}}$ 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 $\tilde{f}_{img}^k$ for the text and image respectively. Formally, the k^{th} co-attention layer is computed by the following formula:

$$h_t^k = relu(W_t^k f_t^{k-1} + b_t^k) \qquad p_t^k = softmax(h_t^k) \tag{12}$$

$$h_{img}^k = relu(W_{img}^k f_{img}^{k-1} + b_{img}^k) \qquad p_{img}^k = softmax(h_{img}^k) \tag{13}$$

Based on the attention distribution p_t^k and p_{img}^k, we can get the updated text features $\tilde{f}_t^k$ and image features $\tilde{f}_{img}^k$ by the following formula:

$$\tilde{f}_t^k = p_{img}^k \circ f_t^{k-1} \qquad \tilde{f}_{img}^k = p_t^k \circ f_{img}^{k-1} \tag{14}$$

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 generally the best choice.

3.5. 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 after k^{th} co-attention layers, i.e. $\tilde{f}_t^k$ and $\tilde{f}_{img}^k$, 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 [12], 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 defined by:

$$Loss = (\sigma^{2+} + \sigma^{2-}) + \lambda_{\max}(0, m - (u^+ - u^-)) \tag{15}$$

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 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.

4. 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.

4.1. Datasets

Experiments are conducted on four English benchmark datasets, i.e. English Wikipedia, NUS-WIDE, Pascal VOC, and TVGraz. To verify the extensibility, 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) [25] contains 2866 image-text pairs divided into 10 classes, where 2173 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 [28]. 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 [38]. For images, we use 500-dimensional bag-of-features. For tags, we construct a graph with 5018 vertices.

Pascal VOC dataset consists of 9963 image-tag pairs belonging to 20 classes. The images containing only one object are selected in our experiments as [27,31,32], obtaining 2808 training and 2841 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.

TVGraz dataset contains 2594 image-text pairs [21]. We choose the texts that have more than 10 words in our experiments and results in 2360 image-text pairs, where 1885 pairs for training and 475 pairs for testing. Each image is represented by a 4,096-dimensional VGG-19 feature and each text is represented by a graph with 8172 vertices.

Chinese wikipedia dataset (Ch-Wiki for short) [23] is collected from Chinese Wikipedia articles. It contains 3103 image-text pairs divided into 9 classes, where 2482 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 9613 vertices.

4.2. Evaluation and implementation

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 [25] is also utilized for evaluation. In our implementation, we set $k = 8$ in k-nearest neighbors for text graph construction.

In this work, positive samples denote the text-image pairs belonging to the same class while negative samples correspond to text-image pairs from different classes. So the ground truth labels are binary denoting whether the input pairs are from the same class or not. 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. Though the experimental results show that no fine-grained labels about the degrees of ”negativity” or ”positivity” of the negative or positive samples are necessary, the training samples are actually not equally ”negative” or ”positive”. Ma et al. [17] have demonstrated that adaptively assigning different degrees of negativity to the negative samples can better capture the discriminate knowledge instead of treating all the negative samples equally. The paper empirically shows the effectiveness of such idea in the tasks of complex event detection. Our future work could explore the discrimination of the samples for fine-grained semantical matching across different modalities and for more accurate retrieval order.

We train the model for 50 epochs with mini-batch size 200. We adopt the dropout ratio of 0.2 at the input of the last FC layer, learning rate 0.001 with an Adam optimisation, and regularisation 0.005. The model is not much sensitive to λ but a little sensitive to m in the loss function. In general, 0.35 for λ and 0.6 for m are the relative best settings for our model. In the last semantic mapping layers of both text path and image path, the reduced dimensions are set to 1024, 500, 256, 1024, 1024 for Eng-Wiki, NUS-WIDE, Pascal, TVGraz, and Ch-Wiki, respectively. The code will be released on github upon the acceptance of the paper.

4.3. Comparison with state-of-the-art methods

We compare our proposed SCANet with a number of state-of-the-art models, including CCA & SCM [25], TCM & w-TCM & c-TCM [23], GMLDA & GMMFA [27], LCFS [31], MvDA [7], LGCFL [8], ml-CCA [24], AUSL [38], JFSSL [30], PLS [26], BLM [27], CDFE [14], CCA-3V [5], CM & SM [21], and TTI [22]. For the same settings with [38], principal component analysis is performed on the original features for CCA, SCM, GMLDA and MvDA.

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

Method	Text query	Image query	Average	Dataset
CCA	0.1872	0.2160	0.2016	Eng-Wiki
SCM	0.2336	0.2759	0.2548	
TCM	0.2930	0.2320	0.2660	
LCFS	0.2043	0.2711	0.2377	
MvDA	0.2319	0.2971	0.2645	
LGCFL	0.3160	0.3775	0.3467	
ml-CCA	0.2873	0.3527	0.3120	
GMLDA	0.2885	0.3159	0.3022	
GMMFA	0.2964	0.3155	0.3060	
AUSL	0.3321	0.3965	0.3643	
JFSSL	0.4102	**0.4670**	0.4386	
SCANet	**0.8081**	0.4603	**0.6342**	
CCA	0.2667	0.2869	0.2768	NUS-WIDE
LCFS	0.3363	0.4742	0.4053	
LGFCL	0.3907	0.4972	0.4440	
ml-CCA	0.3908	0.4689	0.4299	
AUSL	0.4128	**0.5690**	0.4909	
JFSSL	0.3747	0.4035	0.3891	
SCANet	**0.5811**	0.5190	**0.5501**	
PLS	0.1997	0.2757	0.2377	Pascal
BLM	0.2408	0.2667	0.2538	
CCA	0.2215	0.2655	0.2435	
CDFE	0.2211	0.2928	0.2569	
GMLDA	0.2448	0.3094	0.2771	
GMMFA	0.2308	0.3090	0.2699	
CCA3V	0.2562	0.3146	0.2854	
LCFS	0.2674	0.3438	0.3056	
JFSSL	0.2801	**0.3607**	0.3204	
SCANet	**0.4790**	0.3370	**0.4080**	
TTI	0.1530	0.2160	0.1845	
CM	0.4500	0.4600	0.4550	
SM	0.5850	0.6190	0.6020	TVGraz
SCM	0.6960	0.6930	0.6945	
TCM	0.7060	0.6940	0.6950	
SCANet	**0.8241**	**0.8812**	**0.8527**	
w-TCM	0.2980	0.2410	0.2695	Ch-Wiki
c-TCM	0.3170	0.3100	0.3135	
SCANet	**0.4533**	**0.3810**	**0.4155**	

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, ml-CCA, TCM, LCFS, LGCFL, JFSSL, 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 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 on five datasets, for fair comparison, we compare our model to methods on their reported datasets. From Table 1, we can have the following observations:

First, SCANet 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 SCANet are 80.81% and 47.90%, which are about 39.79% and 19.89% higher than the second best result from JFSSL. For the NUS-WIDE dataset, the MAP score of SCANet is 58.11% and 16.83% 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 can effectively leverage 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 mater *word2vec* 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.

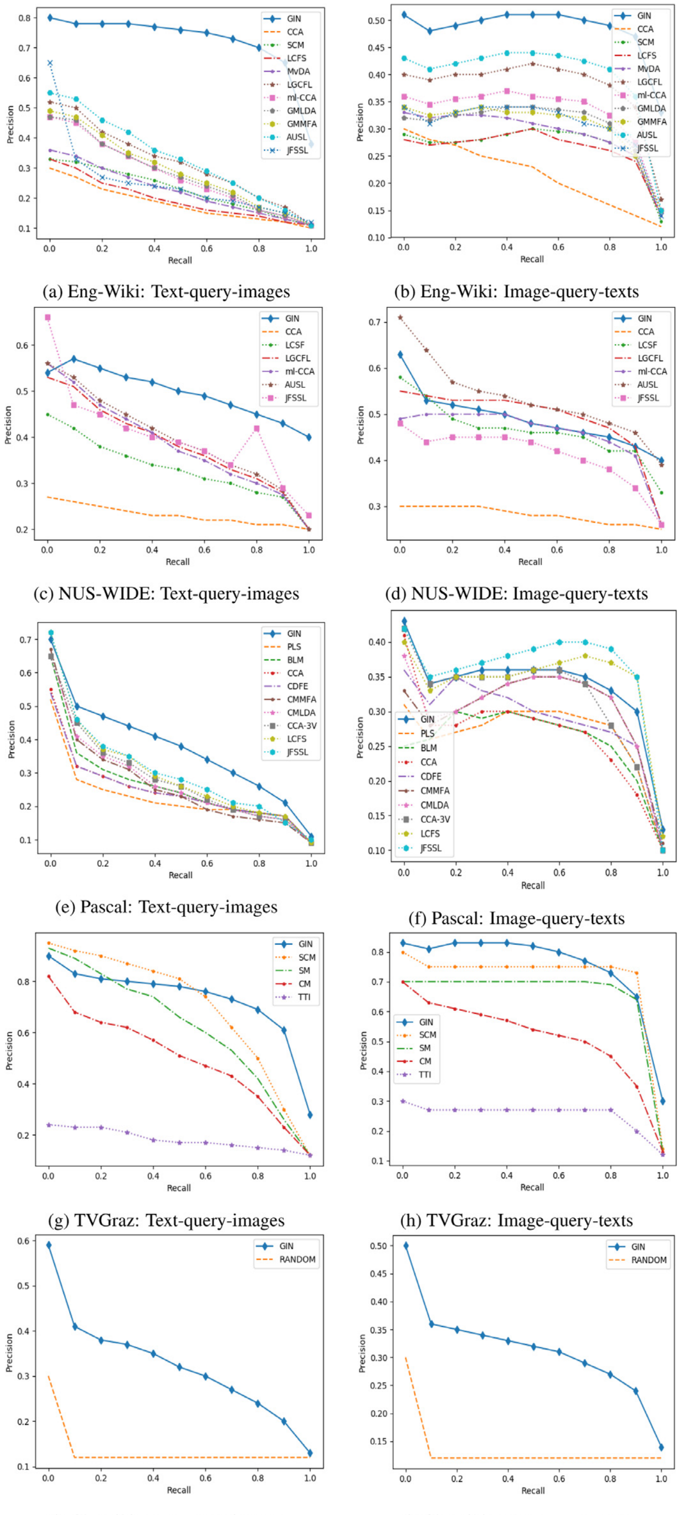
Second, the MAP of SCANet for the image-query-text task is superior to most of the compared methods. SCANet ranks the second best on Eng-Wiki and NUS-WIDE, the third best on Pascal and the best on TVGraz and Chi-Wiki. Table 1 indicates that SCANet is only inferior to JFSSL by 0.67% on Eng-Wiki and 2.37% on Pascal. SCANet is just 5% lower than AUSL on NUS-WIDE. Since SCANet 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, SCANet achieves the best average MAP over all the competitors, especially outperforming the second best method JFSSL by 19.56% 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, SCANet 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, 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, SCANet gains 13.63% and 5.10% 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 Fig. 4. Since the competitive models, i.e. w-TCM and c-TCM, haven't reported PR curves on Ch-Wiki, we compare SCANet with random baseline on this dataset. For JFSSL, we show its best MAP after feature selection (see Table 7 in [30]). Since JFSSL hasn't reported the PR curves corresponding to the best MAP, we use its reported PR curves in [30].

For the text-query-image task, it's obvious that SCANet 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, SCANet outperforms other competitors with almost all the recall rate on Eng-Wiki. For NUS-WIDE dataset, SCANet is only inferior to AUSL and LGCFL. For Pascal dataset, SCANet is just slightly inferior to JFSSL. On the whole, SCANet is comparable with state-of-the-art methods for the image-query-text task.



(a) Eng-Wiki: Text-query-images

(b) Eng-Wiki: Image-query-texts

(c) NUS-WIDE: Text-query-images

(d) NUS-WIDE: Image-query-texts

(e) Pascal: Text-query-images

(f) Pascal: Image-query-texts

(g) TVGraz: Text-query-images

(h) TVGraz: Image-query-texts

(i) Ch-Wiki: Text-query-images

(j) Ch-Wiki: Image-query-texts

Fig. 4. Precision-recall curves on the five datasets.

4.4. Baseline comparisons

Besides our proposed model, we implement another four baseline models to evaluate the influence of inter-word semantics, stacked co-attention, and the variation in text features and image features on the final retrieval performance. All the experiments are conducted on the Eng-Wiki dataset. The retrieval performance of MAP is given in Table 2.

4.4.1. The influence of inter-word semantics

To evaluate the effects of inter-word semantics, we first conduct experiments on our model without the stacked co-attention layers, leaving other modules unchanged. Since our model leverages GCN to extract the inter-word semantics of texts, we implement another two baseline models to replace the GCN module with LSTM [37] and CNN [9] respectively, to prove the effects of

Table 2
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

exploring inter-word semantics. We fix the image features of VGG-19 for all the aforementioned three models. The first three models in Table 2 without attention layers (i.e. #Attention layers = 0) shows the final cross-modal retrieval results. It's obvious that GCN+VGG-19 outperforms the other two models significantly for the text retrieval task. The MAP score of using LSTM module (i.e. LSTM+VGG-19) is inferior to GCN+VGG-19 by 13% while using CNN module performs the worst. Meanwhile, for the image query and average performance, GCN+VGG-19 also outperforms the other two baselines, which indicates the effectiveness of inter-word semantics in the final cross-modal retrieval performance compared with other text modelling methods.

4.4.2. The influence of stacked co-attention

To evaluate the influence of the stacked co-attention, we vary the number of co-attention layers on all the five models in Table 2. Not that, compared with the architecture of SCANet, we fix the text features of GCN and implement the last two models in Table 2 by replacing the image features of VGG-19 with ResNet-50 and CNN with five convolution layers (CNN-5), respectively. For each baseline model, we variate the number of attention layers ranging from 1 to 3 to evaluate the influence of effectiveness of the stacked co-attention.

It's obvious that almost all the models benefit from the stacked co-attention mechanism compared with the corresponding models without co-attention layer (i.e. #Attention layers = 0). Generally speaking, when the number of attention layers changes from 1 to 3, the increase of MAP 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 text-image pairs 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.

4.4.3. The influence of text and image features

For the first three models in Table 2, we fix the image features of 4096-dimentional VGG-19 features. For each number of attention layers, 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 to GCN while CNN performs the worst. For the last three models in Table 2, we fix the text features of GCN-based 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.

4.4.4. Quality analysis of retrieval results

For quality examination, two examples for image-query-text and text-query-image tasks are shown in Figs. 5 and 6, respectively. Fig. 5 shows the corresponding images of the top ten retrieved texts. Most of the images of our model (i.e. GCN+VGG-19, #attn.=2) display figures in the war, which are semantically related to "warfare" and closely related to the query image than other baseline models. In Fig. 6, the text query is a paragraph related to "biology". The corresponding image displays a bird served as ground truth. The top ten retrieved images of our model are semantically related to both the class and the fine-grained content of the query text, compared with other models. Both examples indicate that SCANet is effective in fine-grained correlation learning by considering both inter-word semantics and stacked co-attention.

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

4.5. 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 λ from 0.25 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 and our model 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 m. In general, 0.35 for λ and 0.6 for m are the relative best settings for our model.

4.6. Objective function analysis

The assumption behind the objective function of SCANet is that the similarity scores between matching text-image pairs and non-matching text-image pairs are samples from two different distributions. As formulated in Eq. (15), the global loss aims to minimise the mean value of similarity scores between non-matching pairs and the variance of the two distributions, and maximising the mean value of the similarity score between matching pairs. In this section, we conduct experiments to evaluate the effects of variance and mean value on the final cross-modal retrieval performance. The loss function only containing the parts of variance or mean value are respectively defined by $Loss_{var} = \sigma^{2+} + \sigma^{2-}$ and $Loss_{mean} = \max(0, m - (u^{+} - u^{-}))$. The MAP scores on Eng-Wiki are shown in Table 4. The average MAP score of the model based on $Loss_{mean}$ is 28.4% lower than that of $Loss$-based model. The reason is that the overlap area of the two distributions increases without minimising the corresponding variance, which increasing the

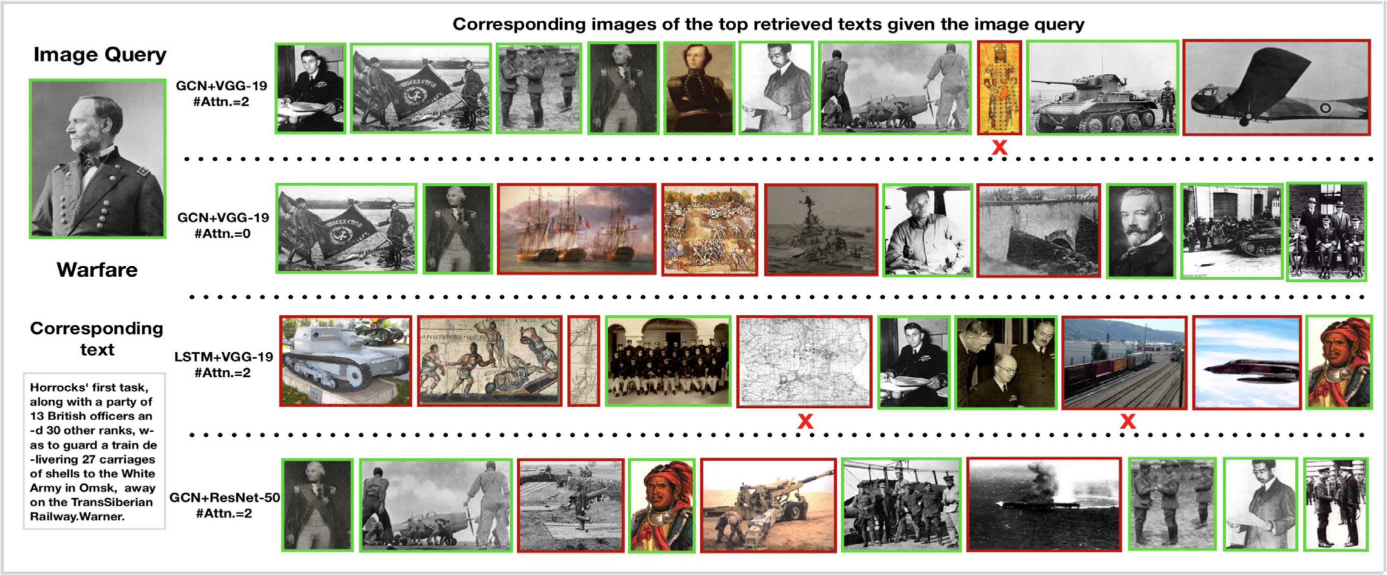

Corresponding images of the top retrieved texts given the image query
Image Query
Warfare
Corresponding text
Horrocks' first task, along with a party of 13 British officers an -d 30 other ranks, w- as to guard a train de -livering 27 carriages of shells to the White Army in Omsk, away on the TransSiberian Railway.Warner.
GCN+VGG-19 #Attn.=2
GCN+VGG-19 #Attn.=0
LSTM+VGG-19 #Attn.=2
GCN+ResNet-50 #Attn.=2
X
X
X

Fig. 5. Retrieval samples of four models for image query task on the Eng-Wiki dataset. Given the query image on the top left, for clear semantic comparisons, we show the corresponding images of the top ten retrieved texts on the right. The green boxes indicate visually related results while the red boxes mean visually irrelated results. The red crosses mark the wrong retrieved results that in different classes with the query. Not that, the right retrieved results, which are in the same class with the query, are not always visually related to the query. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

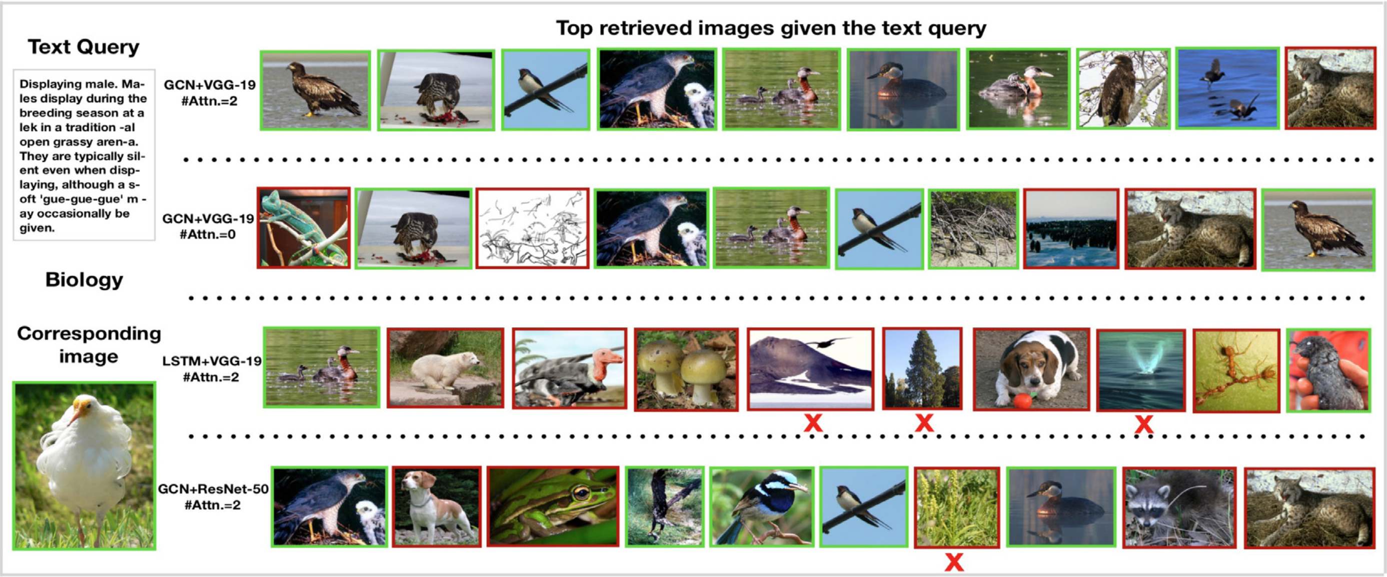

Top retrieved images given the text query
Text Query
Displaying male. Ma- les display during the breeding season at a lek in a tradition -al open grassy aren-a. They are typically sil- ent even when disp- laying, although a s- oft 'gue-gue-gue' m - ay occasionally be given.
Biology
Corresponding image
GCN+VGG-19 #Attn.=2
GCN+VGG-19 #Attn.=0
LSTM+VGG-19 #Attn.=2
GCN+ResNet-50 #Attn.=2
X
X
X
X

Fig. 6. Retrieval samples of four models for text query task on the Eng-Wiki dataset. Given the query image on the top left, for clear semantic comparisons, we show the corresponding image of the query text on the bottom left. The top ten retrieved images are shown on the right. The boxes and crosses have the same meaning as in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The influence of the mean and variance in the objective function.

Objective function	Text query	Image query	Average
$Loss_{mean}$	0.392	0.307	0.350
$Loss_{var}$	0.104	0.103	0.104
$Loss$	0.808	0.460	0.634

proportion of false positive and false negative. The MAP score of the model only based on variance is merely 10.4% on average since that it cannot pushes matching pairs away from non-matching pairs without the constraints between mean values. Therefore, the global loss plays an indispensable role for the superior results of our model due to its better regularization.

5. Conclusion

In this paper, we propose a novel cross-modal retrieval model named as SCANet 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. To deal with the problem of information imbalance between different modalities, stacked co-attention networks are proposed to learn the mutually attended parts from the representations of different modalities and result in fine-grained semantic correlations. A dual-path neural network is trained using a pairwise similarity loss function. Extensive experiments on five benchmark datasets verify that our model considerably outperforms the state-of-the-art models. Besides, our model can be widely used in analyzing heterogeneous data lying on irregular or non-Euclidean domains.

Conflict of interest

The authors declare that they have no conflict of interest.

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] D.M. Blei, M.I. Jordan, Modeling annotated data, in: SIGIR, 2003, pp. 127–134.
[2] D.M. Blei, M.I. Jordan, Modeling anotated data, in: SIGIR, 2003, pp. 127–134.
[3] X. Chang, Y.-L. Yu, Y. Yang, E.P. Xing, Semantic pooling for complex event analysis in untrimmed videos, TPAMI 39 (8) (2017) 1617–1632.
[4] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: NIPS, 2016, pp. 3837–3845.
[5] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for internet images, tags, and their semantics, TPAMI 106 (2) (2014) 210–233.
[6] D.R. Hardoon, S. Szedmák, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[7] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, TPAMI 38 (1) (2016) 188–194.
[8] C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, Learning consistent feature representation for cross-modal multimedia retrieval, TMM 17 (3) (2017) 276–288.
[9] Y. Kim, Convolutional neural networks for sentence classification, arXiv:1408.5882 (2014).
[10] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, ICLR, 2017.
[11] S.I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, D. Rueckert, Distance metric learning using graph convolutional networks: Application to functional brain networks, arXiv:1703.021612017.
[12] V. Kumar BG, G. Carneiro, I. Reid, Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions, in: CVPR, 2016, pp. 5385–5394.
[13] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, IEEE 86 (11) (1998) 2278–2324.
[14] D. Lin, X. Tang, Inter-modality face recognition, in: ECCV, 2006, pp. 13–26.
[15] H. Liu, R. Ji, Y. Wu, G. Hua, Supervised matrix factorization for cross-modality hashing, in: IJCAI, 2016, pp. 1767–1773.
[16] M. Luo, X. Chang, Z. Li, L. Nie, A.G. Hauptmann, Simple to complex cross-modal learning to rank, Comput. Vision Image Understanding (2017).
[17] Z. Ma, X. Chang, Y. Yang, N. Sebe, A.G. Hauptmann, The many shades of negativity, TMM 19 (7) (2017) 1558–1568.
[18] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: ICLR, 2013, pp. 1–12.
[19] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, arXiv:1611.00471 (2016).
[20] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks., in: IJCAI, 2016, pp. 3846–3853.
[21] J.C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, TPAMI 36 (3) (2014) 521–535.
[22] G. Qi, C. Aggarwal, T. Huang, Towards semantic knowledge propagation from text corpus to web images, in: WWW, 2011, pp. 297–306.
[23] Z. Qin, J. Yu, Y. Cong, T. Wan, Topic correlation model for cross-modal multimedia information retrieval, Pattern Anal. Appl. 19 (4) (2016) 1007–1022.
[24] V. Ranjan, N. Rasiwasia, C.V. Jawahar, Multi-label cross-modal retrieval, in: ICCV, 2015, pp. 4094–4102.
[25] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM-MM, 2010, pp. 251–260.
[26] A. Sharma, D.W. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution adn sketch, in: CVPR, 2011, pp. 593–600.
[27] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: CVPR, 2012, pp. 2160–2167.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, ICLR, 2015.
[29] C. Wang, Y. Song, H. Li, M. Zhang, J. Han, Text classification with heterogeneous information network kernels, in: AAAI, 2016, pp. 2130–2136.
[30] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, TPAMI 38 (10) (2016) 2010–2023.
[31] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, in: ICCV, 2013, pp. 2088–2095.
[32] L. Wang, R. He, Z. Sun, T. Tan, Group-invariant cross-modal subspace learning, in: IJCAI, 2016, pp. 1739–1745.
[33] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: CVPR, 2016, pp. 5005–5013.
[34] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: CVPR, 2016, pp. 21–29.
[35] J. Yu, Y. Cong, Z. Qin, T. Wan, Cross-modal topic correlations for multimedia retrieval, in: ICPR, 2012, pp. 246–249.
[36] P. Yuxin, Q. Jinwei, Y. Yuxin, Modality-specific cross-modal similarity measurement with recurrent attention network, arXiv:1708.04776 (2017).
[37] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, arXiv:1409.2329 (2014).
[38] L. Zhang, B. Ma, J. He, G. Li, Q. Huang, Q. Tian, Adaptively unified semi-supervised learning for cross-modal retrieval, in: IJCAI, 2017, pp. 3406–3412.
[39] X. Zhang, S. Zhou, J. Feng, H. Lai, B. Li, Y. Pan, J. Yin, S. Yan, Hashgan: attention-aware deep adversarial hashing for cross modal retrieval, arXiv:1711.09347 (2017b).
[40] Y. Zheng, D.-Y. Yeung, Co-regularized hashing for multimodal data, in: NIPS, 2012, pp. 1376–1384.
[41] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, Y.-D. Shen, Dual-path convolutional image-text embedding, arXiv:1711.05535 (2017).
[42] L. Zhu, Z. Huang, X. Liu, X. He, J. Sun, X. Zhou, Discrete multimodal hashing with canonival views for robust mobile landmark search, TMM 19 (9) (2017) 2066–2079.
[43] L. Zhu, J. Shen, H. Jin, L. Xie, R. Zheng, Landmark classification with hierarchical multi-modal exemplar feature, TMM 17 (7) (2015) 981–993.
[44] L. Zhu, J. Shen, L. Xie, Z. Cheng, Unsupervised topic hypergraph hashing for efficient mobile image retrieval, IEEE Trans. Cybern. 47 (11) (2016) 1–14.