Modality-Specific Cross-Modal Similarity Measurement With Recurrent Attention Network

Yuxin Peng, Jinwei Qi, and Yuxin Yuan

***Abstract*— Nowadays, cross-modal retrieval plays an important role to flexibly find useful information across different modalities of data. Effectively measuring the similarity between different modalities of data is the key of cross-modal retrieval. Different modalities, such as image and text, have imbalanced and complementary relationship, and they contain unequal amount of information when describing the same semantics. For example, images often contain more details that cannot be demonstrated by textual descriptions and vice versa. Existing works based on deep neural network mostly construct one common space for different modalities, to find the latent alignments between them, which lose their exclusive modality-specific characteristics. Therefore, we propose modality-specific cross-modal similarity measurement approach by constructing the independent semantic space for each modality, which adopts an end-to-end framework to directly generate the modality-specific cross-modal similarity without explicit common representation. For each semantic space, modality-specific characteristics within one modality are fully exploited by recurrent attention network, while the data of another modality is projected into this space with attention based joint embedding, which utilizes the learned attention weights for guiding the fine-grained cross-modal correlation learning, and captures the imbalanced and complementary relationship between different modalities. Finally, the complementarity between the semantic spaces for different modalities is explored by adaptive fusion of the modality-specific cross-modal similarities to perform the cross-modal retrieval. Experiments on the widely used Wikipedia, Pascal Sentence, and MS-COCO data sets as well as our constructed large-scale XMediaNet data set verify the effectiveness of our proposed approach, outperforming nine state-of-the-art methods.**

***Index Terms*— Modality-specific cross-modal similarity measurement, recurrent attention network, attention based joint embedding, adaptive fusion.**

I. Introduction

NOWADAYS, multimodal data such as image, video, text and audio, has been widely available on the Internet, which is the fundamental component for promoting artificial intelligence to understand the real world. Faced with the inconsistent representations and distributions of different modalities, some works have been done for breaking the boundaries between them. Cross-modal retrieval has become a highlighted research topic to perform retrieval across different modalities, which is in great demands with many practical applications, such as search engine and digital libraries. An example of cross-modal retrieval is shown in Figure 1. Different from the traditional single-modal retrieval, such as image retrieval [1] and video retrieval [2], which is limited in providing retrieval results of the same single modality with query, cross-modal retrieval is more flexible to retrieve relevant multimodal information by submitting one query of any modality [3].

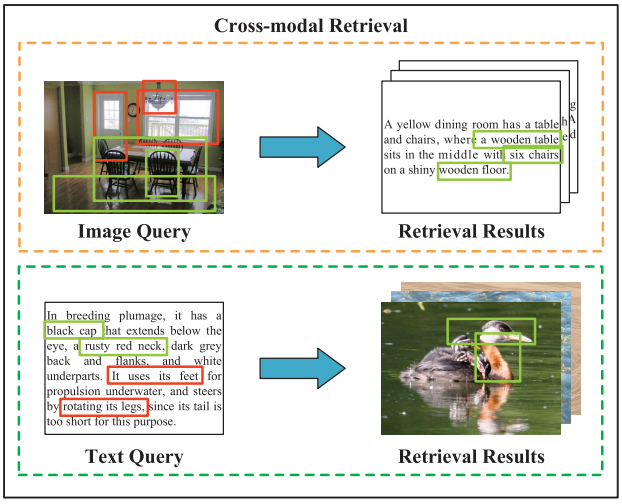


Fig. 1. An example of cross-modal retrieval with image and text, which can present retrieval results with different modalities. We can also see that the instances of different modalities have imbalanced and complementary relationship, where the green boxes indicate the fine-grained information described by both image and text, while the red boxes mean extra information contained in only one modality that is not demonstrated by the other modality.

The main challenge of cross-modal retrieval is to deal with the inconsistency of representations between different modalities and learn their intrinsic correlation. For the fact that data of different modalities has diverse representations and distributions that usually span different feature spaces, such heterogeneous characteristic makes it hard to directly measure the similarity between different modalities, such as an image and an audio clip. To address this issue, some methods [4]–[7] have been proposed to project the features of different modalities into one common space to learn the common representation, by which the cross-modal similarity can be

Manuscript received August 4, 2017; revised March 2, 2018 and June 22, 2018; accepted June 27, 2018. Date of publication July 2, 2018; date of current version August 14, 2018.This work was supported by the National Natural Science Foundation of China under Grant 61771025 and Grant 61532005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chia-Wen Lin. *(Corresponding author: Yuxin Peng.)*

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: pengyuxin@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2852503

calculated to perform retrieval. Traditional methods build the common space by learning mapping matrices to maximize the correlations of variables from different modalities, such as methods based on Canonical Correlation Analysis (CCA) [4], [8], graph regularization [9], [10] and learning to rank [11], [12]. Recently, the dramatic advances in deep learning have inspired researchers to bridge the gap between different modalities with Deep Neural Network (DNN). Such methods like [13]–[15] attempt to exploit the advantages of DNN in modeling nonlinear correlation with multilayer networks.

The aforementioned methods mostly project the data of different modalities from their own feature spaces into one single common space equally to find the latent alignments between them, which means *equal* amount of information is captured from the data of different modalities during cross-modal correlation learning. Generally speaking, different modalities such as image and text have imbalanced and complementary relationship that provides *unequal* amount of information in describing the same semantics, because some modality-specific characteristics within one modality cannot be exactly aligned with other modalities. For example, an image often co-occurs with its corresponding textual descriptions on a web page to describe the same semantics such as objects or events. But not all fine-grained image details can be exactly aligned to the textual descriptions and vice versa. Images often contain more details which cannot be demonstrated by textual descriptions, that is what we usually say "A picture is worth a thousand words". While in other cases, the opposite would happen that textual descriptions contain more semantic information than image when describing some high-level semantics, such as historical events or literature works. As shown in Figure 1, green boxes indicate the appropriate alignments between visual and textual fine-grained information, while the red boxes mean the mis-alignments. Therefore, treating different modalities equally to find fine-grained alignments for constructing one common space loses such exclusive and useful modality-specific characteristics, which cannot fully exploit the intrinsic information within each modality.

Moreover, the imbalanced relationship between image and text with additional information in one modality is also useful for cross-modal retrieval. For example, assume that an image in one image/text pair has more details, which are not demonstrated by its corresponding textual description. On one hand, such additional visual information may be described by other textual descriptions of the same category, which is useful to retrieve other textual results of the same category, and should be preserved. On the other hand, such additional visual information in one image may be also contained in other images, which is helpful for constructing semantic space of image, where those images with similar additional visual information can be clustered together with their corresponding textual descriptions, to describe the latent semantic information. Thus, the relevant image results are easier to be retrieved in this semantic space by the text query.

For addressing the above issues, we aim to preserve the modality-specific characteristics by fully exploiting the fine-grained information within each modality, when learning the cross-modal correlation. Thus, modality-specific measurement is required for each modality instead of only constructing one common space. Recently, attention mechanism has made great advances in DNN, which allows models to concentrate on the necessary fine-grained parts of visual or textual inputs, and has been successfully applied to various multimodal tasks, such as image caption [16] and visual question answering [17]. Inspired by these advances, we attempt to consider the visual and textual attention respectively. It aims to capture the modality-specific characteristics within one modality, and distinguish which parts have larger probabilities to be demonstrated in another modality, instead of exacting alignment between the parts of image and text, which can fully exploit the imbalanced relationship between image and text.

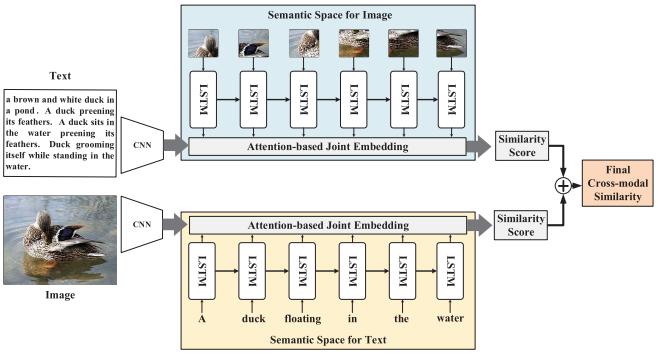


Fig. 2. An overview of the proposed approach, which constructs independent semantic spaces for different modalities, and directly generates modality-specific cross-modal similarities through an end-to-end framework without explicit common representation.

In this paper, we propose modality-specific cross-modal similarity measurement (MCSM) approach, which constructs independent semantic spaces for different modalities, where the modality-specific characteristics can be fully explored by attention mechanism during cross-modal correlation learning. The modality-specific cross-modal similarity is directly generated from each semantic space through the end-to-end framework without explicit common representation. Figure 2 shows an overview of our proposed approach. The main contributions of this paper are presented as follows:

- *Modality-specific cross-modal similarity measurement.* We construct independent semantic space for each modality. In each semantic space, the modality-specific characteristics are fully exploited by modeling the fine-grained information within one modality, while the data of another modality is projected into this semantic space to capture the imbalanced and complementary relationship between different modalities during cross-modal correlation learning.
- *Recurrent attention network with joint embedding.* We design recurrent attention network in each semantic space, which aims to capture the modality-specific characteristics including both fine-grained local and context information by recurrent network with attention mechanism. While an attention based joint embedding

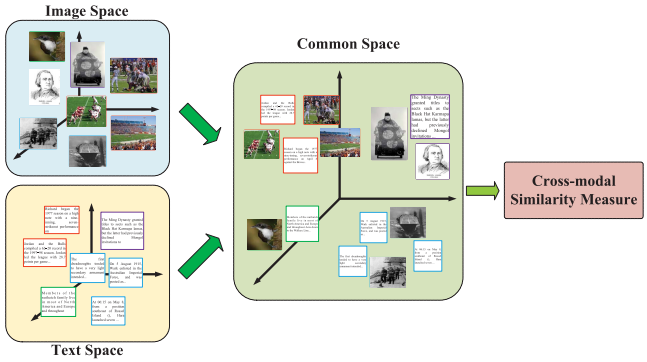


Fig. 3. Illustration of the mainstream framework of the cross-modal retrieval methods, which attempt to represent data of different modalities with the same "feature" type (namely common representation) in one common space, so that the similarity measurement can be directly performed.

loss is proposed to utilize the learned attention weights for guiding the fine-grained cross-modal correlation learning.

- *End-to-end framework with adaptive fusion.* We propose an end-to-end framework in each semantic space to directly generate modality-specific cross-modal similarity, which integrates both representation learning and similarity measurement stages to benefit each other. Furthermore, adaptive fusion is proposed to obtain the final similarity for performing cross-modal retrieval, which can fully explore the complementarity between the semantic spaces for different modalities.

Experiments on the widely-used Wikipedia, Pascal Sentence, MS-COCO datasets as well as our constructed large-scale XMediaNet dataset show that our proposed MCSM approach outperforms 9 state-of-the-art methods, which verifies the effectiveness of MCSM approach.

The remainder of this paper is organized as follows. We briefly review the related works in Section II. In Section III, our proposed MCSM approach is presented in detail. Then Section IV reports the experimental results as well as analyses. Finally, Section V concludes this paper.

II. Related Works

In this section, we briefly review the representative cross-modal retrieval methods with common space learning, as well as the recent advances of attention mechanism in DNN.

A. Common Space Learning for Cross-Modal Retrieval

The current mainstream of cross-modal retrieval methods is to learn an intermediate common space for different modalities, and then the cross-modal similarity can be directly measured in the common space. Figure 3 illustrates the main framework of such common space learning methods. As indicated in [3], we mainly introduce three categories of existing methods as follows, namely traditional statistical correlation analysis methods, cross-modal graph regularization methods and DNN-based methods.

1) Traditional Statistical Correlation Analysis Methods: As the foundation of common space learning methods, these methods mainly optimize the statistical values to learn linear projection matrices, which project features of different modalities into the common space and obtain the common representation. Canonical Correlation Analysis (CCA) [18], as one of the most representative works, is a natural solution to maximize the pairwise correlation between the data of different modalities such as image/text pairs [4]. Furthermore, semantic information can be incorporated to extend CCA for improving the accuracy of cross-modal retrieval. For example Pereira *et al.* [19] integrate semantic labels to improve the performance of CCA. Multi-view CCA [8] is proposed to construct the third view for modeling high-level semantics. Ranjan *et al.* [20] propose multi-label CCA, which considers the high-level semantic information in the form of multi-label annotations. Besides, Cross-modal Factor Analysis (CFA) [21], as one of the alternative methods, is proposed to minimize the Frobenius norm between the data of different modalities after projecting them into one common space.

2) Cross-Modal Graph Regularization Methods: Graph regularization [22] is widely used to construct a partially labeled graph for semi-supervised learning, which aims to enrich the training set with unlabeled data and smooth the solution. Zhai *et al.* [23] are the first to integrate graph regularization into cross-modal retrieval and propose Joint Graph Regularized Heterogeneous Metric Learning (JGRHML), which constructs the joint graph regularization term in the learned metric space. Furthermore, Joint Representation Learning (JRL) [9] is proposed to construct several separate graphs for different modalities and learn projection matrices with the joint consideration of correlation and semantic information. Peng *et al.* [24] further improve the previous works [9], [23] by constructing a unified hypergraph to learn the common space for up to five modalities, which also utilizes the fine-grained information at the same time. Besides, Wang *et al.* [10] also adopt multimodal graph regularization term to preserve inter-modality and intra-modality similarity relationships.

3) DNN-Based Methods: Deep learning has made great advance in multimodal applications, such as image/video classification [25], [26] and object recognition [27]. Researches also adopt DNN to perform common space learning to take the advantage of its considerable ability on modeling highly nonlinear correlation. Most of the DNN-based methods construct two subnetworks for different modalities such as image and text, which are linked at the joint layer to construct the common space for cross-modal correlation modeling. Bimodal Autoencoders (Bimodal AE) [28] is proposed as an extension of Restricted Boltzmann Machine (RBM) to model multiple modalities by minimizing the reconstruction error. Srivastava and Salakhutdinov [29] propose Multimodal Deep Belief Network (Multimodal DBN), which adopts two kinds of DBN for different modalities to model the distributions over their original features, while a joint RBM is adopted on the top of them to model the joint distribution and get the common representation. Correspondence Autoencoder (Corr-AE) [14] and Deep Canonical Correlation Analysis (DCCA) [30] also

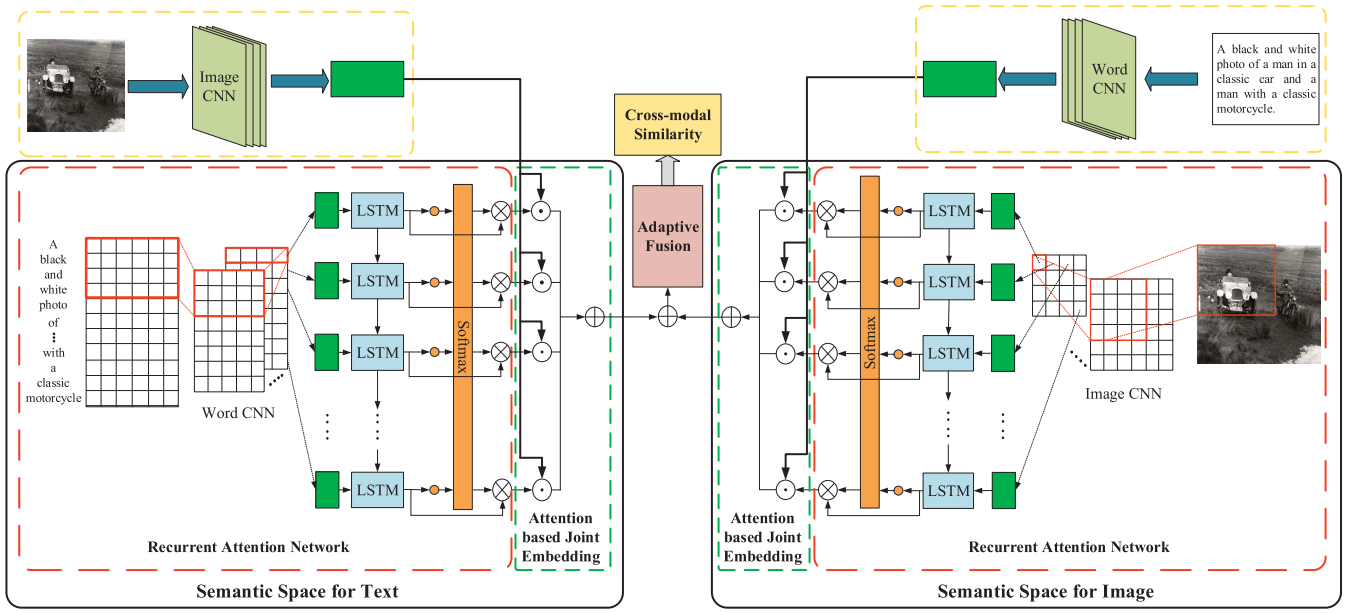


Fig. 4. The overall framework of our MCSM approach, which adopts recurrent attention network with attention based joint embedding loss to construct independent semantic spaces for different modalities and perform cross-modal correlation learning, and the modality-specific cross-modal similarities are directly generated through end-to-end frameworks without explicit common representation.

consist of two subnetworks, while Corr-AE jointly models correlation and reconstruction information, and DCCA combines DNN with CCA to maximize the correlation on the top of two subnetworks. Besides, Peng *et al.* [13] propose Cross-media Multiple Deep Networks (CMDN) to model the intra-modality and inter-modality correlation in a two-stage learning framework. The above works mainly take hand-crafted features as image inputs. Furthermore, Wei *et al.* [31] propose Deep-SM to perform deep semantic matching with Convolutional Neural Network (CNN), which exploits the strong representation ability of CNN features to improve the retrieval accuracy. He *et al.* [32] adopt two convolution-based networks to model the matched and mismatched image/text pairs via deep and bidirectional representation learning. Besides, Peng *et al.* [33] explore the utilization of GANs to realize cross-modal adversarial correlation learning, which constructs a single common space like most prior works to treat data of different modalities equally, and does not involve the exploitation of important hints in fine-grained parts.

The aforementioned methods mostly construct one common space for different modalities to find the latent alignments between them, which lose exclusive modality-specific characteristics. Therefore, we attempt to adopt modality-specific measurement to fully exploit such modality-specific characteristics by modeling the intrinsic fine-grained information within each modality.

B. Attention Mechanism

Attention mechanism, as one of the recent advances in neural network, has been successfully applied to various multimodal tasks, which allows models to focus on the salient fine-grained parts of visual or textual inputs. We mainly introduce two kinds of attention mechanism as follows, namely visual attention and textual attention.

1) Visual Attention: Recently, many methods have adopted visual attention model to promote image processing tasks, which can pay more attention to fine-grained parts in an image. Mnih *et al.* [34] propose an attention based task-driven visual processing framework for image classification, which adopts Recurrent Neural Network (RNN) to adaptively select a sequence of regions. Gregor *et al.* [35] propose a spatial attention mechanism, which designs a sequential variational auto-encoding framework to perform image generation. Yang *et al.* [36] propose Stacked Attention Networks (SANs) for image question answering, which can locate relevant image regions to the question with stacked attention model.

2) Textual Attention: Some related works in Natural Language Processing (NLP) have adopted textual attention model to find semantic alignments with an encoder-decoder network. Rocktäschel *et al.* [37] propose a word-by-word neural attention mechanism to reason over entailments of paired words or phrases. Hermann *et al.* [38] develop a class of attention based deep neural networks, which learn to read and answer complex questions. Rush *et al.* [39] propose a fully data-driven approach, which adopts a local attention based model to generate summarization according to the input sentence.

Inspired by the great progress of attention mechanism, we propose to fully exploit the intrinsic fine-grained information within each modality with attention mechanism, which can preserve the modality-specific characteristics when learning the cross-modal correlation.

III. Our Proposed Approach

As shown in Figure 4, our proposed MCSM approach adopts modality-specific measurement to construct independent semantic spaces by end-to-end framework for image and text respectively. First, *recurrent attention network* is adopted

to fully exploit the fine-grained modality-specific characteristics in each semantic space. Second, the *attention based joint embedding* is employed to capture the imbalanced and complementary relationship between different modalities and perform cross-modal correlation learning. Finally, *adaptive fusion* is proposed to explore the complementarity between different semantic spaces for cross-modal retrieval.

A. Notation

In the beginning, the formal definition of cross-modal retrieval is presented as follows. The two modalities involved in cross-modal retrieval are denoted as I for image and T for text. The multimodal dataset consists of two parts, namely training and testing sets. The training set is denoted as $D_{tr} = \{I_{tr}, T_{tr}\}$. Image training data $I_{tr} = \{i_p\}_{p=1}^{n_{tr}}$ consists of totally n_{tr} instances, and i_p is the p-th image instance. Similarly, the text training data is denoted as $T_{tr} = \{t_p\}_{p=1}^{n_{tr}}$, which has the same number of instances with image for training. Besides, the training data also has its corresponding semantic category labels, which are denoted as $\{c_p^I\}_{p=1}^{n_{tr}}$ and $\{c_p^T\}_{p=1}^{n_{tr}}$. Then, the testing set, which is denoted as $D_{te} = \{I_{te}, T_{te}\}$, includes $I_{te} = \{i_q\}_{q=1}^{n_{te}}$ and $T_{te} = \{t_q\}_{q=1}^{n_{te}}$ both containing n_{te} testing instances. Finally, given a query of any modality, the goal of cross-modal retrieval is to calculate the cross-modal similarity $sim(i_a, t_b)$, and retrieve the relevant instances of another modality in the testing data by the ranking of calculated similarity. In the following subsections, we first introduce the proposed image and text semantic space measurement methods respectively, and then we demonstrate the proposed adaptive fusion approach on different semantic spaces.

B. Image Semantic Space Measurement

1) Recurrent Attention Network for Image: To exploit intrinsic modality-specific characteristics within image data, we design a recurrent network with attention mechanism, which is adopted on the top of CNN hidden layers, to fully model the fine-grained local and context information jointly.

First, each input image i_p is resized to 256×256 and fed into CNNs. Specifically, the network structure is configured the same as the layers before the last pooling layer (pool5) in 19-layer VGGNet [40]. The last pooling layer consists of 49 filters, and we can obtain separate feature vectors for different regions from the response of each filter over a 7×7 mapping of the image, for exploiting the fine-grained local information. Thus, each input image i_p can be represented as $\{v_1^i, ..., v_n^i\}$, where n denotes the total number of image regions and v_n^i is a 512-dimensional feature vector corresponding to the n-th region.

Then, we employ RNN to model the spatial context information among image regions. We compose these regions as a sequence, which can be regarded as the results of eye movement when we glance at the image [41]. Specifically, Long Short Term Memory (LSTM) network [42] is adopted, which is a special kind of RNN, with strong ability to learn long-term dependencies through the memory cell and update gates as well as preserve previous time-steps information at the same time. The architecture of an LSTM unit is shown in Figure 5. Formally, taking a sequence of image regions as input, the LSTM is updated recursively with the following equations:

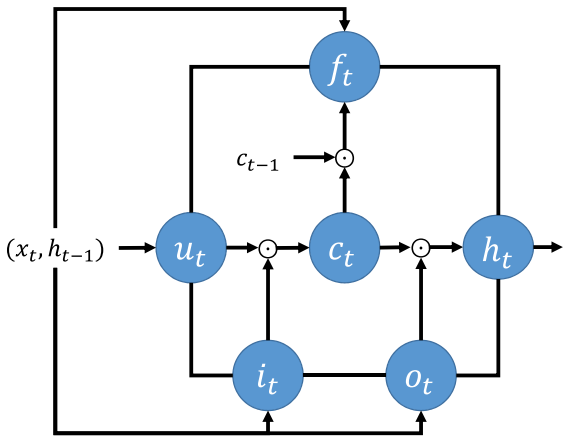


Fig. 5. The architecture of an LSTM unit, which can learn long-term dependencies and retain information of previous time-steps through the memory cell and the update gates.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (1)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (2)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (3)$$
$$u_t = tanh(W_u x_t + U_u h_{t-1} + b_u), \quad (4)$$
$$c_t = u_t \odot i_t + c_{t-1} \odot f_t, \quad (5)$$
$$h_t = o_t \odot tanh(c_t), \quad (6)$$

where the activation vectors of the input, forget, memory cell and output are denoted as i, f, c and o respectively. x is the input region feature and the hidden unit output is denoted as h. While W, U and b are the weight matrices and bias term that need to be trained. $\odot$ denotes the element-wise multiplication. σ is the sigmoid nonlinearity to activate the gate, which is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \quad (7)$$

After that, we apply attention mechanism to allow models focusing on the necessary fine-grained regions within an image. Given the obtained the output sequence $H_i = \{h_1^i, ..., h_n^i\}$ from LSTM, the attention weights a^i can be calculated by a feed-forward neural network with the softmax function as follows:

$$M^i = tanh(W_a^i H_i), \quad (8)$$
$$a^i = softmax(w_{ia}^T M_i), \quad (9)$$

where W_a^i and w_{ia} are the weight parameters for respective layers. a^i denotes the generated attention probabilities for image regions. Thus, the final image vector for the n-th region can be obtained as $a_n^i h_n^i$, which contains both fine-grained local information and spatial context information.

2) Visual Attention Based Joint Embedding: Cross-modal correlation learning is performed to project the text data into the image semantic space, which utilizes the learned visual attention weights from the above recurrent attention network.

We first need to generate the representation for text instance t_p. Each word is represented as a k-dimensional vector extracted by Word2Vec [43] model,[1] which is trained on billions of words in Google News. Thus, a sentence with n words can be represented by an $n \times k$ matrix. And a Word CNN is adopted on the input matrix, which has the same configuration with [44]. The text representation for each sentence is generated from the last fully-connected layer, denoted as q_p^t.

Due to the imbalanced relationship between different modalities, the fine-grained parts of image and text cannot be all exactly aligned with each other. Thus, we aim to measure such imbalanced relationship in each image/text pair on different semantic spaces respectively, and project the text data from its feature space into the constructed semantic space for image. We design the cross-modal similarity sim_i for image semantic space between image i_p and text t_p as follows:

$$sim_i(i_p, t_p) = \sum_{j=1}^{n} a_j^{i_p} h_j^{i_p} \cdot q_p^t, \tag{10}$$

where $h_j^{i_p}$ is the j-th region in the image i_p and $a_j^{i_p}$ is the attention weight for the corresponding image region. With the utilization of attention mechanism, those parts in the image i_p that are exactly demonstrated by the text t_p can be emphasized with larger attention weights, while other parts have relative smaller attention weights to indicate such imbalanced relationship. Besides, each text is projected into the semantic space for image through several fully-connected layers after the convolutional layers, which aims to ensure that the last fully-connected layer has the same dimension with image representation for cross-modal similarity measurement.

Finally, we design an attention based joint embedding loss to perform cross-modal correlation learning, which jointly considers both matched and mismatched image/text pairs with the defined cross-modal similarity based on the learned attention weights. The objective function is defined as follows:

$$L_i = \frac{1}{N} \sum_{n=1}^{N} l_{i1}(i_n^+, t_n^+, t_n^-) + l_{i2}(t_n^+, i_n^+, i_n^-), \tag{11}$$

and the two items in this formula are defined as:

$$\begin{aligned} &l_{i1}(i_n^+, t_n^+, t_n^-) \\ &\quad = max(0, \alpha - sim_i(i_n^+, t_n^+) + sim_i(i_n^+, t_n^-)), \end{aligned} \tag{12}$$

$$\begin{aligned} &l_{i2}(t_n^+, i_n^+, i_n^-) \\ &\quad = max(0, \alpha - sim_i(i_n^+, t_n^+) + sim_i(i_n^-, t_n^+)), \end{aligned} \tag{13}$$

where (i_n^+, t_n^+) denotes the matched image/text pair, while (i_n^+, t_n^-) and (i_n^-, t_n^+) are the mismatched pairs. The margin parameter is set to be α. N is the number of triplet tuples sampled from training set. To train the model, stochastic gradient descent (SGD) is adopted, and the attention weights for image are also updated during end-to-end training, which are not only computed from image, but also guided by the data of text with attention based joint embedding loss. It can capture the modality-specific characteristics within image, and distinguish which image parts have larger probabilities to be demonstrated in textual descriptions.

So far, we have obtained the modality-specific cross-modal similarity sim_i for image semantic space, where both representation learning and similarity measurement to benefit each other, and also fully captures the intrinsic fine-grained clues in image and imbalanced information across different modalities for correlation learning.

C. Text Semantic Space Measurement

1) Recurrent Attention Network for Text: For fully exploiting modality-specific characteristics within text data, we also design a recurrent network with attention mechanism, on the top of CNN hidden layers, which can model both the fine-grained local and context information in textual descriptions.

First, the input text instance t_p, which consists of n words, is represented by an $n \times k$ matrix, where each word is represented as a k-dimensional vector extracted by Word2Vec [43] model. Following [44], we design convolutional networks for text (Word CNN), which are built by several combinations of convolution layer, threshold activation function layer and pooling layer. The Word CNN is similar with the image CNN except the 2D convolution and spatial max-pooling of image CNN are replaced by temporal (1D) convolution and temporal max-pooling. Through the Word CNN network, CNN hidden activation of the last pooling layer is split to generate the features of text fragments.

Second, RNN is adopted on the top of CNN with the sequence of vectors to further model the context information along the input text sequence. Specifically, we also adopt LSTM network [42] to exploit such temporal dependency, which takes the sequence of text fragments as input. LSTM is updated following the equations (1) to (6), where x denotes the input feature of text fragment. Thus, we can obtain the output sequence from LSTM as $H_t = \{h_1^t, ..., h_m^t\}$.

Then, the attention mechanism is applied to capture useful fine-grained fragments in text sequence. The attention weights are denoted as a^t, which are calculated by a feed-forward network with softmax function as follows:

$$M^t = tanh(W_a^t H_t), \tag{14}$$

$$a^t = softmax(w_{ta}^T M_t), \tag{15}$$

where the weight parameters for respective layers are denoted as W_a^t and w_{ia}. a^t denotes the generated attention probabilities for text fragments. Thus, the final text vector for the m-th fragment is calculated as $a_m^t h_m^t$, which captures both fine-grained local and context information in textual description.

2) Textual Attention Based Joint Embedding: To perform cross-modal correlation learning, image data is projected into the text semantic space to utilize the learned textual attention weights from the above recurrent attention network for text. We still need to generate the image representation for each instance i_p, which is also extracted from the last fully-connected layer (fc7) in 19-layer VGGNet [40] with 4,096 dimensions, and denoted as q_p^i.

Then, the image data is projected into the constructed semantic space for text from their own feature space. Thus, we

[1] https://code.google.com/p/word2vec/

compute the cross-modal similarity sim_t for text semantic space between image i_p and text t_p as follows, which aims to explore the imbalanced and complementary relationship between different modalities.

$$sim_t(i_p, t_p) = \sum_{j=1}^{m} a_j^{t_p} h_j^{t_p} \cdot q_p^i, \tag{16}$$

where the j-th fragment of text t_p is denoted as $h_j^{t_p}$, $a_j^{t_p}$ is the attention weight for the corresponding text fragment. With the attention mechanism, we can not only capture the modality-specific characteristics within text, but also distinguish which text parts have larger probabilities to be demonstrated in image for measuring the imbalanced relationship, instead of exacting alignment between the parts of image and text.

Finally, an attention based joint embedding loss is designed similarly for cross-modal correlation learning in the text semantic space, which considers that the difference between the similarities of matched pair and mismatched pair should be as large as possible. Thus, the objective function is defined as follows:

$$L_t = \frac{1}{M} \sum_{n=1}^{M} l_{t1}(t_n^+, i_n^+, i_n^-) + l_{t2}(i_n^+, t_n^+, t_n^-), \tag{17}$$

where $l_{t1}(t_n^+, i_n^+, i_n^-)$ and $l_{t2}(i_n^+, t_n^+, t_n^-)$ are defined similarly as equations (12) and (13) with the cross-modal similarity sim_t as follows:

$$\begin{aligned} &l_{t1}(t_n^+, i_n^+, i_n^-) \\ &\quad = max(0, \beta - sim_t(i_n^+, t_n^+) + sim_t(i_n^-, t_n^+)), \end{aligned} \tag{18}$$

$$\begin{aligned} &l_{t2}(i_n^+, t_n^+, t_n^-) \\ &\quad = max(0, \beta - sim_t(i_n^+, t_n^+) + sim_t(i_n^+, t_n^-)). \end{aligned} \tag{19}$$

Also, the number of triplet tuples sampled from training set is denoted as M, β is the margin parameter. Therefore, the modality-specific cross-modal similarity sim_t for text semantic space can be obtained, with the fully modeling of fine-grained clues in text as well as imbalanced information across different modalities for correlation learning.

D. Adaptive Fusion on Different Semantic Spaces

So far, we have obtained two kinds of modality-specific cross-modal similarities, namely sim_i and sim_t from the semantic spaces for image and text. Inspired by [45], we further attempt to explore the complementarity between different semantic spaces by adaptive fusion.

First, the cross-modal similarity scores obtained from different semantic spaces are min-max normalized to $[0, 1]$ respectively, which aims to overcome the influence of image/text pairs with too large similarity scores, and are defined as follows:

$$r_i(i_p, t_p) = \frac{sim_i(i_p, t_p) - min(sim_i(i, t))}{max(sim_i(i, t)) - min(sim_i(i, t))}, \tag{20}$$

$$r_t(i_p, t_p) = \frac{sim_t(i_p, t_p) - min(sim_t(i, t))}{max(sim_t(i, t)) - min(sim_t(i, t))}. \tag{21}$$

Then, the normalized scores obtained from one semantic space are used as the adaptive weights of the corresponding image/text pair in another semantic space during fusion stage, with the motivation that larger similarity in one semantic space leads to higher importance of the corresponding pair in another semantic space. Finally, the two kinds of modality-specific cross-modal similarities are fused with the following equation:

$$\begin{aligned} &Sim(i_p, t_p) = \\ &\quad r_t(i_p, t_p) \cdot sim_i(i_p, t_p) + r_i(i_p, t_p) \cdot sim_t(i_p, t_p). \end{aligned} \tag{22}$$

Thus, we can obtain the final cross-modal similarity score $Sim(i_p, t_p)$ between image i_p and text t_p, which can fully explore the complementarity between different semantic spaces to boost the performance of cross-modal retrieval.

IV. Experiments

In this section, we conduct experiments on 4 cross-modal datasets, taking 9 state-of-the-art methods for comparison to verify the effectiveness of our proposed approach. Besides, comprehensive experimental analyses are presented including convergence and parameter analyses, as well as baseline experiments to verify the contribution of each component in our approach.

A. Datasets

Here we briefly introduce 4 cross-modal datasets adopted in the experiments, including Wikipedia, Pascal Sentence, MS-COCO and our constructed large-scale XMediaNet datasets. Each dataset is divided into 3 subsets, namely training set, testing set and validation set.

- **Wikipedia dataset** [4], as the most widely-used dataset for cross-modal retrieval, is selected from "featured articles" in Wikipedia[2] with 10 most populated categories, including history, biology and so on. This dataset totally consists of 2,866 image/text pairs. For fair and objective comparison purpose, we exactly follow the dataset partition strategy of [13] and [14] to divide the dataset into 3 subsets: 2,173 pairs in training set, 231 pairs in validation set and 462 pairs in testing set.
- **Pascal Sentence dataset** [46] contains 1,000 images, which is generated from 2008 PASCAL development kit. Each image is annotated via Amazon Mechanical Turk by crowdsourcing to generate 5 independent sentences from different annotators, which form one document. This dataset is categorized into 20 categories, and also following [13] and [14], 800 documents are selected as training set, while 100 documents for testing and 100 documents for validation.
- **MS-COCO dataset** [47] contains 123,287 images and their annotated sentences. Each image is annotated by 5 independent sentences, which are generated by crowdsourcing via Amazon Mechanical Turk with 5 users. Following [48], there are both 5,000 pairs split randomly for testing and validation, while the rest are training set.
- **XMediaNet dataset** is our self-constructed large-scale cross-modal dataset, which consists of 5 modalities, namely text, image, video, audio and 3D model. We select

[2]https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

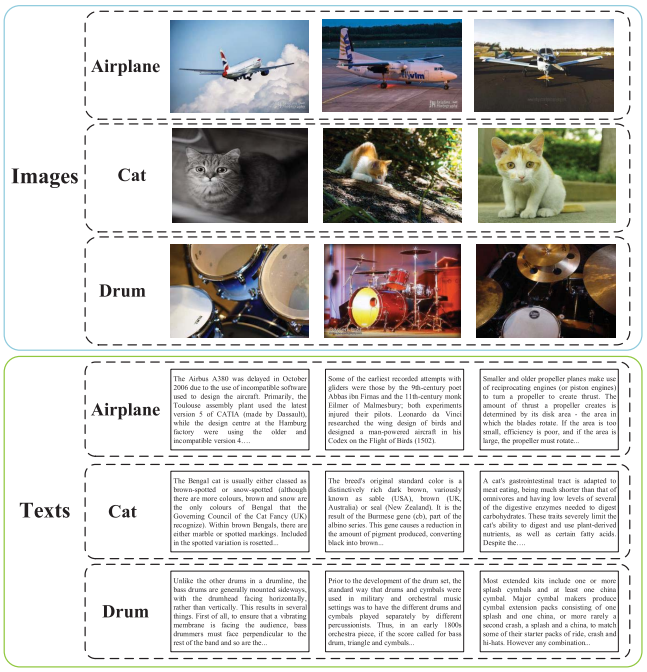


Fig. 6. Image and text examples of 3 categories from XMediaNet dataset, which are airplane, cat and drum.

200 category nodes from WordNet[3] to construct this dataset to ensure the semantic hierarchy structure. These categories can be divided into two main kinds: animals and artifacts. There are 47 species of animal such as elephant, owl, bee and frog as well as 153 types of artifact such as violin, airplane, shotgun, and camera. The total number of instances exceeds 100,000. It is noted that we also use image and text in the experiments, where text paragraphs are extracted from Wikipedia articles whose topics belong to the category, and images including the objects of the category are collected from Flickr.[4] This dataset totally consists of 40,000 image/pairs, and is also divided into 3 subsets, with 32,000 pairs in training set, 4,000 pairs in testing set and 4,000 pairs in validation set. Some examples are shown in Figure 6.

B. Details of the Network

We implement the proposed network by Torch,[5] which is a widely-used scientific computing framework. We introduce the details of the network including data preprocess strategy and network structure as following.

1) Data Preprocess: For image, we resize the original images to 256×256 as inputs. For text, we convert each word into a 300-dimension vector by Word2Vec [43] and generate vector sequences as text inputs. The maximum input length is set as the maximum sequence length in the dataset, and we adopt zero-padding for other sequences beneath this limit.

2) Recurrent Attention Network: The recurrent attention network of each semantic space mainly consists of three parts, namely convolutional network, LSTM network and attention network. For the semantic space for text, the convolutional network consists of three learnable temporal convolution layers, each of which is followed by a ReLU activation function layer and a temporal max-pooling layer. Taking Wikipedia dataset as an example, the first temporal convolution layer has 384 kernels whose widths are 15. The parameter combinations of the remaining convolution layers are (512, 9) and (256, 7). The first parameter of each combination means the number of convolution kernels while the second is the kernel width. The kernel step sizes of all convolution layers are 1. For the other two datasets, the number of kernels on each convolution layer is the same with Wikipedia dataset, but the length of text instances differs greatly between different datasets, thus the kernel widths change according to the lengths of the input sequences. For the semantic space for image, we use the pretrained convolution network in 19-layer VGGNet and treat the output of the last pooling layer as the input of LSTM. The LSTM network has two units in series, and the dimension of its output keeps the same with the input. The LSTM is followed by a fully-connected layer which aims to project the output of LSTM into the target dimension, which is 4,096 dimensions in our case. The attention network is made up of a fully-connected layer and a softmax layer.

Besides, to address the overfitting issue, we have adopted several strategies in training process as follows. First, dropout layers are inserted following the output layer of hidden unit in LSTM network and after the attention layer, which can effectively prevent the overfitting problem as indicated in [49]. We set a default dropout rate of 0.5 for all dropout layers in the experiments, which is a general setting as in most prevalent networks like VGGNet [40]. Second, batch normalization is adopted in the convolutional network to normalize the layer inputs for each training mini-batch, which can regularize the model to avoid overfitting as indicated in [50]. Third, all training data is shuffled, and image data is also augmented with mirrored versions to expand training set following [15].

C. Compared Methods

Totally 9 state-of-the-art methods are compared in the experiments to verify the effectiveness of our proposed approach. There are 5 traditional cross-modal retrieval methods, namely CCA [18], CFA [21], KCCA [51], JRL [9] and LGCFL [52]. While the other 4 methods, Corr-AE [14], DCCA [15], CMDN [13] and Deep-SM [31] are DNN-based methods. Their introductions are presented as follows:

- **CCA** [18] learns project matrices to maximize the correlation between the projected features of different modalities in one common space.
- **CFA** [21] minimizes the Frobenius norm between the data of different modalities after projecting them into one common space.
- **KCCA** [51] uses kernel function to project the features into a higher-dimensional space, and then learns a common space by CCA. In the experiments, we adopt Gaussian kernel as the kernel function.

[3]http://wordnet.princeton.edu/
[4]http://www.flickr.com
[5]http://torch.ch/

- **JRL** [9] learns a common space by using semantic information, with semi-supervised regularization and sparse regularization.
- **LGCFL** [52] jointly learns basis matrices of different modalities, by using a local group based priori in the formulation to fully take advantage of popular block based features.
- **Corr-AE** [14] consists of two autoencoder networks coupled at the code layer to simultaneously model the reconstruction error and correlation loss. It should be noted that Corr-AE has two extensions, namely Corr-Cross-AE and Cross-Full-AE, and in the experiments we compare with the best results of the three models.
- **DCCA** [15] is a nonlinear extension of CCA. The correlation is maximized between the output layers of two separate subnetworks.
- **CMDN** [13] adopts multiple deep networks to generate separate representations and learns the common representation with a stacked network.
- **Deep-SM** [31] adopts convolutional neural network to perform deep semantic matching, which fully exploits the strong power of CNN image feature.

D. Evaluation Metric

We perform two kinds of retrieval tasks on Wikipedia, Pascal Sentence and XMediaNet datasets, which are defined as follows.

- **Image query text** (image→text). Retrieving relevant text instances in the testing set ranked by calculated cross-modal similarity, using a query of image.
- **Text query image** (text→image). Retrieving relevant image instances in the testing set ranked by calculated cross-modal similarity, using a query of text.

As for MS-COCO dataset, two retrieval tasks are conducted following [15]:

- **Image annotation**. Retrieving the groundtruth sentences by a query image.
- **Image retrieval**. Retrieving the groundtruth images by a query text.

It should be noted that our proposed MCSM approach integrates both representation learning and distance metric learning, which takes original image and text as inputs to directly generate the cross-modal similarity score. While other compared methods only learn the common representation taking hand-crafted features as input. Thus, for fair comparison, all compared methods also adopt the same CNN features used in our proposed approach as input on Wikipedia, Pascal Sentence and XMediaNet datasets. Specifically, the CNN feature of image is extracted from fc7 layer in 19-layer VGGNet [40], while the CNN feature of text is extracted by Word CNN with the same configuration of [44]. While for MS-COCO dataset, we exactly follow [53] to extract features as the inputs of all compared methods. We directly adopt the source codes provided by the authors of the compared methods, to fairly evaluate them by the following steps in the experiments.

1) Perform common representation learning using the data in training set to obtain the learned projections or deep models.

TABLE I

THE MAP SCORES OF CROSS-MODAL RETRIEVAL FOR OUR MCSM APPROACH AND 9 COMPARED METHODS ON **WIKIPEDIA** DATASET

Method	MAP scores		
	Image→Text	Text→Image	Average
Our MCSM Approach	**0.516**	**0.458**	**0.487**
CMDN [13]	0.487	0.427	0.457
Deep-SM [31]	0.478	0.422	0.450
LGCFL [52]	0.466	0.431	0.449
JRL [9]	0.479	0.428	0.454
DCCA [15]	0.445	0.399	0.422
Corr-AE [14]	0.442	0.429	0.436
KCCA [51]	0.438	0.389	0.414
CFA [21]	0.319	0.316	0.318
CCA [18]	0.298	0.273	0.286

2) Use the learned projections or deep models to convert the data in testing set into the common representation.

3) Calculate cross-modal similarity between image and text by cosine distance, and then perform cross-modal retrieval.

Mean Average Precision (MAP) score is adopted as the evaluation metric on Wikipedia, Pascal Sentence and our constructed XMediaNet datasets, which is the mean value of Average Precision (AP) of each query. AP is defined as follows:

$$AP = \frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k} \times rel_k, \tag{23}$$

where the testing set contains n instances, which has R relevant instances. R_k is the number of relevant instances in the top k returned results. rel_k is set to be 1 when the k-th returned result is relevant, otherwise, rel_k is set to be 0. MAP score considers the ranking of returned retrieval results as well as precision simultaneously, which is extensively adopted in cross-modal retrieval tasks, such as [4] and [13]. Besides, we further provide the precision-recall curves on Wikipedia, Pascal Sentence and XMediaNet datasets.

It is noted that MS-COCO dataset has no label annotations as other 3 datasets, thus some compared methods cannot be conducted on it, including CMDN [13], Deep-SM [31], LGCFL [52] and JRL [9], which need label annotations for training. Due to the absence of label annotations, MS-COCO dataset takes different evaluation metric. We report the scores of Recall@K following [15], which include recall rate at top 1 result (R@1), top 5 results (R@5) and top 10 results (R@10).

E. Comparisons With State-of-the-Art Methods

The retrieval results on 4 datasets are shown in Tables I to IV, we can see that our proposed approach achieves the best retrieval accuracy compared with 9 state-of-the-art methods. On one hand, among the compared traditional methods, LGCFL achieves the best retrieval accuracy, which is close to CMDN based on DNN. On the other hand, the accuracies of 4 DNN-based methods differ greatly, while some of them are outperformed by the traditional methods, such as the average MAP scores of DCCA and Corr-AE are lower than LGCFL and JRL.

TABLE II

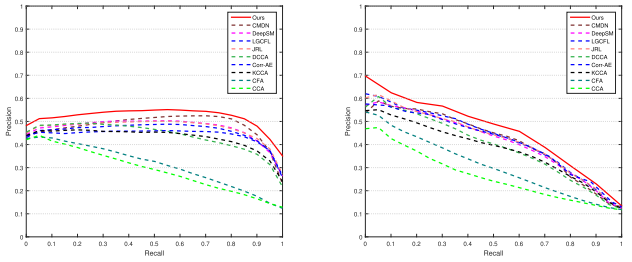
THE MAP SCORES OF CROSS-MODAL RETRIEVAL FOR OUR MCSM APPROACH AND 9 COMPARED METHODS ON **PASCAL SENTENCE** DATASET

Method	MAP scores		
	Image→Text	Text→Image	Average
Our MCSM Approach	**0.598**	**0.598**	**0.598**
CMDN [13]	0.544	0.526	0.535
Deep-SM [31]	0.560	0.539	0.550
LGCFL [52]	0.539	0.503	0.521
JRL [9]	0.563	0.505	0.534
DCCA [15]	0.568	0.509	0.539
Corr-AE [14]	0.532	0.521	0.527
KCCA [51]	0.488	0.446	0.467
CFA [21]	0.476	0.470	0.473
CCA [18]	0.203	0.208	0.206

TABLE III

THE MAP SCORES OF CROSS-MODAL RETRIEVAL FOR OUR MCSM APPROACH AND 9 COMPARED METHODS ON **XMEDIANET** DATASET

Method	MAP scores		
	Image→Text	Text→Image	Average
Our MCSM Approach	**0.540**	**0.550**	**0.545**
CMDN [13]	0.485	0.516	0.501
Deep-SM [31]	0.399	0.342	0.371
LGCFL [52]	0.441	0.509	0.475
JRL [9]	0.488	0.405	0.447
DCCA [15]	0.425	0.433	0.429
Corr-AE [14]	0.469	0.507	0.488
KCCA [51]	0.252	0.270	0.261
CFA [21]	0.252	0.400	0.326
CCA [18]	0.212	0.217	0.215

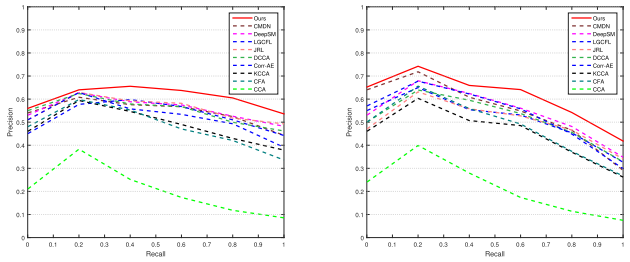


(a) (b)

Fig. 7. Precision-recall curves of cross-modal retrieval on Wikipedia dataset. (a) Image→Text Retrieval. (b) Text→Image Retrieval.

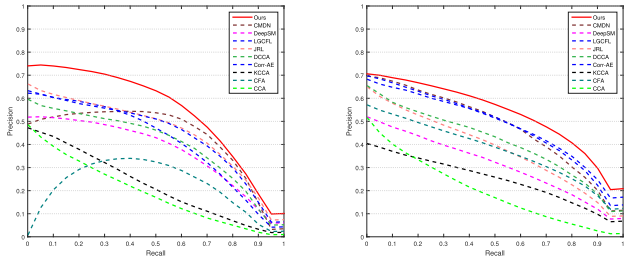
Figures 7 to 9 show the precision-recall curves of two retrieval tasks on Wikipedia, Pascal Sentence and XMediaNet datasets. We can observe that our proposed MSCM approach keeps clear advantages compared with the state-of-the-art methods to verify its effectiveness. Some cross-modal retrieval results on XMediaNet dataset of our proposed MCSM approach and the best DNN-based compared method CMDN are shown in Figure 10.

Next we give the in-depth analyses on the retrieval results of both our proposed MCSM approach and all compared methods. As shown in Tables I to IV, our proposed MCSM approach shows advantage compared with 9 state-of-the-art methods on all 4 datasets. We can also observe that the accuracies of DNN-based methods fail to widen a clear gap with traditional methods. Among the traditional methods,



(a) (b)

Fig. 8. Precision-recall curves of cross-modal retrieval on Pascal Sentence dataset. (a) Image→Text Retrieval. (b) Text→Image Retrieval.



(a) (b)

Fig. 9. Precision-recall curves of cross-modal retrieval on XMediaNet dataset. (a) Image→Text Retrieval. (b) Text→Image Retrieval.

the classical baseline CCA has the worst accuracy for it only models some statistical values, while KCCA extends CCA to achieve better accuracy by the adoption of kernel function with the better ability of modeling nonlinear correlation. CFA has similar results with KCCA, which minimizes the Frobenius norm in the learned common space. JRL and LGCFL are the best two methods among them, while the former utilizes the semi-supervised and sparse regularization, and the latter learns the basis matrices with a local group based priori.

As for DNN-based methods, DCCA and Corr-AE have close accuracies. DCCA maximizes correlation on the outputs of two separate networks to extend CCA, and Corr-AE not only considers the correlation error but also minimizes the reconstruction error. Deep-SM outperforms DCCA and Corr-AE by fully exploring the strong learning power of convolutional neural network with semantic category information. CMDN achieves better accuracy than the above methods, because it takes intra-modality and inter-modality correlation into consideration during both separate representation learning and common representation learning stages.

Compared with all state-of-the-art methods, our proposed MCSM approach achieves the best accuracy for the fact of following 3 aspects: (1) Independent semantic spaces for different modalities with recurrent attention network to fully exploit the modality-specific fine-grained context information. (2) Attention based joint embedding loss to utilize the imbalanced and complementary relationship between different modalities. (3) Adaptive fusion to explore the complementarity between different semantic spaces for cross-modal retrieval.

Furthermore, our proposed MSCM approach can effectively deal with outlier images by modality-specific cross-modal

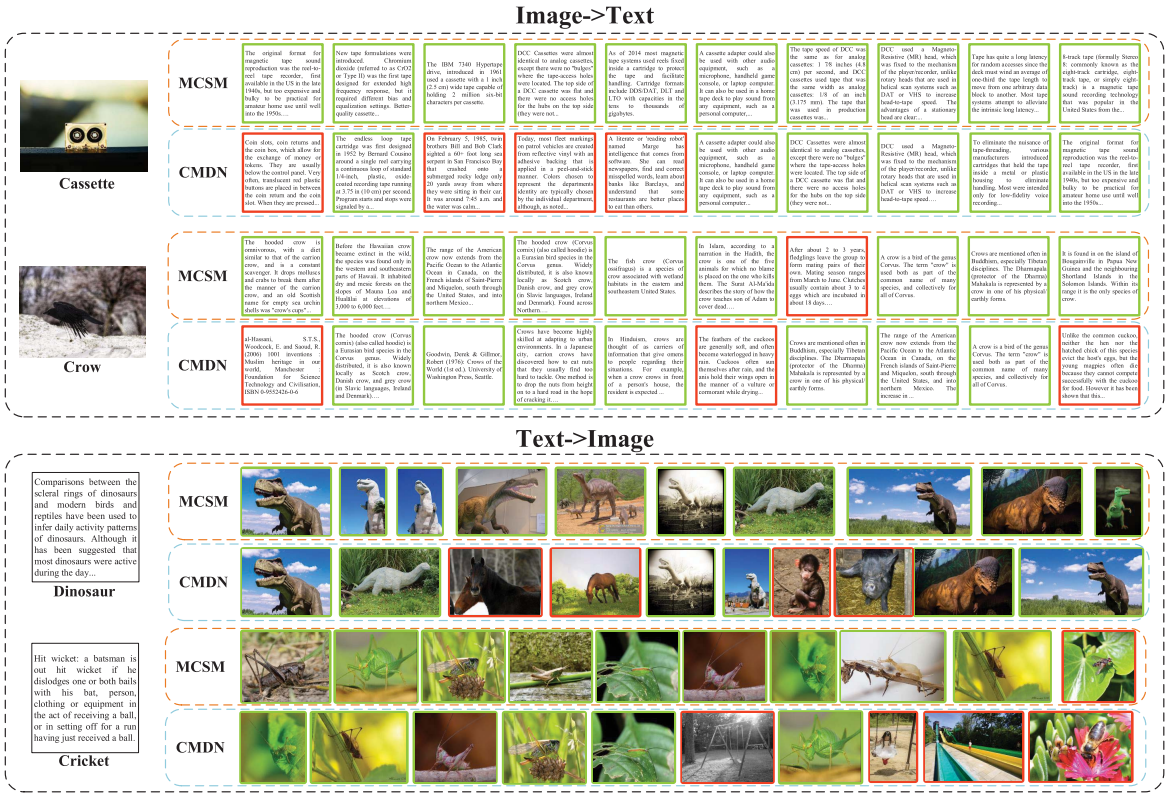


Fig. 10. Examples of the cross-modal retrieval results on XMediaNet dataset by our proposed MCSM approach as well as compared method CMDN [13]. In these examples, the correct retrieval results are with green borders, while the wrong results are with red borders.

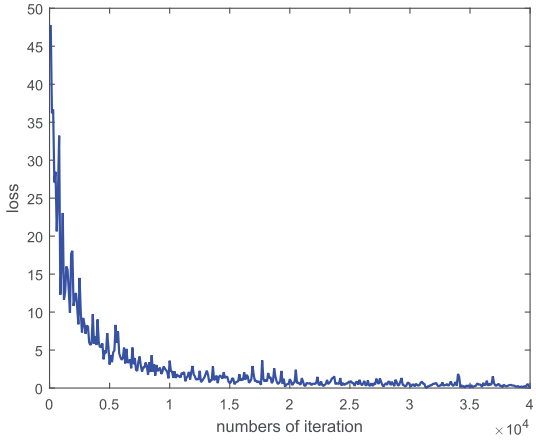


Fig. 11. Convergence experiments of our proposed MCSM approach conducted on the large-scale XMediaNet dataset, which show the curves of downtrend on the loss value.

similarity measurement. On one hand, the outlier images may still contain useful information that can be captured from the semantic space of image by fully fine-grained modeling with attention mechanism, because attention mechanism can allow the model to focus on the relevant fine-grained parts in the outlier image selectively, while ignore those irrelevant parts. On the other hand, the textual descriptions can also provide rich hints from the semantic space of text, which are helpful for understanding the latent semantic information in outlier images. Furthermore, the complementarity between two semantic spaces of image and text can be exploited to reduce the impact of outlier images. Besides, we have given some failure case analyses. As shown in Figure 10, our proposed MSCM approach can effectively reduce the failure cases compared with other methods. While those failures caused by the outlier images are mainly due to the confusion among image instances, where small variance between image instances of different categories leads to wrong retrieval results.

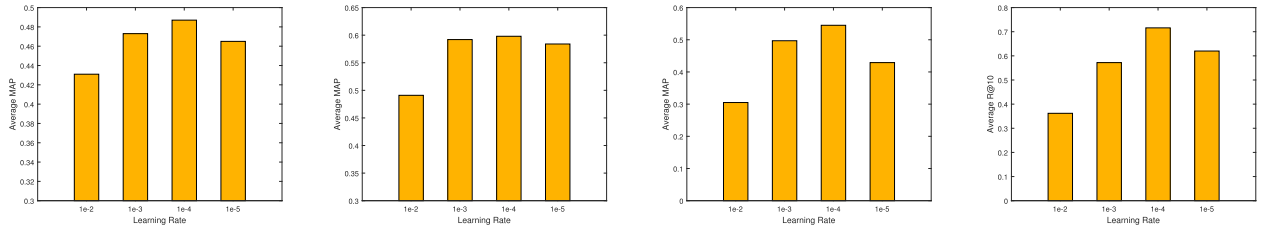
F. Convergence and Parameter Analyses

First, we conduct convergence experiments on XMediaNet dataset. The curve of downtrend on the loss value is shown in Figure 11. We can observe that our proposed approach converges within 15K iterations on the XMediaNet dataset, which shows its efficiency in training stage. Then, we also conduct parameter experiments on the effect of key parameters, including learning rate and margin parameters α and β in equations (12), (13), (18) and (19), which are implemented on all the 4 datasets. For the learning rate, we range the value from 1e-2 to 1e-5, and the results are shown in Figure 12, from which we can see that our proposed approach achieves the best accuracy at the learning rate of 1e-4 on all the 4 datasets, and the accuracy becomes lower at higher learning rates. Then, for the margin parameter, it should be noted that we set the margin parameters α and β with the same value in all loss functions during each experiment. The value is ranged from 0.1 to 0.9. The results are shown in Figure 13, and we can see that the accuracies are not sensitive to the margin parameters.

TABLE IV

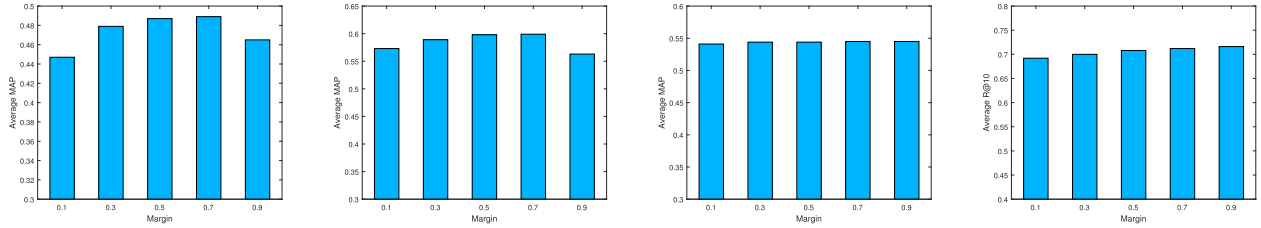
THE RECALL SCORES OF TWO CROSS-MODAL RETRIEVAL TASKS ON MS-COCO DATASET

Method	Image annotation			Image retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Our MSCM Approach	**0.282**	**0.596**	**0.724**	**0.280**	**0.570**	**0.708**
DCCA [15]	0.069	0.211	0.318	0.066	0.209	0.322
Corr-AE [14]	0.154	0.397	0.532	0.138	0.353	0.478
KCCA [51]	0.072	0.202	0.305	0.020	0.074	0.122
CFA [21]	0.086	0.258	0.371	0.150	0.381	0.514
CCA [18]	0.041	0.142	0.226	0.041	0.155	0.251



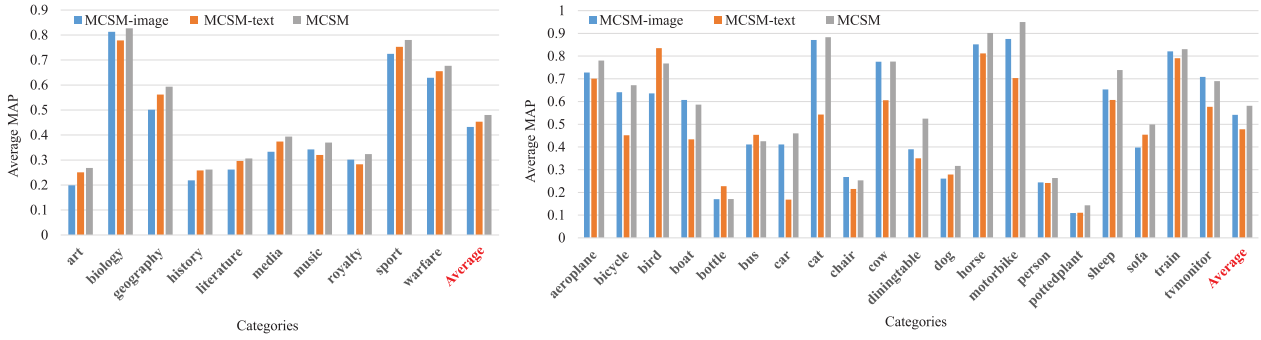
(a) (b) (c) (d)

Fig. 12. Experiments about the impact of learning rate, on Wikipedia, Pascal Sentence, XMediaNet and MS-COCO datasets. It should be noted that we report the average result of two cross-modal retrieval tasks. (a) Wikipedia dataset. (b) Pascal Sentence dataset. (c) XMediaNet dataset. (d) MS-COCO dataset.



(a) (b) (c) (d)

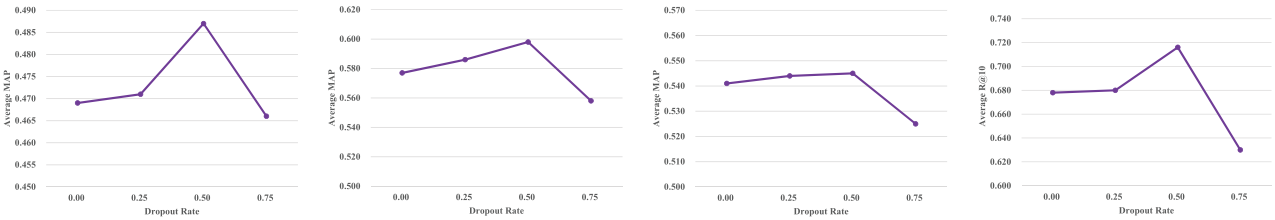
Fig. 13. Experiments about the impact of margin parameter, on Wikipedia, Pascal Sentence, XMediaNet and MS-COCO datasets. It should be noted that we report the average result of two cross-modal retrieval tasks. (a) Wikipedia dataset. (b) Pascal Sentence dataset. (c) XMediaNet dataset. (d) MS-COCO dataset.



(a) (b)

Fig. 14. The respective MAP score of each category on our proposed approach, in Wikipedia dataset and Pascal Sentence dataset. We can see that MCSM-text performs better than MCSM-image in most categories of Wikipedia dataset because of the high-level semantic information in textual description. While MCSM-image outperforms MCSM-text in Pascal Sentence dataset, because of more useful information in image than only 5 sentences annotated on it. Besides, MCSM achieves the best accuracy in final average results, which indicates the complementarity between two semantic spaces. (a) Wikipedia dataset. (b) Pascal Sentence dataset.

We also conduct a parameter experiment to evaluate how dropout rate affects the performance, where the dropout rate is set to be four different values, namely 0, 0.25, 0.5 and 0.75. The results are shown in Figure 15. Our proposed MSCM approach achieves the best accuracy at the dropout rate of 0.5, which verifies the effectiveness of dropout. Besides, we can also observe that the performance decreases quickly at a higher dropout rate of 0.75 on all 4 datasets. It indicates



(a) (b) (c) (d)

Fig. 15. Experiments about the impact of dropout rate, on Wikipedia, Pascal Sentence, XMediaNet and MS-COCO datasets. It should be noted that we report the average result of two cross-modal retrieval tasks. (a) Wikipedia dataset. (b) Pascal Sentence dataset. (c) XMediaNet dataset. (d) MS-COCO dataset.

TABLE V

BASELINE EXPERIMENTS ON **PERFORMANCE OF EACH SEMANTIC SPACE** IN WIKIPEDIA, PASCAL SENTENCE AND XMEDIANET DATASETS, WHERE **MCSM-IMAGE** MEANS TO RETRIEVE ONLY WITH THE CROSS-MODAL SIMILARITY GENERATED FROM IMAGE SEMANTIC SPACE, WHILE **MCSM-TEXT** MEANS TO ONLY USE THE CROSS-MODAL SIMILARITY FROM TEXT SEMANTIC SPACE TO RETRIEVE

Dataset	Method	MAP scores		
		Image→Text	Text→Image	Average
Wikipedia dataset	**Our MCSM Approach**	**0.516**	**0.458**	**0.487**
	MCSM-image	0.448	0.423	0.436
	MCSM-text	0.498	0.438	0.468
Pascal Sentence dataset	**Our MCSM Approach**	**0.598**	**0.598**	**0.598**
	MCSM-image	0.559	0.541	0.550
	MCSM-text	0.500	0.478	0.489
XMediaNet dataset	**Our MCSM Approach**	**0.540**	**0.550**	**0.545**
	MCSM-image	0.453	0.455	0.454
	MCSM-text	0.447	0.417	0.432

TABLE VI

BASELINE EXPERIMENTS ON **PERFORMANCE OF EACH SEMANTIC SPACE** IN MS-COCO DATASET

Method	Image annotation			Image retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Our MSCM Approach	**0.282**	**0.596**	**0.724**	**0.280**	**0.570**	**0.708**
MCSM-image	0.183	0.424	0.569	0.168	0.400	0.543
MCSM-text	0.203	0.491	0.634	0.158	0.420	0.565

the fact that a high dropout rate is easier to cause inadequate training. Therefore, an appropriate dropout rate is necessary to prevent overfitting and improve the retrieval performance.

G. Baseline Comparisons

In this part, we conduct two baseline experiments, to verify the separate contribution of each component in our proposed MCSM approach. Tables V to VIII show the results of our baseline approaches on the following two aspects.

1) Performance of Each Semantic Space: As shown in Tables V and VI, MCSM-image means to perform cross-modal retrieval only by the cross-modal similarity sim_i in equation (10) from image semantic space, while MCSM-text means to use the cross-modal similarity sim_t only in equation (16) from text semantic space. We can observe that MCSM-text has better accuracy than MCSM-image on Wikipedia dataset, while on Pascal Sentence and XMediaNet datasets, MCSM-image outperforms MCSM-text in the average MAP score.

This is because of the imbalanced and complementary relationship between different modalities that contain unequal amount of information. The respective result of each category is reported in Figure 14, taking Wikipedia and Pascal Sentence datasets as examples. Specifically, in Wikipedia dataset, the categories mostly lie in high-level semantics, such as history or literature, where the textual description contains more background information that cannot be presented by its corresponding image. As for the Pascal Sentence and XMediaNet datasets, their categories mostly are specific objects, including animals such as elephant, or artifacts such as airplane. Their corresponding textual descriptions are relatively simple, such as only 5 sentences to describe each image in Pascal Sentence dataset. Under this situation, image contains more useful information than its corresponding text, which leads to higher accuracy of MCSM-image. Besides, MCSM stably outperforms MCSM-image and MCSM-text, which indicates the considerable complementarity between two semantic spaces.

2) Performance of Adaptive Fusion on Different Semantic Spaces: We also present baseline experiments to verify the effectiveness of adaptive fusion on different semantic spaces. We compare our proposed adaptive fusion strategy with late fusion (MCSM-LF), which means to directly average the two kinds of cross-modal similarities generated from different semantic spaces by the following equation:

$$sim_{lf}(i_p, t_p) = \frac{1}{2}(sim_i(i_p, t_p) + sim_t(i_p, t_p)). \quad (24)$$

TABLE VII

BASELINE EXPERIMENTS ON PERFORMANCE OF **ADAPTIVE FUSION** ON DIFFERENT SEMANTIC SPACES IN WIKIPEDIA, PASCAL SENTENCE AND XMEDIANET DATASETS, WHERE **MCSM-LF** MEANS TO ADOPT THE CROSS-MODAL SIMILARITY CALCULATED BY LATE FUSION

Dataset	Method	MAP scores		
		Image→Text	Text→Image	Average
Wikipedia dataset	**Our MCSM Approach**	**0.516**	**0.458**	**0.487**
	MCSM-LF	0.496	0.459	0.478
Pascal Sentence dataset	**Our MCSM Approach**	**0.598**	**0.598**	**0.598**
	MCSM-LF	0.595	0.578	0.587
XMediaNet dataset	**Our MCSM Approach**	**0.540**	**0.550**	**0.545**
	MCSM-LF	0.531	0.518	0.525

TABLE VIII

BASELINE EXPERIMENTS ON PERFORMANCE OF **ADAPTIVE FUSION** ON DIFFERENT SEMANTIC SPACES IN MS-COCO DATASET

Method	Image annotation			Image retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Our MSCM Approach	**0.282**	**0.596**	**0.724**	**0.280**	**0.570**	**0.708**
MCSM-LF	0.266	0.568	0.703	0.214	0.501	0.643

The results on 4 datasets are shown in Tables VII and VIII, where MCSM-LF means the accuracy calculated by late fusion. We can observe that adaptive fusion can further exploit the rich complementary information between the two semantic spaces to boost the accuracy of cross-modal retrieval.

From the above baseline results, the separate contribution of each component in our proposed MCSM approach is verified. First, the imbalanced and complementary relationship between different modalities are fully exploited in different semantic spaces. Second, the complementarity between different semantic spaces is fully captured by the adaptive fusion.

V. CONCLUSION

In this paper, we have proposed a modality-specific cross-modal similarity measurement approach to construct independent semantic spaces for different modalities. First, recurrent attention network with attention based joint embedding loss is adopted to fully model the modality-specific characteristics within each modality, and utilize the imbalanced and complementary relationship between different modalities during correlation learning. Second, end-to-end frameworks are implemented in different semantic spaces to directly generate the cross-modality similarity, integrating both common representation learning and distance metric learning to benefit each other. Third, adaptive fusion is adopted to explore the complementarity between different semantic spaces. Experiments on 4 cross-modal datasets, including widely-used Wikipedia, Pascal Sentence and MS-COCO datasets as well as our constructed large-scale XMediaNet dataset, verify the effectiveness of our proposed approach compared with 9 state-of-the-art methods.

As for the future work, we attempt to extend the current framework to other modalities such as video, audio and so on, for exploring the imbalanced and complementary relationship across the data of more modalities. Besides, we attempt to transfer knowledge from external knowledge base to further boost the performance of cross-modal retrieval.

REFERENCES

[1] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[2] Y. Peng and C.-W. Ngo, "Clip-based similarity measure for query-dependent clip retrieval and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 612–627, May 2006.

[3] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[4] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[5] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[6] D. Wang, X. Gao, X. Wang, L. He, and B. Yuan, "Multimodal discriminative binary embedding for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4540–4554, Oct. 2016.

[7] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin, "Cross-modal subspace learning via pairwise constraints," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5543–5556, Dec. 2015.

[8] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[9] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.

[10] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[11] F. Wu *et al.*, "Cross-modal learning to rank via latent joint representation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1497–1509, May 2015.

[12] F. Wu *et al.*, "Learning of multimodal representations with random walks on the click graph," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 630–642, Feb. 2016.

[13] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3846–3853.

[14] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[15] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3441–3450.

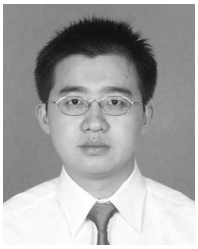
[16] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[17] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Conf. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[18] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[19] J. C. Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[20] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4094–4102.
[21] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 604–611.
[22] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Proc. Int. Conf. Comput. Learn. Theory*, 2004, pp. 624–638.
[23] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1198–1204.
[24] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[26] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 791–800.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.
[29] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2012, pp. 1–8.
[30] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[31] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
[32] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.
[33] Y. Peng, J. Qi, and Y. Yuan. (2017). "CM-GANs: Cross-modal generative adversarial networks for common representation learning." [Online]. Available: https://arxiv.org/abs/1710.05106
[34] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
[35] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–10.
[36] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.
[37] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, and P. Blunsom, "Reasoning about entailment with neural attention," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–9.
[38] K. M. Hermann *et al.*, "Teaching machines to read and comprehend," in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.
[39] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1–59.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.
[41] Y. Lin, Z. Pang, D. Wang, and Y. Zhuang. (2017). "Task-driven visual saliency and attention-based visual question answering." [Online]. Available: https://arxiv.org/abs/1702.06700
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[43] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[44] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1–6.
[45] G. Kong, L. Dong, W. Dong, L. Zheng, and Q. Tian. (2016). "Coarse2Fine: Two-layer fusion for image retrieval." [Online]. Available: https://arxiv.org/abs/1607.00719
[46] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, 2010, pp. 139–147.
[47] T.-Y. Lin *et al.*, "ar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[48] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4437–4446.
[49] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. (2012). "Improving neural networks by preventing co-adaptation of feature detectors." [Online]. Available: https://arxiv.org/abs/1207.0580
[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–9.
[51] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[52] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.
[53] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.



Yuxin Peng received the Ph.D. degree in computer application from Peking University in 2003. He was an Assistant Professor with the Institute of Computer Science and Technology (ICST), Peking University, Beijing, China. He was promoted to an Associate Professor and a Professor at Peking University in 2005 and 2010, respectively, where he is currently a Professor with ICST and the Chief Scientist with the National Hi-Tech Research and Development Program of China (863 Program). In 2006, he was authorized by the Program for New Star in Science and Technology of Beijing and the Program for New Century Excellent Talents in University. He has authored over 100 papers in refereed international journals and conference proceedings, including IJCV, TIP, TMM, TCSVT, PR, ACM MM, ICCV, CVPR, IJCAI, and AAAI. He has submitted 35 patent applications and received 16 of them. His current research interests mainly include cross-media analysis and reasoning, image and video analysis and retrieval, and computer vision. He led his team to participate in the TREC Video Retrieval Evaluation (TRECVID) many times. In 2009, at the TRECVID, his team received four first places on four sub-tasks of the High-Level Feature Extraction task and Search task. In 2012, at the TRECVID, his team received four first places on all four sub-tasks of the Instance Search (INS) task and the Known-Item Search task. In 2014, at the TRECVID, his team received the first place in the Interactive Instance Search task. His team also received first places in both the INS task of TRECVID in 2015, 2016, and 2017, respectively. He received the First Prize of the Beijing Science and Technology Award in 2016 (ranking first).

Jinwei Qi received the B.S. degree in computer science and technology from Peking University, in 2016. He is currently pursuing the M.S. degree with the Institute of Computer Science and Technology, Peking University. His current research interests include cross-media retrieval and machine learning.

Yuxin Yuan received the B.S. degree in electronic science and engineering from Nanjing University, in 2015. He is currently pursuing the M.S. degree with the Institute of Computer Science and Technology, Peking University. His current research interests include cross-media retrieval and machine learning.