Image-Text Cross-Modal Retrieval via Modality-Specific Feature Learning

Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, Chunhong Pan
{jian.wang, yhhe, cckang, smxiang, chpan}@nlpr.ia.ac.cn

ABSTRACT

Cross-modal retrieval extends the ability of search engines to deal with the massive cross-modal data. The goal of image-text cross-modal retrieval is to search images (texts) by using text (image) queries by computing the similarities of images and texts directly. Many existing methods rely on low-level visual features and textual features for cross-modal retrieval, ignoring the characteristics existing in the raw data of different modalities. In this paper, a novel model based on modality-specific feature learning is proposed. Considering the characteristics of different modalities, the model uses two types of convolutional neural networks to map the raw data to the latent space representations for images and texts, respectively. Particularly, the convolution based network used for texts involves word embedding learning, which has been proved effective to extract meaningful textual features for text classification. In the latent space, the mapped features of images and texts form relevant and irrelevant image-text pairs, which are used by the one-vs-more learning scheme. This learning scheme can achieve ranking functionality by allowing for one relevant and more irrelevant pairs. The standard backpropagation technique is employed to update the parameters of two convolutional networks. Extensive cross-modal retrieval experiments are carried out on three challenging datasets that consist of image-document pairs or image-query clickthrough data from a search engine, and the results firmly demonstrate that the proposed model is much more effective.

Keywords

Cross-modal Retrieval; Convolutional Neural Network; Feature Learning; Deep Neural Network

1. INTRODUCTION

Cross-modal retrieval can extend the services of traditional search engines. Text-to-document search and keyword-to-image search are two major applications provided by search engines. Text-to-document search is a task of single modality. Keyword-to-image search is a pseudo "cross-modal" issue, because the query keywords are matched with tags annotated to the images, not the images themselves. The goal of image-text cross-modal retrieval is to search images (texts) by using text (image) queries without any auxiliary information. The main challenges is how to correlate data from different modalities while capturing their respective inner properties.

There are some approaches proposed for cross-modal retrieval [25, 37, 32, 20, 35, 34, 36, 6, 33]. However, most methods do not pay attention to learning modality-specific features. Instead, they directly use hand-crafted visual features, such as SIFT [19] based bag of words feature, Gist [24] feature, MPEG-7 [21] descriptors and color histograms, and popular textural features, such as latent dirichlet allocation (LDA) [2] feature, replicated softmax model (RSM) [10], one-hot feature and word frequency feature. However, these features have some weaknesses, since the artificial features may not contain enough information that are useful for the different modalities. For example, visual features always reflect one aspect property of image contents, and one-hot feature ignores the orders of words. In [6] and [33], auto-encoders are used to relearn the hand-crafted features. The structure of auto-encoders is not delicate enough to grasp the inner properties of different modalities, which prevents these methods from further improvements.

Conventionally, there are some typical image-text cross-modal datasets employed for evaluation. For example, the Wikipedia dataset [25] is widely used for cross-modal retrieval. It contains 2866 image-text pairs from "Wikipedia featured articles", falling into 10 semantic categories. The texts alongside each image are long descriptions, always resulting in paragraphs. This type of "text" is constructed via words, which has strong sequential and structural information. Another well known dataset NUS-WIDE [4] is also used for cross-modal retrieval task. There are 269,648 images collected from Flickr in this dataset. Different from Wikipedia, each image associates with some tags, rather than textual descriptions. These tags have no sequential and structural information. Additionally, the prior knowledge, namely the class labels for image-text pairs, is provided by above two datasets. In our view, such prior information is not easy to obtain in practice. On the Internet, massive web data contains multimedia components, such as texts, images, audio signals and videos, whereas they are seldom classified. Based on these discussions, we focus on using the datasets that have the following properties:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICMR '15, June 23 - 26, 2015, Shanghai, China
Copyright 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00
http://dx.doi.org/10.1145/2671188.2749341.

- The data form is image-text pair, and the text part should be sentences or paragraphs.
- The prior knowledge is not provided.

To satisfy above properties, we use three datasets for the experiments: IAPRTC-12 [7], Attribute Discovery [1] and MSR Bing challenge dataset [12].

In this paper, we propose a novel cross-modal retrieval model based on modality-specific feature learning for images and texts. Two types of convolution based networks are employed. Specifically, conventional convolutional neural network (CNN) [18] is used to model the image modality. CNN has greatly boosted the performance of image classification due to its power of image feature learning [17]. We leverage the benefits of CNN to build up the proposed model. As for the text modeling, we also use a convolution based network (WCNN) [15] which has achieved state-of-the-art performance on text classification. WCNN takes sentences in forms of word embeddings as inputs, and performs convolution operations with different filter widths sequentially to the word embeddings. Finally, max-pooling operations are applied to the outputs of the convolution. It should be noted that WCNN can handle sentences with different lengths and outcome feature vectors with the identical dimension.

By using CNN and WCNN, the proposed model can transform the raw image and text data into a latent space while maintaining the data characteristics. To accomplish the retrieval task, we use cosine similarity to calculate the relevance scores between images and texts in the latent space. The relevant and irrelevant image-text pairs are gathered to form the one-vs-more learning scheme: one relevant and more irrelevant image-text pairs are put to the objective of maximizing posterior likelihood of the relevant pairs. By doing this, the objective function can rank relevant pairs ahead of irrelevant ones. Comprehensively, the proposed model associates the modality-specific feature learning with the goal of cross-modal retrieval. Extensive experiments are conducted on three datasets: IAPRTC-12 [7], Attribute Discovery [1] and MSR Bing challenge dataset [12]. The results consistently demonstrate the effectiveness of the proposed model compared with the existing methods.

2. RELATED WORK

The issue of cross-modal retrieval has attracted much attention from researchers. In the past few years, many methods are proposed. Some of them [25, 32, 35, 36] take advantages of prior knowledge (namely label information) while establishing their models, and some of them [20, 34] leverage ranking information. Since the experiments are conducted on datasets in the form of image-text pairs without any additional information, these methods are hardly applied to such data. Thus, we do not make further discussions of these methods.

Generally, methods that can project heterogeneous data into a latent space are potential to do cross-modal retrieval, since the similarity of heterogeneous data can be computed in the latent space. Canonical correlation analysis (CCA) [8] is one of the candidate algorithms. The method in [25] is developed based on CCA. The key idea of CCA is to learn a common latent space in which the correlations between projected features of two modalities are maximized. Once the heterogeneous data is projected to the latent space, some popular similarities, such as ℓ_2 distance and cosine distance, are used to generate search results. Another method is the partial least squares (PLS) [27], which has been used for heterogeneous face recognition [29]. PLS is a regression model to project data of one modality to another through a latent space. Specifically, two data types are first projected to the latent space instead of a direct mapping. Although CCA and PLS are not designed for cross-modal retrieval, they can project heterogeneous data to a common latent space, which is a key step to address cross-modal retrieval.

Recently, some methods based on deep learning are also proposed for cross-modal tasks. In [30] and [23], deep boltzmann machines [28] and deep auto-encoders [9] are employed to model cross-modal data. They result in generative models and learn unified representations for both modalities, but not aim at retrieval task. Kiros *et al.* [16] proposed to use log-bilinear language model [22] to predict word sequences combined with image information. This model can jointly learn text and image features by using CNN. In recent works [6] and [33], the deep auto-encoders are used to do feature learning with the objective of cross-modal retrieval. In fact, the method in [6] is an extension of [23], in which auto-encoders reconstruct both modalities while integrating the objective of retrieval. Similarly, Wang *et al.* [33] utilized single-modality auto-encoders to separately handle different modalities. Both of them use outputs of middle layers of auto-encoders to represent data of different modalities. And their final objective functions are also similar: minimize the ℓ_2 distance of paired cross-modal data and the reconstruction error, simultaneously. Auto-encoder based methods successfully combine feature learning with the goal of cross-modal retrieval and achieve promising results. However, feature learning in auto-encoders is the relearning of hand-crafted features, which is not able to reflect the inner properties of each modality. On the contrast, in the proposed model, two convolutional networks are used to extract modality-specific features, which is verified to be more effective than auto-encoders.

3. THE PROPOSED MODEL

3.1 Notation

Some important notations are first defined. In this paper, we focus on image-text cross-modal retrieval. Suppose the training set consists of image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, N is the number of pairs, I_i and T_i are raw image and text data respectively. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be the hand-crafted features for images and texts. Additionally, T_i is composed of words from the vocabulary $\mathcal{V}$, and the size of $\mathcal{V}$ is M. Embedding vectors $\{\mathbf{w}_j\}_{j=1}^{M}$ are created for each word. The dimension of $\mathbf{w}_j$ is set manually.

3.2 Feature Learning for Image Modality

Convolutional neural network (CNN) [18] is chosen to handle image modality. In recent years, CNN has demonstrated its outstanding capability in several vision tasks, such as image classification [17], face keypoints detection [31]. A typical CNN structure is shown in Fig. 1. Spatial convolution and pooling operations can preserve inner properties of images. In the proposed model, raw images with RGB channels are taken as inputs of CNN, and the outputs of full connection layer are representations for images:

$$\tilde{\mathbf{x}}_i = CNN(I_i), \tag{1}$$

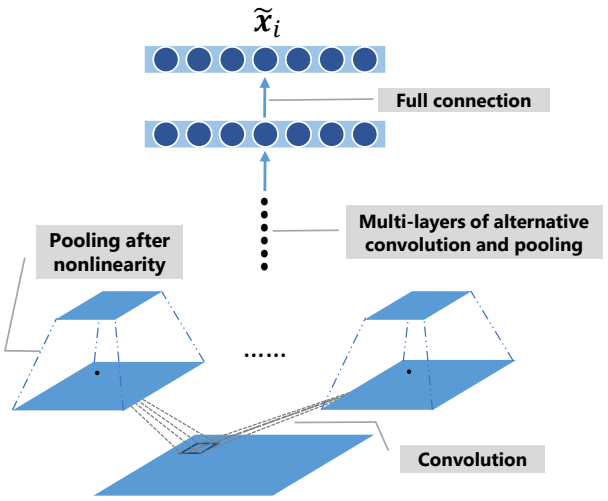

Figure 1: Structure of CNN containing alternative convolution and pooling layers, and full connection layers as well.

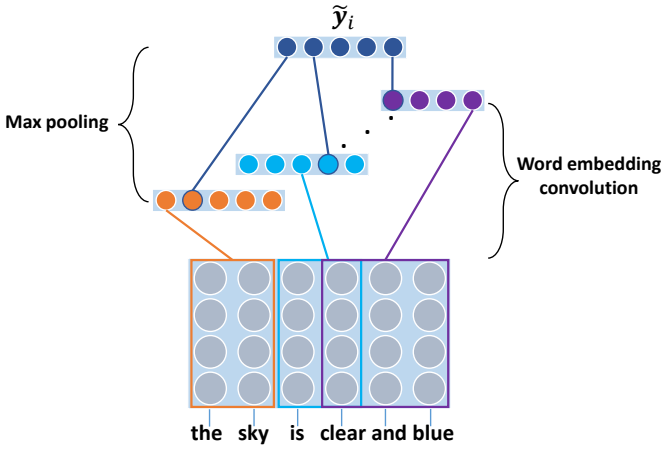

Figure 2: Structure of WCNN containing only one convolutional layer and one max-pooling layer.

where $\tilde{\mathbf{x}}_i$ are image features in the latent space. The details of our CNN structure can be found in Section 4.2.

3.3 Feature Learning for Text Modality

In most previous work, latent dirichlet allocation (LDA) [2] based feature, one-hot feature and word frequency feature are ordinary choices for texts. However, these features have some limitations. LDA relies on corpus to generate topic probabilities. This process needs prior knowledge from corpus and can be influenced by the quality of corpus. As for one-hot feature and word frequency feature, they ignore word sequences and semantic correlations. For example, two words "car" and "vehicle" are semantically similar, and they may be represented as $[\cdots, 1, 0, \cdots]$ and $[\cdots, 0, 1, \cdots]$ in one-hot form. With such one-hot features, the cosine similarity between these two words is zero, which is not rational in practice. What's more, the dimensions of one-hot feature and word frequency feature will linearly increase along with the size of vocabulary. Once the number of unique words is massive, such kinds of features become intractable.

In natural language processing, recent works [15],[3] and [11] propose to model sentences using convolution based structures. The method in [15] achieve state-of-the-art performance on sentence classification with only one convolution layer and one max-pooling layer. We adopt this method (WCNN) to extract text features in the proposed model.

An illustration of WCNN is shown in Fig. 2. The input sentence is represented as a matrix of word embeddings. The word embeddings are aligned to have the same sequence as the words in the sentence. In the phase of convolution, a large number of convolutional filters are applied to the embedding matrix. Filters may have different widths (in Fig. 2, filters in orange and blue have the width of 2, while the width of filter in purple is 3), but they have the same length as the word embedding. It is noticed that the convolution is performed within adjacent word embeddings, which can extract local semantics. The convolutional stride is 1 for all filters, and the outputs of the convolution operation are vectors with different lengths after nonlinearity. Afterwards, each vector is max-pooled to form the final representation of the sentence as illustrated in Fig. 2. The dimension of output feature is equal to the number of filters. This process is notated as:

$$\tilde{\mathbf{y}}_i = WCNN(T_i), \tag{2}$$

where $\tilde{\mathbf{y}}_i$ are text features in the latent space. The detailed calculation of WCNN can be referred in [15].

There are some advantages of WCNN. First, WCNN can naturally handle sentences with variable lengths. The dimension of output features depends on the number of filters, not the length of sentences. Second, the sequential and structural information of sentences can be maintained by convolution and max-pooling operations, which is neglected by one-hot feature and word frequency feature. Third, WCNN is easy to deal with large vocabulary size. The lengths of inputs are related to the lengths of sentences, not the size of the vocabulary.

3.4 Modality-Specific Deep Structure

The structure of the proposed model is illustrated in Fig. 3. The proposed model, named as **M**odality-**S**pecific **D**eep **S**tructure (**MSDS**), is motivated by [13]. Next, we discuss the model in detail.

The one-vs-more learning scheme is adopted here. The model focuses on text modality as queries[1]. For each text query, the text and its matched image form the relevant pair, and the irrelevant pairs are constructed by the text and a few unmatched images. Specifically, the relevant pair is (T_i, I_i). For irrelevant pairs, c random unmatched images are selected $\{(T_i, I_{ij}^-)\}_{j=1}^c$ (c is set to 4 in the experiments). Notice that, in the training process, unmatched images selected for the text T_i are different in each iteration. The input text and images are transformed to the latent space

[1]Actually, the goal of the model is to score the image-text pairs. No matter what the query modality is, the model can be sufficiently learned and then used to do cross-modal retrieval. Here, we use the text modality as queries.

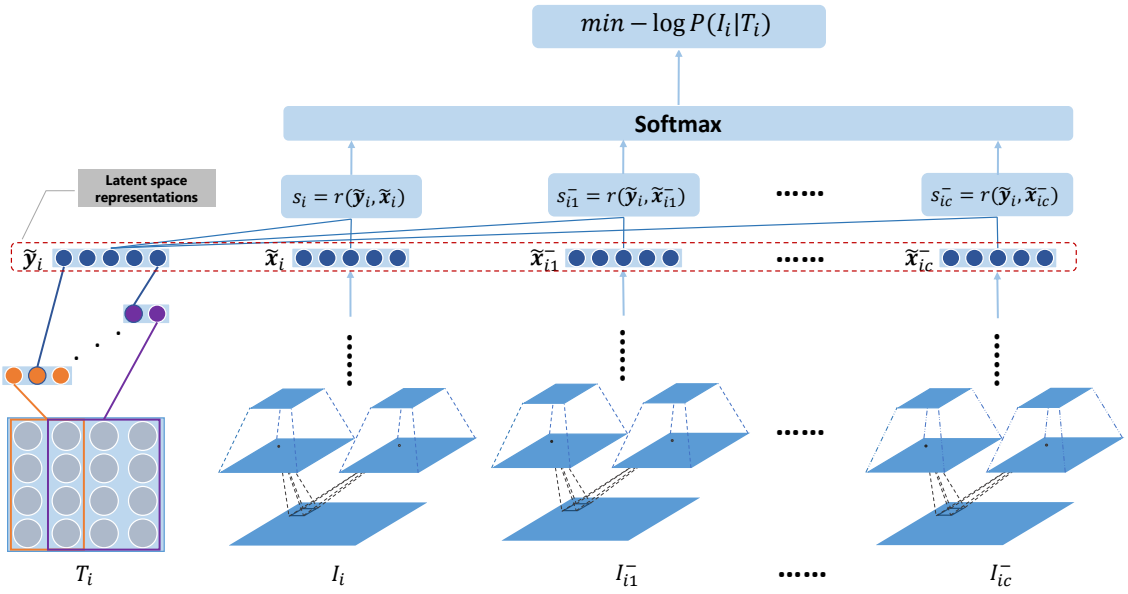

Figure 3: Structure of the proposed model. The one-vs-more scheme can involve one relevant and multiple irrelevant image-text pairs. Image modality and text modality are processed by CNN and WCNN, respectively. The relevance scores of image-text pairs are calculated using cosine similarity. The posterior probability of relevant image given the text query is estimated by softmax function. Finally, a maximum likelihood framework is adopted to optimize the model parameters.

using WCNN and CNNs. In the latent space, the relevance scores of image-text pairs are estimated by cosine similarity:

$$s_i = r(\tilde{\mathbf{y}}_i, \tilde{\mathbf{x}}_i) = \frac{\tilde{\mathbf{y}}_i^T \tilde{\mathbf{x}}_i}{\|\tilde{\mathbf{y}}_i\| \|\tilde{\mathbf{x}}_i\|}. \tag{3}$$

Our objective is to enlarge the relevance scores of relevant pairs while suppressing the relevance scores of irrelevant pairs. To achieve this goal, a maximum likelihood framework is employed. First, the relevance scores of the relevant and irrelevant pairs are gathered to compute the posterior probability of the relevant image given the text query via the softmax function:

$$P(I_i|T_i) = \frac{exp(s_i)}{exp(s_i) + \sum_{j=1}^{c} exp(s_{ij}^-)}. \tag{4}$$

Next, minimizing the negative logarithmic likelihood of the posterior probability of the relevant images over the training set is formulated as:

$$\min_{\boldsymbol{\theta}} -\log\{\prod_{i=1}^{N} P(I_i|T_i)\}, \tag{5}$$

where $\boldsymbol{\theta}$ is the model parameter, including WCNN parameters and CNN parameters (each CNN substructure in Fig. 3 share the same parameters).

From the description above, the architecture in the proposed model can engage both relevant and irrelevant pairs. This design can make the model sufficiently explore the interactions between relevant and irrelevant pairs, which is not considered in [6, 33].

In test phase, the features for images and texts are first extracted by trained CNN and WCNN. Next, the relevance scores of image-text pairs are computed using Eq. 3. Finally, images (texts) are ordered into the ranking lists according to their relevance scores.

3.5 Variations of the Proposed Structure

Furthermore, two trivial variations of the original structure are studied. The first variation (**MSDS-v1**) takes the hand-crafted features as inputs. WCNN and CNN are replaced by deep forward neural networks (DNN). DNNs are ordinary methods of feature learning without considering the characteristics of a specific modality. This process of feature learning is formulated as:

$$\tilde{\mathbf{x}}_i = DNN_I(\mathbf{x}_i), \tag{6}$$

$$\tilde{\mathbf{y}}_i = DNN_T(\mathbf{y}_i), \tag{7}$$

where $DNN_I(\cdot)$ and $DNN_T(\cdot)$ are DNNs for images and texts, respectively. The comparison between **MSDS** and **MSDS-v1** will provide insights of the power of modality-specific feature learning. The second variation (**MSDS-v2**) only replaces CNNs with DNNs, preserving WCNN. This modification is to prove that features extracted by WCNN are more effective than hand-crafted features, such as LDA based feature and word frequency feature.

The model and its variations can be optimized using standard stochastic gradient descend (SGD) and backpropagation algorithm.

4. EXPERIMENT

In this section, the proposed model is compared with some representative methods on three datasets. The datasets are first introduced in subsection 4.1. Then, the implementation details of the proposed model are described. Next, compared

methods and evaluation metric are briefly introduced. Subsequently, the experimental results are shown and discussed.

4.1 Datasets

IAPRTC-12. This dataset is initially released by Grubinger *et al.* [7] for cross-lingual retrieval. There are 19,627 images and each image is attached with several descriptive sentences. The vocabulary size is 4,576. The dataset is split into two subsets: 17,627 images-text pairs for training and 2000 pairs for test. The texts in this dataset are grammatical with little noise. The language in the sentences is well organized and the content of the sentences is closely related to images. As for the hand-crafted features, we use word frequency feature for text modality (the feature dimension is equal to the size of vocabulary) and CNN feature[2] for image modality (the feature dimension is 4096). The same hand-crafted features are extracted for other two datasets.

Attribute Discovery. This dataset is created for visual attribute discovering from noisy text descriptions by Berg *et al.* [1], which consists of 37,794 images. Dissimilar to IAPRTC-12, the texts in Attribute Discovery are in form of sentences but less informative to images with much noisy. The vocabulary size is 27,570. 32,794 image-text pairs are used for training and 5,000 pairs for test. Notably, the content of images is very simple, since each image only depicts one product without background clutter.

MSR-Bing Challenge Dataset. This dataset is originally used for image retrieval challenge [12]. The data is collected from a practical search engine—Bing[3], with the form of triplets: <query, image, click count>. The meaning of a triplet is that an image is clicked several times by users under a text query. In this dataset, there are 23M triplets including 1M images and 11.7M queries, and the click count varies from zero to thousands. In the experiments, image-query pairs with large click counts are selected, since they are highly correlated. The final dataset results in 40,000 image-query pairs for training and 5,000 pairs for test. The vocabulary size is 24078. A difference between this dataset and other datasets is that user queries are short and arbitrary. We notice that the queries suffer typos, near-duplications and reversed order. Additionally, user queries are less structural than sentences, which make it a big challenge.

4.2 Implement Details

In this section, we describe the settings of the proposed structure including WCNN, CNN and DNN in detail.

WCNN. The structure of WCNN is determined by three factors: the embedding size, the number of filters and filter widths. For IAPRTC-12 and Attribute Discovery, the embedding size is set to 25, filter widths are $\{3, 4, 5, 6, 7, 8\}$ and 50 filters are created for each width. For MSR-Bing Dataset, the embedding size is also set to 25, filter widths are $\{1, 2, 3, 4, 5\}$ and there are 60 filters for each width. Thus, for all datasets, the output dimension of WCNN is 300. Additionally, all word embeddings and filters are initialized with random variables. These WCNN settings are applied to MSDS and MSDS-v2.

CNN. The CNN used in the proposed model has the same structure as that in Caffe (details can be found in http://caffe.berkeleyvision.org/). Furthermore, we use the trained model on Imagenet to initialize our model, which is known as supervised pre-training [26]. The output dimension of CNN is 4096 that is not equal to the output dimension of WCNN, thus two extra fully connected layers with sizes of 1024 and 300 are sequently added to the output layer of CNN. The parameters of these additional layers are randomly initialized. This CNN setting is applied in three datasets for MSDS.

DNN. Image DNN and text DNN are used in MSDS-v1 and MSDS-v2. The input dimension of text DNN depends on the vocabulary size, since word frequency feature is used. For IAPRTC-12 and Attribute Discovery, text DNN has two hidden layers (vocabulary_size-512-512), and image DNN also has two hidden layers (4096-1024-512). For MSR-Bing Dataset, text DNN has the structure of vocabulary_size-1024-512, and image DNN is the same as the other two datasets. All parameters for DNNs are initialized with random variables.

Input images for CNN are centered and input features for DNNs are processed to have zero mean and unit variance. The learning rates for WCNN, CNN and DNN are 0.01, 0.001 and 0.001, respectively. Some other techniques, such as ℓ_2 decay and momentum, are also utilized. Actually, the one-vs-more scheme can naturally generate much more data. For example, if there are 100 matched image-text pairs, the number of inputs can reach to $C_{100}^1 \times C_{99}^4$. This may partially overcome the overfitting.

4.3 Compared Methods and Evaluation Metric

Some comparative methods are listed below:

- **CCA**. CCA takes the CNN feature and the word frequency feature as inputs to calculate the projection matrices. In the latent space, we use ℓ_2 distance to measure the similarity between two modalities.
- **PLS**. Similar with CCA, PLS also project the hand-crafted features of both modalities to the latent space in which the similarity between two modalities is measured by ℓ_2 distance.
- **corr-AE**. This method is proposed in [6]. Two auto-encoders are used to reconstruct the input hand-crafted features. Each auto-encoder in this method is for single modality, namely the input modality is consistent with the output modality. Particularly, the input layer of text auto-encoder is modeled using replicated softmax [10]. The middle layer is used for features in latent space, and the modality similarity is measured by ℓ_2 distance.
- **cross-corr-AE**. This method is also proposed in [6]. Different from corr-AE, cross-corr-AE uses cross-modal auto-encoders that take one modality as input to reconstruct another modality.

The cross-modal retrieval has two tasks: search images by text queries and search texts by image queries. In previous works [25, 37, 32, 20, 35, 34, 36, 6, 33], the evaluation metric widely used is the mean average precision (MAP). However, in practice, users always care the top ranked results. Thus, we suggest to use top@k precision instead of MAP.

[2]CNN features are extracted using the tool of Caffe [14]. This CNN model is trained based on Imagenet dataset [5]. The output of the $6th$ layer is used for features.
[3]www.bing.com

Table 1: Cross-modal retrieval results on three datasets. Two tasks, searching images by text queries and searching texts by image queries, are evaluated.

(a) IAPRTC-12 (search images by text queries)

Method	top@1	top@2	top@4	top@10	top@20	top@40	top@100	top@200	top@500	top@1000
Random	0.0005	0.0010	0.0020	0.0050	0.0100	0.0200	0.0500	0.1000	0.2500	0.5000
CCA	0.0724	0.0943	0.1254	0.1692	0.2171	0.2691	0.3502	0.4511	0.6157	0.8119
PLS	0.1677	0.2569	0.3435	0.5051	0.6152	0.7222	0.8287	0.8899	0.9506	0.9806
corr-AE	0.0285	0.0510	0.0765	0.1412	0.2136	0.3038	0.4730	0.6203	0.8344	0.9429
cross-corr-AE	0.0958	0.1463	0.2171	0.3333	0.4475	0.5647	0.7105	0.8211	0.9363	0.9913
MSDS-v1	0.1750	0.2700	0.3800	0.5590	0.6970	0.8120	0.9170	0.9590	**0.9920**	0.9980
MSDS-v2	0.1860	0.2800	0.3920	0.5680	0.6990	0.8230	**0.9250**	**0.9660**	**0.9920**	**0.9990**
MSDS	**0.2560**	**0.3640**	**0.4710**	**0.6260**	**0.7400**	**0.8360**	0.9210	0.9600	0.9900	0.9970

(b) IAPRTC-12 (search texts by image queries)

Method	top@1	top@2	top@4	top@10	top@20	top@40	top@100	top@200	top@500	top@1000
Random	0.0005	0.0010	0.0020	0.0050	0.0100	0.0200	0.0500	0.1000	0.2500	0.5000
CCA	0.0938	0.1407	0.1942	0.2773	0.3440	0.4225	0.5082	0.5780	0.6779	0.7895
PLS	0.1035	0.1529	0.2095	0.2992	0.3710	0.4383	0.5423	0.6147	0.7263	0.8405
corr-AE	0.1570	0.2385	0.3502	0.5117	0.6284	0.7554	0.8879	0.9439	0.9913	0.9985
cross-corr-AE	0.1009	0.1483	0.2171	0.3491	0.4664	0.5928	0.7416	0.8405	0.9414	0.9837
MSDS-v1	0.1620	0.2600	0.3810	0.5530	0.6870	0.8120	0.9150	0.9620	0.9800	0.9990
MSDS-v2	0.1680	0.2530	0.3710	0.5300	0.6710	0.7950	0.9160	**0.9640**	**0.9920**	**1.0000**
MSDS	**0.2550**	**0.3620**	**0.4790**	**0.6330**	**0.7420**	**0.8320**	**0.9190**	0.9610	0.9880	0.9970

(c) Attribute Discovery (search images by text queries)

Method	top@1	top@5	top@10	top@25	top@50	top@100	top@500	top@1000	top@2000	top@3000
Random	0.0002	0.0010	0.0020	0.0050	0.0100	0.0200	0.1000	0.2000	0.4000	0.6000
CCA	**0.3454**	0.4608	0.4958	0.5362	0.5608	0.5782	0.6112	0.6260	0.6502	0.6832
PLS	0.2044	0.3636	0.4320	0.5250	0.6070	0.6792	0.8392	0.8990	0.9444	0.9638
corr-AE	0.0284	0.1036	0.1592	0.2498	0.3422	0.4478	0.7446	0.8528	0.9274	0.9642
cross-corr-AE	0.1066	0.2514	0.3360	0.4686	0.5818	0.6960	0.9168	0.9706	0.9948	**0.9978**
MSDS-v1	0.1114	0.3100	0.4272	0.5948	0.7208	0.8376	**0.9750**	**0.9896**	**0.9954**	0.9976
MSDS-v2	0.2688	0.5152	0.6184	0.7388	0.8092	0.8696	0.9606	0.9794	0.9920	0.9960
MSDS	0.3298	**0.5748**	**0.6760**	**0.7904**	**0.8536**	**0.9024**	0.9618	0.9828	0.9934	0.9968

(d) Attribute Discovery (search texts by image queries)

Method	top@1	top@5	top@10	top@25	top@50	top@100	top@500	top@1000	top@2000	top@3000
Random	0.0002	0.0010	0.0020	0.0050	0.0100	0.0200	0.1000	0.2000	0.4000	0.6000
CCA	**0.3286**	**0.4376**	0.4746	0.5172	0.5562	0.5874	0.6750	0.7266	0.8060	0.8748
PLS	0.0982	0.1988	0.2530	0.3368	0.4008	0.4810	0.6648	0.7466	0.8468	0.9112
corr-AE	0.0886	0.2464	0.3456	0.4892	0.6038	0.7170	0.9384	0.9810	**0.9964**	0.9986
cross-corr-AE	0.0560	0.1790	0.2674	0.4120	0.5418	0.6724	0.9308	0.9758	0.9952	0.9978
MSDS-v1	0.1122	0.2978	0.4154	0.5816	0.7124	0.8286	**0.9736**	**0.9908**	0.9962	**0.9988**
MSDS-v2	0.1988	0.4358	**0.5528**	**0.6928**	**0.7886**	**0.8560**	0.9582	0.9810	0.9928	0.9968
MSDS	0.2986	0.4134	0.5138	0.6548	0.7502	0.8244	0.8968	0.9330	0.9632	0.9802

(e) MSR-Bing Dataset (search images by text queries)

Method	top@1	top@5	top@10	top@25	top@50	top@100	top@500	top@1000	top@2000	top@3000
Random	0.0002	0.0010	0.0020	0.0050	0.0100	0.0200	0.1000	0.2000	0.4000	0.6000
CCA	0.0060	0.0126	0.0188	0.0312	0.0484	0.0730	0.1986	0.3200	0.5154	0.7014
PLS	**0.0240**	**0.0650**	0.0954	0.1518	0.2072	0.2776	0.5000	0.6226	0.7598	0.8558
corr-AE	0.0010	0.0038	0.0072	0.0154	0.0256	0.0486	0.1638	0.2826	0.4792	0.6674
cross-corr-AE	0.0008	0.0026	0.0060	0.0126	0.0204	0.0420	0.1620	0.2892	0.4958	0.6832
MSDS-v1	0.0106	0.0322	0.0528	0.0966	0.1652	0.2566	0.5332	0.6720	0.8144	0.9090
MSDS-v2	0.0094	0.0326	0.0564	0.1112	0.1776	0.2666	0.5410	0.6870	0.8420	0.9230
MSDS	0.0198	0.0618	**0.1002**	**0.1778**	**0.2602**	**0.3634**	**0.6208**	**0.7488**	**0.8748**	**0.9386**

(f) MSR-Bing Dataset (search texts by image queries)

Method	top@1	top@5	top@10	top@25	top@50	top@100	top@500	top@1000	top@2000	top@3000
Random	0.0002	0.0010	0.0020	0.0050	0.0100	0.0200	0.1000	0.2000	0.4000	0.6000
CCA	0.0086	0.0236	0.0346	0.0598	0.0834	0.1070	0.2154	0.3266	0.5068	0.6626
PLS	0.0162	0.0494	0.0726	0.1128	0.1494	0.1922	0.3174	0.4586	0.6364	0.7834
corr-AE	0.0042	0.0196	0.0370	0.0774	0.1234	0.1858	0.4166	0.5654	0.7528	0.8816
cross-corr-AE	0.0030	0.0086	0.0128	0.0238	0.0352	0.0564	0.1822	0.3110	0.5094	0.6916
MSDS-v1	0.0104	0.0330	0.0676	0.1404	0.2190	0.2568	0.5169	0.6883	0.8536	0.9230
MSDS-v2	0.0100	0.0346	0.0586	0.1170	0.1780	0.2636	0.5272	0.6772	0.8336	0.9162
MSDS	**0.0222**	**0.0712**	**0.1072**	**0.1798**	**0.2614**	**0.3634**	**0.6176**	**0.7418**	**0.8734**	**0.9378**

4.4 Results and Discussions

We present the results of cross-modal retrieval in Table 1. For each dataset, searching images by text queries (the first task) and searching texts by image queries (the second task) are performed. Notice that, there is only one matched image (text) for each text (image). The evaluation metric top@k covers a wide range of k values. In practice, users only concentrate on top ranked items. So the discussions are mainly for top 10% (IAPRTC-12: $\leq$top@200, Attribute Discovery and MSR-Bing Dataset: $\leq$top@500).

4.4.1 On IAPRTC-12 (Table 1(a)-(b))

On IAPRTC-12, MSDS significantly outperforms other methods including its two variations in both tasks. Particularly, top@1 performance of MSDS is 25.6% which is much larger than the second 18.6% obtained by MSDS-v2. MSDS-v2 is slightly better than MSDS-v1, which indicates that WCNN offers limited promotion compared with word frequency feature in the scenario of "clean" texts and small vocabulary size. PLS performs surprisingly well. As for the auto-encoder based methods, they achieve unsatisfactory results on the first task (namely searching images by text queries), particularly when k is small. But they perform better on the second task. IAPRTC-12 is the ideal one among three datasets. The good performance of MSDS relies on the modality-specific feature learning that extract good representations for input modalities.

4.4.2 On Attribute Discovery (Table 1(c)-(d))

On Attribute Discovery, the CCA achieves the best performance for top@1 on both tasks. Recall that the images in this dataset are very simple, this may lead to easy correspondences between images and texts, which is considered by CCA. PLS can only obtain good performance on the first task. Although the proposed method MSDS and MSDS-v2 are marginally lower than CCA at top@1 and top@5, they consistently perform well when $k \leq 100$. MSDS-v1 can not get a good performance when $k \leq 10$. This is due to the poor feature learning ability of DNN for text modality. Text descriptions in this dataset are noisy and less informative, and the vocabulary size is larger than that in IAPRTC-12. Thus word frequency feature is less functional. Noticeably, MSDS-v2 can perform much better than MSDS-v1 with WCNN for text feature learning. Auto-encoder based methods do not perform well when k is small, but when $k \geq 500$, their performances are beyond CCA and PLS except corr-AE on the first task.

4.4.3 On MSR-Bing Dataset (Table 1(e)-(f))

As aforementioned, MSR-Bing Dataset comes from a real-world search engine, making it a big challenge. Overall, the average performances of all methods are relatively lower than those in the other two datasets. PLS achieves satisfactory results on both tasks. MSDS still obtains the best performances in most cases. Whereas, MSDS-v1 and MSDS-v2 do not obtain the expected performances. Furthermore, MSDS-v2 dose not get performance gain compared with MSDS-v1 due to the short and noisy user queries, since WCNN is suitable for sequential and structural sentences. In this dataset, auto-encoder based methods are failed to give ordinary performances.

In summary, MSDS is constantly superior than other methods, and its two variations are proved to be effective as well. The good performances of MSDS and its variations firmly demonstrate the effectiveness of modality-specific feature learning and the one-vs-more learning scheme. Although CCA and PLS show their potential to do cross-modal retrieval, they need in-memory computation, which limits them to scale to large datasets. Whereas, the proposed model uses minibatch based stochastic gradient descend optimization that is easy to be applied to large datasets. As for auto-encoder based methods, they do not exhibit the good potential for cross-modal retrieval.

5. CONCLUSIONS

In this paper, a novel model based on modality-specific feature learning is proposed. We adopted two convolution based neural networks, namely CNN and WCNN, to fulfill feature learning for image and text modalities, respectively. CNN and WCNN take the modality-specific characteristics into consideration, and thus they can extract better features from raw data. As a result, data from two modalities are transformed to a latent space where the inter-modality similarity is calculated via cosine distance. Subsequently, the one-vs-more learning scheme with maximum posterior likelihood objective function is used to optimize the model parameters. In summary, we combine the modality-specific feature learning with the goal of cross-modal retrieval. Extensive experiments are carried out on three datasets with different properties, and the results firmly demonstrate the effectiveness of the proposed model compared with state-of-the-art methods.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, and the National Natural Science Foundation of China under Grants 91338202, 61203277 and 91438105.

7. REFERENCES

[1] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Proceddings of European Conference on Computer Vision*, pages 663–676. 2010.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] P. Blunsom, E. Grefenstette, and N. Kalchbrenner. A convolutional neural network for modelling sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

[4] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, page 48, 2009.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *ACM International Conference on Multimedia*, pages 7–16, 2014.

[7] M. Grubinger, P. Clough, H. Muller, and T. Deselaers. The iapr tc-12 benchmark: a new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006.

[8] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[9] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[10] G. Hinton and R. Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems*, pages 1607–1614, 2009.

[11] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, pages 2042–2050, 2014.

[12] X. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In *ACM International Conference on Multimedia*, pages 243–252, 2013.

[13] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *International Conference on Information and Knowledge Management*, pages 2333–2338, 2013.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678, 2014.

[15] Y. Kim. Convolutional neural networks for sentence slassification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014.

[16] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *International Conference on Machine Learning*, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[20] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang. A low rank structural large margin method for cross-modal ranking. In *ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 433–442, 2013.

[21] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[22] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In *International Conference on Machine Learning*, pages 641–648, 2007.

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, 2011.

[24] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[25] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia*, pages 251–260, 2010.

[26] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[27] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, pages 34–51. 2006.

[28] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[29] A. Sharma and D. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 593–600, 2011.

[30] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2012.

[31] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[32] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *IEEE International Conference on Computer Vision*, pages 2088–2095, 2013.

[33] W. Wang, B. Ooi, X. Yang, D. Zhang, and Y. Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. *Proceedings of the VLDB Endowment*, 7(8), 2014.

[34] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. Cross-media semantic representation via bi-directional learning to rank. In *ACM International Conference on Multimedia*, pages 877–886, 2013.

[35] L. Xie, P. Pan, and Y. Lu. A semantic model for cross-modal and multi-modal retrieval. In *ACM International Conference on Multimedia Retrieval*, pages 175–182, 2013.

[36] X. Zhai, Y. Peng, and J. Xiao. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval. In *Proceedings of Advances in Multimedia Modeling*, pages 312–322, 2012.

[37] Y. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.