Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks

Yuxin Peng*, Xin Huang and Jinwei Qi
Institute of Computer Science and Technology,
Peking University, Beijing 100871, China
{pengyuxin, huangxin_14, qijinwei}@pku.edu.cn

Abstract

Inspired by the progress of deep neural network (DNN) in single-media retrieval, the researchers have applied the DNN to cross-media retrieval. These methods are mainly two-stage learning: the first stage is to generate the **separate representation** for each media type, and the existing methods only model the intra-media information but ignore the inter-media correlation with the rich complementary context to the intra-media information. The second stage is to get the **shared representation** by learning the cross-media correlation, and the existing methods learn the shared representation through a shallow network structure, which cannot fully capture the complex cross-media correlation. For addressing the above problems, we propose the cross-media multiple deep network (CMDN) to exploit the complex cross-media correlation by hierarchical learning. In the first stage, CMDN jointly models the intra-media and inter-media information for getting the complementary **separate representation** of each media type. In the second stage, CMDN hierarchically combines the inter-media and intra-media representations to further learn the rich cross-media correlation by a deeper two-level network strategy, and finally get the **shared representation** by a stacked network style. Experiment results show that CMDN achieves better performance comparing with several state-of-the-art methods on 3 extensively used cross-media datasets.

1 Introduction

With the rapid growth of multimedia information, the cross-media data, like image, text, video and audio, has been the main form of big data, and the demand of cross-media retrieval is greatly spurred. For example, if users are on a visit to Buckingham Palace, by submitting a photo of it, they can get the relevant images, videos and text at the same time. In the last decades, content-based information retrieval has been widely studied [Lew *et al.*, 2006]. However, much research effort is devoted to the single-media retrieval, like image retrieval, video retrieval and audio retrieval. To further model multimedia data, some methods are proposed to combine different media types of data as [Znaidia *et al.*, 2012; Liu *et al.*, 2010], which can provide diverse information, and users can get results with more than one media type. In these methods, the retrieval results and the user query must share the same media type. For instance, users submit an image/text pair, and get the relevant pairs as the results, which restricts the flexibility of information retrieval. As digital media can be found and generated everywhere, users would like to submit queries of any media type, and get the relevant results with all media types.

Under this situation, the cross-media retrieval has become increasingly important, which makes it possible for users to submit any media types at hand, and get the relevant results with different media types. The traditional cross-media retrieval methods mainly lie on the common space learning. A representative method is the canonical correlation analysis (CCA) [Hotelling, 1936], which learns a subspace to maximize the correlation between data of different media types, and is widely used for modeling multimedia data [Hardoon *et al.*, 2004; Bredin and Chollet, 2007; Klein *et al.*, 2015]. Some CCA-based methods attempt to combine CCA with other information, such as semantic categories [Rasiwasia *et al.*, 2010]. An alternative method is Cross-modal Factor Analysis (CFA) approach [Li *et al.*, 2003]. This method finds the projection functions for different media types, by which it minimizes the Frobenius norm between the pairwise data in the common space. Zhai et al. [Zhai *et al.*, 2013] propose to learn projection functions by the metric learning, and this method is further improved as Joint Representation Learning (JRL) [Zhai *et al.*, 2014] by adding other information such as semantic categories and semi-supervised information. Some methods for image annotation as [Weston *et al.*, 2011] also learn a common space. The above methods obtain promising improvement, but they are mostly based on linear projection, which cannot fully model the intrinsic correlation of cross-media data. Another kind of method aims to extend the single-modal topic model for the joint distribution of topics in different media types, such as Correspondence LDA (corr-LDA) [Blei and Jordan, 2003], but they mostly take strong assumptions on the topic

*Corresponding author.

distribution, which may not be satisfied under the real condition.

Inspired by the progress of DNN in single-media retrieval and classification such as image classification [Krizhevsky A, 2012], the DNN has been applied to cross-media retrieval for converting the cross-media data to the shared representation, which is used to measure the similarity of cross-media data. Ngiam et al. [Ngiam *et al.*, 2011] apply an extension of Restricted Boltzmann Machine (RBM) to get the shared representation. In this work, Bimodal Autoencoders (Bimodal AE) is proposed, in which the inputs of different media types pass through a shared code layer to get the shared representation. Following this idea, some similar network structures are proposed, and achieve progress in modeling the cross-media data [Zhang *et al.*, 2014; Kim *et al.*, 2012; Srivastava and Salakhutdinov, 2012b; Wang *et al.*, 2015]. DeViSE [Frome *et al.*, 2013] uses a linear projection layer to project the image representations to the text representations from a pre-trained visual model and a language model. Deep CCA [Andrew *et al.*, 2013; Yan and Mikolajczyk, 2015] is a non-linear extension of CCA, which learns two separate correlated deep encodings. Feng et al. [Feng *et al.*, 2014] propose Correspondence Autoencoder (Corr-AE) to simultaneously model the reconstruction error and the correlation loss, and Wang et al. [Wang *et al.*, 2014] propose to use Stacked Autoencoders (SAE) for cross-media retrieval, which has two coupled subnetworks like [Feng *et al.*, 2014].

These methods are mainly two-stage learning methods: the first stage is to generate the separate representation for each media type, and the existing methods only model the intra-media information but ignore the inter-media correlation as [Srivastava and Salakhutdinov, 2012a; Feng *et al.*, 2014] with the rich complementary context to the intra-media information. The second stage is to get the shared representation by learning the cross-media correlation, and the existing methods learn the shared representation through a shallow network structure, which cannot fully capture the complex cross-media correlation. For addressing the above problems, we propose the cross-media multiple deep network (CMDN) to exploit the complex and rich cross-media correlation by hierarchical learning. In the first stage, CMDN jointly learns two kinds of complementary separate representation for each media type, instead of only intra-media separate representation of the previous work. Cross-media retrieval focuses on the correlation between different media types, so the inter-media representation can provide important hints and should be preserved. In the second stage, as there are two complementary separate representations for each media type, we hierarchically combine the separate representations in a deeper two-level network so that the inter-media and intra-media information can be jointly modeled to generate the shared representation. Compared to our approach, the existing methods only adopt a single-level network with only intra-media information as input. In addition, we learn the shared representations in a stacked network style to fully mine the complex cross-media correlation, which has better learning ability than only a shallow network structure of the existing methods. Experiment results show that the proposed CMDN model achieves better performance comparing with 7 state-of-the-art methods on 3 extensively used datasets (Wikipedia, NUS-WIDE-10k and Pascal Sentences).

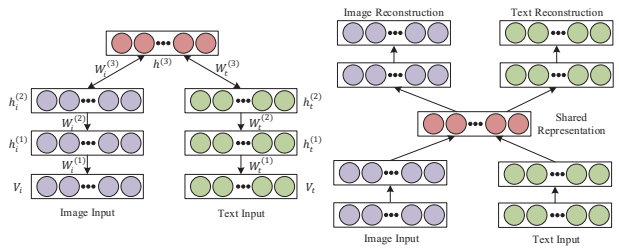

Figure 1: The Multimodal DBN and Bimodal Autoencoders.

2 Background

Various deep networks have been applied to cross-media retrieval. Considering the generality of cross-media retrieval, we adopt relatively general models for the features of different media types. Deep Belief Network (DBN) and autoencoders are the basic models chosen in CMDN according to our motivation of separate representation learning and hierarchical combination. Specifically, CMDN uses Multimodal DBN [Srivastava and Salakhutdinov, 2012a] and Stacked Autoencoders (SAE) [Vincent *et al.*, 2008] to model the intra-media information and inter-media correlation for learning complementary separate representations. Then CMDN applies joint RBM to combine the inter-media and intra-media separate representations, and uses Bimodal Autoencoder (Bimodal AE) [Ngiam *et al.*, 2011] in a stacked style to generate the shared representations for cross-media retrieval. In this section, we review these models briefly, which form the basis of our proposed CMDN model in Section 3.

The Multimodal Deep Belief Network (Multimodal DBN) has been widely used for learning a shared representation of multimodal data. It models data of each media type with a separate two-layer DBN, using image and text feature as input to model the distribution by Gaussian Restricted Boltzmann Machine (RBM) and Replicated Softmax model[Salakhutdinov and Hinton, 2009], which are widely used in cross-media retrieval [Srivastava and Salakhutdinov, 2012a; Feng *et al.*, 2014]. RBM is an undirected graphical model with visible units v and hidden units h connected to each other. An energy function and the joint distribution are defined as follows:

$$E(v,h;\theta) = -a^{\mathrm{T}}v - b^{\mathrm{T}}h - v^{\mathrm{T}}Wh \quad (1)$$

$$P(v,h;\theta) = \frac{1}{Z(\theta)}exp(-E(v,h;\theta)) \quad (2)$$

where θ contains three parameters a, b, w and $Z(\theta)$ is the normalizing constant. To form a Multimodal DBN, it combines the two DBN by learning a joint RBM on the top of them to get a shared representation. As shown in the left subfigure of Figure 1, the Multimodal DBN can model the joint distribution over data of multiple media types, which makes it possible to capture the inter-media correlation and will be used for the inter-media representation learning in the first stage of our CMDN model.

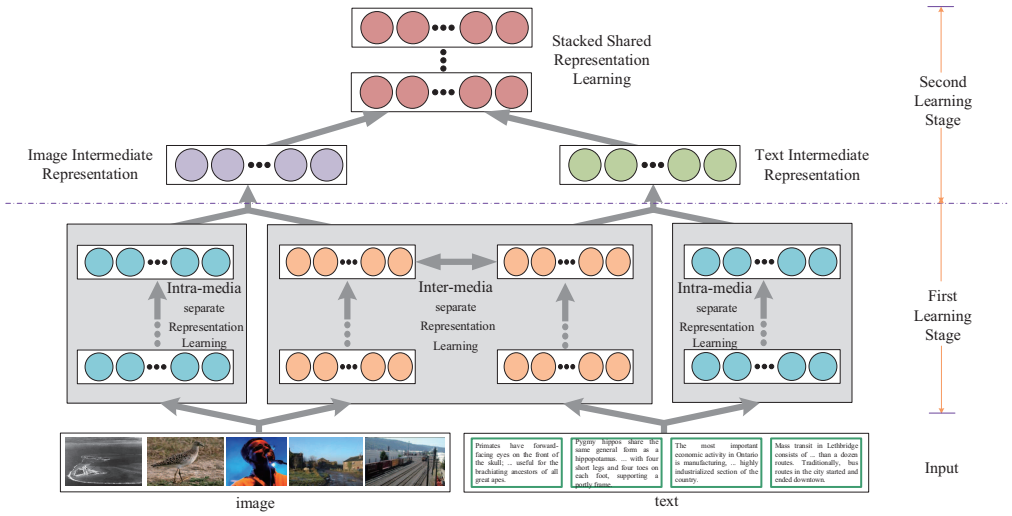

Figure 2: An overview of our CMDN model.

The Stacked Autoencoders (SAE) is a neural network which consists of multiple layers of autoencoders and requires less prior knowledge from the training data. SAE has several autoencoders which are trained in a bottom-up and layer-wise manner. The bottom autoencoder uses the original media features as input, and the high layer autoencoder uses the output generated from the bottom autoencoder. All these autoencoders are trained in turn as a pre-training stage, and then the whole neural network is fine-tuned based on a pre-trained model. SAE can get the high-level semantic representations and will be used for the intra-media representation learning in the first stage of our CMDN model.

The Bimodal Autoencoders (Bimodal AE) is a deep autoencoder network, as shown in the right subfigure of Figure 1, which takes multiple media types as input and has a middle layer generating the shared representation. The network aims to reconstruct both media types such as image and text, which minimizes the reconstruction error between the input features and the reconstruction representation. Bimodal AE can efficiently learn the higher-order correlations between different media types, and preserve the reconstruction information within each media type, which is useful for the shared representation learning in the second stage of our CMDN model.

3 Our CMDN Model

As shown in Figure 2, our CMDN model can be divided into two stages: in the first learning stage, we use Multimodal DBN to model the inter-media separate representation and SAE to model the intra-media separate representation for each media type. In the second learning stage, we use the two-level network including the joint RBM and Bimodal AE to get the final shared representation for each media type of the cross-media data.

Formally, given a dataset $D = \left\{D^{(i)}, D^{(t)}\right\}$ with the labeled multimedia content consists of $m+n$ media objects of two media types which are image and text. Here $D^{(i)} = \left\{x_p^{(i)}, y_p^{(i)}\right\}_{p=1}^{m}$ denotes the image data, and $D^{(t)} = \left\{x_q^{(t)}, y_q^{(t)}\right\}_{q=1}^{n}$ denotes the text data. $x_p^{(i)} \in \mathbb{R}^{d^{(i)}}$ is the p-th image data, and $x_q^{(t)} \in \mathbb{R}^{d^{(t)}}$ is the q-th text data, which are labeled as $y_p^{(i)}$ and $y_q^{(t)}$. And $d^{(i)}$, $d^{(t)}$ are the dimension of the image and text feature. The cross-media retrieval aims to retrieve relevant text y given the query of image x in the unlabeled dataset of text $T^{(t)} = \{y_1, ..., y_l\}$ and vice-versa.

3.1 Inter-media and Intra-media Separate Representation Learning

Multiple deep networks are used for the complementary inter-media and intra-media representation learning in the first stage.

Inter-media separate representation learning. We use a Multimodal DBN [Srivastava and Salakhutdinov, 2012a] for inter-media separate representation learning. First, we model each media type with a separate two-layer DBN, where Gaussian RBM is used to model the distribution over image features $X^{(i)} = \left\{x_p^{(i)}\right\}$, and Replicated Softmax is used to model the distribution over text features $X^{(t)} = \left\{x_p^{(t)}\right\}$. Then, to capture the inter-media correlation, we learn a joint RBM on the top of them to combine the two separate DBN, which can model the joint distribution over the data of two media types. We can perform alternating Gibbs sampling through the joint layer with the following conditional distributions:

$$P(h_t^{(3)} \mid h_i^{(2)}) = \sigma(W_i^{(3)} h_i^{(2)} + a_t) \quad (3)$$

$$P(h_i^{(3)} \mid h_t^{(2)}) = \sigma(W_t^{(3)} h_t^{(2)} + a_i) \quad (4)$$

where $\sigma(x) = 1/(1+e^{-x})$. The sample $h_t^{(3)}$ and $h_i^{(3)}$ can be used to generate a distribution over each media type, the output of which would be denoted as X_{inter}^{i} and X_{inter}^{t} used to be the inter-media separate representation.

Intra-media separate representation learning. We adopt the SAE [Vincent *et al.*, 2008] for intra-media separate representation learning. We obtain one SAE for each media type

and train each SAE independently. The input features $X^{(i)}$ and $X^{(t)}$ are the same with the Multimodal DBN, while $X_{2h}^{(i)}$ and $X_{2h}^{(t)}$ are the reconstruction of $X^{(i)}$ and $X^{(t)}$. The SAE for image and text data which consists of h layers of autoencoders can be trained separately by minimizing the objective function as follows:

$$L(X^{(i)}) = L_r(X^{(i)}, X_{2h}^{(i)}) + \alpha \sum_{p=1}^{h} (\|W_{ie}^p\|_2^2 + \|W_{id}^p\|_2^2) \tag{5}$$

$$L(X^{(t)}) = L_r(X^{(t)}, X_{2h}^{(t)}) + \beta \sum_{p=1}^{h} (\|W_{te}^p\|_2^2 + \|W_{td}^p\|_2^2) \tag{6}$$

where the average reconstruction error is denoted as $L_r(X^{(i)}, X_{2h}^{(i)})$ and $L_r(X^{(t)}, X_{2h}^{(t)})$, while W_{ie}, W_{id} and W_{te}, W_{td} are the parameters of the activation function of the encoder and decoder. By minimizing the reconstruction error, we can get the latent features to be the intra-media separate representation X_{intra}^i for image, and X_{intra}^t for text, which can preserve the original characteristic of each media type and get the high-level semantic representation.

3.2 Cross-media Shared Representation Learning with Two-level Networks

In the second learning stage, we have already obtained the multiple complementary separate representations X_{inter}^i, X_{inter}^t and X_{intra}^i, X_{intra}^t for each media type, which capture both the inter-media and intra-media information in the first learning stage. To get the shared representation, we hierarchically combine the separate representations by a new learning method which contains a deeper two-level network. It can also be divided into inter-media level and intra-media level.

On the first level of the network, we consider combining the inter-media and intra-media representations of each media type with a joint RBM. It jointly models the distribution over the representation captured from Multimodal DBN and SAE of one media. The joint distribution can be defined as follows:

$$P(v_1, v_2) = \sum_{h_1^{(1)}, h_2^{(1)}, h^{(2)}} P(h_1^{(1)}, h_2^{(1)}, h^{(2)}) \times \sum_{h_1^{(1)}} P(v_i \mid h_1^{(1)}) \times \sum_{h_2^{(1)}} P(v_t \mid h_2^{(1)}). \tag{7}$$

where v_1 denotes the inter-media separate representation X_{inter}^i while v_2 denotes X_{intra}^i for image. And as for the text media, this joint distribution are adopted on the inter-media separated representation X_{inter}^t and X_{intra}^t for text. We can collect these joint distributions as the intermediate representation of each media type, which are denoted as $Y^{(i)}$ for image and $Y^{(t)}$ for text, and they will be used as the input of the next level in the network.

On the second level of the network, we need to learn the shared representation for different media types. We use several Bimodal AE [Ngiam *et al.*, 2011] which can model the cross-media correlation at joint layer as well as the reconstruction information at the top layer. For training the network, we propose a stacked learning method having n (can be adjusted dynamically) Bimodal AE trained in a bottom-up method following [Ngiam *et al.*, 2011], and also added the additional label information. We use the intermediate representation $Y^{(i)}$and $Y^{(t)}$ as input for the bottom Bimodal AE, and its output $Z_{(1)}^{(i)}$ and $Z_{(1)}^{(t)}$ will be used as the input to propagate to the higher network so as to get $Z_{(2)}^{(i)}$, $Z_{(2)}^{(t)}$ as output and reduce the dimension to the half of the input at the same time, until we get $Z_{(n)}^{(i)}$ and $Z_{(n)}^{(t)}$ to be the final shared representation. The amount n of the network to be stacked in learning process can be adjusted according to the validation set.

We get the final shared representation through n stacked Bimodal AE which has better learning ability than only one Bimodal AE, so that the complementary inter-media and intra-media information can be jointly modeled to mine the complex cross-media correlation.

4 Experiment

In this section, we will introduce our experiments conducted on 3 cross-media datasets (Wikipedia, NUS-WDIE-10k and Pascal Sentences) with 7 state-of-the-art methods. In this paper, image and text are chosen for experiments because all the 3 datasets which are widely-used in multi-modal and cross-media research only include these two media types. However, our CMDN can perform the cross-media retrieval across various media types, such as video, audio and 3D model. To objectively and fully evaluate the results, we conduct two retrieval tasks: retrieving text by image (Image→Text) and retrieving image by text (Text→Image). For further verifying the effectiveness of our CMDN model, we also conduct experiments with two baseline methods: CMDN with only intra-media separate representation learning and CMDN with only inter-media separate representation learning. We can see from the experimental results that our method achieves consistently inspiring improvement on all retrieval tasks with all 3 datasets, which shows the generality of our method.

4.1 Datasets

Now the 3 datasets will be briefly introduced as follows. It should be noted that for fair comparison, in our experiments, the feature extraction and dataset partition of the training, testing and validation set are strictly according to [Feng *et al.*, 2014], and are also exactly the same with our approach and all the compared methods in the experiments.

Wikipedia dataset [Rasiwasia *et al.*, 2010]. Wikipedia dataset is from "feature articles" of Wikipedia with 10 semantic categories. The dataset contains a total of 2,866 image/text pairs, and is randomly split into a training set of 2,173 documents, a testing set of 462 documents and a validation set of 231 documents following [Feng *et al.*, 2014]. The image features are the concatenation of three parts: 1000 dimensional Pyramid Histogram of Words, 512 dimensional GIST, and 784 dimensional MPEG-7. The text representation is 3000 dimensional bag of words vector. The

dataset has been extensively used in the cross-media retrieval as [Rasiwasia *et al.*, 2010; Zhai *et al.*, 2013; 2014; Feng *et al.*, 2014].

NUS-WIDE-10k dataset [Chua *et al.*, 2009]. NUS-WIDE-10k is a subset of NUS-WIDE dataset. NUS-WIDE has about 270,000 images and their corresponding tags, categorized into 81 classes. NUS-WIDE-10k is constructed by selecting 10000 image/text pairs evenly from the 10 largest categories. In [Feng *et al.*, 2014], they randomly select 8000 documents for training, 1000 documents for testing and 1000 documents for validation evenly from the 10 categories, and the same partition is adopted in our experiment. The image features are the concatenation of 64 dimensional color histogram, 144 dimensional color correlogram, 73 dimensional edge direction histogram, 128 dimensional wavelet texture, 225 dimensional block-wise color moments and 500 dimensional SIFT-based bag of words features. The text is represented by a 1000 dimensional bag of words vector.

Pascal Sentences [Farhadi *et al.*, 2010]. Pascal Sentences dataset is an image/text dataset selected from 2008 PASCAL development kit. In this dataset, each image is described by 5 sentences. It has 1000 image/text pairs belonging to 20 categories totally. From each category we randomly select 40 documents for training, 5 documents for validation and 5 documents for testing following [Feng *et al.*, 2014]. The image representation is the same with Wikipedia dataset, and the text representation is 1000 dimensional bag of words vector.

4.2 Details of the Deep Architecture

Here we will introduce the details of our architecture in the experiment. The implementation of our CMDN model is based on deepnet [1]. As shown in Figure 2, our CMDN model mainly has three components and none of them is special for a single media. After feature extraction, any media can be represented as feature vectors and serve as the input of CMDN for cross-media retrieval.

In the inter-media separate representation learning, Multimodal DBN is used to model the inter-media correlation. For image input, there is a two-layer DBN which has 2048 hidden units on the first layer and 1024 hidden units on the second layer. For text input, we also use a two-layer DBN with 1024 hidden units on both layers. The joint layer is a joint RBM with 2048 hidden units taking the output of the two separate DBN as input. In the intra-media separate representation learning, we take the SAE model with three layers of autoencoder. For each media type, we first pre-train a 1024 dimensional single-layer autoencoder. The pre-trained autoencoder is used to initialize the first and third layer of our SAE model, and then the middle layer is initialized by the constant value of 512 dimensional number. As for the shared representation learning, on the first level, there are two joint RBM with 1024 hidden units for combining the inter-media and intra-media separate representations. On the top of each joint RBM, there is a three-layer feed-forward neural net with a Softmax layer, and the dimensional number of each layer is 1024. On the second level, we use BAE for getting the final shared representation. The reconstruction layers have the same dimensional numbers with the input. The dimensional number of the joint layer is half of the input, from which we get the final shared representation. In addition, there is also a Softmax layer connected to the joint layer for further optimization. As mentioned in Section 3, the number of BAE is adjusted according to the validation set.

In the training stage, the text and image input data should be organized as the pairs, but in the test stage they are actually independent. In addition, the numbers of the hidden units or dimensional numbers mentioned above are for the Wikipedia dataset as an example, which has 2296 dimensional image representation and 3000 dimensional text representation, and they need to be adjusted for other datasets according to the dimensional number of inputs. Except the above details, the other parameters remain the same with the already existing network implementation in deepnet.

4.3 Compared Methods and Evaluation Metrics

As the 3 datasets all have 2 media types (image and text), two retrieval tasks are conducted: retrieving text by image query (Image→Text) and retrieving image by text query (Text→Image). For example, in the Image→Text task, we take each image in the test set to retrieve all the text in the test set. For comparison purpose, we adopt 7 state-of-the-art cross-media retrieval methods, namely CCA, CFA, KCCA, Bimodal AE, Multimodal DBN, Corr-AE and JRL. The CCA, KCCA and CFA are the classical baselines. Multimodal DBN, Bimodal AE and Corr-AE are the DNN-based cross-media retrieval methods proposed recently. It should be noted that the source codes of the 3 DNN-based methods (Multimodal DBN, Bimodal AE and Corr-AE) are all from [Feng *et al.*, 2014] and they need validation set as input. JRL is the state-of-the-art method based on linear projection. Additionally, we use mean average precision (MAP) for all results for fully comprehensive evaluation, instead of MAP for the top 50 results in [Feng *et al.*, 2014]. The 7 compared methods are briefly introduced as follows:

- **CCA** [Hotelling, 1936]. CCA learns a common subspace for different media types, which is able to maximize the correlation of them.
- **CFA** [Li *et al.*, 2003]. CFA learns linear projection functions to project the cross-media data to one common space, which minimizes the Frobenius norm between the pairwise cross-media data.
- **KCCA** [Hardoon *et al.*, 2004]. KCCA holds the idea of first projecting the data into a higher-dimensional feature space and then performing CCA. In our experiments, the kernel functions used are polynomial kernel (Poly) and radial basis function (RBF).
- **Bimodal AE** [Ngiam *et al.*, 2011]. Bimodal AE is composed by a deep autoencoder network which takes multiple media types as input. It has a middle layer for the shared representation and is also required to reconstruct both media types.
- **Multimodal DBN** [Srivastava and Salakhutdinov, 2012a]. Multimodal DBN learns a joint representation of multimodal data, which models each media type with

[1] https://github.com/nitishsrivastava/deepnet

a separate two-layer DBN, and combines the two networks by learning a joint RBM on the top of them.

- **Corr-AE** [Feng *et al.*, 2014]. Corr-AE simultaneously models the reconstruction error and correlation loss by two subnetworks coupled at their code layers. Corr-AE has two extensions: Corr-Cross-AE and Corr-Full-AE. In our experiments, we adopt the best performance of the three models for comparison on MAP scores.
- **JRL** [Zhai *et al.*, 2014]. JRL simultaneously learns linear projections for different media types with semantic information, semi-supervised regularization and sparse regularization.

After obtaining the cross-media shared representation by our method and the compared methods, we get the ranking list of all the result with the cosine distance metric. The results will be evaluated by MAP scores and PR (precision-recall) curves, which can fairly and comprehensively evaluate the performance of our method and the compared methods. For fair and comprehensive evaluation, it should be noted that we adopt the MAP score and PR curves to compute *all the returned results* for our approach and all compared methods in the experiments. In [Feng *et al.*, 2014], they only take *the first 50 returned results* for evaluation, while the rest returned results are not considered. In our experiments, we use the source codes provided by [Feng *et al.*, 2014] for Bimodal AE, Multimodal DBN and Corr-AE, and adopt the MAP score and PR curves of *all the returned results* for them, which is the same with other compared methods and our CMDN model.

4.4 Experimental Results

Table 1 shows the MAP scores of our CMDN model and the compared methods on the 3 datasets. On the Wikipedia dataset, compared with the state-of-the-art method JRL, CMDN achieves inspiring MAP improvement from 0.311 to 0.359. NUS-WIDE-10k dataset is relatively large in the 3 datasets with 10000 data, and CMDN achieves the best results so far, improving the MAP score from 0.294 to 0.374. On the Pascal Sentences dataset, the 3 DNN-based compared methods have much better performance than that on Wikipedia dataset and NUS-WIDE-10k dataset, and our proposed CMDN model remains the best. Although the 3 datasets all have image and text, they have different data, feature dimensions and kinds. CMDN achieves consistently improvement on all retrieval tasks with all 3 datasets, which shows the generality of our method. Figures 3, 4 and 5 show the PR curves on the 3 datasets. Our proposed CMDN shows clear advantage over the compared methods in all figures, which demonstrates the effectiveness of hierarchy learning with CMDN model.

Table 2 shows the MAP scores of our baselines and the complete CMDN model. Intra-CMDN means CMDN with only intra-media separate representation learning, and Inter-CMDN means CMDN with only inter-media separate representation learning. In these two baselines, the Intermediate Representation in Figure 2 is the inter-media or intra-media representation, and the other architectures remain the same for fair comparison. The compared methods (DNN-based methods as [Feng *et al.*, 2014]) also adopt network combination and nonlinear projection, but our CMDN still outperforms the compared methods, showing that the combination of two complementary representations learning is the key element for result improvement. It should be noted that the Intra-media and Inter-media CMDN also outperform most of the compared methods, which demonstrates the effectiveness of the stacked network style in the shared representation learning.

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	CCA	0.124	0.120	0.122
	CFA	0.236	0.211	0.224
	KCCA(Poly)	0.200	0.185	0.193
	KCCA(RBF)	0.245	0.219	0.232
	Bimodal AE	0.236	0.208	0.222
	Multimodal DBN	0.149	0.150	0.150
	Corr-AE	0.280	0.242	0.261
	JRL	0.344	0.277	0.311
	our CMDN Model	**0.393**	**0.325**	**0.359**
NUS-WIDE -10k dataset	CCA	0.120	0.120	0.120
	CFA	0.211	0.188	0.200
	KCCA(Poly)	0.150	0.149	0.150
	KCCA(RBF)	0.232	0.213	0.223
	Bimodal AE	0.159	0.172	0.166
	Multimodal DBN	0.158	0.130	0.144
	Corr-AE	0.223	0.227	0.225
	JRL	0.324	0.263	0.294
	our CMDN Model	**0.391**	**0.357**	**0.374**
Pascal Sentences dataset	CCA	0.099	0.097	0.098
	CFA	0.187	0.216	0.202
	KCCA(Poly)	0.207	0.191	0.199
	KCCA(RBF)	0.233	0.249	0.241
	Bimodal AE	0.245	0.256	0.251
	Multimodal DBN	0.197	0.183	0.190
	Corr-AE	0.268	0.273	0.271
	JRL	0.300	0.286	0.293
	our CMDN Model	**0.334**	**0.333**	**0.334**

Table 1: The MAP scores of our CMDN model and the compared methods.

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	Intra-CMDN	0.303	0.284	0.294
	Inter-CMDN	0.303	0.294	0.299
	our CMDN Model	**0.393**	**0.325**	**0.359**
NUS-WIDE -10k dataset	Intra-CMDN	0.360	0.317	0.339
	Inter-CMDN	0.334	0.330	0.332
	our CMDN Model	**0.391**	**0.357**	**0.374**
Pascal Sentences dataset	Intra-CMDN	0.280	0.242	0.261
	Inter-CMDN	0.242	0.218	0.230
	our CMDN Model	**0.334**	**0.333**	**0.334**

Table 2: The MAP scores of our CMDN model and the baseline methods.

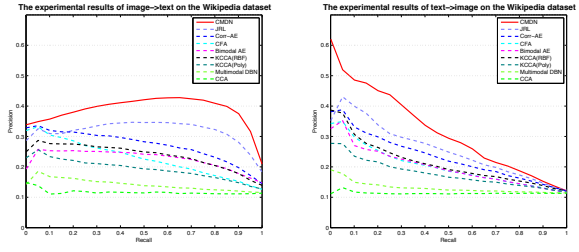

Figure 3: The PR curves on Wikipedia dataset.

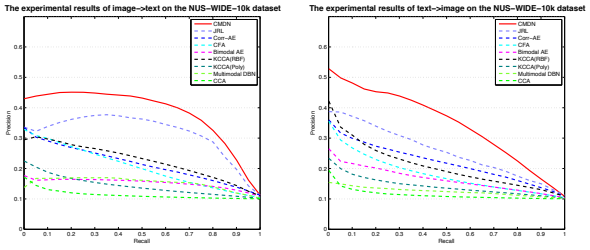

Figure 4: The PR curves on NUS-WIDE-10k dataset.

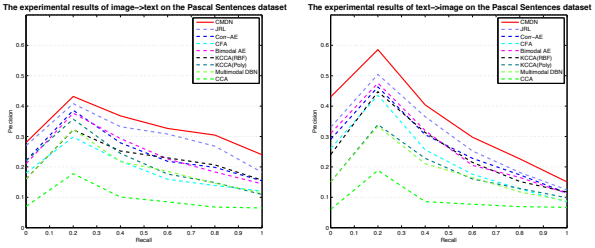

Figure 5: The PR curves on Pascal Sentences dataset.

From the above results, we can see that our proposed CMDN model has stable advantages. Compared to the methods without DNN (as CCA and CFA), our CMDN can effectively model the intrinsic correlations of cross-media data for the better learning ability of DNN. As for the DNN-based method (as Bimodal AE and Multimodal DBN), they only model the intra-media information in the first stage and construct the shallow networks for shared representation learning in the second stage, which limits their performance of cross-media retrieval. In addition, although JRL is based on linear projection, it can effectively incorporate semantic information with semi-supervised and sparse regularization, so achieves relatively high accuracy. However, our CMDN model outperforms JRL. This is because in the first stage, CMDN jointly models the intra-media and inter-media information for getting the complementary separate representation of each media type. In the second stage, CMDN hierarchically combines the inter-media and intra-media representations to further learn the intrinsic and rich cross-media correlation by a two-level network strategy, and finally get the shared representation by a stacked network style. Compared with the existing methods, CMDN can fully exploit the intra-media and inter-media information, so it can improve the retrieval accuracy.

5 Conclusion

In this paper, we have proposed CMDN model to get the cross-media shared representation by hierarchical learning. In the first learning stage, CMDN jointly models the intra-media and inter-media information for getting the complementary separate representation of each media type. In the second learning stage, CMDN hierarchically combines the inter-media and intra-media representations to further learn the rich cross-media correlation by a deeper two-level network strategy, and finally get the shared representation. Experiment results show the effectiveness of our method compared with state-of-the-art methods on 3 datasets. The future work lies in two aspects. First, we intend to incorporate semi-supervised information into our learning process, which will be helpful to get better high-level semantic representations to enrich the training data. Second, different representations can be obtained by different networks, we will still focus on applying and combining other deep networks to improve the retrieval results.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61371128 and 61532005, and National Hi-Tech Research and Development Program of China (863 Program) under Grant 2014AA015102.

References

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 1247–1255, 2013.

[Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling annotated data. In *International Conference on Research on Development in Information Retrieval (SIGIR)*, pages 127–134, 2003.

[Bredin and Chollet, 2007] Hervé Bredin and Gérard Chollet. Audio-visual speech synchrony measure for talking-face identity verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 233–236, 2007.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval (ACM-CIVR)*, 2009.

[Farhadi *et al.*, 2010] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision (ECCV)*, pages 15–29, 2010.

[Feng *et al.*, 2014] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACM international conference on Multimedia (ACM-MM)*, pages 7–16, 2014.

[Frome *et al.*, 2013] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2121–2129, 2013.

[Hardoon *et al.*, 2004] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[Hotelling, 1936] Harold Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[Kim *et al.*, 2012] Jungi Kim, Jinseok Nam, and Iryna Gurevych. Learning semantics with deep belief network for cross-language information retrieval. In *International Conference on Computational Linguistics (COLING)*, pages 579–588, 2012.

[Klein *et al.*, 2015] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446, 2015.

[Krizhevsky A, 2012] Hinton G Krizhevsky A, Sutskever I. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, 2012.

[Lew *et al.*, 2006] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.

[Li *et al.*, 2003] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. In *ACM international conference on Multimedia (ACM-MM)*, pages 604–611, 2003.

[Liu *et al.*, 2010] Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, and Han-Qing Lu. Coherent bag-of audio words model for efficient large-scale video copy detection. In *ACM International Conference on Image and Video Retrieval (ACM-CIVR)*, pages 89–96, 2010.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM international conference on Multimedia (ACM-MM)*, pages 251–260, 2010.

[Salakhutdinov and Hinton, 2009] Ruslan Salakhutdinov and Geoffrey E. Hinton. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1607–1614, 2009.

[Srivastava and Salakhutdinov, 2012a] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International Conference on Machine Learning (ICML)*, 2012.

[Srivastava and Salakhutdinov, 2012b] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2222–2230, 2012.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning (ICML)*, pages 1096–1103, 2008.

[Wang *et al.*, 2014] Wei Wang, Beng Chin Ooit, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. Effective multimodal retrieval based on stacked autoencoders. In *Conference on Very Large Data Bases (VLDB)*, pages 649–660, 2014.

[Wang *et al.*, 2015] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2291–2297, 2015.

[Weston *et al.*, 2011] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2764–2770, 2011.

[Yan and Mikolajczyk, 2015] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3441–3450, 2015.

[Zhai *et al.*, 2013] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1198–1204, 2013.

[Zhai *et al.*, 2014] Xiaohua Zhai, YuXin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semi-supervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 24:965–978, 2014.

[Zhang *et al.*, 2014] Hanwang Zhang, Yang Yang, Huanbo Luan, Shuicheng Yang, and Tat-Seng Chua. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *ACM international conference on Multimedia (ACM-MM)*, pages 187–196, 2014.

[Znaidia *et al.*, 2012] Amel Znaidia, Aymen Shabou, Hervé Le Borgne, Céline Hudelot, and Nikos Paragios. Bag-of-multimedia-words for image classification. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 1509–1512, 2012.