

Cross-modal Common Representation Learning by Hybrid Transfer Network

Xin Huang, Yuxin Peng*, and Mingkuan Yuan

Institute of Computer Science and Technology, Peking University, Beijing 100871, China
pengyuxin@pku.edu.cn

Abstract

DNN-based cross-modal retrieval is a research hotspot to retrieve across different modalities as image and text, but existing methods often face the challenge of insufficient cross-modal training data. In single-modal scenario, similar problem is usually relieved by transferring knowledge from large-scale auxiliary datasets (as ImageNet). Knowledge from such single-modal datasets is also very useful for cross-modal retrieval, which can provide rich general semantic information that can be shared across different modalities. However, it is challenging to transfer useful knowledge from *single-modal* (as image) source domain to *cross-modal* (as image/text) target domain. Knowledge in source domain cannot be directly transferred to both two different modalities in target domain, and the inherent cross-modal correlation contained in target domain provides key hints for cross-modal retrieval which should be preserved during transfer process. This paper proposes Cross-modal Hybrid Transfer Network (CHTN) with two subnetworks: *Modal-sharing transfer subnetwork* utilizes the modality in both source and target domains as a bridge, for transferring knowledge to both two modalities simultaneously; *Layer-sharing correlation subnetwork* preserves the inherent cross-modal semantic correlation to further adapt to cross-modal retrieval task. Cross-modal data can be converted to common representation by CHTN for retrieval, and comprehensive experiments on 3 datasets show its effectiveness.

1 Introduction

With the rapid progress of human civilization and technology, multimodal data as image, text, video, and audio has been rapidly increasing on the Internet, and gradually becomes the main form of information. In this situation, cross-modal retrieval has become an important application of artificial intelligence, which is a novel paradigm where the retrieval results and user query are relevant in semantics but of differ-

ent modalities. For example, user can submit an image to retrieve relevant text documents, and vice versa. Different from single-modal analysis as [Peng, 2015; Tang *et al.*, 2015], cross-modal retrieval faces the challenge that different modalities have inconsistent representations, and the existing mainstream is to learn common representation for them. Recent years, cross-modal retrieval based on Deep Neural Network (DNN) has become an active research topic [Ngiam *et al.*, 2011; Srivastava and Salakhutdinov, 2012; Feng *et al.*, 2014; Peng *et al.*, 2016], which aims to fulfill the DNN's strong ability of abstraction for dealing with complex cross-modal correlation.

Training data is important for the performance of DNN-based methods, but there is often insufficient training data for a specific task. In single-modal scenario, the problem of insufficient training data is usually relieved by the idea of transfer learning [Pan and Yang, 2010; Samdani and Yih, 2011; Chen and Zhang, 2013], which can transfer the knowledge of large-scale training data in source domain to target domain. For example, CNN model pre-trained on a subset of ImageNet [Deng *et al.*, 2009] with over 1,200,000 labeled images usually acts as the basic model for many problems in computer vision [Krizhevsky *et al.*, 2012], which is from ImageNet large-scale visual recognition challenge (ILSVRC) 2012. However, the knowledge transfer is usually performed within the same modality. Knowledge contained in such large-scale single-modal datasets is also very valuable for cross-modal retrieval because it can provide rich general semantic information, which can be shared by different modalities to facilitate cross-modal semantic learning. But it is a challenging problem to effectively transfer useful knowledge from such *single-modal source domain* to *cross-modal target domain*. For example, we aim to train a cross-modal retrieval model on a small image/text dataset as target domain, and have a large-scale image dataset ImageNet as source domain. On the one hand, because there are only labeled images in ImageNet, the knowledge contained in source domain cannot be directly transferred to both image and text modalities in the target domain. On the other hand, the inherent cross-modal semantic correlation contained in the image/text dataset provides key hints for cross-modal retrieval and should be preserved in the transfer process.

For addressing the above problems, this paper proposes Cross-modal Hybrid Transfer Network (CHTN), which is a

*Corresponding author.

unified architecture consisting of two subnetworks: modal-sharing transfer subnetwork and layer-sharing correlation subnetwork. *Modal-sharing transfer subnetwork* utilizes the modality contained in both source and target domains as a bridge, for propagating knowledge to both the two modalities in target domain simultaneously by single-modal transfer and cross-modal transfer. *Layer-sharing correlation subnetwork* focuses on preserving the inherent cross-modal semantic correlation to further adapt to the cross-modal retrieval task in the target domain. CHTN performs knowledge transfer and correlation learning at the same time, and can generate common representation for different modalities to perform cross-modal retrieval. For verifying the effectiveness of CHTN, we conduct cross-modal retrieval experiments on 3 widely-used datasets: Wikipedia, NUS-WIDE-10k, and Pascal Sentences. The experimental results show CHTN achieves the best accuracy among 10 state-of-the-art methods.

2 Related Work

2.1 Cross-modal Retrieval

Cross-modal retrieval is the basic research topic of this paper, which performs retrieval across different modalities as image and text. The current mainstream is common representation learning, which projects data of different modalities into a common space, and then the cross-modal similarity can be directly computed by distance measurement. The learning process is usually guided by co-existence (mainly pairwise correlation in the existing works) [Hotelling, 1936; Li *et al.*, 2003; Ngiam *et al.*, 2011; Feng *et al.*, 2014] and semantic correlation [Rasiwasia *et al.*, 2010; Kang *et al.*, 2015; Peng *et al.*, 2016]. According to the difference of basic models, existing works can be classified into traditional methods and DNN-based methods.

Traditional methods mainly learn linear projections for cross-modal common representation. Some representative methods are Canonical Correlation Analysis (CCA) [Hotelling, 1936], Cross-modal Factor Analysis (CFA) [Li *et al.*, 2003]. A few recent works take other information into consideration as semi-supervised regularization [Zhai *et al.*, 2014] and local group based priori [Kang *et al.*, 2015], but they are still based on linear projection. Recent years, inspired by the successful application of DNN in many single-modal tasks such as image classification, DNN-based methods for cross-modal retrieval have become an active research topic, which learn common representation with deep network as [Ngiam *et al.*, 2011; Kim *et al.*, 2012; Wang *et al.*, 2015; Peng *et al.*, 2016; ?].

Existing DNN-based cross-modal methods mostly take hand-crafted feature vector as input, and are trained only with the target cross-modal datasets, so often face the challenge of insufficient training data. Some recent works as [Wei *et al.*, 2017] adopt CNN as a component in their network, where the CNN is pre-trained with a large-scale dataset (ImageNet here), and fine-tuned with the images in cross-modal dataset. That is to say, the knowledge from single-modal dataset (ImageNet) is only utilized to improve the image representation, but the knowledge transfer between different modalities is not involved. Our CHTN aims to make full advantage

of the single-modal auxiliary dataset, by transferring semantic knowledge not only within the same modality, but also across different modalities simultaneously, so as to learn better cross-modal common representation and facilitate the performance of cross-modal retrieval.

2.2 Transfer Learning

Transfer learning [Pan and Yang, 2010] aims to propagate knowledge from source domain to target domain for relieving the problem of insufficient labeled training data, and has achieved success in wide applications [Samdani and Yih, 2011; Chen and Zhang, 2013]. The idea of transfer learning plays a key role in DNN-based methods, because the performance of DNN-based methods often relies on training data size, but there is usually insufficient training data for specific tasks. For addressing this, there are some DNN-based transfer learning methods as [Glorot *et al.*, 2011; Long *et al.*, 2015].

All the mentioned transfer learning technologies are limited in single-modal scenario, i.e., the source domain and target domain share the same single modality. Beyond these, some works have been proposed involving more than one modality. For example, the method of [Tsai *et al.*, 2016] proposes heterogeneous transfer from one modality to another, which is still a one-to-one transfer paradigm. Some methods as [Yang *et al.*, 2015] perform knowledge transfer between two domains with two modalities, but they assume that the two domains both share the two modalities. The method proposed by [Gupta *et al.*, 2016] aims to transfer knowledge from a large-scale labeled image dataset to a dataset of paired images, where there are two different kinds of images (specifically, RGB image and Depth image).

Our CHTN aims to address the knowledge transfer problem from *single-modal source domain* (as image dataset) to *cross-modal target domain* (as image/text dataset), which is different from the above mentioned works. Note that a recent work transitive hashing network (THN) [Cao *et al.*, 2016] has similar setting with our CHTN involving the knowledge transfer between single-modal and cross-modal datasets. However, THN and our CHTN aim at different problems with different focuses of method design: THN learns from an auxiliary cross-modal dataset to bridge two modalities from single-modal datasets; CHTN uses an auxiliary single-modal dataset to promote common representation learning in one cross-modal dataset. This transfer paradigm can effectively utilize useful knowledge from single-modal dataset to relieve the problem of insufficient labeled cross-modal data for benefiting applications as cross-modal retrieval.

3 Cross-modal Hybrid Transfer Network

The overview of CHTN is as shown in Figure 1, which is a unified architecture consisting of two subnetworks: *modal-sharing transfer subnetwork* and *layer-sharing correlation subnetwork*. In this paper, we take knowledge transfer from a large-scale image dataset (ImageNet) to an image/text dataset as an example to describe our methods.

We denote the single-modal source domain as $D^s = \{I^s\}$ with the labeled images $I^s = \{i_r^s, y_r^s\}_{r=1}^m$, i_r^s is the r -th image in source domain and its label is y_r^s . The cross-modal

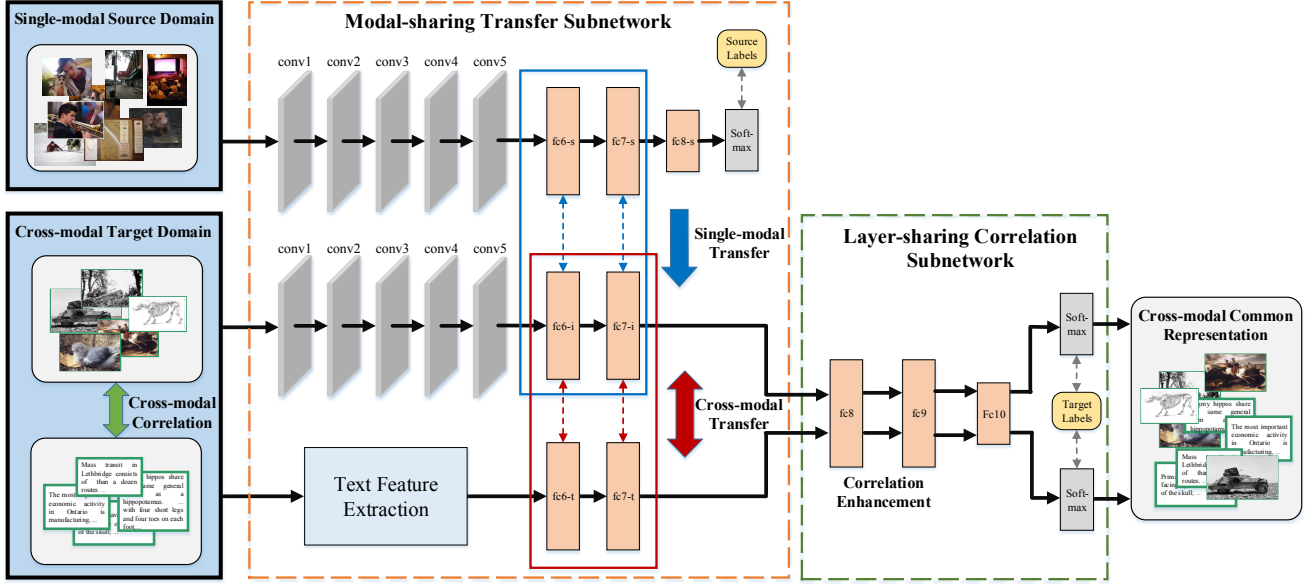


Figure 1: An overview of our Cross-modal Hybrid Transfer Network (CHTN).

target domain is denoted as $D^t = \{I^t, T^t, I_u^t, T_u^t\}$ with labeled images/text pairs $\{I^t, T^t\}$ and unlabeled image/text pairs $\{I_u^t, T_u^t\}$. Similarly, $I^s = \{i_p^s, y_p^s\}_{p=1}^{n_l}$ denotes the labeled images in target domain, and $T^s = \{t_q^s, y_q^s\}_{q=1}^{n_l}$ denotes the labeled text in the target domain. Correspondingly, the unlabeled images and text in target domain are denoted as $I_u^s = \{i_p^s\}_{p=n_l+1}^n$, and $T_u^s = \{t_q^s\}_{q=n_l+1}^n$. The aim is to transfer knowledge from D^s to D^t , and learn cross-modal common representation as $R^I = \{r_p^I\}_{p=n_l+1}^n$ for unlabeled images and $R^T = \{r_q^T\}_{q=n_l+1}^n$ for unlabeled text in target domain.

3.1 Modal-sharing Transfer Subnetwork

Modal-sharing transfer subnetwork is a hybrid architecture for transferring knowledge from image source domain to image/text target domain. It firstly uses convolutional layers to produce convolutional feature maps and receive representation of text (as BoW vector). Then the image feature maps and text representation will pass through fully-connected layers, where the knowledge transfer is performed. This subnetwork can be further viewed as two parts: single-modal knowledge transfer and cross-modal knowledge transfer, where the image pathway for target domain is shared by both the two parts, acting as a bridge for performing knowledge transfer from images in source domain to both the images and text in target domain.

Single-modal knowledge transfer. It has been shown in many literatures that the model of DNN has considerable generalization ability. However, when we use DNN in a new domain, the domain discrepancy still exists [Yosinski *et al.*, 2014]. To perform single-modal knowledge transfer from images in source domain to images in target domain, the key

problem is the domain discrepancy between them.

Following [Long *et al.*, 2015], we use feature adaptation method [Gretton *et al.*, 2012] which aims to minimize the Maximum Mean Discrepancy (MMD) of the same modality between source and target domains. By minimizing MMD, we can let a transferred model match the target domain distributions effectively, and so achieve the knowledge transfer within the same modality. We denote MMD between distribution a of images from source domain $\{i^s\}$ and distribution b of images from target domain $\{i^t\}$ by $d_k(a, b)$, so the squared formulation of MMD between a and b in reproducing kernel Hilbert space (RKHS) \mathcal{H}_k can be defined as:

$$d_k^2(a, b) \triangleq \|\mathbf{E}_a[\phi(i^s)] - \mathbf{E}_b[\phi(i^t)]\|_{\mathcal{H}_k}^2 \quad (1)$$

where ϕ refers to the representation of a certain layer in deep neural network and the mean embedding of distribution a in \mathcal{H}_k is $\mu_k(a)$, so that $\mathbf{E}_{X \sim a} f(X) = \langle f(X), \mu_k(a) \rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$. So the single-modal transfer loss can be written as:

$$Loss_{Single} = \sum_{l=l_6}^{l_7} d_k^2(I^s, I^t) \quad (2)$$

where the single-modal transfer loss at one layer is denoted as $d_k^2(I^s, I^t)$. l_6 and l_7 are corresponding layers in the network (i.e., fc6-s/fc6-i and fc7-s/fc7-i).

Moreover, the layers of source domain require fine-tuning itself on the source labeled instances. In this way, the labels of source domain can be used to provide supplementary supervision information and guide the single-modal knowledge transfer. The source domain supervision loss is as follows:

$$Loss_{Source} = -\frac{1}{m} \sum_{r=1}^m f_s(i_r^s, y_r^s, \theta^s) \quad (3)$$

where $f_s(i_r^s, y_r^s, \theta^s)$ is the softmax loss function as:

$$f_s(x, y, \theta) = \sum_{j=1}^c 1\{y=j\} \log \frac{e^{\theta_j x}}{\sum_{l=1}^c e^{\theta_l x}} \quad (4)$$

where θ are parameters of the network. y denotes the label of instance x and c is the number of all possible classes of instance x . $1\{y=j\}$ is an indicator function, which is equal to 1 if $y=j$, otherwise 0.

By minimizing the single-modal transfer loss and the source domain supervision loss, we can effectively decrease the domain discrepancy of image modality between the two domains, and perform single-modal knowledge transfer.

Cross-modal knowledge transfer. The aforementioned single-modal knowledge transfer can only propagate knowledge from images in source domain to images in target domain. But because there exists cross-modal correlation between image and text in the target domain, it is possible to further achieve cross-modal knowledge transfer between image and text.

The image and text in an image/text pair are closely corresponded with each other in cross-modal dataset, so the knowledge can be effectively shared within each pair. The main idea of cross-modal knowledge transfer is to let the representations from high-level layers of pairwise data similar to each other, which is an intuitive idea as [Li *et al.*, 2003; Feng *et al.*, 2014]. Specifically, we adopt Euclidean distance between cross-modal high-level layers (i.e., fc6-i/fc6-t and fc7-i/fc7-t) as the risk of cross-modal knowledge transfer, and denote the cross-modal discrepancy between an image instance i_p^t and its paired text instance t_p^t both from target domain as:

$$d_c^2(i_p^t, t_p^t) = \|\phi(i_p^t) - \phi(t_p^t)\|^2 \quad (5)$$

Then we get the cross-modal transfer loss as:

$$Loss_{Cross} = \sum_{l=l_6}^{l_7} \sum_{p=1}^{n_l} d_c^2(i_p^t, t_p^t) \quad (6)$$

By minimizing this cross-modal transfer loss, we can achieve cross-modal knowledge transfer by reducing the discrepancy between representations of pairwise cross-modal data from high-level layers.

It should be noted that the modality shared by both source domain and target domain (image here) acts as a bridge to link the single-modal and cross-modal knowledge transfer. So these two kinds of transfer can be considered simultaneously, which leads to a hybrid knowledge transfer style. By optimization with Stochastic Gradient Descent (SGD), we can minimize these loss functions and perform knowledge transfer from single-modal source domain to cross-modal target domain in training stage.

3.2 Layer-sharing Correlation Subnetwork

The inherent semantic correlation in cross-modal target domain is the key hint for bridging different modalities, which is the essential property of cross-modal dataset. Because our aim is to perform cross-modal retrieval, we further design a

layer-sharing correlation subnetwork to enhance such correlation, and make the model more adapted to cross-modal retrieval task.

This subnetwork is a simple but effective task-specific structure for cross-modal retrieval, with layers shared by both image and text. As shown in Figure 1, we use two shared fully-connected layers (fc8 and fc9) to construct it. Both the representations of image and text output from modal-sharing transfer subnetwork will pass through the two fully-connected layers, and then there is a common classification layer.

Because the parameters of fc8 and fc9 are shared by both image and text, we can use the supervision information in cross-modal target domain to ensure the semantic correlation of different modalities. Given the labels of two paired modalities in target domain, the correlation loss is:

$$Loss_{Correlation} = -\frac{1}{n_l} \sum_{p=1}^{n_l} (f_s(i_p^t, y_p^t, \theta^t) + f_s(t_p^t, y_p^t, \theta^t)) \quad (7)$$

where f_s is the softmax loss function in Equation 4. $f_s(i_p^t, y_p^t, \theta^t)$ is the target image supervision term, while $f_s(t_p^t, y_p^t, \theta^t)$ is the corresponding target text supervision term. After standard back-propagation with multiple iterations, we can minimize this correlation loss, and the semantic information in cross-modal target domain will be fully exploited for guiding cross-modal retrieval task.

It should be noted that the modal-sharing subnetwork and layer-sharing correlation subnetwork are in a unified architecture, so the two subnetworks can be trained jointly and boost each other. Especially, the training of modal-sharing transfer network can also be guided by correlation loss, which makes the whole knowledge transfer processing further adapt to cross-modal retrieval task in target domain. In test stage, we take the predicted probability vectors as the final common representation R^I and R^T of image and text, for performing cross-modal retrieval.

4 Experiments

4.1 Details of the Deep Architecture

In this section, we present the details of our CHTN in the experiments. In the modal-sharing transfer subnetwork, there are three pathways: source image pathway, target image pathway, and target text pathway. In the source image pathway and target image pathway, we take five convolutional layers (conv1-conv5) of AlexNet [Krizhevsky *et al.*, 2012] pre-trained on ImageNet from the Caffe¹ Model Zoo. It receives the images resized as 256×256 and generates convolutional feature maps (pool5) for images, and the five convolution layers are regarded as general layers to be frozen in training stage. All the three pathways of modal-sharing transfer subnetwork have two fully-connected layers, and all of them have $4,096 \times 4,096$ units. The base learning rates of all the fully-connected layers are set to be 0.01.

The MMD loss layers are implemented following [Long *et al.*, 2015] between the two corresponding fully-connected

¹<http://caffe.berkeleyvision.org>

layers of source image pathway and target image pathway, which performs single-modal knowledge transfer. Similarly, there are two contrastive loss layers from Caffe between the two fully-connected layers of target image pathway and target text pathway to realize cross-modal knowledge transfer. Moreover, because the magnitude of $Loss_{Cross}$ is much larger than those of $Loss_{Single}$, $Loss_{Source}$, and $Loss_{Correlation}$ (about 1,000 times), we set its weight as 0.001, and those of $Loss_{Single}$, $Loss_{Source}$, and $Loss_{Correlation}$ are all 1. These parameter settings can be easily adjusted for other datasets in the implementation of network. After an fc8-s layer and a softmax loss layer of source image pathway, we can calculate softmax loss function for images of source domain in training stage.

Layer-sharing correlation subnetwork consists of two fully-connected layers (fc8 and fc9), for which the base learning rates are set to be 0.01 and have $4,096 \times 4,096$ units. The two fully-connected layers receive the outputs from the last fully-connected layer of both target image pathway and target text pathway in modal-sharing transfer subnetwork. Finally, after a fully-connected layer and a softmax layer, we can calculate softmax loss function for images and text of target domain in training stage, and get the probability vector as common representation in test stage.

4.2 Dataset Introduction

In the experiments, ImageNet serves as the single-modal source domain, and we adopt a widely-used subset with over 1,200,000 labeled images [Krizhevsky *et al.*, 2012] from ILSVRC 2012. We perform knowledge transfer from ImageNet to 3 cross-modal datasets as target domains respectively, and conduct cross-modal retrieval on them, namely Wikipedia, NUS-WIDE-10k and Pascal Sentences. For fair comparison, we strictly take the same dataset partition according to [Feng *et al.*, 2014; Peng *et al.*, 2016] for our CHTN and all the compared methods in the experiments.

Wikipedia dataset [Rasiwasia *et al.*, 2010] is the most widely-used dataset for cross-modal retrieval evaluation as [Zhai *et al.*, 2014; Feng *et al.*, 2014; Peng *et al.*, 2016], which is constructed from the Wikipedia “featured articles”. This dataset has 2,866 image/text pairs classified into 10 high-level semantic categories: art, biology, geography, history, literature, media, music, royalty, sport, and warfare. In each image/text pair, the text contains several paragraphs of description for the image, so they have close correlation, and each pair exclusively belongs to one of the 10 categories. The dataset is randomly split into training set with 2,173 pairs, test set with 462 pairs, and validation set with 231 pairs.

NUS-WIDE-10k dataset [Feng *et al.*, 2014] is a subset of NUS-WIDE dataset [Chua *et al.*, 2009]. NUS-WIDE dataset is a relatively large-scale image/tag dataset with about 270,000 images. Each image has several corresponding text tags, which are used as the text modality in the experiments. NUS-WIDE dataset has 81 categories, but there exist overlaps among the categories. NUS-WIDE-10k dataset is constructed by selecting 10,000 image/text pairs evenly from 10 largest categories of NUS-WIDE dataset, under conditions that each pair exclusively belongs to one of the 10 categories, so there are 1,000 pairs for each category. The dataset is randomly

split into training set with 8,000 pairs, test set with 1,000 pairs, and validation set with 1,000 pairs evenly from the 10 categories.

Pascal Sentences dataset [Rashtchian *et al.*, 2010] is also an image/text dataset which is selected from 2008 PASCAL development kit. This dataset has 1,000 images evenly belonging to 20 categories, and each image has 5 description sentences. Similar to NUS-WIDE-10k dataset, Pascal Sentences dataset is randomly split into training set with 800 pairs, test set with 100 pairs, and validation set with 100 pairs evenly from the 20 categories.

It should be noted that among the compared methods, CMDN, Corr-AE, Bimodal AE, and Multimodal DBN need validation set for parameter adjustment, while the other methods including our CHTN don’t. So for the methods except CMDN, Corr-AE, Bimodal AE, and Multimodal DBN, validation set is not used for the whole training and test stages.

4.3 Evaluation Metrics

In the experiments, two retrieval tasks are conducted: retrieving text by image query (Image→Text) and retrieving images by text query (Text→Image). Specifically, we first obtain the common representation for the images and text in the test set with CHTN and all the compared methods. Then we take one of the images in test set as query, compute the cross-modal similarity with all text in test set by cosine distance, and evaluate the ranking list by mean average precision (MAP), and vice-versa. It should be noted that the MAP score is computed for *all the retrieval results* for comprehensive evaluation.

The MAP scores are computed as the mean of average precision (AP) for all queries, and AP is computed as:

$$AP = \frac{1}{R} \sum_{k=1}^n \frac{R_k}{k} \times rel_k \quad (8)$$

where R denotes relevant item number in test set (according to the label in our experiments), R_k denotes the relevant item number in top k results, n denotes the test set size, and $rel_k = 1$ means the k -th result is relevant, and 0 otherwise.

4.4 Compared Methods and Input Settings

Totally 10 state-of-the-art methods are compared in the experiments: CCA [Hotelling, 1936], CFA [Li *et al.*, 2003], KCCA (with Gaussian kernel and polynomial kernel) [Hardoon *et al.*, 2004], Bimodal AE [Ngiam *et al.*, 2011], Multimodal DBN [Srivastava and Salakhutdinov, 2012], Corr-AE [Feng *et al.*, 2014], JRL [Zhai *et al.*, 2014], LGCFL [Kang *et al.*, 2015], CMDN [Peng *et al.*, 2016] and Deep-SM [Wei *et al.*, 2017]. Among these, CCA, CFA, KCCA, JRL, LGCFL are traditional methods, and Bimodal-AE, Multimodal DBN, Corr-AE, CMDN and Deep-SM are DNN-based methods.

For image, CHTN has the end-to-end processing ability and directly takes the image pixels as input. However, all the compared methods except Deep-SM and our CHTN can only take extracted feature vector as input. So for fair comparison, we use the same AlexNet adopted by CHTN, and further fine-tuned to convergence with the images in each dataset as feature extractor for them. For text, we



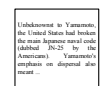

| Query | Top-5 Retrieved Results | | | |
|--|---|---|---|---|
|  <p>Category: Music</p> | <p>Marshaled her had a stirring... Although most (it) hands in... The first problem the... "Shake a fist" was written... Following the lead of...</p> | <p>The 12th-century historian... Henry of Huntingdon, in his... Writers of the nineteenth... English was daily... An expedition... George III himself hoped...</p> | <p>Unbeknownst to Yamamoto... the United States had broken... Category: Warfare</p> | <p>The track has just... the inception of 1990 East... in the 12th East... the name the first... since Graham Gough in 1990... in... a century... recessive East... English... March... 2007... 11...</p> |
|  <p>Category: Royalty</p> | <p>Unbeknownst to Yamamoto... the United States had broken... Category: Warfare</p> | <p>The track has just... the inception of 1990 East... in the 12th East... the name the first... since Graham Gough in 1990... in... a century... recessive East... English... March... 2007... 11...</p> | <p>Unbeknownst to Yamamoto... the United States had broken... Category: Warfare</p> | |
|  <p>Category: Warfare</p> | <p>The track has just... the inception of 1990 East... in the 12th East... the name the first... since Graham Gough in 1990... in... a century... recessive East... English... March... 2007... 11...</p> | <p>Unbeknownst to Yamamoto... the United States had broken... Category: Warfare</p> | <p>The track has just... the inception of 1990 East... in the 12th East... the name the first... since Graham Gough in 1990... in... a century... recessive East... English... March... 2007... 11...</p> | |
|  <p>Category: Sport</p> | <p>Unbeknownst to Yamamoto... the United States had broken... Category: Warfare</p> | <p>The track has just... the inception of 1990 East... in the 12th East... the name the first... since Graham Gough in 1990... in... a century... recessive East... English... March... 2007... 11...</p> | <p>Unbeknownst to Yamamoto... the United States had broken... Category: Warfare</p> | |

Figure 2: Some examples of cross-modal retrieval results on Wikipedia dataset obtained by CHTN. All the top-5 retrieval results are correct.

take exactly the same input feature as [Feng *et al.*, 2014; Peng *et al.*, 2016] for all compared methods. In detail, the input feature is 3,000 dimensional BoW vector for Wikipedia dataset, and 1,000 dimensional BoW vector for both NUS-WIDE-10k and Pascal Sentences datasets.

4.5 Experimental Results

Table 1 shows the MAP scores of our CHTN and the compared methods on the 3 datasets. It can be seen that on Wikipedia dataset, CHTN achieves an inspiring accuracy improvement from 0.402 to 0.470, compared with the state-of-the-art method Deep-SM. Figure 2 shows some examples of cross-modal retrieval results on Wikipedia dataset. On NUS-WIDE-10k dataset, CHTN keeps the best and achieves the MAP score of 0.517. The result trend on Pascal Sentences dataset is much different from Wikipedia and NUS-WIDE-10k datasets. On Pascal Sentences dataset, the compared methods Bimodal AE and Multimodal DBN achieve similar accuracy with Deep-SM, and CMDN obtains the best accuracy among all compared methods. However, the accuracy of CHTN is still the highest. It can be also seen that the performance of compared methods is not stable. For example, on Wikipedia dataset CMDN has a very clear advantage over Multimodal DBN, but on Pascal Sentences dataset they are comparable. Our CHTN keeps the best on the 3 datasets, which shows the hybrid transfer paradigm has generality, and it effectively transfers the knowledge from single-modal source domain to cross-modal target domain for learning better common representation and improving the accuracy of cross-modal retrieval.

Table 2 shows the MAP scores of our baselines and the complete CHTN. CHTN (OnlyCross) means we remove the source image pathway in modal-sharing transfer subnetwork, and there is no layer-sharing correlation subnetwork (the fc7-i and fc7-t layers are directly connected to classification layers). This is a very basic and intuitive way to perform cross-modal retrieval learning with pairwise correlation and seman-

| Dataset | Method | Task | | |
|--------------------------|----------------|--------------|--------------|---------|
| | | Image→Text | Text→Image | Average |
| Wikipedia dataset | CCA | 0.176 | 0.178 | 0.177 |
| | CFA | 0.330 | 0.306 | 0.318 |
| | KCCA(Poly) | 0.230 | 0.224 | 0.227 |
| | KCCA(Gaussian) | 0.357 | 0.328 | 0.343 |
| | Bimodal AE | 0.301 | 0.267 | 0.284 |
| | Multimodal DBN | 0.204 | 0.145 | 0.175 |
| | Corr-AE | 0.373 | 0.357 | 0.365 |
| | JRL | 0.408 | 0.353 | 0.381 |
| | LGCFL | 0.416 | 0.360 | 0.388 |
| | CMDN | 0.409 | 0.364 | 0.387 |
| | Deep-SM | 0.458 | 0.345 | 0.402 |
| our CHTN | 0.508 | 0.432 | 0.470 | |
| NUS-WIDE -10k dataset | CCA | 0.159 | 0.189 | 0.174 |
| | CFA | 0.299 | 0.301 | 0.300 |
| | KCCA(Poly) | 0.129 | 0.157 | 0.143 |
| | KCCA(Gaussian) | 0.295 | 0.162 | 0.229 |
| | Bimodal AE | 0.234 | 0.376 | 0.305 |
| | Multimodal DBN | 0.178 | 0.144 | 0.161 |
| | Corr-AE | 0.306 | 0.340 | 0.323 |
| | JRL | 0.410 | 0.444 | 0.427 |
| | LGCFL | 0.408 | 0.374 | 0.391 |
| | CMDN | 0.410 | 0.450 | 0.430 |
| | Deep-SM | 0.389 | 0.496 | 0.443 |
| our CHTN | 0.518 | 0.516 | 0.517 | |
| Pascal Sentences dataset | CCA | 0.110 | 0.116 | 0.113 |
| | CFA | 0.341 | 0.308 | 0.325 |
| | KCCA(Poly) | 0.271 | 0.280 | 0.276 |
| | KCCA(Gaussian) | 0.312 | 0.329 | 0.321 |
| | Bimodal AE | 0.404 | 0.447 | 0.426 |
| | Multimodal DBN | 0.438 | 0.363 | 0.401 |
| | Corr-AE | 0.411 | 0.475 | 0.443 |
| | JRL | 0.416 | 0.377 | 0.397 |
| | LGCFL | 0.381 | 0.435 | 0.408 |
| | CMDN | 0.458 | 0.444 | 0.451 |
| | Deep-SM | 0.440 | 0.414 | 0.427 |
| our CHTN | 0.467 | 0.477 | 0.472 | |

Table 1: MAP scores of our CHTN and compared methods.

| Dataset | Method | Task | | |
|--------------------------|------------------|--------------|--------------|--------------|
| | | Image→Text | Text→Image | Average |
| Wikipedia dataset | CHTN (OnlyCross) | 0.465 | 0.407 | 0.436 |
| | CHTN (NoShare) | 0.483 | 0.422 | 0.453 |
| | CHTN (NoSrcSp) | 0.489 | 0.415 | 0.452 |
| | our CHTN | 0.508 | 0.432 | 0.470 |
| NUS-WIDE -10k dataset | CHTN (OnlyCross) | 0.360 | 0.406 | 0.383 |
| | CHTN (NoShare) | 0.488 | 0.442 | 0.465 |
| | CHTN (NoSrcSp) | 0.487 | 0.495 | 0.491 |
| | our CHTN | 0.518 | 0.516 | 0.517 |
| Pascal Sentences dataset | CHTN (OnlyCross) | 0.433 | 0.443 | 0.438 |
| | CHTN (NoShare) | 0.446 | 0.467 | 0.457 |
| | CHTN (NoSrcSp) | 0.435 | 0.457 | 0.446 |
| | our CHTN | 0.467 | 0.477 | 0.472 |

Table 2: MAP scores of our CHTN and the baselines.

tic information. CHTN (NoShare) means there is no layer-sharing correlation subnetwork, and the fc7-i and fc7-t are directly connected to classification layers. CHTN (NoSrcSp) means there is no source supervision loss of Eqn. (3), which is designed to verify if the supervision information of source domain can provide complementary semantic information for the target domain. Except for the above differences, the rest parts of the three baselines keep the same with complete CHTN.

By comparing CHTN (OnlyCross) and CHTN (NoShare), we can see the single-modal transfer part provides important supplementary information to cross-modal transfer part, which leads to better accuracy. By comparing CHTN (NoShare) with complete CHTN, we can see the preserving of inherent cross-modal semantic correlation helps the deep model further adapt to cross-modal retrieval task in target domain. By comparing CHTN (NoSrcSp) with complete CHTN, we can see that although label space and task

of the two domains may be different, they can share relevant high-level semantic knowledge for improving retrieval accuracy. The above baseline experiments show the modal-sharing transfer architecture and layer-sharing correlation subnetwork both play important role in our CHTN, and they can boost each other for better accuracy.

5 Conclusion

In this paper, we have proposed a Cross-modal Hybrid Transfer Network (CHTN), for addressing the problem of knowledge transfer from single-modal source domain (as image) to cross-modal target domain (as image/text). CHTN can be viewed as a unified model with two subnetworks: Modal-sharing transfer subnetwork utilizes the modality in both source and target domains as a bridge, for transferring knowledge to both two modalities simultaneously; Layer-sharing correlation subnetwork focuses on preserving the inherent cross-modal semantic correlation to further adapt to cross-modal retrieval task in target domain. Experiments show that CHTN effectively utilizes the large-scale single-modal dataset to improve the cross-modal retrieval accuracy. The future works lie in two aspects. First, we intend to apply CHTN to other cross-modal applications like image caption to verify its generality. Second, we will focus on how to transfer knowledge from single-modal source domain to target domain with more than two modalities.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61371128 and 61532005.

References

- [Cao *et al.*, 2016] Zhangjie Cao, Mingsheng Long, and Qiang Yang. Transitive hashing network for heterogeneous multimedia retrieval. *arXiv*, abs/1608.04307, 2016.
- [Chen and Zhang, 2013] Zheng Chen and Weixiong Zhang. Domain adaptation with topical correspondence learning. *IJCAI*, pages 1280–1286, 2013.
- [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. *ACM-CIVR*, 2009.
- [Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. *CVPR*, pages 248–255, 2009.
- [Feng *et al.*, 2014] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. *ACM-MM*, pages 7–16, 2014.
- [Glorot *et al.*, 2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. *ICML*, pages 513–520, 2011.
- [Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.
- [Gupta *et al.*, 2016] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. *CVPR*, pages 2827–2836, 2016.
- [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.
- [Kang *et al.*, 2015] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia (TMM)*, 17(3):370–381, 2015.
- [Kim *et al.*, 2012] Jungi Kim, Jinseok Nam, and Iryna Gurevych. Learning semantics with deep belief network for cross-language information retrieval. *International Committee on Computational Linguistic (ICCL)*, pages 579–588, 2012.
- [Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, pages 1106–1114, 2012.
- [Li *et al.*, 2003] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. *ACM-MM*, pages 604–611, 2003.
- [Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. *ICML*, pages 97–105, 2015.
- [Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. *ICML*, pages 689–696, 2011.
- [Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010.
- [Peng *et al.*, 2016] Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. *IJCAI*, pages 3846–3853, 2016.
- [Peng, 2015] Yuxin Peng. Adaptive sampling with optimal cost for class-imbalance learning. *AAAI*, pages 2921–2927, 2015.
- [Rashtchian *et al.*, 2010] C. Rashtchian, P. Young, M. Hoshino, and J. Hockenmaier. Collecting image annotations using amazon’s mechanical turk. *North American Chapter of the Association for Computational Linguistics*, pages 139–147, 2010.
- [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. *ACM-MM*, pages 251–260, 2010.

- [Samdani and Yih, 2011] Rajhans Samdani and Wen-tau Yih. Domain adaptation with ensemble of feature groups. *IJCAI*, pages 1458–1464, 2011.
- [Srivastava and Salakhutdinov, 2012] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multi-modal data with deep belief nets. *ICML Workshop*, 2012.
- [Tang *et al.*, 2015] Jinhui Tang, Zechao Li, Meng Wang, and Ruizhen Zhao. Neighborhood discriminant hashing for large-scale image retrieval. *IEEE Transactions on Image Processing (TIP)*, 24(9):2827–2840, 2015.
- [Tsai *et al.*, 2016] Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Learning cross-domain landmarks for heterogeneous domain adaptation. *CVPR*, pages 5081–5090, 2016.
- [Wang *et al.*, 2015] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. *IJCAI*, pages 2291–2297, 2015.
- [Wei *et al.*, 2017] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics (TCYB)*, 47(2):449–460, 2017.
- [Yang *et al.*, 2015] Xiaoshan Yang, Tianzhu Zhang, and Changsheng Xu. Cross-domain feature learning in multimedia. *IEEE Transactions on Multimedia (TMM)*, 17(1):64–78, 2015.
- [Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *NIPS*, pages 3320–3328, 2014.
- [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(6):965–978, 2014.