Deep Cross-media Knowledge Transfer

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

Abstract

Cross-media retrieval is a research hotspot in multimedia area, which aims to perform retrieval across different media types such as image and text. The performance of existing methods usually relies on labeled data for model training. However, cross-media data is very labor consuming to collect and label, so how to transfer valuable knowledge in **existing data** *to* **new data** *is a key problem towards application. For achieving the goal, this paper proposes deep cross-media knowledge transfer (DCKT) approach, which transfers knowledge from a large-scale cross-media dataset to promote the model training on another small-scale cross-media dataset. The main contributions of DCKT are: (1)* **Two-level transfer architecture** *is proposed to jointly minimize the media-level and correlation-level domain discrepancies, which allows two important and complementary aspects of knowledge to be transferred: intra-media semantic and inter-media correlation knowledge. It can enrich the training information and boost the retrieval accuracy. (2)* **Progressive transfer mechanism** *is proposed to iteratively select training samples with ascending transfer difficulties, via the metric of cross-media domain consistency with adaptive feedback. It can drive the transfer process to gradually reduce vast cross-media domain discrepancy, so as to enhance the robustness of model training. For verifying the effectiveness of DCKT, we take the large-scale dataset XMediaNet as source domain, and 3 widely-used datasets as target domain for cross-media retrieval. Experimental results show that DCKT achieves promising improvement on retrieval accuracy.*

1. Introduction

With the rapid development of computer and digital transition technology, multimedia data such as image, text, video and audio can be found everywhere and exists as a whole to reshape our lives. Human can naturally receive information from different sensory channels, such as vision and auditory. However, it has been indicated that the importances of sensory channels differ among people, resulting in different learning styles [6]. For example, when students take in information with all the senses, such as seeing pictures and reading texts, they will have the highest efficiency of studying [6]. If relevant multimedia data can be conveniently retrieved and provided, it will be very helpful to increase the efficiency of information acquisition for human.

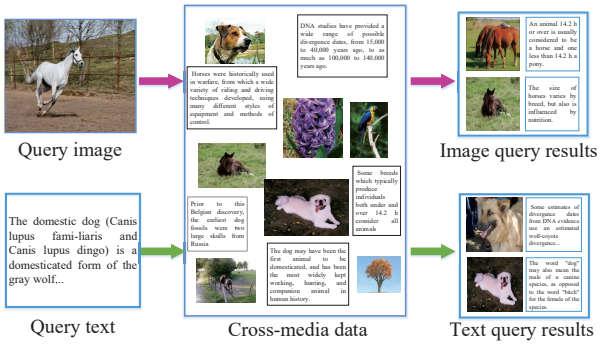

Figure 1: An example of cross-media retrieval.

Cross-media retrieval [28] is such a kind of technique to flexibly provide data of different media types, with one query of any media type. Figure 1 shows an example of cross-media retrieval, which includes two media types: image and text. As a highlighting research hotspot, cross-media retrieval has the advantage for realizing the coordination of different media types compared with traditional single-media retrieval. To perform cross-media retrieval, we have to deal with "heterogeneity gap". This means that different media types have inconsistent representation forms, so the similarity of them cannot be directly measured in their original feature spaces. Intuitively, the mainstream methods of cross-media retrieval are common representation learning, which aim to project data of different media types into an intermediate common space [10, 27, 35, 44]. Among them, deep neural network (DNN) based methods have currently become an active topic, which take DNN

*Corresponding author.

as basic model to perform common representation projection [5, 12, 25, 27, 38].

Cross-media retrieval is still a challenging problem, and the performance of existing methods usually relies on labeled data for model training. However, insufficient training data is a common and severe challenge, especially for DNN-based methods. From the view of *model training requirement*, because cross-media correlation is very complex and diverse, high-quality labeled data is crucial to provide cues for training "good" DNN models. Insufficient data limits the training performance and easily leads to overfitting. From the view of *human labor*, it is extremely labor-consuming to collect and label cross-media data. For example, if we want to collect data for "water", we need to see the images, read the texts, watch the videos, and even listen to the audio, and carefully judge whether the data is actually relevant to each other.

In this situation, the idea of transfer learning [21, 22, 26] becomes significant, which exploits general knowledge from source domain (usually a large-scale dataset) for relieving the problem of insufficient data. As known, cross-media data is quite labor consuming to collect and label, so existing labeled cross-media data is precious and valuable. It is a key problem towards application to distill knowledge from ***existing data*** for boosting retrieval performance on ***new data***. Nevertheless, existing transfer methods pay little attention to transfer between a large-scale cross-media dataset and a small-scale one. They also usually assume the domains share the same label space, which is often not satisfied due to the challenge of collecting cross-media data with the same semantic across domains. So we consider the following problem: ***How can we fully transfer knowledge from a large-scale cross-media dataset to promote the model training on another small-scale dataset, where they may have different label spaces?*** For addressing this problem, this paper proposes deep cross-media knowledge transfer (DCKT) approach. The main contributions of DCKT can be summarized as follows:

- ***Two-level transfer architecture*** is proposed to jointly minimize the media-level and correlation-level domain discrepancies, which allows two important and complementary aspects of knowledge to be transferred: intra-media semantic and inter-media correlation knowledge. It can enrich the training information and boost the retrieval accuracy on target domain.
- ***Progressive transfer mechanism*** is proposed to iteratively select training samples with ascending transfer difficulties in target domain, via the metric of cross-media domain consistency with adaptive feedback. It can gradually reduce the vast cross-media domain discrepancy to enhance the robustness of model training.

For performing knowledge transfer, a high-quality source domain is indispensable. In the experiment, we take a large-scale dataset XMediaNet as source domain, containing more than 100,000 labeled data with 200 distinct semantic categories. For target domain, we adopt 3 widely-used datasets: Wikipedia, NUS-WIDE-10k and Pascal Sentences. Experimental results show that DCKT achieves promising improvement on cross-media retrieval accuracy.

The following sections are organized as follows: Section 2 gives a brief review of related work. Section 3 presents the network architecture of DCKT, and Section 4 introduces the progressive transfer mechanism of DCKT. The experimental results and discussion are presented in Section 5, and finally Section 6 concludes this paper.

2. Related Work

2.1. Cross-media Retrieval

The current mainstream of cross-media retrieval is common representation learning, and the existing methods can be summarized as two main categories: shallow learning methods and DNN-based methods. Shallow learning methods usually take linear projections to convert cross-media data to common representation. A representative method is canonical correlation analysis (CCA) [10], which is a classical solution and extended by following works as [33, 35]. Besides CCA, there are also many methods which incorporate various information to learn projection matrices as [11, 14, 20, 44]. Furthermore, link information can also be an important source of cross-media correlation, which has been used for clustering heterogeneous social media objects [32].

DNN-based cross-media retrieval methods are the currently active direction [1, 15, 25, 27, 41, 42]. Bimodal deep autoencoder [25] is a representative method, which is an extension of restricted Boltzmann machine (RBM). It can be seen as two autoencoders sharing the same code layer, where the common representation is obtained. Deep canonical correlation analysis (DCCA) [1, 42] is a non-linear extension of CCA, which can learn the complex non-linear transformations for two modalities. Cross-media multiple deep networks (CMDN) [27] jointly preserve the intra-media and inter-media information and then hierarchically combine them for improving the retrieval accuracy.

However, insufficient training data is a common and severe problem for existing methods. Inspired by the common use of large-scale single-media datasets like ImageNet [18], we intend to address this problem by exploiting a large-scale cross-media dataset *XMediaNet* with general knowledge and transfer knowledge from it. This is useful towards real-world application where it is usually very hard to collect and label enough cross-media data.

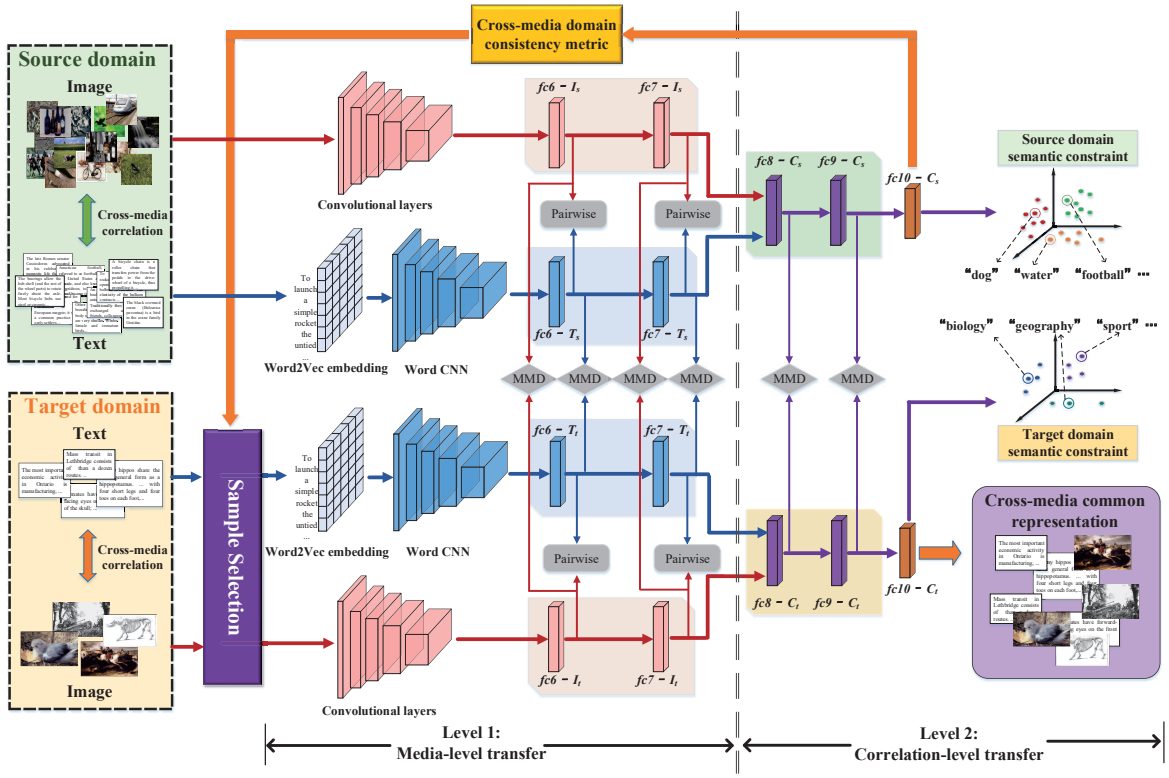

Figure 2: The overview of proposed deep cross-media knowledge transfer (DCKT) approach.

2.2. Transfer Learning

It is natural that human can adapt the knowledge from already learned tasks to new tasks. Transfer learning [26] aims to simulate such mechanism, and relieve the problem of insufficient training data for a specific task. The focus of transfer learning is to reduce the domain discrepancy, which is widely used in DNN-based methods [17, 21, 22] for relieving the problem of insufficient training data, but mainly deals with single-media scenario. Besides, some works are proposed to perform transfer between different feature spaces [39, 45] and multimedia domains [43]. Transitive hashing network [3] is proposed to learn from an auxiliary cross-media dataset to bridge two separate single-media datasets. Some works as [31, 36] also propose to effectively transfer knowledge from text to image. Besides, Cross-media hybrid transfer network (CHTN) [12] aims to transfer from a single-media source domain to cross-media target domain. Different from the above works, this paper aims at transferring from a source domain with large-scale cross-media dataset to a target domain with small-scale cross-media dataset, where the label spaces are different. It is a challenging task because the intra-media semantic information, inter-media intrinsic correlation and vast domain discrepancy should be jointly considered.

2.3. Curriculum Learning

The idea of progressive learning in this paper is inspired by curriculum learning (CL). The motivation of CL is simple: to first learn from easy samples, and gradually learn from harder samples [2], which aims to reduce the negative effects brought by noisy data in early period of training. It can be also applied for deciding learning order of tasks [30]. Self-paced learning (SPL) is based on CL, which designs a weighted loss term on all samples in the learning objective [19], and can be regarded as CL's implementation as indicated in [7]. CL has been applied in many problems like image classification [7] and object tracking [13].

This paper adopts the idea of CL to assign samples with different transfer difficulties by metric of cross-media domain consistency. This is an iterative process with adaptive feedback, which gradually reduces the discrepancy between cross-media domains to enhance the robustness of model training, and improve retrieval accuracy on cross-media target domain.

3. Network Architecture of DCKT

This section will introduce the network architecture of DCKT in Figure 2. The training process of progressive transfer, including the domain consistency metric and sam-

ple selection, will be further introduced in Section 4.

This paper focuses on the scenarios where source and target domains both have two media types (i.e., image and text), but DCKT can be simply extended to more than two media types by adding pathways. The end-to-end architecture of DCKT can be seen as two levels: media-level transfer and correlation-level transfer. We denote the source domain as $Src = \{(i_s^p, t_s^p), y_s^p\}_{p=1}^P$, where (i_s^p, t_s^p) is the p-th image/text pair with label y_s^p. Similarly, the target domain includes training set $Tar_{tr} = \{(i_t^q, t_t^q), y_t^q\}_{q=1}^Q$ and testing set $Tar_{te} = \{(i_t^m, t_t^m)\}_{m=1}^M$. The aim of DCKT is to exploit both Src and Tar_{tr} to train the model for generating common representation of Tar_{te}, which is $c_t(I)^m$ and $c_t(T)^m$ for each image and text. After this, the cross-media retrieval can be performed by distance computing with common representation.

3.1. Level 1: Media-level Transfer

As the two domains both have two media types, the domain discrepancy can come from two aspects: (1) ***Media-level discrepancy***, which means the intra-media semantic information in two domains has discrepancy; (2) ***Correlation-level discrepancy***, which means the inter-media correlation information in two domains has discrepancy. Media-level transfer aims to address the media-level discrepancy by feature adaptation of the same media type between two domains.

For each domain, we have two pathways for image and text respectively, and the two domains have the same architecture. For image pathway, we take widely-used VGG19 [37] as basic model. We keep all the layers of VGG19 except the last fully-connected layer, and each input image is converted to 4,096-d representations via $fc6-I_s/fc7-I_s$ for source domain, and $fc6-I_t/fc7-I_t$ for target domain. For text pathway, we first embed each word into a vector via Word2Vec model [24], and then generate the 300-d input feature vector of each text with Word CNN [16]. Similar to image pathway, the input text feature will pass through two fully connected layers, namely $fc6-T_s/fc7-T_s$ and $fc6-T_t/fc7-T_t$.

Between the two domains, we achieve media-level transfer by feature adaptation [21] via minimizing the maximum mean discrepancy (MMD) [8] of the same media type. Taking image as an example, we use $I_s = \{i_s\}$ and $I_t = \{i_t\}$ to denote the distributions of images in Src and Tar_{tr}. $\mu_k(a)$ denotes the mean embedding of a in reproducing kernel Hibert space (RKHS) $\mathcal{H}_k$, and $\mathbf{E}_{x\sim a} f(x) = \langle f(x), \mu_k(a)\rangle_{\mathcal{H}_k}$ for $f \in \mathcal{H}_k$. So the squared MMD $m_k^2(I_s, I_t)$ is denoted as follows:

$$m_k^2(I_s, I_t) \triangleq \left\| \mathbf{E}_{I_s}[\phi(i_s, \theta_{I_s})] - \mathbf{E}_{I_t}[\phi(i_t, \theta_{I_t})] \right\|_{\mathcal{H}_k}^2 \quad (1)$$

where ϕ denotes a network layer's output, and θ_x denotes the network parameters for each pathway. For example, θ_{I_s} means parameters of **I**mage pathway in **s**ource domain, and θ_{I_t} means those of **I**mage pathway in **t**arget domain.

MMD is computed in a layer-wise style, which is between the corresponding layers of two domains, i.e., $fc6-I_s/fc6-I_t$, $fc7-I_s/fc7-I_t$ for image, and $fc6-T_s/fc6-T_t$, $fc7-T_s/fc7-T_t$ for text. By minimizing MMD, the media-level domain discrepancy can be reduced, which can align the single-media representation of two domains for knowledge transfer. The *MMD loss* functions of image and text can be defined as:

$$Loss_{MMD_I} = \sum_{l=l_6}^{l_7} m_k^2(I_s, I_t) \quad (2)$$

$$Loss_{MMD_T} = \sum_{l=l_6}^{l_7} m_k^2(T_s, T_t) \quad (3)$$

where $Loss_{MMD_I}$ and $Loss_{MMD_T}$ mean MMD loss functions for two media types.

Besides, in two domains, each pair of image and text as (i_s^p, t_s^p) and (i_t^q, t_t^q) exists together to represent closely relevant semantic, which is an important coexistence cue for cross-media retrieval. We preserve such pairwise constraint during transfer process via reducing the representation difference of each pair, which is a commonly-used criterion in cross-media retrieval [5, 12]. Specifically, we use Euclidean distance as measurement, denoted as:

$$d^2(i_s^p, t_s^p) = \left\| \phi(i_s^p, \theta_{I_s}) - \phi(t_s^p, \theta_{T_s}) \right\|^2 \quad (4)$$

$$d^2(i_t^q, t_t^q) = \left\| \phi(i_t^q, \theta_{I_t}) - \phi(t_t^q, \theta_{T_t}) \right\|^2 \quad (5)$$

Similar to what we have in Equation 1, θ_x denotes the network parameters for each pathway. Then we get the *pairwise constraint loss* for two domains as:

$$Loss_{Pair_s} = \sum_{l=l_6}^{l_7} \sum_{p=1}^{P} d^2(i_s^p, t_s^p) \quad (6)$$

$$Loss_{Pair_t} = \sum_{l=l_6}^{l_7} \sum_{q=1}^{Q} d^2(i_t^q, t_t^q) \quad (7)$$

where $Loss_{Pair_s}$ and $Loss_{Pair_t}$ mean pairwise constraint loss for two domains, which are also computed in a layer-wise style between corresponding layers $fc6-I_s/fc6-T_s$, $fc7-I_s/fc7-T_s$ for source domain, and $fc6-I_t/fc6-T_t$, $fc7-I_t/fc7-T_t$ for target domain. By minimizing the MMD loss and pairwise constraint loss, we can transfer the intra-media semantic information from source domain to target domain, as well as avoiding damaging the data-coexistence relationship.

3.2. Level 2: Correlation-level Transfer

Cross-media domain discrepancy not only lies in the difference within each media type, but also in the correlation patterns for them to be correlated with each other.

Correlation-level transfer aims to align the inter-media correlation of the two domains. For capturing the cross-media correlation in each domain, we adopt the strategy of shared layers to generate the common representation for different media types as [12].

In the two domains, both image and text pathways will share two fully-connected layers. So the parameters of shared layers can fit the semantic learning of both two media types, which has the ability to capture inter-media correlation. We add MMD loss function between the shared layers for correlation-level transfer. Similar to media-level transfer, we compute the *MMD loss* function as follows:

$$Loss_{MMD_C} = \sum_{l=l_8}^{l_9} m_k^2(C_s, C_t) \tag{8}$$

where $l_{8/9}$ means the corresponding shared layers in two domains, i.e., $fc8{-}C_s/fc8{-}C_t$ and $fc9{-}C_s/fc9{-}C_t$ in Figure 2, and C_s and C_t mean the output of shared layers of two domains. By minimizing $Loss_{MMD_C}$, the correlation-level domain discrepancy can be reduced, which aligns the inter-media correlation of two domains for knowledge transfer.

Besides, we should preserve the semantic information to maintain the semantically discriminative ability of common representation. This is intuitively achieved by semantic constraints with *semantic loss* functions as follows:

$$Loss_{S_{e_s}} = \sum_{p=1}^{P} (f_{sm}(i_s^p, y_s^p, \theta_{C_s}) + f_{sm}(t_s^p, y_s^p, \theta_{C_s})) \tag{9}$$

$$Loss_{S_{e_t}} = \sum_{q=1}^{Q} (f_{sm}(i_t^q, y_t^q, \theta_{C_t}) + f_{sm}(t_t^q, y_t^q, \theta_{C_t})) \tag{10}$$

where θ_{C_s} and θ_{C_t} are the network parameters for pathways of **s**ource and **t**arget domains, and f_{sm} is the softmax loss function.

The architecture of DCKT is end-to-end, so the two levels of transfer can be jointly performed to mutually boost. It comprehensively allows the knowledge from cross-media source domain to be propagated to target domain. In this way, DCKT can enrich the training information with supplementary information of both intra-media semantic and inter-media correlation knowledge, thus promoting the model training performance and improve retrieval accuracy.

4. Progressive Transfer Mechanism

All the introduced loss functions are able to be minimized by Stochastic Gradient Descent (SGD), so DCKT can be simply trained by simultaneously optimizing all of them with all data in Src and Tar_{tr} as input. However, because the discrepancy of two cross-media domains is usually quite vast with different label spaces, it may bring much noise and mislead the model training, especially for "empty" models.

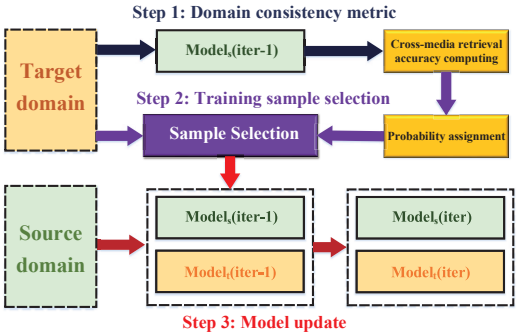

Figure 3: Process in each iteration of progressive transfer.

So we propose a progressive transfer mechanism to gradually reduce the cross-media domain discrepancy.

To start from a "safe" point, we first pre-train the model for each domain separately, removing all the MMD loss linking the two domains. For convenience, we denote the networks for two domains as $Model_s$ and $Model_t$. Then we progressively transfer the knowledge from source domain to target domain, which is an iterative process shown as Figure 3. Because source domain is relatively large-scale and reliable, we take $Model_s$ as reference model to perform sample selection in target domain. The motivation is intuitive: In early period of training, we choose "easy" samples in Tar_{tr} whose cross-media correlation can be successful molded by $Model_s$, which are of high consistency with source domain. For example, although the label spaces are different, some categories such as "sport" and "football" have strong consistency. In late period of training when the model is stable, we can incorporate "harder" samples with low domain consistency to further adapt to target domain.

In each iteration *iter*, we generate common representation (class probability vector) for Tar_{tr} as C_s by $Model_s(iter)$, including $C_s(I)$ and $C_s(T)$. Next, we perform bi-directional cross-media retrieval and evaluate domain consistency according to the accuracy, which is Image→Text and Text→Image. Taking Image→Text as an example, we compute the cosine distance between each image $c_s(I)^q$ and every text in $C_s(T)$, and then rank them to get the AP score of $c_s(I)^q$ as:

$$AP(I)^q = \frac{1}{R} \sum_{k=1}^{Q} \frac{R_k}{k} \times rel_k \tag{11}$$

where R is the number of text with the same label of $c_s(I)^q$, R_k is the number of relevant text in top-k results. rel_k indicates whether $c_s(I)^q$ and k-th result have the same label.

A high $AP(I)^q$ means $Model_s(iter)$ successfully captures the cross-media correlation of i_t^q, i.e., the source domain contains closely relevant knowledge of i_t^q, so it can be regarded as an "easy" transfer sample. Similarly we have

Algorithm 1 : Progressive Transfer

Require: Training data Src and Tar_{tr}, maximal iteration number MI, and epoch number in each iteration Ep.

1: Pre-train $Model_s$ and $Model_t$ separately, denoted as $Model_s(0)$ and $Model_t(0)$. Set $iter = 1$.
2: **repeat**
3: Generate the common representation for Tar_{tr} with $Model_s(iter-1)$ as $C_s(I)$ and $C_s(T)$.
4: Compute $AP(I)^q$ and $AP(T)^q$ for all $(i_t^q, t_t^q) \in Tar_{tr}$.
5: Compute AP^q via Equation 12.
6: Estimate $Prob(q)$ via Equation 13, and select training sample set $Tar_{tr}(iter)$.
7: Train model for Ep epochs with Src and $Tar_{tr}(iter)$, to get $Model_s(iter)$ and $Model_t(iter)$.
8: $iter = iter + 1$.
9: **until** $iter = MI$.
10: **return** $Model_s(MI)$ and $Model_t(MI)$.

$AP(T)^q$ and obtain:

$$AP^q = AP(I)^q + AP(T)^q \tag{12}$$

where AP^q can be used to estimate domain consistency of a pair (i_t^q, t_t^q). During the training process, $Model_s$ is also iteratively updated, so AP^q should be computed in each iteration. A high AP^q means q-th pair is proper to be a bridge of the two domains. We assign the probability to be selected for each pair as:

$$Prob(q) = \alpha[1 - \log_2(\frac{max(AP) - AP^q}{max(AP) \times iter} + 1)] \tag{13}$$

where $max(AP)$ is the maximal value of AP^q, and $\alpha \in (0, 1]$ is the upper bound of $Prob(q)$. α prevents the "easiest" samples from always being selected, which leads to the risk of overfitting. When $iter$ increases, the value of item $(max(AP) - AP^q)/(max(AP) \times iter)$ will turn small, which means the selection will gradually become random sampling. The above process can be summarized as Algorithm 1.

After training, each testing data can be converted as common representation (actually class probability vector), and then the cross-media retrieval can be performed by distance metric. Note that in testing stage, the image and text data can be input separately, whose labels and pairwise correlation are not used at all. This setting is widely adopted in cross-media retrieval as [12, 41].

5. Experiments

5.1. Details of Implementation

The architecture of DCKT is easy to implement, and the parts of two domains share the same architecture. For image we use VGG19 [37] as basic model to generate convolutional feature maps of pool5, which is pre-trained by ImageNet [18] of ImageNet large-scale visual recognition challenge (ILSVRC) 2012. For text we first embed each word into a vector via Word2Vec model [24], and then generate 300-d text feature following [16]. The classification layers $fc10 - C_s$ and $fc10 - C_t$ are fully-connected layers of the same unit number with the semantic categories in each domain. All the other layers are fully-connected layers of 4,096 units, including $fc6 - I_{s/t}$, $fc7 - I_{s/t}$, $fc6 - T_{s/t}$, $fc7 - T_{s/t}$, $fc8 - C_{s/t}$, and $fc9 - C_{s/t}$. The *pairwise constraint loss* functions are implemented by contrastive loss layers from Caffe[1]. The *MMD loss* functions are implemented following [21], by which the knowledge transfer of the two domains is actually performed. As for network parameters, we set the initial learning rates as 0.01, and the weight decay 0.0005. In the mechanism of progressive training in Algorithm 1, we set α as 0.2, Ep as 1, and MI as 10. These parameters will be further analyzed in Section 5.5.3.

5.2. Datasets

5.2.1 Source Domain

To serve as the source domain, the dataset should be large-scale, high-quality, and of general knowledge like ImageNet [18] and Google News corpus [23], so that the knowledge is proper to be adapted to other domains.

XMediaNet [28] dataset is adopted to serve as the source domain. It is a large-scale dataset with 5 media types, which has more than 100,000 media instances of text, image, audio, video and 3D model. All the instances are manually collected and labeled from famous websites such as Wikipedia, Flickr, Youtube, Findsounds, Freesound, and Yobi3D. It includes 200 distinct semantic categories based on wordNet hierarchy to avoid semantic confusion, including 47 animal species like "dog" and 153 artifact species like "airplane". In this paper, we focus on the scenario of image and text, so we choose the training set of image and text data from XMediaNet with 32,000 pairs.

5.2.2 Target Domain

For target domain, we adopt 3 widely-used datasets to conduct cross-media retrieval, namely Wikipedia, NUS-WIDE-10k and Pascal Sentences. They all have two media types image and text. The dataset split is strictly according to [5, 12, 27], shown as Table 1.

Dataset	Split		
	Training	Testing	Validation
Wikipedia [35]	2,173	462	231
NUS-WIDE-10k [4, 5]	8,000	1,000	1,000
Pascal Sentences [34]	800	100	100

Table 1: The size and split of each dataset as target domain.

[1] http://caffe.berkeleyvision.org

5.3. Compared Methods

We compare our proposed DCKT approach with totally 12 state-of-the-art methods with source codes from the authors of original papers, namely CCA [10], CFA [20], KCCA (with Gaussian kernel) [9], Corr-AE [5], JRL [44], LGCFL [14], DCCA [42], CMDN [27] Deep-SM [41], CHTN [12], ACMR [40], and CCL [29].

Due to the wide range of comparison methods, their original papers adopt different input settings. For example, CHTN and Deep-SM are based on AlexNet and take original image pixels as input, while others like CCL take feature vectors as input. For fair comparison, we replace the AlexNet of CHTN and Deep-SM with VGG19, and use the 4,096-d VGG19 image feature for methods which need feature vector as input. As for text, we use the same 300-d Word CNN text features for all the methods, which is the same with our DCKT.

5.4. Evaluation Metrics

We conduct cross-media retrieval task with two directions: Image→Text and Text→Image. Taking Image→Text as an example, the retrieval process is conducted as follows: (1) Get the common representation for all images and texts in testing set. (2) Take one image as query, and compute the cosine distance between the common representation of query image and all texts. (3) Rank all the texts in testing set with similarities according to the distances.

The metric adopted for evaluating the retrieval results is mean average precision (MAP) score, which is the mean value of average precision (AP) scores of all queries. AP is computed as Equation 11. *All retrieval results* will be considered for the computation of MAP score following [12, 29, 41], instead of *top-50 results* as [5, 40].

5.5. Experimental Results

5.5.1 Comparison with State-of-the-art methods

Table 2 shows the retrieval accuracy of DCKT and compared methods. On Wikipedia dataset, DCKT gains the improvement from 0.492 to 0.511, compared with the method with highest MAP score CHTN. Among the compared methods, we can see that the shallow learning method JRL achieves comparable accuracy with DNN-based methods, and even outperforms Corr-AE, Deep-SM, and DCCA. This is probably because that the small scale of Wikipedia dataset is insufficient for deep network to get ideal training performance. On NUS-WIDE-10k and Pascal Sentences datasets, our DCKT achieves the best MAP scores, too. The above results show the stable advantage of DCKT compared with existing methods. This is because the two-level transfer network architecture and progressive transfer mechanism allow the intra-media semantic and inter-media correlation knowledge to be propagated to the target domain, improving training effectiveness on cross-media target domain.

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	**our DCKT**	**0.537**	**0.485**	**0.511**
	CCL [29]	0.505	0.457	0.481
	ACMR [40]	0.468	0.412	0.440
	CHTN [12]	0.523	0.460	0.492
	Deep-SM [41]	0.478	0.422	0.450
	CMDN [27]	0.487	0.427	0.457
	DCCA [42]	0.445	0.399	0.422
	LGCFL [14]	0.466	0.431	0.449
	JRL [44]	0.479	0.428	0.454
	Corr-AE [5]	0.442	0.429	0.436
	KCCA [9]	0.438	0.389	0.414
	CFA [20]	0.319	0.316	0.318
	CCA [10]	0.298	0.273	0.286
NUS-WIDE -10k dataset	**our DCKT**	**0.556**	**0.584**	**0.570**
	CCL [29]	0.481	0.520	0.501
	ACMR [40]	0.519	0.542	0.531
	CHTN [12]	0.537	0.562	0.550
	Deep-SM [41]	0.497	0.478	0.488
	CMDN [27]	0.492	0.542	0.517
	DCCA [42]	0.452	0.465	0.459
	LGCFL [14]	0.453	0.485	0.469
	JRL [44]	0.466	0.499	0.483
	Corr-AE [5]	0.441	0.494	0.468
	KCCA [9]	0.351	0.356	0.354
	CFA [20]	0.406	0.435	0.421
	CCA [10]	0.167	0.181	0.174
Pascal Sentences dataset	**our DCKT**	**0.582**	**0.587**	**0.585**
	CCL [29]	0.576	0.561	0.569
	ACMR [40]	0.538	0.544	0.541
	CHTN [12]	0.556	0.534	0.545
	Deep-SM [41]	0.560	0.539	0.550
	CMDN [27]	0.544	0.526	0.535
	DCCA [42]	0.568	0.509	0.539
	LGCFL [14]	0.539	0.503	0.521
	JRL [44]	0.563	0.505	0.534
	Corr-AE [5]	0.532	0.521	0.527
	KCCA [9]	0.488	0.446	0.467
	CFA [20]	0.476	0.470	0.473
	CCA [10]	0.203	0.208	0.206

Table 2: MAP scores of our DCKT and compared methods. *All retrieval results* are evaluated for comprehensive comparison, instead of *top-50 results* as [5, 40].

It should be noted that CHTN is also a transfer learning based method, which transfers knowledge from single-media source domain (ImageNet) to cross-media target domain. By comparing the MAP scores of DCKT and CHTN, it can be seen that it is helpful to transfer from a cross-media source domain, because the cross-media source domain has not only media-level knowledge, but also rich correlation-level knowledge.

5.5.2 Baseline Experiment

To further analyze the performance of DCKT, we conduct baseline experiments on 3 datasets. The results are shown in Table 3. Due to the page limitation, we show the **average MAP scores** of retrieval in 2 directions. The basic idea of this paper is knowledge transfer, so the first question is: Is the knowledge transfer process actually helpful? To verify this, we perform retrieval with the separately pre-trained model of each dataset, i.e., $Model_t(0)$. We denote the com-

plete DCKT model as $DCKT_{Full}$. By comparing $Model_t(0)$ with $DCKT_{Full}$ in Table 3, we can see that the transfer process achieves inspiring improvement.

Then we verify the effectiveness of two key strategies of DCKT: *Two-level transfer* and *progressive transfer*. For *two-level transfer*, we design 2 baselines: only with media-level transfer (Transfer_1 in Table 3) or correlation-level transfer (Transfer_2 in Table 3), and keep other parts unchanged. From Table 3 we can see that the combination of the two levels can achieve more improvement than either of them, which shows that the two levels of knowledge are complementary for cross-media retrieval.

For *progressive transfer*, we design 2 baselines: $DCKT_{All}$ means that in each iteration, we use all data in Tar_{tr}. $DCKT_{Random}$ means that we select samples randomly. It can be seen that although knowledge transfer is helpful, the domain discrepancy is vast in cross-media scenario, so $DCKT_{All}$ and $DCKT_{Random}$ both achieve lower MAP scores than $DCKT_{Full}$. We also observe that $DCKT_{Random}$ is slightly lower than $DCKT_{All}$, which is because that by arbitrary sampling, the model cannot have the whole view in each iteration, which brings negative effects than $DCKT_{All}$.

Besides, there may exist category overlaps between the source and target domains. Wikipedia has no category overlap with XMediaNet dataset (0 of totally 10), while NUS-WIDE-10k has minor overlap (3 of 10), and Pascal Sentences has large overlap (12 of 20). $DCKT_{No\ overlap}$ means that we remove the overlap categories in XMediaNet dataset with NUS-WIDE-10k and Pascal Sentences datasets, respectively. The results are not sensitive to overlap, which shows our DCKT is robust for different label spaces.

Method	Dataset		
	Wikipedia	NUS-WIDE-10k	Pascal Sentences
$DCKT_{Full}$	**0.511**	**0.570**	**0.585**
$Model_t(0)$	0.459	0.527	0.529
Transfer_1	0.491	0.555	0.565
Transfer_2	0.487	0.553	0.569
$DCKT_{All}$	0.498	0.560	0.574
$DCKT_{Random}$	0.494	0.553	0.573
$DCKT_{No\ overlap}$	–	0.566	0.579

Table 3: Average MAP scores of baseline experiments.

5.5.3 Parameter Analysis

In this section we analyze the settings of parameters MI, Ep, and α in Algorithm 1. In our experiment, because the sizes of Src and Tar_{tr} are different, for ensuring in each iteration Src can be processed throughout, we set $Ep = 1$ for it. Correspondingly, for Tar_{tr} the epoch number is P/Q. As for MI, we set it as 10 in our experiment, and the performance will tend to be stable. They can also be intuitively adjusted according to validation set.

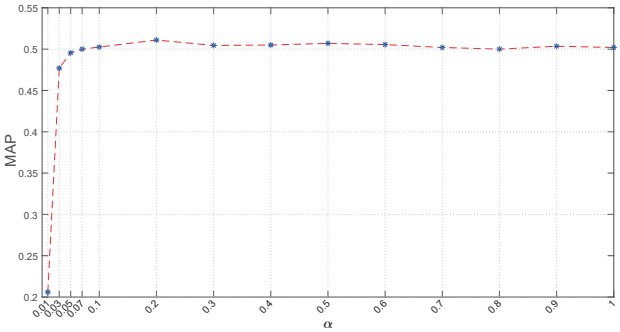

Figure 4: Impact of α on MAP score of Wikipedia dataset.

Next, α determines how many samples we can select in an iteration of progressive transfer. For investigating the impact of α, we conduct DCKT with different α values. The impact is shown as Figure 4. We can see that although we perform transfer based on pre-trained model $Model_t(0)$, the performance is seriously damaged with very small α. When α increases, the MAP score will increase apparently until 0.2. Then the MAP scores are generally stable but tend to be lower. This shows that a large α means the "easy" samples are always selected, which can lead to the risk of overfitting.

6. Conclusion

This paper has proposed deep cross-media knowledge transfer (DCKT) approach, which transfers knowledge from a large labeled cross-media dataset as source domain to promote the performance of model training on target domain. DCKT is a two-level transfer network to allow the intra-media and inter-media knowledge to be propagated to the target domain, which can enrich the training information and boost the retrieval accuracy on target domain. For addressing the vast domain gap, we propose progressive transfer mechanism to iteratively select training samples with ascending transfer difficulties in target domain, which can drive the cross-media transfer process to gradually reduce the vast cross-media domain discrepancy, and enhance the robustness. In the experiments, we take the large-scale dataset XMediaNet as source domain, and 3 widely-used datasets as target domain for cross-media retrieval. Experimental results show that DCKT achieves promising improvement on retrieval accuracy. For the future work, we intend to propose more effective strategy for sample selection, and extend DCKT for unsupervised transfer scenario, i.e, the semantic labels of target domain are unknown, which will further save the human labor of labeling data.

7. Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61771025 and 61532005.

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 3408–3415, 2010.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, pages 41–48, 2009.

[3] Z. Cao, M. Long, J. Wang, and Q. Yang. Transitive hashing network for heterogeneous multimedia retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 81–87, 2017.

[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 (CIVR)*, 2009.

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

[6] A. P. Gilakjani. Visual, auditory, kinaesthetic learning styles and their impacts on english language teaching. *Journal of Studies in Education*, 2:104–113, 2012.

[7] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing (TIP)*, 25(7):3249–3260, 2016.

[8] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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

[10] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[11] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang. Cross-modal correlation learning by adaptive hierarchical semantic aggregation. *IEEE Transactions on Multimedia (TMM)*, 18(6):1201–1216, 2016.

[12] X. Huang, Y. Peng, and M. Yuan. Cross-modal common representation learning by hybrid transfer network. *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1893–1900, 2017.

[13] J. S. S. III and D. Ramanan. Self-paced learning for long-term tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2379–2386, 2013.

[14] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia (TMM)*, 17(3):370–381, 2015.

[15] J. Kim, J. Nam, and I. Gurevych. Learning semantics with deep belief network for cross-language information retrieval. In *International Committee on Computational Linguistic (ICCL)*, pages 579–588, 2012.

[16] Y. Kim. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

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

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

[19] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1189–1197, 2010.

[20] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *ACM International Conference on Multimedia (ACM MM)*, pages 604–611, 2003.

[21] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, pages 97–105, 2015.

[22] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 136–144, 2016.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*, 2013.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.

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

[26] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010.

[27] Y. Peng, X. Huang, and J. Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3846–3853, 2016.

[28] Y. Peng, X. Huang, and Y. Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2017. Early Access.

[29] Y. Peng, J. Qi, X. Huang, and Y. Yuan. CCL: cross-modal correlation learning with multi-grained fusion by hierarchical network. *IEEE Transactions on Multimedia (TMM)*, 20(2):405–420, 2018.

[30] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5492–5500, 2015.

[31] G. Qi, C. C. Aggarwal, and T. S. Huang. Towards semantic knowledge propagation from text corpus to web images. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 297–306, 2011.

[32] G. Qi, C. C. Aggarwal, and T. S. Huang. On clustering heterogeneous social media objects with outlier links. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining (WSDM)*, pages 553–562, 2012.

[33] V. Ranjan, N. Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4094–4102, 2015.

[34] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. *North American Chapter of the Association for Computational Linguistics*, pages 139–147, 2010.

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

[36] X. Shu, G. Qi, J. Tang, and J. Wang. Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In *ACM International Conference on Multimedia (ACM MM)*, pages 35–44, 2015.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*, 2014.

[38] N. Srivastava and R. Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International Conference on Machine Learning (ICML) Workshop*, 2012.

[39] Y. H. Tsai, Y. Yeh, and Y. F. Wang. Learning cross-domain landmarks for heterogeneous domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5081–5090, 2016.

[40] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In *ACM International Conference on Multimedia (ACM MM)*, pages 154–162, 2017.

[41] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics (TCYB)*, 47(2):449–460, 2017.

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

[43] X. Yang, T. Zhang, and C. Xu. Cross-domain feature learning in multimedia. *IEEE Transactions on Multimedia (TMM)*, 17(1):64–78, 2015.

[44] X. Zhai, Y. Peng, and J. 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.

[45] J. Zhang, Y. Han, J. Tang, Q. Hu, and J. Jiang. Semi-supervised image-to-video adaptation for video action recognition. *IEEE Transactions on Cybernetics (TCYB)*, 47(4):960–973, 2017.