

Transfer Learning of Distance Metrics by Cross-Domain Metric Sampling across Heterogeneous Spaces

Guo-Jun Qi*

Charu Aggarwal†

Thomas Huang‡

Abstract

The problem of transfer learning has recently been of great interest in a variety of machine learning applications. In this paper, we examine a new angle to the transfer learning problem, where we examine the problem of distance function learning. Specifically, we focus on the problem of how our knowledge of distance functions in one domain can be transferred to a new domain. A good semantic understanding of the feature space is critical in providing the domain specific understanding for setting up good distance functions. Unfortunately, not all domains have feature representations which are equally interpretable. For example, in some domains such as text, the semantics of the feature representation are clear, as a result of which it is easy for a domain expert to set up distance functions for specific kinds of semantics. In the case of image data, the features are semantically harder to interpret, and it is harder to set up distance functions, especially for particular semantic criteria. In this paper, we focus on the problem of transfer learning as a way to close the semantic gap between different domains, and show how to use correspondence information between two domains in order to set up distance functions for the semantically more challenging domain.

1 Introduction

The problem of transfer learning [15][14][8][23][24] has seen a revival in recent years because of the tremendous amounts of heterogeneous data which are available in a wide variety of networks and content-based applications. Different domains provide a different level of ease in data collection and processing. Therefore, it is useful to somehow transfer the knowledge from one domain to the other. For example, in cross-lingual learn-

ing, labeled English text is widely available, whereas it is much harder to obtain labeled Chinese documents. Therefore, the focus of transfer learning in this example is to use the natural correspondence between the feature spaces of the two domains in order to create an automated learner for Chinese documents. The focus of most transfer learning problems is on aspects which involve the *unavailability of sufficient data* for learning purposes. The transfer learning model is used as a way to learn cases in which sufficient data is not available to create the classification model.

In this paper, we examine a different angle to the transfer learning problem, by exploring the varying **semantic gap** [6] in different feature spaces. An understanding of the semantics of a feature space is critical in setting up key operations in that space. One such example is the problem of distance function design. Distance function design is a key problem for many fundamental applications such as similarity search [1][22][3][16][21][9] and retrieval [13].

Distance functions can be set up much more easily in a feature space, when the semantics of that space are easy to interpret. This is especially true for applications in which the distance function needs to be designed with specific criteria in mind. For example, in the text domain, a distance function which is discriminatory between certain kinds of topics can be easily set up by restricting the feature space to words which belong to the set of topics at hand. On the other hand, this is much harder to achieve in a domain such as image data in which the features cannot be naturally interpreted in terms of the different criteria, and the distance function design is far more challenging.

In this paper, we focus on the problem of transfer learning as a way to link the different domains. As an example, we assume that the only input to the process is a set of images with corresponding text in the learning phase. We would like to explore this correspondence between the two domains in order to set up a distance function which uses **only** the image features, even in a different collection of images **which do not have** corresponding text. We also note that in some cases, the metric information in original target domain may

*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 1406 W. Green St., Urbana, IL 61801-2918. Email: qi4@illinois.edu

†IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532. Email: charu@ibm.com

‡Department of Electrical and Computer Engineering, the University of Illinois at Urbana-Champaign, 1406 W. Green St., Urbana, IL 61801-2918. Email: huang@ifp.uiuc.edu

be available in order to further improve the accuracy of the transfer learning process.

The remainder of this paper is organized as follows. In Section 2, we formally define the problem of transfer learning of distance functions across heterogeneous domains. Section 3 formulates and solves the optimization problem of learning the distance function through a transfer learning process. We relate the proposed method with existing work in the literature in Section 4. In Section 5, we present experiments on real-world data sets and show the advantages of the proposed algorithm. The conclusions are presented in Section 6.

2 Problem Definition and Target Metric

Let \mathbb{R}^s and \mathbb{R}^t be the source and target feature spaces with dimensionalities of s and t respectively. Each instance in the source space is represented by a feature vector $\mathbf{y} \in \mathbb{R}^s$, and the target instances are represented by feature vectors \mathbf{x} in the target space \mathbb{R}^t . In order to transfer the metric structure from source domain to target domain, we define a random variable $\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y})$ to indicate the cross-domain relevance between a target instance \mathbf{x} and a source instance \mathbf{y} . We define a transfer function $T(\mathbf{x}, \mathbf{y})$ to measure the probability of \mathbf{x} and \mathbf{y} being relevant to each other, over $\mathbb{R}^s \times \mathbb{R}^t$ as

$$(2.1) \quad T : \mathbb{R}^s \times \mathbb{R}^t \rightarrow [0, 1], (\mathbf{x}, \mathbf{y}) \mapsto T(\mathbf{x}, \mathbf{y})$$

Then the cross-domain relevance variable $\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y})$ follows the Bernoulli distribution $\mathbb{B}(T(\mathbf{x}, \mathbf{y}))$ parameterized by the transfer function, i.e., $p(\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y}) = 1) = T(\mathbf{x}, \mathbf{y})$ and $p(\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y}) = 0) = 1 - T(\mathbf{x}, \mathbf{y})$.

Additionally, to capture the metric structure in source domain, the source space may use a particular kind of similarity function, which is the most effective for processing in that domain. For example, the cosine similarity function is likely to be quite effective in the text domain. We use a kernel function $k(\mathbf{y}, \tilde{\mathbf{y}})$ in order to encode this metric structure in the source space, which measures the similarity of \mathbf{y} and $\tilde{\mathbf{y}}$ in the source space. Any Mercer kernel which satisfies the positive semi-definite property[17] in source space can be used here. In the meantime, we assume all the source instances are sampled from a true distribution $p(\mathbf{y})$. Then the kernel similarity together with $p(\mathbf{y})$ completely describes the metric structure between source instances.

Now given the kernel structure in *source* space, with the help of transfer function T we can define the metric structure in target space by exploring the metric structure in source space. Specifically, we depict the following cross-domain metric sampling process to compute the similarity between the target instances \mathbf{x} and $\tilde{\mathbf{x}}$:

1. Sample a pair of source instances \mathbf{y} and $\tilde{\mathbf{y}}$ from

$p(\mathbf{y})$.

2. Sample $\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y}) \sim \mathbb{B}(T(\mathbf{x}, \mathbf{y}))$ and $\mathbb{I}_{\text{Rel}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \sim \mathbb{B}(T(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}))$ to decide whether \mathbf{y} and $\tilde{\mathbf{y}}$ are relevant to \mathbf{x} and $\tilde{\mathbf{x}}$, respectively.

3. If both are relevant, i.e., $\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y}) \cdot \mathbb{I}_{\text{Rel}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = 1$, output $k(\mathbf{y}, \tilde{\mathbf{y}})$ as the target similarity between \mathbf{x} and $\tilde{\mathbf{x}}$; otherwise, output 0 which means in terms of the sampled source instances \mathbf{y} and $\tilde{\mathbf{y}}$ no evidence shows the target instances \mathbf{x} and $\tilde{\mathbf{x}}$ are similar.

Based on the above sampling process, we define the target similarity as the *expected* output of the target similarity over $p(\mathbf{y})$:

$$(2.2) \quad \begin{aligned} s(\mathbf{x}, \tilde{\mathbf{x}}) &\triangleq \mathbb{E}_{\mathbf{y}, \tilde{\mathbf{y}} \sim p(\mathbf{y})} [\mathbb{E} [\mathbb{I}_{\text{Rel}}(\mathbf{x}, \mathbf{y}) \cdot \mathbb{I}_{\text{Rel}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) k(\mathbf{y}, \tilde{\mathbf{y}}) | \mathbf{y}, \tilde{\mathbf{y}}]] \\ &= \mathbb{E}_{\mathbf{y}, \tilde{\mathbf{y}} \sim p(\mathbf{y})} [T(\mathbf{x}, \mathbf{y}) T(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) k(\mathbf{y}, \tilde{\mathbf{y}})] \\ &= \int_{\Delta \times \Delta} T(\mathbf{x}, \mathbf{y}) T(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) k(\mathbf{y}, \tilde{\mathbf{y}}) p(\mathbf{y}) p(\tilde{\mathbf{y}}) d\mathbf{y} d\tilde{\mathbf{y}} \end{aligned}$$

where Δ is the support of the distribution $p(\mathbf{y})$. It computes the target similarity metric by taking expectation of the source similarity $k(\mathbf{y}, \tilde{\mathbf{y}})$ transferred by T with respect to $p(\mathbf{y})$.

Figure 1 illustrates this idea by demonstrating how (target) image similarity is computed from the relevant (source) text documents. The images are linked to the relevant text documents by sampling the cross-domain relevance variables. The transfer function is used to link the images to the relevant text documents. Then the target similarity between images is obtained by accumulating the similarities of the relevant text documents weighted by the transfer function. If the two text documents are relevant to the target images based on sampled relevance indicator variables, their similarity will be accumulated for computing the image similarity; otherwise the text similarity will be neglected since they describe irrelevant content to the images.

It can be proved that the above target similarity is a valid Mercer kernel function, which is the positive semi-definite by the Mercer theorem:

THEOREM 1. *Given a positive semi-definite source kernel k , $s(\mathbf{x}, \tilde{\mathbf{x}})$ in Eq. (2.2) is a valid Mercer kernel.*

Proof. We show that s is a positive semi-definite kernel. For a set of finite target instances $\{\mathbf{x}_i, 1 \leq i \leq l\}$ and corresponding coefficients $\{\alpha_i, 1 \leq i \leq l\}$, we have

$$(2.3) \quad \begin{aligned} \sum_{i,j=1}^l \alpha_i \alpha_j s(\mathbf{x}_i, \mathbf{x}_j) &= \int_{\Delta \times \Delta} \sum_{i,j=1}^l \alpha_i \alpha_j T(\mathbf{x}_i, \mathbf{y}) T(\mathbf{x}_j, \tilde{\mathbf{y}}) \\ &\cdot k(\mathbf{y}, \tilde{\mathbf{y}}) p(\mathbf{y}) p(\tilde{\mathbf{y}}) d\mathbf{y} d\tilde{\mathbf{y}} = \int_{\Delta \times \Delta} \left(\sum_{i=1}^l \alpha_i T(\mathbf{x}_i, \mathbf{y}) p(\mathbf{y}) \right) \\ &\cdot \left(\sum_{i=1}^l \alpha_i T(\mathbf{x}_i, \tilde{\mathbf{y}}) p(\tilde{\mathbf{y}}) \right) k(\mathbf{y}, \tilde{\mathbf{y}}) d\mathbf{y} d\tilde{\mathbf{y}} \\ &= \int_{\Delta \times \Delta} \beta(\mathbf{y}) \beta(\tilde{\mathbf{y}}) k(\mathbf{y}, \tilde{\mathbf{y}}) d\mathbf{y} d\tilde{\mathbf{y}} \geq 0 \end{aligned}$$

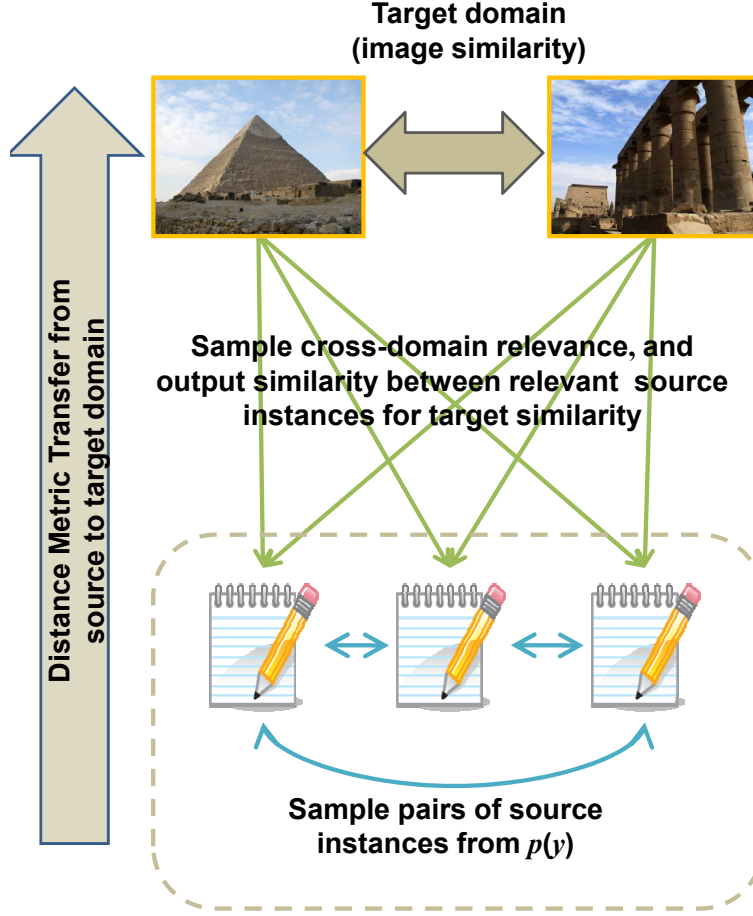


Figure 1: Illustration of computing the target image similarity from the relevant text documents sampled from cross-domain metric sampling process. Although the pyramid (the left) and Luxor Temple (the right) images look visually different, both of them are semantically related in the context of text documents introducing Egypt architecture.

where $\beta(\mathbf{y}) = \sum_{i=1}^m \alpha_i T(\mathbf{x}_i, \mathbf{y}) p(\mathbf{y})$ and the last inequality follows from the semi-definite positivity of the kernel k . Thus $s(\mathbf{x}, \tilde{\mathbf{x}})$ is a valid Mercer kernel.

According to the definition of the Mercer kernel, there exists a function $\phi(\mathbf{x})$ that maps each target instance \mathbf{x} to $\phi(\mathbf{x})$ in an output feature space, in which the inner product is implicitly given by $s(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}}), \cdot \rangle$. Hence, the (squared) distance between two target instances can be computed as

$$(2.4) \quad \begin{aligned} d_{\text{tgt}}(\mathbf{x}, \tilde{\mathbf{x}}) &= \langle \phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}}), \phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}}) \rangle \\ &= s(\mathbf{x}, \mathbf{x}) + s(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - 2s(\mathbf{x}, \tilde{\mathbf{x}}) \end{aligned}$$

This distance function formally satisfies the mathematical properties of a *metric*, i.e., this distance metric in the target space is symmetric, nonnegative and satisfying the triangle inequality.

We define the target similarity in terms of a population expectation w.r.t. the true distribution $p(\mathbf{y})$ in Eq. (2.2). However, in reality the underlying $p(\mathbf{y})$ is unknown beforehand. Alternatively, we can consider the empirical version of the *true* target similarity. Given a set of source instances $\mathbf{y}_i, 1 \leq i \leq n$ i.i.d. sampled from $p(\mathbf{y})$, the empirical distribution is $p_n(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \delta[\mathbf{y} - \mathbf{y}_i]$ with the Dirac's delta function $\delta[\cdot]$. Substituting $p(\mathbf{y})$ with $p_n(\mathbf{y})$, we obtain the following *empirical* target similarity

$$(2.5) \quad \begin{aligned} s_n(\mathbf{x}, \tilde{\mathbf{x}}) &= \int_{\Delta \times \Delta} T(\mathbf{x}, \mathbf{y}) T(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) k(\mathbf{y}, \tilde{\mathbf{y}}) p_n(\mathbf{y}) p_n(\tilde{\mathbf{y}}) d\mathbf{y} d\tilde{\mathbf{y}} \\ &= \frac{1}{n^2} \sum_{i,j=1}^n \{T(\mathbf{x}, \mathbf{y}_i) T(\tilde{\mathbf{x}}, \mathbf{y}_j) k(\mathbf{y}_i, \mathbf{y}_j)\} \end{aligned}$$

Note that in the cross-domain metric sampling process

the *pairs* of source instances are sampled independently. However, in $s_n(\mathbf{x}, \tilde{\mathbf{x}})$ the pairs of $(\mathbf{y}_i, \mathbf{y}_j)$ are not statistically independent although \mathbf{y}_i 's are independently sampled from $p(\mathbf{y})$. The conventional analysis tools for i.i.d. samples do not apply in this case, and instead we apply McDiarmid inequality [5][11] to bound the difference between the true and empirical target similarity. We show that $s_n(\mathbf{x}, \tilde{\mathbf{x}})$ asymptotically converges to $s(\mathbf{x}, \tilde{\mathbf{x}})$ at rate $O(\frac{1}{\sqrt{n}})$:

THEOREM 2. *Given any two target instances \mathbf{x} and $\tilde{\mathbf{x}}$, with probability at least $1 - \mu$, we have*

$$(2.6) \quad |s_n(\mathbf{x}, \tilde{\mathbf{x}}) - s(\mathbf{x}, \tilde{\mathbf{x}})| \leq \frac{1}{n} |\varrho(\mathbf{x}, \tilde{\mathbf{x}})| + B \sqrt{\frac{2}{n} \ln \frac{2}{\mu}}$$

where B is the upper bound of the kernel function, i.e., $|k(\mathbf{y}, \mathbf{z})| < B$ for any \mathbf{y} and \mathbf{z} ; and

$$(2.7) \quad \varrho(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [T(\mathbf{x}, \mathbf{y}) T(\tilde{\mathbf{x}}, \mathbf{y}) k(\mathbf{y}, \mathbf{y})] - s(\mathbf{x}, \tilde{\mathbf{x}})$$

REMARK 1. *Here, a bounded kernel function is a rather mild condition as most of kernels have finite upper bound, e.g., the absolute value of the cosine kernel is always less than one and the linear kernel is bounded as long as the support Δ of $p(\mathbf{y})$ is compact.*

We leave the proof of the theorem in the appendix.

The empirical target similarity function s_n can be rewritten in a compact matrix form as

$$(2.8) \quad \begin{aligned} s_n(\mathbf{x}, \tilde{\mathbf{x}}) &= \sum_{i,j=1}^n \{T(\mathbf{x}, \mathbf{y}_i) T(\tilde{\mathbf{x}}, \mathbf{y}_j) k(\mathbf{y}_i, \mathbf{y}_j)\} \\ &= \mathbf{v}_T(\mathbf{x})^T K \mathbf{v}_T(\tilde{\mathbf{x}}) \end{aligned}$$

where K is a $n \times n$ kernel matrix with $K = [k(\mathbf{y}_i, \mathbf{y}_j)]_{n \times n}$, and the corresponding distance metric d_{tgt} is

$$(2.9) \quad \begin{aligned} d_{\text{tgt}}(\mathbf{x}, \tilde{\mathbf{x}}) &= s_n(\mathbf{x}, \mathbf{x}) + s_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - 2s_n(\mathbf{x}, \tilde{\mathbf{x}}) \\ &= (\mathbf{v}_T(\mathbf{x}) - \mathbf{v}_T(\tilde{\mathbf{x}}))^T K (\mathbf{v}_T(\mathbf{x}) - \mathbf{v}_T(\tilde{\mathbf{x}})) \end{aligned}$$

where $\mathbf{v}_T(\cdot)$ defines a mapping

$$(2.10) \quad \mathbf{v}_T : \mathbb{R}^t \rightarrow \mathbb{R}^n, \mathbf{x} \mapsto \mathbf{v}_T(\mathbf{x})$$

from the target space \mathbb{R}^t to a n dimensional vector space \mathbb{R}^n :

$$(2.11) \quad \mathbf{v}_T(\mathbf{x}) = [T(\mathbf{x}, \mathbf{y}_1) \quad T(\mathbf{x}, \mathbf{y}_2) \quad \cdots \quad T(\mathbf{x}, \mathbf{y}_n)]^T$$

These n source instances $\mathbf{y}_i, 1 \leq i \leq n$ can be seen as ‘‘landmark’’ instances in the source space, and this mapping summarizes the relevance of the target instance \mathbf{x} to these landmark instances. It asymptotically captures the target metric structure as $n \rightarrow +\infty$ by Theorem 2. Note that for ease of notation we discard the constant factor $\frac{1}{n^2}$ in Eq. (2.5) here.

3 Transfer Learning of Distance Functions

The transfer function $T(\cdot, \cdot)$ plays the central role in connecting the metric structures in target and source spaces as shown in Eq. (2.5) (2.9). To learn the transfer function, two aspects can be explored to reveal the intrinsic distance structure in the target space.

The most direct component which provides the connection between the source and target domains is a set $\mathcal{C} = \{(\mathbf{x}_k, \mathbf{y}_k)\}$ of observed pairs of relevant instances between the two domains. For example, this can be images and their surrounding text; or the equivalent English translation to a Chinese document. This provides the bridge needed for transfer learning of metrics across heterogeneous spaces.

In the cross-domain metric sampling process, only source similarity is sampled to compute the target similarity. On the other hand, a priori information about the structure of the target distance is directly available in the *original target* space. We refer to this as *structural information* about the target space. The learned distance should inherit the metric structures of the original target space as well. Specifically, given a set of target instances, let $Q_{p,q}$ denote the similarity between two instance \mathbf{x}_p and \mathbf{x}_q , $1 \leq p, q \leq m$ in the original target space. Then they can be utilized to make the target distance (2.9) consistent with the metric structure of the original target space. Moreover, aligning source and target metric structures also maximizes the cross-domain correlations, which equivalently imposes a global consistency prior to link the relevant instances in heterogeneous domains. We will reveal this connection in the later.

Now we propose an algorithm in order to optimize the distance transfer process between the two spaces. The optimization problem over the transfer function T is defined as follows:

$$(3.12) \quad \min_T \gamma \mathcal{L}_\epsilon(T, \mathcal{C}) + \frac{\eta}{2} \sum_{p,q=1}^m g(Q_{p,q}, d_{\text{tgt}}(\mathbf{x}_p, \mathbf{x}_q)) + \Omega(T)$$

The expression in Eq. (3.12) measures the effectiveness of the distance transfer process, with the corresponding balancing parameters γ and η .

- The first term encodes how the source and target spaces are linked by T in \mathcal{C} . As aforementioned, the transfer function T measures the probability of source and target instances being relevant to each other. Based on this probabilistic explanation, we choose the negative logistic loss to estimate the transfer function by maximizing the likelihood over

the pairs of the relevant instances in \mathcal{C} :

$$(3.13) \quad \begin{aligned} \mathcal{L}_\varepsilon(T, \mathcal{C}) &= \sum_{\mathcal{C}} -\log \{(1 - \varepsilon)T(\mathbf{x}_k, \mathbf{y}_k) + \varepsilon(1 - T(\mathbf{x}_k, \mathbf{y}_k))\} \end{aligned}$$

Here we consider the noise in \mathcal{C} , which flips a pair of irrelevant source and target instances to relevant one in \mathcal{C} with probability $\varepsilon \in [0, 1]$. By minimizing the objective function in Eq. (3.12) alternately between ε and T in a coordinate descent manner, they can be simultaneously inferred. When fixing T , minimizing w.r.t. ε is a standard convex optimization problem. In Section 3.2, we will present the optimization of T with fixed ε . Minimizing this term makes the output of the transfer learning process consistent with observations of the paired source and target samples, so that the transfer function has larger output on a pair of target and source instances in \mathcal{C} .

- The second term measures the consistency of the target distance with the structural information about the original target metric space. We choose the loss function $g(Q_{p,q}, d_{\text{tgt}}(\mathbf{x}_p, \mathbf{x}_q)) = Q_{p,q}d_{\text{tgt}}(\mathbf{x}_p, \mathbf{x}_q)$ in this paper. If two target instances are similar according to $Q_{p,q}$, their target distances are minimized; otherwise, their distances will be maximized.
- The last term $\Omega(T)$ regularizes learning of the transfer learning process, which will be extended in the following section when establishing the transfer function.

We note that the expression above contains several terms, the most important of which correspond to the effects of the co-occurrence data and auxiliary data in the effectiveness of the distance function. The relative importance of co-occurrence data and auxiliary data in the objective function are regulated by the balancing parameters γ and η . The expression discussed above is an optimization problem designed to determine the best translator function T . However, in order to determine this optimum function, we need to further express it in the form of other simplified semantic topic space matrices. This results in a closed form description of the translator function, whose parameters can be optimized. The decomposition of T into semantic topic spaces will be discussed in the next section.

3.1 Designing the Transfer Function The source and target spaces are quite different in terms of their feature representation. To establish their connection, we must discover a common structure which can link them together. It is possible to discover some common factors

to describe the heterogeneous instances simultaneously. For example, a text document usually contains several topics which describe different aspects of the underlying concepts at a higher level. In a web page depicting *bird*, the related topics, such as the head, body and tail, are described in its textual part. Meanwhile, there is a corresponding *bird* image illustrating them. By aligning the topics of the text (i.e., the source instances) and images (i.e., the target instances) in a space with several unspecified topics, they can be semantically linked together by investigating their co-occurrence data. For this purpose, we construct two transformation matrices U and V to map the source and target instances into a common space with r unspecified factors to link heterogeneous domains as follows. This dimensionality is essentially the number of topics, because each dimension in this space represents a latent topic for semantic correspondence. We will show that the translator function can be expressed in terms of these topic spaces, and therefore the key to finding an optimal translator function T is to determine the optimal translation matrices U and V (or, as we will see later, an appropriate function of them). The matrices U and V are defined as follows.

$$(3.14) \quad \begin{aligned} U &\in \mathbb{R}^{r \times s} : \mathbb{R}^s \rightarrow \mathbb{R}^r, \mathbf{y} \mapsto U\mathbf{y}, \\ V &\in \mathbb{R}^{r \times t} : \mathbb{R}^t \rightarrow \mathbb{R}^r, \mathbf{x} \mapsto V\mathbf{x} \end{aligned}$$

Then, the transfer function T is a function of the source and target instances as

$$(3.15) \quad T(\mathbf{x}, \mathbf{y}) = f(\langle V\mathbf{x}, U\mathbf{y} \rangle) = f(\mathbf{x}^T V^T U \mathbf{y}) = f(\mathbf{x}^T S \mathbf{y})$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and the matrix S is used to briefly denote $V^T U$; f is the activation function acting on $\mathbf{x}^T S \mathbf{y}$. We choose the logistic sigmoid function as f , i.e., $f(\theta) = \frac{1}{1 + e^{-\theta}}$. It is differentiable and real-valued in the interval $[0, 1]$. In this case, $T(\mathbf{x}, \mathbf{y})$ outputs the probability that \mathbf{x} and \mathbf{y} are a pair of the relevant target and source instances.

We can use the conventional squared norm $\Omega(T) = \frac{1}{2} (\|U\|_F^2 + \|V\|_F^2)$ to regularize the transfer function T on two transformations respectively, where $\|\cdot\|_F$ is the Frobenius norm. However, since this $\Omega(T)$ is not convex, the global minima cannot be guaranteed by a solution. Fortunately, it is possible to learn S directly by the trace norm as in [18] [2]. It is defined as follows

$$(3.16) \quad \|S\|_\Sigma = \inf_{S=U^T V} \frac{1}{2} (\|U\|_F^2 + \|V\|_F^2)$$

The trace norm is a convex function of S , and can be computed as the sum of its singular values. It is a surrogate of matrix rank [7], and minimizing it can

limit the dimensionality r of the latent factor space. In other words, *minimizing the trace norm results in the fewest topics to explain the correspondence between text and images*. This regularizes the transfer function by the preference to a small size of intermediate topics to link heterogeneous domains as stated in the information bottleneck method [19].

3.2 Implementation Details We use the second term in (3.12) to leverage the similarity structure in original target space. The loss function penalizes the large distance between similar instances. We can rewrite this term as

$$\begin{aligned}
(3.17) \quad & \frac{1}{2} \sum_{p,q=1}^m g(Q_{p,q}, d_{\text{tgt}}(\mathbf{x}_p, \mathbf{x}_q)) = \frac{1}{2} \sum_{p,q=1}^m Q_{p,q} d_{\text{tgt}}(\mathbf{x}_p, \mathbf{x}_q) \\
& = \sum_{p,q=1}^m Q_{p,q} \mathbf{v}_T(\mathbf{x}_p)^T K \mathbf{v}_T(\mathbf{x}_q) \\
& - \sum_{p,q=1}^m Q_{p,q} \mathbf{v}_T(\mathbf{x}_p)^T K \mathbf{v}_T(\mathbf{x}_q) \\
& = \text{tr}(\Xi(S)^T K \Xi(S) D) - \text{tr}(\Xi(S)^T K \Xi(S) Q) \\
& = \text{tr}(K \Xi(S) L \Xi(S)^T)
\end{aligned}$$

where $\Xi(S) = [\mathbf{v}_T(\mathbf{x}_1), \mathbf{v}_T(\mathbf{x}_2), \dots, \mathbf{v}_T(\mathbf{x}_m)]$ is a $n \times m$ matrix dependent on S , and tr denotes the trace operation of a matrix. D is a diagonal $m \times m$ matrix with each diagonal element being the corresponding row summation of Q , and $L = D - Q$ is the Laplacian matrix. Then the objective function in Equation (3.12) with fixed ε can be rewritten as

$$\begin{aligned}
(3.18) \quad & \min_S \gamma \sum_C -\log \{(1 - \varepsilon) f(\mathbf{x}_k^T S \mathbf{y}_k) + \varepsilon (1 - f(\mathbf{x}_k^T S \mathbf{y}_k))\} \\
& + \eta \text{tr}(K \Xi(S) L \Xi(S)^T) + \|S\|_\Sigma
\end{aligned}$$

The objective function of Eq. (3.18) contains non-differentiable trace norm regularizer and a differentiable part. In order to represent the objective function of Eq. (3.18) more succinctly, we introduce the differentiable part $F(S)$ as

$$\begin{aligned}
(3.19) \quad & F(S) \\
& = \gamma \sum_C -\log \{(1 - \varepsilon) f(\mathbf{x}_k^T S \mathbf{y}_k) + \varepsilon (1 - f(\mathbf{x}_k^T S \mathbf{y}_k))\} \\
& + \eta \text{trace}(K \Xi(S) L \Xi(S)^T)
\end{aligned}$$

Then, the objective function of Eq. (3.19) can be rewritten as $F(S) + \|S\|_\Sigma$. For the differentiable part $F(S)$, its gradient $\nabla F(S)$ can be computed as

$$\begin{aligned}
(3.20) \quad & \nabla F(S) = \gamma \sum_C \left\{ -\frac{(1 - 2\varepsilon) f'(a_k)}{(1 - \varepsilon) f(a_k) + \varepsilon (1 - f(a_k))} \mathbf{x}_k \mathbf{y}_k^T \right\} \\
& + \eta \Gamma
\end{aligned}$$

where f' is the derivative of f , $a_k = \mathbf{x}_k^T S \mathbf{y}_k$, and Γ is the $t \times s$ gradient matrix of $\text{tr}(K \Xi(S) L \Xi(S)^T)$ w.r.t.

S , whose (u, v) th element can be computed as

$$\begin{aligned}
(3.21) \quad & \Gamma_{uv} = \frac{\partial \text{tr}(K \Xi(S) L \Xi(S)^T)}{\partial S_{uv}} = 2 \text{tr} \left[(K \Xi(S) L)^T \frac{\partial \Xi(S)}{\partial S_{uv}} \right]
\end{aligned}$$

Here $\frac{\partial \Xi(S)}{\partial S_{uv}}$ is a $n \times m$ matrix, and its (i, j) th element is

$$(3.22) \quad \left[\frac{\partial \Xi(S)}{\partial S_{uv}} \right]_{ij} = f'(\mathbf{x}_j^T S \mathbf{y}_i) X_{ju} Y_{iv},$$

Denote $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]^T$ and $Y = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]^T$ are $m \times t$ and $n \times s$ data matrices, then X_{ju} and Y_{iv} are the u th and v th dimensional features in \mathbf{x}_j and \mathbf{y}_i , respectively. Combining Eq. (3.21) and (3.22), with some algebraic operations, the gradient matrix Γ can be rewritten in a compact form as

$$(3.23) \quad \Gamma = X^T (K \Xi(S) L \circ H)^T Y$$

where \circ denotes the element-wise product of two matrices, and H is a $n \times m$ matrix with its elements as $H_{ij} = f'(\mathbf{x}_j^T S \mathbf{y}_i)$.

We apply the proximal gradient method [20] to minimize the loss function with trace norm regularizer. In order to optimize this objective function, the proximal gradient method quadratically approximates it by Taylor expansion at current S_τ and Lipschitz coefficient α as follows

$$\begin{aligned}
(3.24) \quad & Q(S, S_\tau) = \frac{\alpha}{2} \|S - G_\tau\|_F^2 + \|S\|_\Sigma + F(S_\tau) \\
& - \frac{1}{2\alpha} \|\nabla F(S_\tau)\|_F^2
\end{aligned}$$

and

$$(3.25) \quad G_\tau = S_\tau - \alpha^{-1} \nabla F(S_\tau)$$

Algorithm 1 summarizes the proximal gradient based method to optimize the expression in Eq. (3.18). As shown, S can be updated by minimizing $Q(S, S_\tau)$ with the fixed S_τ iteratively. This can be solved by singular value thresholding [7] in line 4 in Algorithm 1. As pointed out in [20], the convergence of the proximal gradient algorithm in loop 2-5 can be accelerated by making an initial estimate of α and increasing it by a constant factor λ until $F(S_{\tau+1}) + \|S_{\tau+1}\|_\Sigma \leq Q(S_{\tau+1}, S_\tau)$.

4 Related Work

Various methods have been proposed to learn distance metric by leveraging the correspondence knowledge across heterogeneous domains [13][24]. In [13], *multi-label distance metric learning* (ML-DML) is proposed

Algorithm 1 Proximal Gradient Solver for (3.18) with fixed ε .

input Correspondence set \mathcal{C} , source kernel matrix K , and Laplacian matrix L , balancing parameters γ and η .

1 Initialize $S_\tau \leftarrow 0$ and $\tau \leftarrow 0$.

repeat

repeat

2 Initialize $\alpha \leftarrow \alpha_0$.

3 Set $G_\tau = S_\tau - \alpha^{-1} \nabla F(S_\tau)$.

4 Update $S_{\tau+1} \leftarrow U \text{diag} \left(\sigma - \frac{\gamma}{\alpha} \right)_+ V^T$. Here

$U \text{diag}(\sigma) V^T$ gives the SVD of G_τ .

5 Set $\alpha \leftarrow \lambda \alpha$

until $F(S_{\tau+1}) + \|S_{\tau+1}\|_\Sigma \leq Q(S_{\tau+1}, S_\tau)$.

6 $\tau \leftarrow \tau + 1$.

until Convergence or maximum iteration number achieves.

to learn a distance metric on the target space from the observed occurrence between source and target instances. It explores the semantic correlation of images and the keywords in the associated text documents, and learns a Mahalanobis metric in closed form. The problem of learning the distance metric in target spaces can also be seen as a kind of transfer learning from heterogeneous data in different feature spaces. The work in [23] proposes *heterogeneous transfer learning* (HTL) algorithm, which uses both text and visual words as source information to extract a new latent feature representation for each image, which could be used to compute a new distance metric in the target image space. However, both of these algorithms do not explore the problem of transfer learning of distance metrics. As mentioned in the introduction section, we assume the metric structure has a smaller semantic gap between the low-level features and high-level semantic concepts. The goal of this paper is to transfer this metric structure into the target space, which can result in more effective distance functions in the target space. To the best of our knowledge, the method in this paper is one of the first to demonstrate how to “translate” distance structures across heterogeneous domains and show the results for the case of a practical problem.

Finally, we distinguish the proposed translator function from other latent models. Previous latent methods, such as Latent Semantic Analysis [12], Probabilistic Latent Semantic Analysis [10] and Latent Dirichlet Allocation [4], are restricted to latent factor discovery from the co-occurrence observations. On the contrary, in this paper, the goal of our approach is to establish the correspondence between the underlying distance metrics in the source and target space so that in the target space the obtained target feature space has a tractable semantic gap. To the best of our knowledge, it is one

of the first algorithms to address such a heterogeneous distance transfer problem.

5 Experiments

In this section, we compare the proposed distance metrics derived from the transfer learning process to other natural distance metrics which are typically used for a variety of applications. We will show that our approach provides superior results to the other methods.

One challenge is to design a method for qualitative evaluation of the distance metrics. Since distance metrics are inherently semantic functions which are used as subroutines in the context of different kinds of applications, it is natural to test the effectiveness of using different kinds of distance functions on a particular application in order to measure its quality. For example, one can test the effectiveness of a nearest neighbor classifier with the use of different kinds of distance metrics. The idea is that a distance function which retains the most meaningful aspects of the feature space, and adjusts for the most noisy aspects is most likely to work effectively within the context of an application such as classification. In general, for unsupervised problems such as clustering and distance function design, qualitative tests on real data are generally intended to be designed in an evidentiary way, so as to provide an understanding of the advantages of using a particular kind of approach for distance function design.

5.1 Data Sets In order to test our approach we needed paired image and text documents. Furthermore, since we used classification as our base application, we also needed some class labels on the images in order to test the effectiveness of the distance function learning process. The data sets consist of the Corel image data set and a collection of Flickr web pages. We use 10 categories to evaluate the effectiveness on the image classification task. To collect paired image and text collections for experiments, the names of these 10 categories are used as query keywords to crawl web pages from the Flickr web site and Wikipedia. Table 1 shows the number of crawled web pages for each category. Flickr is an image sharing web site, where the users can share images with their friends and other users, and make textual comments and tags on the shared images. In each crawled web page, the images and the corresponding text documents are used to establish correspondence between text and images. The textual parts of the crawled web pages are used as source instances for metric transfer, and the images are used as auxiliary images in training set.

For images, visual features are extracted in order to construct a multi-dimensional representation. These

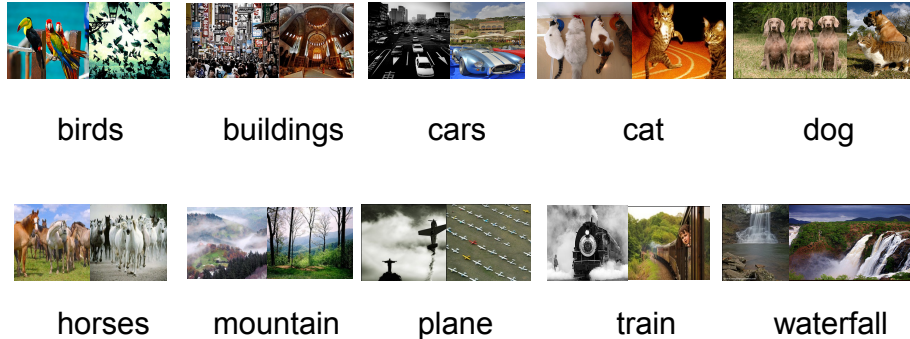


Figure 2: Illustration of example images in the data set.

Table 1: The number of the crawled web pages by each query. By using the category names as query keywords, the returned web pages are crawled. The images in these web pages are also collected.

Category	Crawled web pages	Category	Crawled web pages
birds	930	horses	654
buildings	9216	mountain	4153
cars	728	plane	1356
cat	229	train	457
dog	486	waterfall	22006

Table 2: The number of images in each category for performance evaluation. For performance evaluation, all the images are manually annotated with ground truth by human for evaluation purpose.

Category	Number of positive examples	Number of negative examples	Category	Number of positive examples	Number of negative examples
birds	338	349	horses	263	268
buildings	2301	2388	mountain	927	1065
cars	120	125	plane	509	549
cat	67	72	train	52	53
dog	132	142	waterfall	5153	5737

include 500 dimensional bag-of-word feature representation quantized from SIFT descriptor. χ^2 similarity between the target instances is used as $Q_{p,q}$ to provide metric information in the original target space. For text documents, all the tokens are extracted and stemmed, and the remaining term frequencies are used as textual features in experiments. For each category, the images are manually annotated by human to collect the ground truth labels for evaluation purpose as shown in Table 2. Nearly the same number of images are collected as the negative examples. These images contain the objects of the different categories. These categories are not exclusive which means one image can be annotated to be positive examples by more than one category. Accordingly the following experiments are conducted in such

multi-label case with binary labels for each category.

5.2 Compared Algorithms We use the following algorithms and baselines in order to test the effectiveness of our distance-transfer process.

- As the baseline, we directly compute the Euclidean distance between images based on their visual features. We refer to this metric as **ED**. This method does not use any of the additional information available in corresponding text in order to improve the quality of the distance function.
- The Kernel Multi-Label Distance Metric Learning [13] algorithm computes the image distance from the co-occurrence between image and text

instances. We refer to this algorithm as **KML-DML**. The Gaussian kernel on the image domain is used here.

- The Heterogeneous Transfer Learning [24] method is a classification algorithm across heterogeneous spaces. Relational matrix between images and text documents is factorized to extract the implicit representation of target instances, based on which the distance function can be set up. We refer to this method as **HTL**.
- Finally, we test the proposed method in this paper with two kinds of text similarity measures k . One uses the linear similarity of inner product of text vectors and the other is the typical cosine similarity between text vectors. They are two of the most effective kernel similarities used for text corpus. We denote the distance translators associated with these two text similarity measures by “DT-Lin” and “DT-Cos”, respectively. We refer to this method as **DT**, with specific instantiations as **DT-Lin** and **DT-Cos** respectively.

The nearest neighbor (NN) classifier is applied to classify the images based on the above learned distances to compare their performance in classifying the images. For each image category, ten positive examples and ten negative images are randomly selected as labeled instances for the classifiers, and the remaining are used for testing. This process is repeated five times. The error rate and the associated standard deviation for each category is reported. We also use varying number of text documents as landmark source instances to construct the distance, and compare the corresponding results with related algorithms. All the parameters are tuned based on a two-fold cross-validation procedure on the selected training set, and the parameters with the best performance are selected to train the models.

5.3 Results Next, we present the error rates of the classifiers with the use of this nearest neighbor metric. Table 3 compares different algorithms in terms of their classification error rates. In this case, we used 2,000 associated images and text documents in order to learn the distance metric in the image space. Later on, we will compare the average error rates by using different numbers of text documents. From this result, we can find that among all ten categories, the proposed distance transfer, both DT-Lin and DT-Cos, performs the best on seven categories as compared with the existing methods, respectively. Moreover, as illustrated in Figure 3, in terms of average rates, both DT-Lin and DT-Cos gain a significant improvement compared with other algorithms.

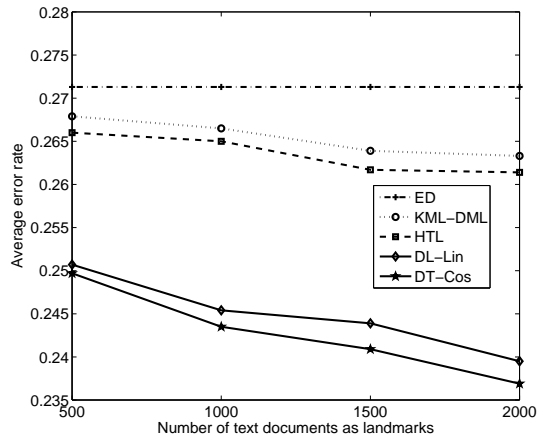


Figure 3: Average error rates of compared algorithms with varying number of text documents as landmark source instances.

As stated in Section 2, the text documents play an important landmark role of embedding the image instances by the distance transfer learning process. With more landmark source instances, the empirical target metric asymptotically converges to the true one. Therefore, it is instructive to examine the effect of increasing the number of such landmarks. In Figure 3, we illustrate the effectiveness of different algorithms with varying number of text documents. The number of documents is illustrated on the X-axis, whereas the error rate is illustrated on the Y-axis. As we can see, the error rates of the DT-Lin and DT-Cos algorithms are reduced with an increasing number of documents in the source space since more information about source metric structure is transferred to the target space. We also note that their improvements are more significant than other algorithms when more text documents are involved. This suggests that there is a real gain in the quality of the distance function through the process of transfer learning from text to images.

5.4 Computing Time Finally, we compare the computational efficiency of the different algorithms for learning the target distance metric. All the algorithms are conducted on the same computing platform with 2.10 GHz Intel CPU and 3 GB physical memory. Since the Euclidean metric is directly available without any learning process, we omit its computing time here. Table 4 shows the computing time with 2,000 text documents for learning the distance. DT-Lin and DT-Cos are much faster than HTL but slower than KML-DML, since KML-DML has a closed-form solution when learning the metric and involves only one matrix inversion

Table 3: Comparison of error rates and the deviations of the proposed distance transfer algorithms (DT-Lin and DT-Cos) compared with the other state-of-the-art transfer methods over ten categories. Our results in bold achieve smaller error rates than the other existing algorithms.

Category	ED	KML-DML	HTL	DT-Lin	DT-Cos
birds	0.2639±0.0012	0.2481±0.0008	0.2619±0.0015	0.2421±0.0010	0.2559±0.0011
buildings	0.2856±0.0002	0.2625±0.0004	0.2707±0.0021	0.2157±0.0000	0.2145±0.0004
cars	0.3027±0.0073	0.2414±0.0054	0.3065±0.0030	0.2107±0.0044	0.2031±0.0026
cat	0.2755±0.0043	0.3333±0.0040	0.2525±0.0038	0.3131±0.0084	0.2929±0.0053
dog	0.2252±0.0039	0.1802±0.0057	0.2343±0.0037	0.1802±0.0027	0.1712±0.0031
horses	0.2667±0.0019	0.3000±0.0015	0.2500±0.0021	0.2517±0.0014	0.2467±0.0018
mountain	0.3176±0.0010	0.2974±0.0008	0.3097±0.0003	0.2974±0.0005	0.2952±0.0005
plane	0.2667±0.0009	0.2633±0.0011	0.2133±0.0008	0.2633±0.0009	0.2617±0.0005
train	0.2716±0.0029	0.2593±0.0068	0.2716±0.0118	0.1924±0.0058	0.1852±0.0049
waterfall	0.2611±0.0008	0.2476±0.0015	0.2435±0.0009	0.2409±0.0002	0.2425±0.0001

Table 4: Comparison of computing time (in seconds) of different algorithms for learning the target distance metric.

Category	Computing Time
ED	N/A
KML-DML	562.52
HTL	4536.07
DT-Lin	678.93
DT-Cos	719.25

operation [13].

6 Conclusion

In this paper, we propose a transfer learning process for distance metrics, which can effectively transfer the metric information in source domain to learn an effective metric structure in target domain. For this purpose, as a bridge, we learn the distance transfer by exploring the correspondence information between the source and target spaces. The distance metric in the target space can then be constructed by embedding the target instances into a new feature vector space by a set of landmarks in the source space. The proposed method is compared with existing metric learning algorithms, and the competitive results are achieved.

Appendix

Proof of Theorem 2 Here we prove the Theorem 2. We first prove the following two lemmas.

LEMMA 1.

$$\mathbb{E}s_n(\mathbf{x}, \tilde{\mathbf{x}}) = s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n} \varrho(\mathbf{x}, \tilde{\mathbf{x}})$$

where

$$\varrho(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{E}_{\mathbf{y}} [T(\mathbf{x}, \mathbf{y}) T(\tilde{\mathbf{x}}, \mathbf{y}) k(\mathbf{y}, \mathbf{y})] - s(\mathbf{x}, \tilde{\mathbf{x}})$$

Proof.

$$\begin{aligned} \mathbb{E}s_n(\mathbf{x}, \tilde{\mathbf{x}}) &= \mathbb{E} \frac{1}{n^2} \sum_{i,j=1}^n T(\mathbf{x}, \mathbf{y}_i) T(\mathbf{x}, \mathbf{y}_j) k(\mathbf{y}_i, \mathbf{y}_j) \\ &= \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E} [T(\mathbf{x}, \mathbf{y}_i) T(\mathbf{x}, \mathbf{y}_j) k(\mathbf{y}_i, \mathbf{y}_j)] \\ &= \frac{1}{n^2} \sum_{i \neq j, i,j=1}^n \mathbb{E} [T(\mathbf{x}, \mathbf{y}_i) T(\mathbf{x}, \mathbf{y}_j) k(\mathbf{y}_i, \mathbf{y}_j)] \\ &\quad + \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{\mathbf{y}_i} [T(\mathbf{x}, \mathbf{y}_i) T(\mathbf{x}, \mathbf{y}_i) k(\mathbf{y}_i, \mathbf{y}_i)] \\ &= \frac{n(n-1)}{n^2} s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{\mathbf{y}} [T(\mathbf{x}, \mathbf{y}) T(\mathbf{x}, \mathbf{y}) k(\mathbf{y}, \mathbf{y})] \\ &= s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n} \{ \mathbb{E}_{\mathbf{y}} [T(\mathbf{x}, \mathbf{y}) T(\mathbf{x}, \mathbf{y}) k(\mathbf{y}, \mathbf{y})] - s(\mathbf{x}, \tilde{\mathbf{x}}) \} \\ &= s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n} \varrho(\mathbf{x}, \tilde{\mathbf{x}}) \end{aligned}$$

This lemma shows that $\mathbb{E}s_n(\mathbf{x}, \tilde{\mathbf{x}}) \rightarrow s(\mathbf{x}, \tilde{\mathbf{x}})$ as $n \rightarrow +\infty$.

LEMMA 2. Let $s_n^{i,\mathbf{z}}(\mathbf{x}, \tilde{\mathbf{x}})$ be the empirical estimator of s with the i th source instance \mathbf{y}_i replaced with \mathbf{z} . Then we have

$$|s_n^{i,\mathbf{z}}(\mathbf{x}, \tilde{\mathbf{x}}) - s_n(\mathbf{x}, \tilde{\mathbf{x}})| \leq \frac{2B}{n}$$

where B is the upper bound of the kernel function, i.e., $|k(\mathbf{y}, \mathbf{z})| < B$ for any \mathbf{y}, \mathbf{z} .

Proof.

$$\begin{aligned}
& |s_n^{i,z}(\mathbf{x}, \tilde{\mathbf{x}}) - s_n(\mathbf{x}, \tilde{\mathbf{x}})| \\
&= \left| \frac{1}{n^2} \sum_{j=1}^n T(\mathbf{x}, \mathbf{z}) T(\mathbf{x}, \mathbf{y}_j) k(\mathbf{z}, \mathbf{y}_j) \right. \\
&\quad \left. - \frac{1}{n^2} \sum_{j=1}^n T(\mathbf{x}, \mathbf{y}_i) T(\mathbf{x}, \mathbf{y}_j) k(\mathbf{y}_i, \mathbf{y}_j) \right| \\
&= \frac{1}{n^2} \left| \sum_{j=1}^n T(\mathbf{x}, \mathbf{y}_j) \{T(\mathbf{x}, \mathbf{z}) k(\mathbf{z}, \mathbf{y}_j) \right. \\
&\quad \left. - T(\mathbf{x}, \mathbf{y}_i) k(\mathbf{y}_i, \mathbf{y}_j)\} \right| \\
&\leq \frac{1}{n^2} \sum_{j=1}^n |T(\mathbf{x}, \mathbf{y}_j)| \{|T(\mathbf{x}, \mathbf{z}) k(\mathbf{z}, \mathbf{y}_j)| \\
&\quad + |T(\mathbf{x}, \mathbf{y}_i) k(\mathbf{y}_i, \mathbf{y}_j)|\} \\
&\leq \frac{1}{n^2} \sum_{j=1}^n \{|k(\mathbf{z}, \mathbf{y}_j)| + |k(\mathbf{y}_i, \mathbf{y}_j)|\} \\
&\leq \frac{1}{n^2} \cdot 2nB = \frac{2B}{n}
\end{aligned}$$

The second inequality applies the fact that $T(\mathbf{x}, \mathbf{y}) \leq 1$.

Now we revisit McDiarmid inequality [5] here.

THEOREM 3. (*McDiarmid Inequality*) *Given random variables $\{\mathbf{y}_i, 1 \leq i \leq n\}$, z , and a function $F(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$ which satisfies*

$$\begin{aligned}
& \sup_{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n, z} |F(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) \\
& \quad - F(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{i-1}, z, \mathbf{y}_{i+1}, \dots, \mathbf{y}_n)| \leq c_i
\end{aligned}$$

then the following inequality holds

$$\begin{aligned}
& p(|F(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbb{E}F(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)| > \varepsilon) \\
& \leq 2 \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)
\end{aligned}$$

Combining Lemma 1 and Lemma 2, applying McDiarmid inequality, we obtain the following theorem

THEOREM 4.

$$\begin{aligned}
& p\left(\left|s_n(\mathbf{x}, \tilde{\mathbf{x}}) - \left(s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n}\varrho(\mathbf{x}, \tilde{\mathbf{x}})\right)\right| > \varepsilon\right) \\
& \leq 2 \exp\left(-\frac{\varepsilon^2 n}{2B^2}\right)
\end{aligned}$$

Now we prove the Theorem 2 in the main draft. Let $\mu = 2 \exp\left(-\frac{\varepsilon^2 n}{2B^2}\right)$, we have $\varepsilon = B\sqrt{\frac{2}{n} \ln \frac{2}{\mu}}$. Then

$$\begin{aligned}
& p\left(\left|s_n(\mathbf{x}, \tilde{\mathbf{x}}) - \left(s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n}\varrho(\mathbf{x}, \tilde{\mathbf{x}})\right)\right| \leq \varepsilon\right) \\
&= 1 - p\left(\left|s_n(\mathbf{x}, \tilde{\mathbf{x}}) - \left(s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n}\varrho(\mathbf{x}, \tilde{\mathbf{x}})\right)\right| > \varepsilon\right) \\
&> 1 - 2 \exp\left(-\frac{\varepsilon^2 n}{2B^2}\right) \\
&= 1 - \mu
\end{aligned}$$

Thus with probability at least $1 - \mu$,

$$\begin{aligned}
& |s_n(\mathbf{x}, \tilde{\mathbf{x}}) - s(\mathbf{x}, \tilde{\mathbf{x}})| - \frac{1}{n} |\varrho(\mathbf{x}, \tilde{\mathbf{x}})| \\
& \leq \left|s_n(\mathbf{x}, \tilde{\mathbf{x}}) - \left(s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{1}{n}\varrho(\mathbf{x}, \tilde{\mathbf{x}})\right)\right| \leq \varepsilon = B\sqrt{\frac{2}{n} \ln \frac{2}{\mu}}
\end{aligned}$$

That is,

$$|s_n(\mathbf{x}, \tilde{\mathbf{x}}) - s(\mathbf{x}, \tilde{\mathbf{x}})| \leq \frac{1}{n} |\varrho(\mathbf{x}, \tilde{\mathbf{x}})| + B\sqrt{\frac{2}{n} \ln \frac{2}{\mu}}$$

As $n \rightarrow +\infty$, $s_n(\mathbf{x}, \tilde{\mathbf{x}})$ will converge in probability at rate $O\left(\frac{1}{\sqrt{n}}\right)$ to $s(\mathbf{x}, \tilde{\mathbf{x}})$.

Acknowledgment

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. Guo-Jun Qi is also partly supported by the IBM PhD fellowship.

References

- [1] C. AGGARWAL, *Towards systematic design of distance functions for data mining applications*, in Proceedings of the ACM KDD Conference, 2003.
- [2] Y. AMIT, M. FINK, N. SREBRO, AND S. ULLMAN, *Uncovering shared structures in multiclass classification*, in Proceedings of International Conference on Machine Learning, 2007.
- [3] A. BAR-HILLEL, T. HERTZ, N. SHENTAL, AND D. WEINSHALL, *Learning distance functions using equivalence relations*, in Proc. of International Conference on Machine Learning, 2003.
- [4] D. M. BLEI, A. Y. NG, AND M. I. JORDAN, *Latent dirichlet allocation*, Journal of Machine Learning Research, (2003), pp. 993–1022.
- [5] O. BOUSQUET AND A. ELISSEEFF, *Stability and generalization*, Journal of Machine Learning Research, (2002), pp. 499–526.
- [6] S. V. C. DORAI, *Bridging the semantic gap with computational media aesthetics*, IEEE MultiMedia, 10 (2003), pp. 15–17.
- [7] J.-F. CAI, E. CANDÉS, AND Z. SHEN, *A singular value thresholding algorithm for matrix completion*, September 2008.
- [8] W. DAI, Y. CHEN, G.-R. XUE, Q. YANG, AND Y. YU, *Translated learning: Transfer learning across different feature spaces*, in Proceedings of Advances in Neural Information Processing Systems, 2008.

- [9] J. V. DAVIS, B. KULIS, P. JAIN, S. SRA, AND I. S. DHILLON, *Information-theoretic metric learning*, in Proc. of International Conference on Machine Learning, 2007.
- [10] T. HOFMANN, *Probabilistic latent semantic analysis*, in Uncertainty in Artificial Intelligence, 1999.
- [11] R. JIN, S. WANG, AND Y. ZHOU, *Regularized distance metric learning: Theory and algorithm*, in Proc. of NIPS, 2009.
- [12] T. K. LANDAUER, P. W. FOLTZ, AND D. LAHAM, *An introduction to latent semantic analysis*, Discourse Processes, 25 (1998), pp. 259–284.
- [13] G.-J. QI, X.-S. HUA, AND H.-J. ZHANG, *Learning semantic distance from community-tagged media collection*, in Proc. of International ACM Conference on Multimedia, 2009.
- [14] R. RAINA, A. BATTLE, H. LEE, B. PACKER, AND A. NG, *Self-taught learning: Transfer learning from unlabeled data*, in Proceedings of International Conference on Machine Learning, 2007.
- [15] R. RAINA, A. NG, AND D. KOLLER, *Constructing informative priors using transfer learning*, in Proceedings of International Conference on Machine Learning, 2006.
- [16] M. SCHULTZ AND T. JOACHIMS, *Learning a distance metric from relative comparisons*, in Proc. of Advanced Neural Information Processing System, 2004.
- [17] J. SHAWE-TAYLOR AND N. CRISTIANINI, *Kernel Methods for Pattern Recognition*, Cambridge University Press, 2004.
- [18] N. SREBRO, J. RENNIE, AND T. JAAKKOLA, *Maximum margin matrix factorization*, in Proceedings of Advances in Neural Information Processing Systems, 2005.
- [19] N. TISHBY, F. PEREIRA, AND W. BIALEK, *The information bottleneck method*, in Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing, 1999, pp. 368 – 377.
- [20] K. C. TOH AND S. YUN, *An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems*. Preprint on Optimization Online, April 2009.
- [21] K. Q. WEINBERGER, J. BLITZER, AND L. K. SAUL, *Distance metric learning for large margin nearest neighbor classification*, in Proc. of NIPS, 2005.
- [22] E. P. XING, A. Y. NG, M. I. JORDAN, AND S. RUSSELL, *Distance metric learning, with application to clustering with side-information*, in Proc. of Advanced Neural Information Processing System, 2003.
- [23] Q. YANG, Y. CHEN, G. R. XUE, W. DAI, AND Y. YU, *Heterogeneous transfer learning for image clustering via the social web*, in Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore, August 2009, pp. 1–9.
- [24] Y. ZHU, S. J. PAN, Y. CHEN, G.-R. XUE, Q. YANG, AND Y. YU, *Heterogeneous transfer learning for image classification*, in Special Track on AI and the Web, associated with The Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.