Guojun Qi (UIUC)
Charu C. Aggarwal (IBM)
Thomas S. Huang (UIUC)

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

SDM Conference, 2012

Introduction

- Distance metrics are often more easily designed in some data domains than others:
  - Some domains may have more semantically well-defined features than others eg. text vs images
  - More training data may be available in some domains
- **Goal:** Use of semantic knowledge propagation for text to image distance learning.

Semantic Challenges

- The semantic challenges of image features are evident, when we attempt to recognize complex abstract concepts.
  - The visual features often fail to discriminate such concepts.
- Distance functions naturally work better with features that have semantic interpretability.
  - Similarity is usually designed on the basis of application-specific semantic criteria.
- Text features are inherently friendly to the similarity computation process in a way that is often a challenge for image representations.

Observations in the Context of Web and Social Networks

- In many real web and social media applications, it is possible to obtain *co-occurrence information* between text and images.

- Tremendous amount of linkage between text and images on the web, social media and information networks
  - In web pages, the images co-occur with text on the same web page.
  - Comments in image sharing sites.
  - Posts in social networks.

Learning from Semantic Bridges

- The copious availability of bridging relationships between text and images in the context of web and social network data can be leveraged for better learning models.
  - The goal is to learn similarity in one domain with the use of knowledge from another
- It is reasonable to assume that the content of the text and the images are highly correlated in both scenarios.
- The relationships between text and images can be used in order to facilitate the learning process.

Modeling with Topic Spaces

- Develop a mathematical model for the functional relationships between text and image features, so as to *indirectly transfer semantic knowledge through feature transformations*.

- This feature transformation is accomplished by mapping instances from different domains into a common space of unspecified topics.

- This is used as a bridge to semantically connect the two heterogeneous spaces.

Broad Approach

- Design a transfer function which represents the functional relationships between images and text (from the common topic space).

- Both the correspondence information and auxiliary image training set are used to learn the transfer function.
  - Links the instances across heterogeneous text and image spaces.
  - Follow the principle of parsimony and encode as few topics as possible.

- After the transfer function is learned, the similarity knowledge can be propagated from one domain to the other.

Notations and Definitions

- Let $\mathbb{R}^s$ and $\mathbb{R}^t$ be the source and target feature spaces, with dimensionalities 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 $\mathbb{R}^t$.

- The source space may use a particular kind of similarity function, which is the most effective for processing in that domain.

  – Eg. Cosine in text domain

- The connection between source and target domains is provided by a set $\mathcal{C} = \{(\mathbf{x}_k, \mathbf{y}_k)\}$ of observed pairs of relevant instances between the two domains.

Source Similarity Kernel Function

- 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.

- Assume all the source instances are sampled from a true distribution $p(\mathbf{y})$.

- The kernel similarity together with $p(\mathbf{y})$ completely describes the metric structure between source instances.

Transfer Function Definition

- 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

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

- In order to transfer the metric structure from source domain to target domain, we define a random variable $1\!\mathrm{I}_{\mathsf{Rel}}(\mathbf{x}, \mathbf{y})$ to indicate the cross-domain relevance between a target instance $\mathbf{x}$ and a source instance $\mathbf{y}$.

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

Leveraging the Transfer Function

- Use the cross-domain metric sampling process to compute the similarity between the target instances $\mathbf{x}$ and $\tilde{\mathbf{x}}$, and take expectation over multiple samples:
  - Sample a pair of source instances $\mathbf{y}$ and $\tilde{\mathbf{y}}$ from $p(\mathbf{y})$.
  - Sample $1\!\!\mathrm{I}_{\mathsf{Rel}}(\mathbf{x}, \mathbf{y}) \sim \mathbb{B}(T(\mathbf{x}, \mathbf{y}))$ and $1\!\!\mathrm{I}_{\mathsf{Rel}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \sim \mathbb{B}(T(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}))$ to decide whether $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are relevant to $\mathbf{x}$ and $\tilde{\mathbf{x}}$, respectively.
  - If both are relevant, i.e., $1\!\!\mathrm{I}_{\mathsf{Rel}}(\mathbf{x}, \mathbf{y}) \cdot 1\!\!\mathrm{I}_{\mathsf{Rel}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = 1$, output $k(\mathbf{y}, \tilde{\mathbf{y}})$ as the target similarity between $\mathbf{x}$ and $\tilde{\mathbf{x}}$; otherwise, output 0.

Estimating the Source Distribution

- The underlying $p(\mathbf{y})$ of source instances is unknown beforehand.

- We use 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:

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

Learning the Transfer Function

- The key to an effective transfer learning process is to learn the function T.

- We need to formulate an optimization problem which maximizes the correspondence between the two spaces.

- Set up a *canonical form* for the transfer function in the form of matrices which represent topic spaces.

- The parameters of this canonical form will be optimized in order to learn the transfer function

Learning the Transfer Function

- We propose to optimize the following problem to learn the semantic transfer function:

$$\min_{T} \gamma \mathcal{L}_{\varepsilon}(T, \mathcal{C}) + \frac{\eta}{2} \sum_{p,q=1}^{m} g\left(Q_{p,q}, d_{\mathrm{tgt}}\left(\mathbf{x}_p, \mathbf{x}_q\right)\right) + \Omega\left(T\right) \qquad (3)$$

- η is a balancing parameter

- $Q(p, q)$ measures the similarity of x_p and x_q in original target space

Co-Occurrence Term

- We choose the negative logistic loss to estimate the transfer function by maximizing the likelihood over the pairs of the relevant instances in $\mathcal{C}$:

$$\mathcal{L}_{\varepsilon}(T, \mathcal{C}) = \sum_{\mathcal{C}} - \log \{(1 - \varepsilon) T(\mathbf{x}_k, \mathbf{y}_k) + \varepsilon(1 - T(\mathbf{x}_k, \mathbf{y}_k))\} \tag{4}$$

- Minimizing this term makes the output of the transfer learning process consistent with observations of the paired source and target samples.

Designing the Transfer Function

- We will design the canonical form of the transfer function in terms of underlying *topic spaces*.

- This provides a closed form to our transfer function, which can be effectively optimized.

- Topic spaces provide a natural intermediate representation which can semantically link the information between the two domains

Designing the Transfer Function

- Topic spaces are represented by transformation matrices.
  $U \in \mathbb{R}^{r \times s} : \mathbb{R}^s \to \mathbb{R}^r, y \mapsto Uy$
  $V \in \mathbb{R}^{r \times t} : \mathbb{R}^t \to \mathbb{R}^r, x \mapsto Vx$

- The transfer function is defined as a function of the source and target instances by computing the inner product in our hypothetical topic space, which is implied by these transformation matrices:

$$T(x, y) = f(\langle Vx, Uy \rangle) = f(x^T V^T U y) = f(x^T S y)$$

- The function $f(\cdot)$ is the logistic sigmoid function:

$$f(\theta) = 1/(1 + e^{-\theta}) \tag{5}$$

Observations

- The transfer function maps to $[0, 1]$ because of the use of the logistic sigmoid function

- The choice of the transformation matrices (or rather the product matrix $V^T U$) impacts the transfer function T directly.

- We will use the notation S in order to briefly denote the matrix $V^T U$.

- It suffices to learn this product matrix S rather than the two transformation matrices separately.

Regularization

- Use conventional squared norm for regularization.

- $\Omega(T) = \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$

- Use trace-norm as a substitute to force convexity

- It is defined as follows:

$$\|S\|_\Sigma = \inf_{S=V^TU} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$$

Objective Function after Regularization

- The regularized objective function can be rewritten as follows:

$$\begin{aligned}&\min_{S}\gamma\sum_{\mathcal{C}}-\log\left\{(1-\varepsilon)f(\mathbf{x}_k^T S\mathbf{y}_k)+\varepsilon(1-f(\mathbf{x}_k^T S\mathbf{y}_k))\right\}\\&+\eta\,\mathrm{tr}\left(K\Xi(S)L\Xi(S)^T\right)+\|S\|_{\Sigma}\end{aligned} \quad (6)$$

- $\Xi(S)=[\mathbf{v}_T(\mathbf{x}_1),\mathbf{v}_T(\mathbf{x}_2),\cdots,\mathbf{v}_T(\mathbf{x}_m)]$ is a $n\times m$ matrix

- L is the Laplacian of the similarity matrix Q

- Objective function has been rewritten after regularization and simplification of second term

Objective Function Decomposition

- Objective function contains a differentiable part and non-differentiable part

- Separate out into differentiable and non-differentiable components

$$O = F(S) + \|S\|_{\Sigma}$$

- Differentiable part is:

$$\begin{aligned} & F(S) \\ & = \gamma \sum_{\mathcal{C}} -\log\left\{(1-\varepsilon) f(\mathbf{x}_k^T S \mathbf{y}_k) + \varepsilon(1 - f(\mathbf{x}_k^T S \mathbf{y}_k))\right\} \\ & + \eta \,\mathrm{trace}\left(K \Xi(S) L \Xi(S)^T\right) \end{aligned} \tag{7}$$

Objective Function Gradient

- The gradient of the function needs to be evaluated in order to enable the iterative method

- The gradient $\nabla F(S_\tau)$ can be computed as follows:

$$\nabla F(S) = \gamma \sum_{\mathcal{C}} \left\{ -\frac{(1-2\varepsilon)f'(a_k)}{(1-\varepsilon)f(a_k) + \varepsilon(1-f(a_k))} \mathbf{x}_k \mathbf{y}_k^T \right\} \quad (8)$$
$$+\eta\Gamma$$

- Γ is the $t \times s$ gradient matrix of $\mathrm{tr}\left(K\Xi(S)L\Xi(S)^T\right)$ w.r.t. S

Proximal Gradient Method

- 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

$$
\begin{aligned}
& Q(S, S_\tau) = \frac{\alpha}{2} \|S - G_\tau\|_F^2 + \|S\|_\Sigma + F(S_\tau) \\
& -\frac{1}{2\alpha} \|\nabla F(S_\tau)\|_F^2
\end{aligned} \tag{9}
$$

- Where G_τ is as follows:

$$
G_\tau = S_\tau - \alpha^{-1} \nabla F(S_\tau) \tag{10}
$$

- S can be updated by minimizing $Q(S, S_\tau)$ with the fixed S_τ iteratively.
  - Can be solved by singular value thresholding

Evaluation

- Need to design a method for qualitative evaluation of the distance metrics.

- Distance metrics are often used as subroutines in the context of different kinds of applications.
  - One can test the effectiveness of a nearest neighbor classifier with the use of different kinds of distance metrics.
  - Indirect measure of quality.

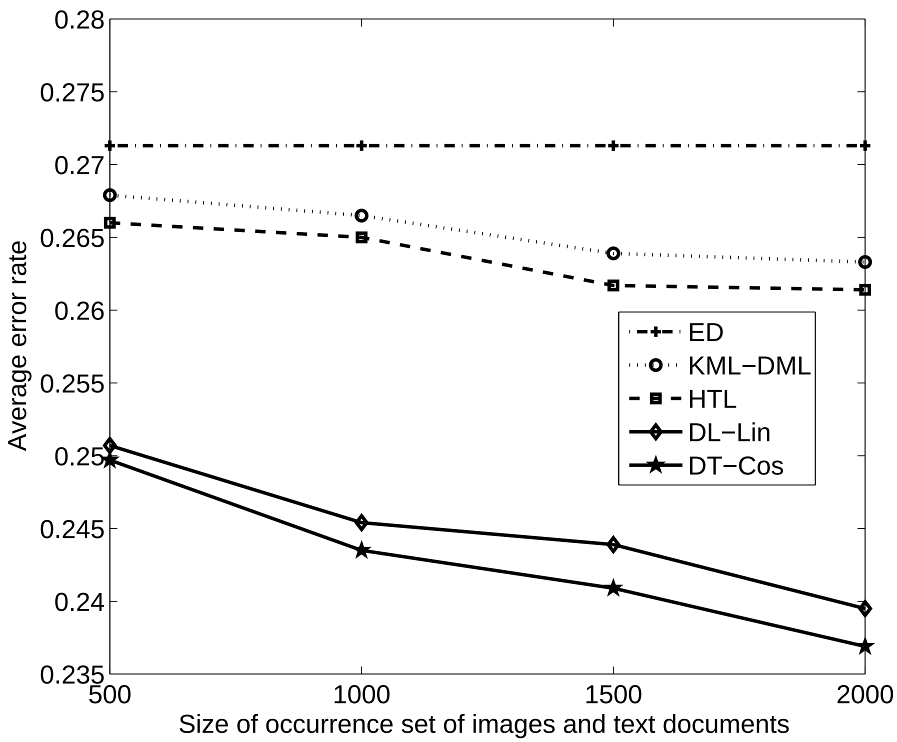
Data Sets

- Tested the method on number of real data sets.
- Use Wikipedia and Flickr data for text and associated images
- Used Corel data set for images.
- 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.

Error Rates of Different Methods

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**

Error with Varying Co-Occurrence Set Size



- Error with Varying Co-Occurrence Set Size

Computational Time

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

Conclusions and Summary

- New method for similarity transfer learning between text and web images
- Uses co-occurrence data as a bridge for the transfer process
- Builds new topic space based on co-occurrence data
- Leverages topic space for similarity transfer
- Experimental results show advantages over competing methods