Guojun Qi (University of Illinois at Urbana-Champaign)
Charu C. Aggarwal (IBM T. J. Watson Research Center)
Thomas Huang (University of Illinois at Urbana-Champaign)

Towards Semantic Knowledge Propagation from Text to Web Images

WWW Conference, Hyderabad, India 2011

Introduction

- The problem of image classification is challenging in many scenarios:
  - Labeled image data may be scarce in many settings.
  - The image features are not directly related to semantic concepts inherent in class labels.
- The combination of the above factors can be challenging.
- **Goal:** To address these challenges with the use of semantic knowledge propagation.

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.
- Classifiers naturally work better with features that have semantic interpretability.
  - Class labels are also usually designed on the basis of application-specific semantic criteria.
- Text features are inherently friendly to the classification 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.

- 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.

Observations

- The co-occurrence data provides a *raw semantic bridge* which needs to be further learned and refined in functional form.

- This learned bridge is then leveraged in order to translate the semantic information in the text features into the image domain.

- Related to the problem of *transfer learning*.

  – It is applied to the multimedia domain in order to leverage the semantic labels in text corpora to annotate image corpora with scarce labels.

- Co-occurrence information is noisy, but sufficiently rich on an aggregate basis.

Modeling

- 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.

- We evaluate our knowledge transfer techniques on an image classification task with labeled text corpora.

Broad Approach

- Design a translator 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 translator.
  - Links the instances across heterogeneous text and image spaces.
  - Follow the principle of parsimony and encode as few topics as possible.

- After the translator function is learned, the semantic labels can be propagated from any labeled text corpus to any new image by a process of cross-domain label propagation.

Notations and Definitions

- Let $\mathbb{R}^a$ and $\mathbb{R}^b$ be the source and target feature spaces.
- In the source (text) space, we have a set of $n^{(s)}$ text documents in $\mathbb{R}^a$.
- Each text document is represented by a feature vector $x_i^{(s)} \in \mathbb{R}^a, 1 \leq i \leq n^{(s)}$.
- This text corpus has already been annotated with class labels $\mathcal{A}^{(s)} = \left\{ \left(x_i^{(s)}, y_i^{(s)} \right) | 1 \leq i \leq n^{(s)} \right\}$, where $y_i^{(s)} \in \{+1, -1\}$ is the binary label.
- Binary assumption is without loss of generality, because the extension to the multi-class case is straightforward.

Notations and Definitions

- The images are represented by feature vectors $x^{(t)}$ in the target space $\mathbb{R}^b$

- A key component which provides such bridging information about the relationship between the text and image feature spaces is a *co-occurrence set* $\mathcal{C} = \left\{\left(\bar{x}_k^{(s)}, \bar{x}_l^{(t)}, c\left(\bar{x}_k^{(s)}, \bar{x}_l^{(t)}\right)\right)\right\}$.

- For the text document $\bar{x}_k^{(s)}$ its corresponding image feature vector $\bar{x}_l^{(t)}$ in the co-occurrence set, we denote the co-occurrence frequency by $c\left(\bar{x}_k^{(s)}, \bar{x}_l^{(t)}\right)$.

- For brevity, we use $c_{k,l}$ to denote $c\left(\bar{x}_k^{(s)}, \bar{x}_l^{(t)}\right)$.

Notations and Definitions

- Besides the linkage-based co-occurrence set, we sometimes also have a small set $\mathcal{A}^{(t)} = \left\{ \left(x_j^{(t)}, y_j^{(t)} \right) | 1 \leq j \leq n^{(t)} \right\}$ of labeled target instances.

- This is an auxiliary set of labeled target instances, and its size is usually much smaller than that of the set of labeled source examples.

  - $n^{(t)} \ll n^{(s)}$.

- The auxiliary set is used in order to enhance the accuracy of the transfer learning process.

Formal Approach with Translator Function

- One of the key intermediate steps during this process is the design of a *translator function* between text and images.
- This translator function serves as a conduit to measure the linking strength between text and image features.
- The translator T is a function defined on text space $\mathbb{R}^a$ as well as image space $\mathbb{R}^b$ as $T : \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$.
- It assigns a real value to each pair of text and image instances to weigh their linking strength.
  - Value can be either positive or negative, representing either positive or negative linkages.

Formal Approach with Translator Function

- Given a new image $x^{(t)}$, its label is determined by a discriminant function as a linear combination of the class labels in $\mathcal{A}^{(s)}$ weighted by the corresponding translator functions:

$$f_T\left(x^{(t)}\right) = \sum_{i=1}^{n^{(s)}} y_i^{(s)} T\left(x_i^{(s)}, x^{(t)}\right)$$

- The sign of $f_T\left(x^{(t)}\right)$ provides the class label of $x^{(t)}$.

- The key to translating from text to images is to learn a translator which can properly explain the correspondence between text and image spaces.

Learning the Translator 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 translator function in the form of matrices which represent topic spaces.
- The parameters of this canonical form will be optimized in order to learn the translator function

Learning the Translator Function

- We propose to optimize the following problem to learn the semantic translator:

$$\min_{T} \gamma \sum_{j=1}^{n^{(t)}} \ell\left(y_j^{(t)} f_T(x_j^{(t)})\right) + \lambda \sum_{\mathcal{C}} \chi\left(c_{k,l} \cdot T(\bar{x}_k^{(s)}, \bar{x}_l^{(t)})\right) + \Omega\left(T\right)$$

- γ and λ are balancing parameters

- Uses both co-occurrence data and auxiliary data

Designing the Translator Function

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

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

- Topic spaces provide a natural intermediate representation which can semantically link the information between the text and images.

Designing the Translator Function

- Topic spaces are represented by transformation matrices.

$$W^{(s)} \in \mathbb{R}^{p \times a} : \mathbb{R}^a \to \mathbb{R}^p, x_i^{(s)} \mapsto W^{(s)} x_i^{(s)}$$

$$W^{(t)} \in \mathbb{R}^{p \times b} : \mathbb{R}^b \to \mathbb{R}^p, x_j^{(t)} \mapsto W^{(t)} x_j^{(t)}$$

- The translator 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:

$$T\left(x_i^{(s)}, x_j^{(t)}\right) = \left\langle W^{(s)} x_i^{(s)}, W^{(t)} x_i^{(t)} \right\rangle = x_i^{(s)\prime} W^{(s)\prime} W^{(t)} x_j^{(t)} = x_i^{(s)\prime} S x_j^{(t)}$$

Observations

- The choice of the transformation matrices (or rather the product matrix $W^{(s)\prime}W^{(t)}$) impacts the translator function T directly.

- We will use the notation S in order to briefly denote the matrix $W^{(s)\prime}W^{(t)}$.

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

- The above definition of the matrix S can be used to rewrite the discriminant function as follows:

$$f_S\left(x^{(t)}\right) = \sum_{i=1}^{n^{(s)}} y_i^{(s)} x_i^{(s)\prime} S x_j^{(t)}$$

Regularization

- Use conventional squared norm for regularization.
- $\Omega(T) = \frac{1}{2}\left(\left\|W^{(s)}\right\|_F^2 + \left\|W^{(t)}\right\|_F^2\right)$
- Use trace-norm as a substitute to force convexity
- It is defined as follows:

$$\|S\|_\Sigma = \inf_{S=W^{(s)\prime}W^{(t)}} \frac{1}{2}\left(\left\|W^{(s)}\right\|_F^2 + \left\|W^{(t)}\right\|_F^2\right)$$

Objective Function after Regularization

- The regularized objective function can be rewritten as follows:

$$\min_{S} \gamma \sum_{j=1}^{n^{(t)}} \ell\left(y_j^{(t)} f_S(x_j^{(t)})\right) + \lambda \sum_{\mathcal{C}} \chi\left(c_{k,l} \cdot \bar{x}_k^{(s)\prime} S \bar{x}_l^{(t)}\right) + \|S\|_{\Sigma}$$

- Convex function which can be optimized with gradient descent method.

Choice of Loss Function

- Need to decide which functions are used for the loss functions $\ell(\cdot)$ and $\chi(\cdot)$.

- These functions measure compliance with the observed co-occurrence and the margin of discriminant functions $f_S(\cdot)$ on the auxiliary data set, respectively.

  - Use the well known logistic loss function $\ell(z) = \log\{1 + \exp(-z)\}$ for the first function.

  - Use the exponentially decreasing function $\chi(z) = \exp(-z)$ for the second.

Optimization Strategy

- After performing the afore-mentioned substitutions the objective function is non-linear.

- One possibility for optimizing an objective function of the form represented can be solved by using semi-definate programming on the dual formulation (Srebro et al).

  - Approach does not scale well with the size of the problem

- General approach is to adapt a proximal gradient method to minimize such non-linear objective functions with the use of a trace norm regularizer (Toh & Yun'09).

Proximal Gradient Method

- In order to represent the objective function more succinctly, we introduce the function $F(S)$:

$$F(S) = \gamma \sum_{j=1}^{n^{(t)}} \ell\left(y_j^{(t)} f_S\left(x_j^{(t)}\right)\right) + \lambda \sum_{\mathcal{C}} \chi\left(c_{k,l} \cdot x_k^{(s)\prime} S x_l^{(t)}\right)$$

- The optimization objective function can be rewritten as $F(S) + \|S\|_\Sigma$.

- In order to optimize this objective function, the proximal gradient method quadratically approximates it by Taylor expansion at current value of $S = S_\tau$ and Lipschitz coefficient α as follows:

$$Q(S, S_\tau) = F(S_\tau) + \langle \nabla F(S_\tau), S - S_\tau \rangle + \frac{\alpha}{2} \|S - S_\tau\|_F^2 + \|S\|_\Sigma$$

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:

$$\nabla F(S_\tau) = \gamma \sum_{i=1}^{n^{(s)}} y_i^{(s)} x_i^{(s)} \cdot \sum_{j=1}^{n^{(t)}} \ell' \left(y_j^{(t)} f_{S_\tau} \left(x_j^{(t)} \right) \right) y_j^{(t)} x_j^{(t)\prime} +$$

$$+ \lambda \sum_{\mathcal{C}} \left\{ \chi' \left(c_{k,l} \cdot \bar{x}_k^{(s)\prime} S_\tau \bar{x}_l^{(t)} \right) c_{k,l} \; \bar{x}_k^{(s)} \bar{x}_l^{(t)\prime} \right\}$$

- $\ell'(z) = \dfrac{-e^{-z}}{1+e^{-z}}$ and $\chi'(z) = -e^{-z}$ are the derivatives of $\ell(z)$ and $\chi(z)$ with respect to z.

Proximal Gradient Method

- S can be updated by minimizing $Q(S, S_\tau)$ with fixed S_τ iteratively.

- This can be solved by singular value thresholding (Cai, Candes, Shen 08)

- The convergence of the proximal gradient algorithm can be accelerated by making an initial estimate of α and increasing it by a constant factor η (Toh and Yun 09).

- Approach continued until $F\left(S_{\tau+1}\right) + \|S_{\tau+1}\|_\Sigma \leq Q\left(S_{\tau+1}, S_\tau\right)$.

- At this point, it is deemed that we are sufficiently close to an optimum solution, and the algorithm terminates.

Experimental Results

- Tested the method on number of real data sets.
- Use Wikipedia and Flickr data for text and associated images
  - Use class labels as query keywords to obtain text and associated images
- Compared our technique to:
  - SVM based image classifier
  - Transfer Learning Methods (TLRisk, HTL)

Ten Auxiliary Training Images

Category	Image only	HTL	TLRisk	TTI
birds	0.2639±0.0012	0.2619±0.0015	0.2546±0.0018	**0.252±0.0008**
buildings	0.2856±0.0002	0.2707±0.0021	0.2555±0.0014	**0.2303±0.0017**
cars	0.3027±0.0073	0.3065±0.0030	0.2543±0.0029	**0.2299±0.0031**
cat	0.2755±0.0043	0.2525±0.0038	0.2553±0.0028	**0.2424±0.0026**
dog	0.2252±0.0039	0.2343±0.0037	0.2545±0.0031	**0.2162±0.0027**
horses	0.2667±0.0019	0.2500±0.0021	0.2551±0.0016	**0.2383±0.0013**
mountain	0.3176±0.0010	0.3097±0.0003	**0.2541±0.0011**	0.2626±0.0007
plane	0.2667±0.0009	**0.2133±0.0008**	0.2546±0.0005	0.2567±0.0012
train	0.2624±0.0029	0.2716±0.0118	0.2552±0.0025	**0.2346±0.0031**
waterfall	0.2611±0.0008	**0.2435±0.0009**	0.2555±0.0016	0.2546±0.0007

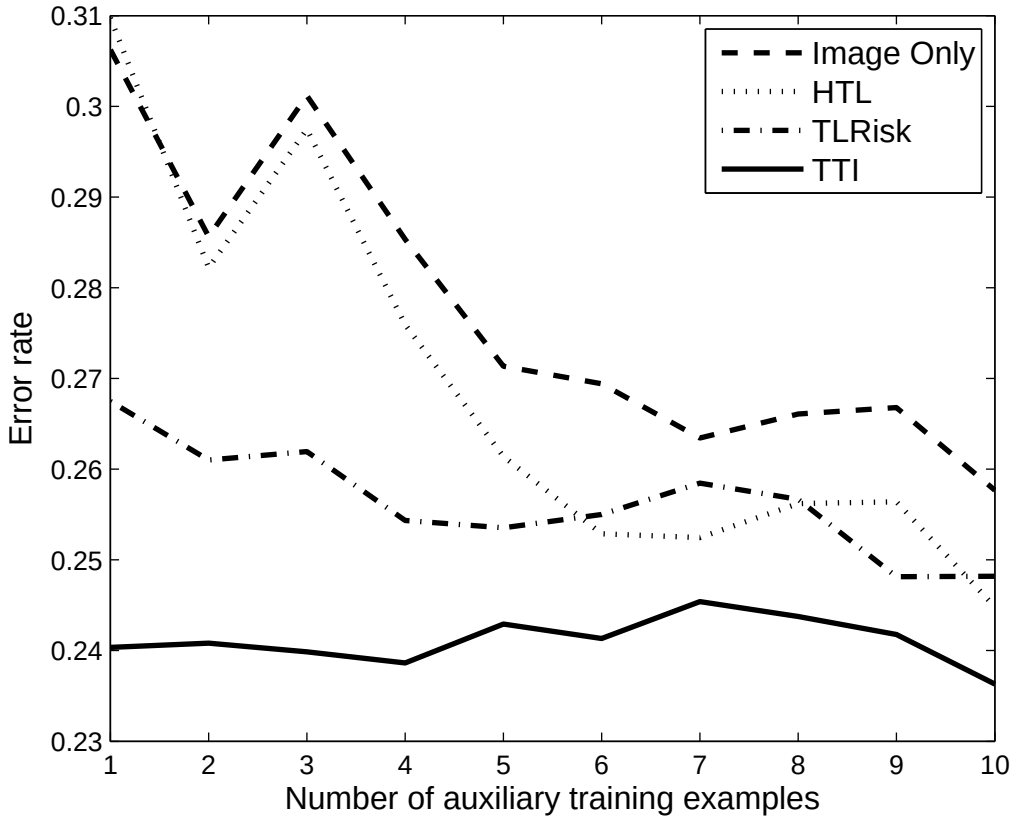
Two Auxiliary Training Images

Category	Image only	HTL	TLRisk	TTI
birds	0.3293±0.0105	0.3293±0.0124	0.2817±0.0097	**0.2738±0.0080**
buildings	0.3272±0.0061	0.3295±0.0041	0.2758±0.0023	**0.2329**±0.0032
cars	0.2529±0.0059	0.2759±0.0048	0.2639±0.0032	**0.1647**±0.0058
cat	0.3333±0.0071	0.3333±0.0060	**0.2480±0.0109**	0.2525±0.0083
dog	0.3694±0.0031	0.3694±0.0087	0.2793±0.0161	**0.252±0.0092**
horses	0.25±0.0087	0.3±0.0050	0.2679±0.0069	**0.2±0.0015**
mountain	0.3311±0.0016	0.3322±0.0009	0.2817±0.0021	**0.2699±0.0004**
plane	0.2667±0.0019	**0.225±0.0006**	0.2758±0.0006	0.2517±0.0011
train	0.3333±0.0084	0.3333±0.0068	0.2738±0.0105	**0.2099±0.0060**
waterfall	0.2693±0.0009	0.2694±0.0016	0.2659±0.0020	**0.257±0.0007**

Rank of Topic Matrix

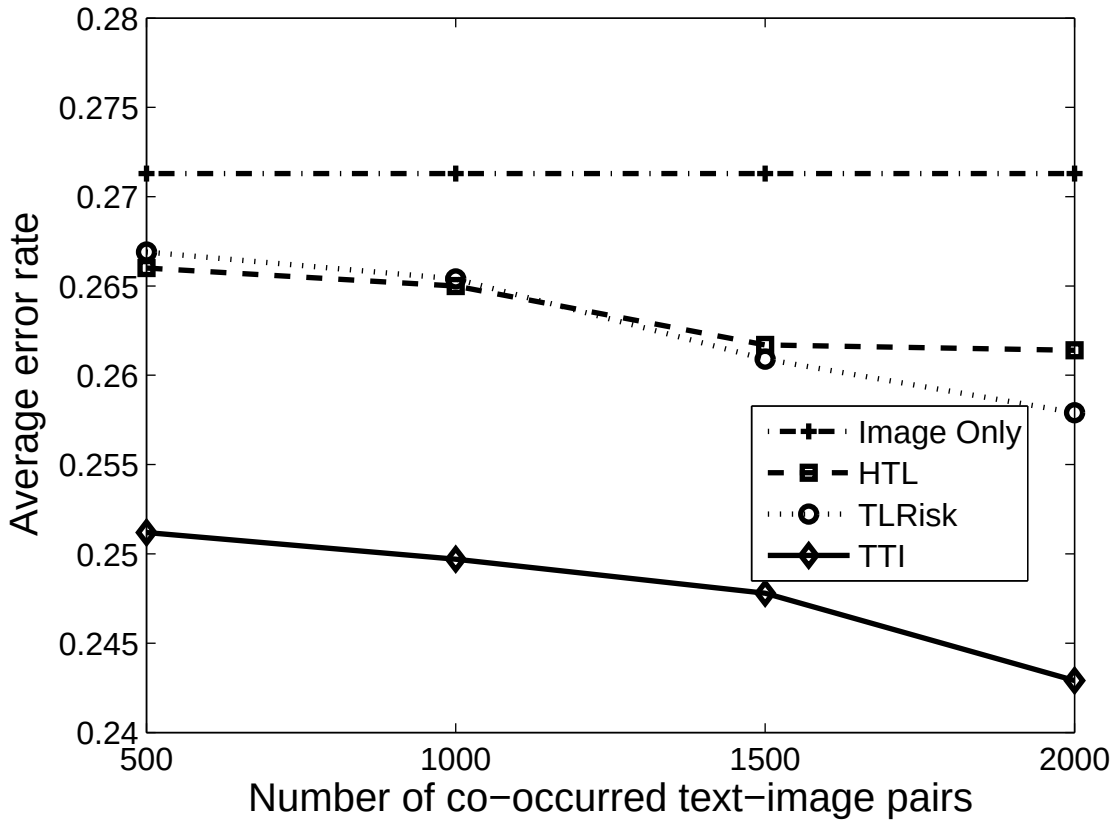
Category	Two trn. ex.	Ten trn. ex.
birds	11	9
buildings	88	102
cars	19	3
cat	18	2
dog	7	5
horses	4	1
mountain	6	1
plane	15	25
train	6	3
waterfall	21	26

Sensitivity to Varying Number of Training Images



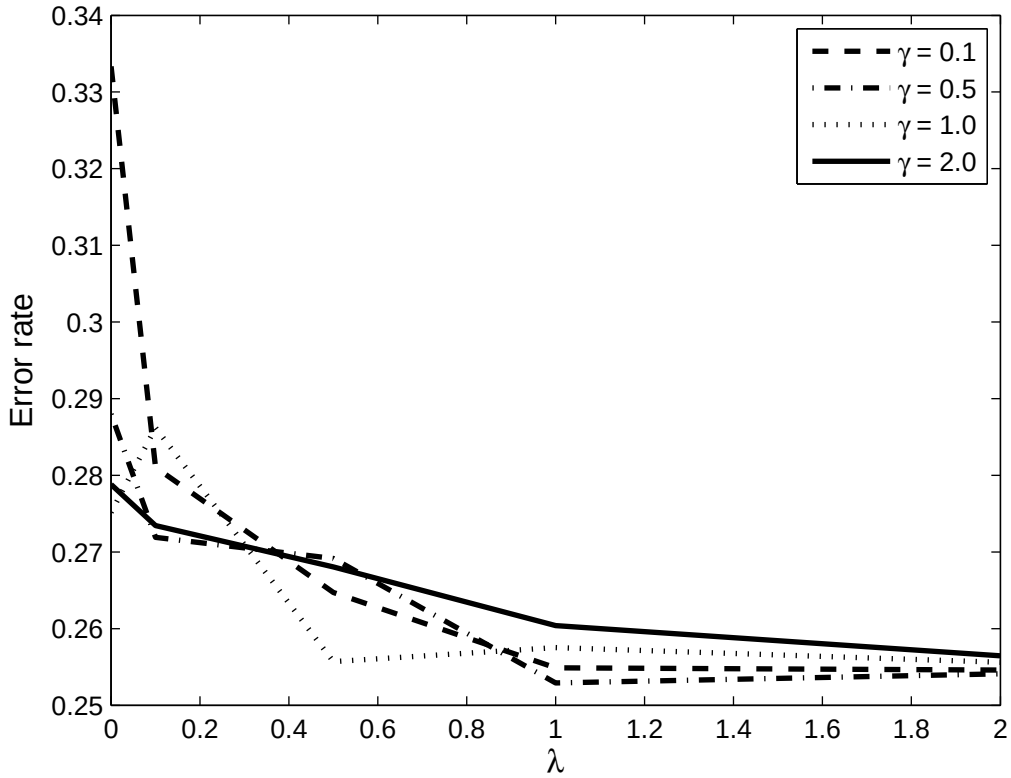
- Sensitivity to number of Training Images

Sensitivity to Number of Co-occurring examples



- Sensitivity to number of co-occurring examples

Sensitivity to Parameters



- Sensitivity to Parameter Values

Conclusions and Summary

- New method for 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 classification
- Experimental results show advantages over competing methods