Heterogeneous Transfer Learning for Image Classification

Yin Zhu[†], Yuqiang Chen[‡], Zhongqi Lu[†], Sinno Jialin Pan[*], Gui-Rong Xue[‡], Yong Yu[‡], and Qiang Yang[†]
[†]Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[‡]Shanghai Jiao Tong University, Shanghai, China
[*]Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
[†]{yinz, cs_lzxaa, qyang}@cse.ust.hk, [‡]{yuqiangchen, grxue, yyu}@sjtu.edu.cn, [*] jspan@i2r.a-star.edu.sg

Abstract

Transfer learning as a new machine learning paradigm has gained increasing attention lately. In situations where the training data in a target domain are not sufficient to learn predictive models effectively, transfer learning leverages auxiliary source data from other related auxiliary domains for learning. While most of the existing works in this area are only focused on using the source data with the same representational structure as the target data, in this paper, we push this boundary further by extending a *heterogeneous transfer learning* framework for knowledge transfer between text and images. We observe that for a target-domain classification problem, some annotated images can be found on many social Web sites, which can serve as a bridge to transfer knowledge from the abundant text documents available over the Web. A key question is how to effectively transfer the knowledge in the source data even though the text documents are arbitrary. Our solution is to enrich the representation of the target images with semantic concepts extracted from the auxiliary source data through matrix factorization, and to use the latent semantic features generated by the auxiliary data to build a better image classifier. We empirically verify the effectiveness of our algorithm on the Caltech-256 image dataset.

Introduction

Image classification has found many applications ranging from Web search to multimedia information delivery. In the past, image classification has met two major difficulties. First, the labeled images for training are often in short supply, and labeling new images incur much human labor. Second, images are usually ambiguous; e.g. an image can have multiple explanations. How to effectively overcome these difficulties and build a good classifier therefore becomes a challenging research problem. While labeled images are expensive, abundant unlabeled text data are easier to obtain. This motivates us to find a way to use the abundantly available text data to help improve the image classification performance.

In the past, several approaches have been proposed to solve the 'lack of labeled data' problem in supervised learning; e.g. semi-supervised learning methods (Zhu 2009) are proposed to utilize some unlabeled data under the assumption that the labeled and unlabeled data are from the same domain and drawn from the same distribution. Recently, transfer learning methods have been proposed to use knowledge from auxiliary data in a different but related domain to help learn the target tasks (Wu and Dietterich 2004; Mihalkova *et al.* 2007; Quattoni *et al.* 2008; Daumé 2007). However, a commonality among most transfer learning methods so far is that the data from different domains have the same feature space.

In some scenarios, given a target task, one may easily collect much auxiliary data that are represented in a different feature space. For example, suppose our task is to classify some images of dolphins into 'yes' or 'no' labels. Suppose that we have only a few labeled images for training. Suppose also that we can easily collect a large amount of text documents from the Web. In this case, we can model the image classification task as the target task, where we have a few labeled data and some unlabeled data. In the target domain, the data are both represented in pixels. Also in our case, the auxiliary domain, or the source domain, is the text domain, which are unlabeled text documents. Now, we ask: Is it possible to use the *cheap* auxiliary data to help imporve the performance of the image classification task? This is an interesting and difficult question, since the relationship between text and images is not explicitly given. This problem has been referred to as a *Heterogeneous Transfer Learning* problem (Yang *et al.* 2009)[1]. In this paper, we focus on heterogeneous transfer learning for image classification by exploring knowledge transfer from auxiliary unlabeled images and text data.

In image classification, if the labeled data are limited, classifiers trained on the original feature representation such as image pixels may have very poor performance. A key issue for us to address is to discover a new and improved representation, so that high level features such as edges and angles can be used to boost the classification performance. In this paper, we investigate how to obtain the high-level features for image classification tasks from both auxiliary data that contain both additional images and text documents.. Although images and text are represented in different feature spaces, they share a latent semantic space when they are re-

[1]Heterogeneous transfer learning can be defined for learning when auxiliary data have different features or different outputs. In this paper, we focus on the 'different features' version.

lated, which can be used to provide a better representation for images. We apply collective matrix factorization (CMF) techniques (Singh and Gordon 2008) on the auxiliary image and text data to discover the semantic space underlying the image and text domains. The traditional version of CMF assumes that correspondence exists between images and text data, an assumption that may not hold in our problem. To address this issue, we make use of tagged images that are available on the *social Web*, such as Flickr, to construct a connection between images and text. A semantic space is then learned to better represent the images.

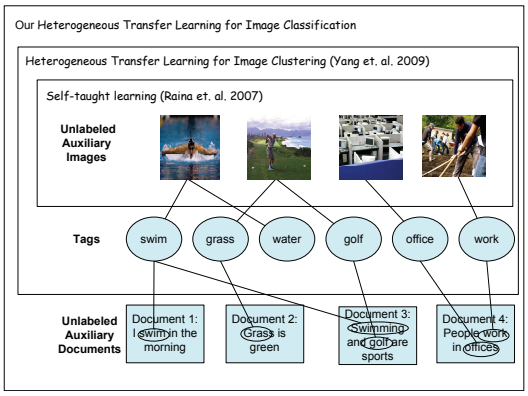


Figure 1: Source data used for different transfer learning algorithms. *Self-taught learning* only uses unlabeled auxiliary images, *heterogeneous transfer learning for image clustering* uses images and their annotations, while our proposed *heterogeneous transfer learning for image classification* takes all three information sources as inputs.

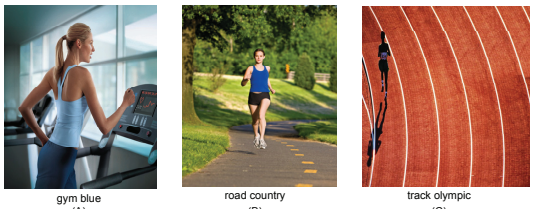


Figure 2: Three pictures different in pixel representation, but have the same the semantic meaning: running.

We illustrate the overall framework in Figure 1. Compared to self-taught learning (Raina *et al.* 2007), our approach can not only use the auxiliary images that have the same feature representation as the source data, but also use a different feature representation (i.e., text) for transfer learning. Compared to translated learning (Dai *et al.* 2008), our approach can use both labeled and unlabeld auxiliary data to help improve the target-domain learning performance.

Motivation and Problem Formulation

Before describing our proposed method in detail, we first illustrate a motivating example and give a problem statement.

A Motivating Example

Why can unlabeled text data help improve the classification performance of images? We give an illustrative example here. As shown in Figure 2, we may have three different pictures of people running, and by these pictures alone we cannot relate them well. As noted before, treating these images differently increases the sparsity in the image data and degrades the classification performance. However, suppose that we can find their corresponding tags in social Web data; e.g., the image C has a tag "track" associated with it, and image A has "gym" associated with it. By looking at the words alone, we still cannot connect them. However, by using some additional auxiliary text documents where these tags co-occur frequently, we may establish a strong similarity between these tags, and thus a relationship between their corresponding images. This allows us to find out that the images in fact share some common semantic level representations. As a result, we may reduce the data sparsity in the image domain.

Table 1: Problem formulation

Learning objective	Make predictions on target test images
Target image classification	Training images: $\mathbf{X}=\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ Testing images: $\mathbf{X}^*=\{\mathbf{x}_i^*, y_i^*\}_{i=n+1}^{n+m}$
Auxiliary source data	Unlabeled annotated images: $\mathbf{I}=\{\mathbf{z}_i, \mathbf{t}_i\}_{i=1}^{l}$ Unlabeled text documents: $\mathbf{D}=\{\mathbf{d}_i\}_{i=1}^{k}$

Problem Definition

Suppose that we are given a few labeled image data instances $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ and some test images $\mathbf{X}^* = \{\mathbf{x}_i^*, y_i^*\}_{i=n+1}^{n+m}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is an input vector of image features and y_i is the corresponding label of image i. Using "bag-of-words" (Csurka *et al.* 2004) to represent image features, we can assume that the feature values are nonnegative. n and m are the numbers of training and testing instances, respectively. In addition, we also have a set of auxiliary tagged images $\mathbf{I} = \{\mathbf{z}_i, \mathbf{t}_i\}_{i=1}^{l}$ and a set of unlabeled text documents $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^{k}$, where each $\mathbf{z}_i \in \mathbb{R}^d$ is an image instance represented by a feature vector as $\mathbf{x}_i$, $\mathbf{t}_i \in \mathbb{R}^h$ is its corresponding vector of tags, and h is the number of tags. For example, if an image $\mathbf{z}_i$ is annotated by tags α and β with $\alpha, \beta \in \{1, ..., h\}$, then $\mathbf{t}_i = [0, ..., 1, ..., 1, ...0]$ is a vector of dimensionality h with all zeros and one's in the α and β positions. $\mathbf{d}_i \in \mathbb{R}^m$ is a document represented by a vector of bag-of-words, and l and k are the numbers of auxiliary images and documents, respectively. Our goal is to learn an accurate image classifier $f(\cdot)$ from $\mathbf{X}$, $\mathbf{I}$ and $\mathbf{D}$, that can predict the testing data accurately onto $\mathbf{X}^*$; we denote the classifier as $f(\mathbf{X}^*)$. We summarize the problem definition in Table 1. For convenience, we denote $\mathbf{Z}=\{\mathbf{z}_i\}_{i=1}^{l} \in \mathbb{R}^{l\times d}$ and $\mathbf{T}=\{\mathbf{t}_i\}_{i=1}^{l} \in \mathbb{R}^{l\times h}$ the image features and text tags of the auxiliary images separately. Furthermore, we abuse the notation $\mathbf{X}$, $\mathbf{X}^*$, $\mathbf{Z}$, and $\mathbf{T}$ to represent the data matrices with instances $\mathbf{x}_i$, $\mathbf{x}_i^*$, $\mathbf{z}_i$ and $\mathbf{t}_i$ being row vectors in them.

Algorithm Description

In this section, we describe the details of our proposed method. We first introduce how to build a connection between the auxiliary images and text data. We then show how to apply the collective matrix factorization method to learn

high-level features behind the connection. Finally, we describe how to construct a new feature presentation for target images, on which standard classifiers can perform well.

Bridging Images and Text

Given a set of auxiliary images $\mathbf{Z} \in \mathbb{R}^{l\times d}$ with their corresponding tags $\mathbf{T} \in \mathbb{R}^{l\times h}$, and a set of unlabeled documents $\mathbf{D} \in \mathbb{R}^{k\times m}$, we wish build a connection between images and text documents. As illustrated in Figure 1, we can construct a two-layer bipartite graph among the images, tags and text documents. More specifically, the top layer of the bipartite graph is used to represent the relationship between images and tags. Each image can be annotated by tags, and some images may share one or multiple tags. If two images are annotated by shared tags, they tend to be related to each other semantically. Similarly, if two tags co-occur in annotations of shared images, they tend to be related to each other. This image-tag bipartite graph is represented via the tag matrix $\mathbf{T}$. The bottom layer of the bipartite graph is used to represent the relationship between tags and documents. If a tag, more precisely, the text word of the tag, occurs in a document, then there is an edge connecting the tag and the document. We use a matrix $\mathbf{F} \in \mathbb{R}^{k\times h}$ to represent the document-tag bipartite graph, where $\mathbf{F}_{ij}$=1 if there is an edge between the i^{th} document and the j^{th} tag, otherwise 0.

Learning Semantic Features for Images

So far, we have built a connection between images and text through annotating tags. In this section, we learn the semantic features for images by exploiting the relationship between images and text from the auxiliary sources. Recall that we have a matrix of images with low-level image features $\mathbf{Z}$ and a relational matrix between images and annotations $\mathbf{T}$. We first define a new matrix $\mathbf{G} = \mathbf{Z}^\top\mathbf{T} \in \mathbb{R}^{d\times h}$ to denote the correlation between low-level image features and annotations which can be referred to as high-level concepts. Note that $\mathbf{G}_{ij} = \sum_k \mathbf{z}_{ik} \cdot \mathbf{t}_{kj}$, where $\mathbf{z}_{ik} \geq 0$ is the value of the i^{th} *visual word* in the k^{th} image, and $n_j^{(i)} = \sum_k \mathbf{t}_{kj}$ is the number of images that are annotated by the j^{th} tag and whose i^{th} visual word is observed at the same time. $\mathbf{G}_{ij}$ is large when $n_j^{(i)}$ is large or some of the values of the i^{th} visual word in the images with the j^{th} tag annotation are large. This implies that if $\mathbf{G}_{ij}$ is large, then the i^{th} image feature and the j^{th} tag may have strong correlation.

Motivated by the well-known Latent Semantic Analysis (LSA) (Deerwester *et al.* 1990), we proceed to extract the latent semantic features for each low-level image feature. We accomplish this by a matrix factorization that decomposes $\mathbf{G}$ into latent factor matrices as

$$\mathbf{G} = \mathbf{U}\mathbf{V}_1^\top,$$

where $\mathbf{U} \in \mathbb{R}^{d\times g}$, $\mathbf{V}_1 \in \mathbb{R}^{h\times g}$, and g is the number of latent factors. Each vector $\mathbf{u}_i$ is a latent semantic representation of the i^{th} image low-level feature, and $\mathbf{v}_{1j}$ is a latent semantic representation of j^{th} tag. Note that the matrix $\mathbf{G}$ may be sparse, resulting in the decomposition on $\mathbf{G}$ being imprecise.

The text documents and the tags also define a matrix $\mathbf{F} \in \mathbb{R}^{k\times h}$. We can decompose it as well:

$$\mathbf{F} = \mathbf{W}\mathbf{V}_2^\top,$$

In the above, $\mathbf{W} \in \mathbb{R}^{k\times g}$, $\mathbf{V}_2 \in \mathbb{R}^{h\times g}$. Then, $\mathbf{w}_i$ can be treated as a latent semantic representation of a text document $\mathbf{d}_i$, and $\mathbf{v}_{2j}$ can be treated as a latent semantic representation of the j^{th} tag. Our motivation is to use the results of the decomposition on $\mathbf{F}$ to help improve the decomposition on $\mathbf{G}$, in order to learn a more precise matrix $\mathbf{U}$. Note that if we can decompose $\mathbf{G}$ and $\mathbf{F}$ perfectly, then we may get $\mathbf{V}_1 = \mathbf{V}_2$, because the tags in the two sides should have the same latent semantic meanings.

Motivated by this observation, we propose to learn the latent semantic representation $\mathbf{U}$ by decomposing $\mathbf{G}$ and $\mathbf{F}$ jointly with the constraint $\mathbf{V}_1 = \mathbf{V}_2$. This is called collective matrix factorization (CMF), which was proposed by Singh and Gordon (2008). It has been shown that when relational matrices are sparse, decomposing them simultaneously can give better performance than decomposing them individually.

Hence, our objective can be written as follows,

$$\min_{\mathbf{U},\mathbf{V},\mathbf{W}} \lambda \left\|\mathbf{G} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 + (1-\lambda)\left\|\mathbf{F} - \mathbf{W}\mathbf{V}^\top\right\|_F^2 + R(\mathbf{U},\mathbf{V},\mathbf{W}), \tag{1}$$

where $0 \leq \lambda \leq 1$ is a tradeoff parameter to control the decomposition error between the two matrix factorizations, $||\cdot||_F$ denotes the Frobenius norm of matrix, and $R(\mathbf{U},\mathbf{V},\mathbf{W})$ is the regularization function to control the complexity of the latent matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$. In this paper, we define the regularization function as

$$R(\mathbf{U},\mathbf{V},\mathbf{W}) = \gamma_1 \|\mathbf{U}\|_F^2 + \gamma_2 \|\mathbf{V}\|_F^2 + \gamma_3 \|\mathbf{W}\|_F^2,$$

where γ_1, γ_2 and γ_3 are nonnegative parameters to control the responding regularization terms. In this paper, we set $\gamma_1 = \gamma_2 = \gamma_3 = 1$.

The optimization problem in (1) is an unconstrained nonconvex optimization problem with three matrix variables $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, thus only has local optimal solutions. However, (1) is convex with respect to any one of the three matrices while fixing the other two. A common technique for solving this optimization problem is to fix two of the matrices and optimize the remaining one iteratively, until the results converge. The algorithm is summarized in Algorithm 1.

Constructing New Representations

So far, we have described how to learn a semantic view $\mathbf{U}$ for the low-level image features. In this section, we show how to map target images $\mathbf{X}$ to semantic feature space for image classification. This is accomplished by first transforming each target image $\mathbf{x}_i$ into its semantic space as $\widetilde{\mathbf{x}}_i = \mathbf{x}_i\mathbf{U}$. After constructing a new representation for target images, we can then train a standard classification model on $\{\widetilde{\mathbf{x}}_i, y_i\}$'s. Finally, we can use the model to make predictions on the test images $\widetilde{\mathbf{X}}^*$.

Putting it together, our overall heterogeneous transfer learning algorithm is referred to as *HTLIC*, which stands for HTL for Image Classification.

Algorithm 1 Image Semantic View Learning via CMF

Input: A auxiliary image matrix $\mathbf{Z}$ with its corresponding annotation matrix $\mathbf{T}$, a document-tag relational matrix $\mathbf{F}$, a parameter λ, and the number of latent factors g.

Output: A new representation $\mathbf{U}$ for images $\mathbf{Z}$.

Compute $\mathbf{G} = \mathbf{Z}^\top \mathbf{T}$ and randomly initialize matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$.
repeat
  Fix $\mathbf{U}$ and $\mathbf{V}$, apply conjugate gradient descent (CGD) (Shewchuk 1994) on (1) to update $\mathbf{W}$;
  Fix $\mathbf{U}$ and $\mathbf{W}$, apply CGD on (1) to update $\mathbf{V}$;
  Fix $\mathbf{W}$ and $\mathbf{V}$, apply CGD on (1) to update $\mathbf{U}$;
until $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are convergent.

Experiments

Our experiments are designed to demonstrate that the effectiveness of exploiting unlabeled text in our heterogeneous learning algorithm.

Dataset and Processing

We use a benchmark dataset, Caltech-256 (Griffin *et al.* 2007), as our target images. The auxiliary annotated images and the text documents are crawled from the online photo sharing website Flickr and Google search engine respectively.

The Caltech-256 image dataset consists of 256 categories of images, where each category has hundreds of images. We randomly selected 19 categories from the 256 categories, and built $\binom{19}{2} = 171$ pairs of binary classification tasks. The selected 19 categories and the corresponding number of images in each category are : *tennis-racket* (298), *american-flag* (299), *school-bus* (361), *cake* (96), *cd* (300), *chessboard* (299), *greyhound* (299), *fried-egg* (300), *dog* (359), *light-house* (242), *llama* (300), *minaret* (300), *motorbike* (300), *rainbow* (300), *sheet-music* (300), *smokestack* (300), *starfish* (300), *watermelon* (300), `zebra` (299).

The auxiliary annotated images from Flickr were crawled during December 2009. We collected $5,700$ images and $64,021$ related tags, among which $2,795$ tags were distinct. Each of these tags is a single word. These Flickr images are relevant to the image categories described above. For example, for the image category "dog", we collect Flickr images with tags "dog", "greyhound" or "doggy". In order to obtain auxiliary text data, we use the Google search engine to crawl documents from the Web. For each tag, we search the tag name via Google search engine and get the first 100 resulting webpages as the text documents. Each resulting webpage is treated as an auxiliary document. We collect $279,500$ documents in total. Note that one can also use other data sources, e.g., articles and images from Wikipedia. In this paper, we focus on how to use auxiliary data sources to help on target image classification tasks. In our experiments, we use bag-of-words to represent images (Csurka *et al.* 2004). More specifically, for the target and auxiliary images from Flickr, we use SIFT descriptors (Lowe 2004) to identify interesting points. We then use the K-means clustering algorithm to group all the interesting points into 512 clusters as a codebook. In this way, each cluster is treated as a feature. For auxiliary documents and the tags associated to the auxiliary images, we do stemming on them, and build a tag-document co-occurrence matrix.

Evaluation and Baseline Methods

We use the prediction *accuracy* on the test data as our evaluation criterion:

$$\mathrm{ACC}(f, \mathbf{X}^*, \mathbf{Y}^*) = \frac{\sum_{x_i^* \in \mathbf{X}^*} I[f(\mathbf{x}_i^*) = y_i^*]}{|\mathbf{X}^*|}, \quad (2)$$

where f is the trained classifier, I is an indicator function.

For each binary classification task, there are hundreds of images that can serve as the training or testing data. We randomly selected 5 of the images as the training data, and the rest as test instances. We repeat this 30 times and report the average results. We use linear Support Vector Machines (SVMs)[2] as a base classifier. In all experiments, we set the trade off parameter C of linear SVMs to 10.

We compare our proposed method with three different baselines with different feature presentations for image classification. The three baselines and our proposed method are summarized as follows,

Orig. This baseline only uses the SIFT image features of the target images without considering to use any auxiliary sources to enrich the feature representation.

PCA. In this baseline, we first apply Principal Component Analysis (PCA) on the auxiliary images to learn some latent factors, and use the latent factors as features to represent the target images for classification. This method is also reported in (Raina *et al.* 2007), which obtains promising performance for image classification.

Tag. We implemented the method proposed in (Wang *et al.* 2009) as another baseline, which builds a text view for target images by using some auxiliary annotated images. For each target image, this method finds the K most similar images from the annotated image set and aggregate all the tags associated to these similar images as a text representation. Here, K is set to 100 in our experiments.

HTLIC. This denotes our proposed method, which uses all the auxiliary data including annotated images and unlabeled documents. The parameter setting is discussed in the following section.

For each classification task, **PCA**, **Tag** and **HTLIC** use the same set of annotated images, that are images relevant to two categories in the task.

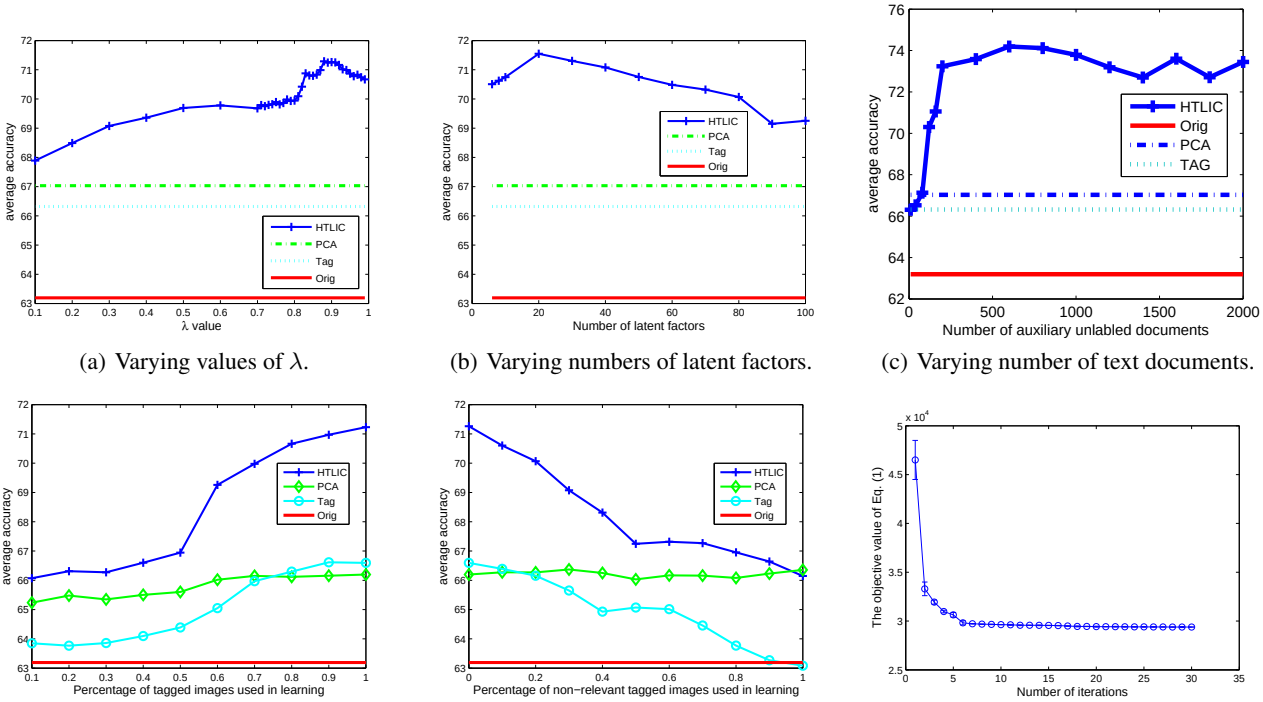
Experimental Results

In the first experiment, we compare our method with three baselines on all the classification tasks. Because of the limited space, we are not able to report the results of all 171 tasks. To show results on some representative tasks, we first rank all tasks based on the improvement of **HTLIC** compared to **Orig** in terms of classification accuracy. We then select 4 tasks with largest improvement and 3 ones with smallest improvement as shown in table 2. Note that the value

[2]We use LibSVM that is available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Table 2: Comparison with baselines.

Tasks	**Orig**	**PCA**	**Tag**	**HTLIC**
watermelon vs *sheet-music*	64.66 ± 9.99	70.28 ± 11.33	78.13 ± 14.40	85.29 ± 11.94
fried-egg vs *american-flag*	59.19 ± 7.80	60.54 ± 9.28	63.70 ± 12.54	78.80 ± 12.21
fried-egg vs *school-bus*	65.42 ± 10.72	66.73 ± 11.01	75.58 ± 14.56	83.74 ± 11.88
zebra vs *motorbikes*	69.95 ± 11.74	70.55 ± 12.37	85.74 ± 13.72	86.66 ± 12.32
minaret vs *lighthouse*	53.67 ± 7.62	53.61 ± 6.18	52.71 ± 7.03	53.32 ± 6.38
llama vs *greyhound*	51.48 ± 7.11	52.65 ± 5.58	50.79 ± 5.53	51.94 ± 5.40
cd vs *cake*	62.85 ± 10.45	65.20 ± 11.87	54.98 ± 5.33	57.71 ± 8.35
Average	63.1925	67.0312	66.3192	71.5493



(a) Varying values of λ. (b) Varying numbers of latent factors. (c) Varying number of text documents.

(d) Varying number of annotated images. (e) Varying noise in the tags. (f) Varying number of iterations.

Figure 3: Empirical evaluation of HTLIC.

of improvement can be negative if **HTLIC** performs worse than **Orig**. The last row in the table shows the average results over **all** the 171 classification tasks in term of accuracy. In this experiment, for **HTLIC**, we set the tradeoff parameter λ in (1) to 0.85. As we can see from table 2, our proposed **HTLIC**, which only uses semantic features to represent the target image for classification, outperforms other baselines. This implies that the semantic features learned by our proposed method is powerful for image classification.

In the second experiment, we study the parameter sensitivity of λ on the overall performance of **HTLIC** in image classification. Figure 3(a) shows the average classification accuracy of **HTLIC** over the all 171 image classification tasks under varying values of λ. We can find that **HTLIC** performs best and steadily when λ falls in the range from 0.8 to 0.95, which implies the jointly decomposition on the auxiliary document-tag matrix can indeed help learning a more precise latent factor matrix $\mathbf{U}$ for low-level image features.

In the third experiment, we study the parameter sensitivity of g, the number of the latent factors in the matrix factorization, on the overall performance of **HTLIC** in image classification. Figure 3(b) shows the average classification accuracy of **HTLIC** over all image classification tasks under varying numbers of the latent factors g. We can find that **HTLIC** performs best when g falls in the range $[10\ 30]$.

We also analyze the impact of the number of auxiliary text documents on the overall performance of **HTLIC** in image classification. The experimental results are shown in Figure 3(c). As we can see, when the number of the auxiliary documents increases, the performance of **HTLIC** increases as well, up to a point when the improvement stops. The reason is that when the number of documents is larger, the document-tag matrix $\mathbf{F}$ may be denser, which makes the decomposed matrix $\mathbf{V}$ more precise. This results in the decomposition on $\mathbf{G}$ being more precise. However, the improvement saturates when the number of documents reach a certain point (200), after which additional documents are not useful anymore.

We also vary the number of annotated images in the social Web. As shown in Figure 3(d), varying auxiliary images can affect the results for all the methods that use these images. **HTLIC** and **Tag** show improvement with more auxiliary images, while **PCA** improves much slower. We also did experiments to show how the quality of annotated images affect the performance of these methods. As shown in Figure 3(e), when the auxiliary images are gradually substituted by non-relevant images, which are just random images from Flickr, the result of **HTLIC** and **Tag** have the clear drop, while **PCA** is quite stable in its performance. Note that our method performs close to **PCA** when there is no relevant images at all in the auxiliary image set.

The last experiment is to measure the convergence of the collective matrix factorization algorithm in Algorithm 1. Figure 3(f) shows the average objective value of Eq. (1) over 30 random initializations when doing the CMF for task *watermelon* vs *sheet-music*. As can be seen in the figure, after 10 iterations the objective value converges.

Related Work

Transfer learning emphasizes the transferring of knowledge across different domains or tasks. For example, Wu and Dietteirch (2004) investigated methods for improving SVM classifiers with auxiliary training data. Raina *et al.* (2007) proposed a learning strategy known as self-taught learning which utilizes irrelevant unlabeled data to enhance the classification performance. Pan and Yang (2010) surveyed the filed of transfer learning. Recently, Yang *et al.* (2009) proposed a heterogenous transfer learning algorithm for image clustering by levering auxiliary annotated images. We also aim to levering auxiliary annotated images for target image classification. The difference between our work and theirs is that other than using the annotated images, we also try to utilize unlabeled text data for further boosting the performance in image classification. Translated learning (Dai *et al.* 2008) utilizes the labeled text data to help classify images, while in our work the auxiliary text data are unlabeled. Our work also relates to multimedia area, especially works using text and image together, e.g. leveraging image content for Web search (Zhou and Dai 2007). Our work is also related to works on tagged images, e.g. (Wu *et al.* 2011).

Conclusions

In this paper, we explore heterogeneous transfer learning for image classification by using unlabeled auxiliary text to help learning. We show that the performance of image classification can be improved by utilizing textual information and tagged image data from a social Web. A common semantic view is established by using text, tags and images via collective matrix factorization. Experimental results show that our method outperforms other baselines when the labeled data in the target domain are short in supply.

In the future, we would like to continue to work on the other heterogeneous transfer scenario where the training and testing labels are different in the target domain. We will consider other types of auxiliary data as well as more than one data source.

Acknowledgement

We thank the support of Hong Kong RGC/NSFC projects N_HKUST624/09 and 60910123. Gui-Rong Xue thanks support from NSFC project (NO. 60873211), RGC/NSFCproject (NO. 60910123). We also thank Nathan N. Liu, Evan W. Xiang and Fan Zhou for discussions.

References

Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008.

Hal Daumé. Frustratingly easy domain adaptation. In *ACL*, 2007.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, Technical Report, California Institute of Technology., 2007.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Lilyana Mihalkova, Tuyen N. Huynh, and Raymond J. Mooney. Mapping and revising markov logic networks for transfer learning. In *AAAI*, pages 608–614, 2007.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, pages 759–766, 2007.

Jonathan Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, 1994.

Ajit Paul Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *KDD*, pages 650–658, 2008.

Gang Wang, Derek Hoiem, and David A. Forsyth. Building text features for object image classification. In *CVPR*, pages 1367–1374, 2009.

Pengcheng Wu and Thomas G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *ICML*, 2004.

Lei Wu, Steven C. H. Hoi, Rong Jin, Jianke Zhu, and Nenghai Yu. Distance metric learning from uncertain side information for automated photo tagging. *ACM TIST*, 2(2):13, 2011.

Qiang Yang, Yuqiang Chen, Gui-Rong Xue, Wenyuan Dai, and Yong Yu. Heterogeneous transfer learning for image clustering via the social web. In *ACL*, 2009.

Zhi-Hua Zhou and Hong-Bin Dai. Exploiting image contents in web search. In *IJCAI*, pages 2922–2927, 2007.

Xiaojin Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin Madison, 2009.