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 source domains for learning. While most of the existing works in this area only focused on using the source data with the same structure as the target data, in this paper, we push this boundary further by proposing 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 can be arbitrarily found. Our solution is to enrich the representation of the target images with semantic concepts extracted from the auxiliary source data through a novel matrix factorization method. By using the latent semantic features generated by the auxiliary data, we are able to build a better integrated image classifier. We empirically demonstrate the effectiveness of our algorithm on the Caltech-256 image dataset.

Introduction

Image classification has found many applications ranging from Web search engines to multimedia information delivery. However, it has 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 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 label problem in supervised learning, e.g. semi-supervised learning methods (Zhu 2009) are proposed to utilize unlabeled data assuming that the labeled and unlabeled data are from the same domain and drawn from the same distribution. Recently, transfer learning methods (Wu and Dietterich 2004; Mihalkova *et al.* 2007; Quattoni *et al.* 2008; Daumé 2007) are proposed to transfer knowledge from auxiliary data in a different but related domain to help on target tasks. But a commonality among most transfer learning methods is that data from different domains are required to be in the same feature space.

In some scenarios, given a target task, one may easily collect a lot of auxiliary data that are represented in a different feature space. For example, suppose our task is to classify *dolphin* pictures (e.g. yes or no). We have only a few labeled images for training. Besides, we can collect a large amount of text documents from the Web easily. Here, the target domain is the image domain, where we have a few labeled data and some unlabeled data, which are both represented by pixels. The auxiliary domain or source domain is the text domain, where we have large amount of unlabeled textual documents. Is it possible to use these *cheap* auxiliary data to help the image classification task? This is an interesting and difficult problem, since the relationship between text and images is not given. This can be also 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 extremely limited, classifiers trained on the original feature representation (e.g. pixels) directly may get very bad performance. One key issue is to discover a new *powerful* representation, such as high level features beyond pixels (e.g. edge, angle), to boost the performance. In this paper, we are interested in discovering the high-level features for images from both auxiliary image and text data. Although images and text are represented in different feature spaces, they are supposed to share a latent semantic space, which can be used to represent images when it is well learned. We propose to 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. CMF techniques assume some correspondences between images

[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.

and text data, which may not hold in our problem. To solve this problem, we make use of the tagged images available on the social Web, such as tagged images from Flickr, to construct connections between image and text data. After a semantic space is learnt by CMF using the auxiliary images and text documents, a new feature representation called semantic view is created by mapping the target images into this semantic space.

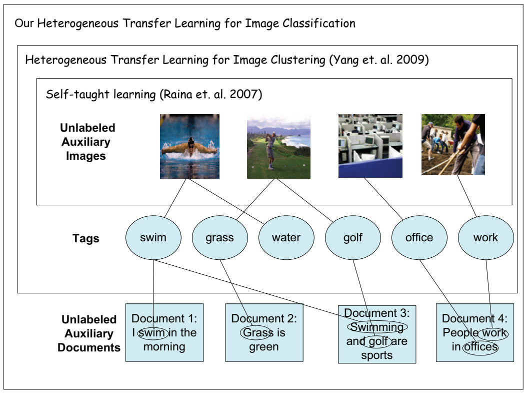

Our Heterogeneous Transfer Learning for Image Classification
Heterogeneous Transfer Learning for Image Clustering (Yang et. al. 2009)
Self-taught learning (Raina et. al. 2007)
Unlabeled Auxiliary Images
Tags
swim
grass
water
golf
office
work
Unlabeled Auxiliary Documents
Document 1: I swim in the morning
Document 2: Grass is green
Document 3: Swimming and golf are sports
Document 4: People work in offices

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.

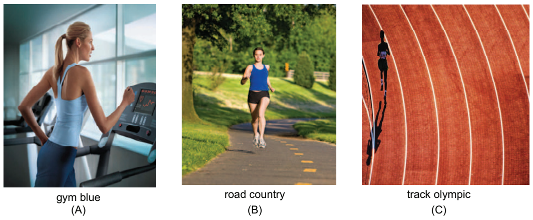


Figure 2: Three pictures different in pixel representation, but have the same the semantic meaning: running.

The main contribution of our work is that we utilize three kinds of source data, namely auxiliary images, their tags and auxiliary text documents and build a good latent semantic space using these data. The structure of the source data is clearly shown in Figure 1, where self-taught learning can only utilize the auxiliary images, which have the same feature representation with the source data, a previous heterogeneous transfer learning work for clustering (Yang *et al.* 2009) uses annotated images, while our method uses all the data. The advantage of our method over the two previous methods is that by leveraging the knowledge in the abundant text documents, we are able to build a better feature space with semantic meaning.

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

Figure 2 shows three pictures with people running. In some classification tasks, e.g. a working-or-running classification problem, these three pictures should be classified as the same class. But they are quite dissimilar in the pixel level representation, thus any classifier based on pixel level representation would fail the classification task. However, they look similar in their semantic space. First, tags are found for an image by comparing it with all tagged auxiliary images, selecting the top similar images and aggregating their tags. We can find that some of these tags are quite relevant to the picture, e.g. image (B) has top tags “road” and “country”. By further exploring more text documents, the similarity between the three images can be found as their tags “road”, “track” and “gym” have similar latent meanings in the text.

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 we are given a few labeled image instances $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ and testing instances $\mathbf{X}^* = \{\mathbf{x}_i^*, y_i^*\}_{i=n+1}^{n+m}$, 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. In this paper, we use “bag-of-words” (Csurka *et al.* 2004) to represent image features, whose values are nonnegative. n and m are the numbers of training and testing instances respectively. In addition, we are also given a set of auxiliary annotated images $\mathbf{I} = \{\mathbf{z}_i, \mathbf{t}_i\}_{i=1}^{l}$ and a set of documents $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^{k}$, where $\mathbf{z}_i \in \mathbb{R}^d$ is an image 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 1'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 $\mathbf{D}$ to make predictions on $\mathbf{X}^*$, $f(\mathbf{X}^*)$. 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.

Heterogenous Transfer Learning for Image Classification

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 now show how to build connections between them. As illustrated in Figure 1, we can construct a two layer Bi-partite graph among images, tags and documents. More specifically, the top layer Bi-partite 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 Bi-partite graph has been represented by the tag matrix $\mathbf{T}$. The bottom layer Bi-partite 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, there is an edge connecting the tag and the document. We define a matrix $\mathbf{F} \in \mathbb{R}^{k \times h}$ to represent the document-tag Bi-partite 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 try to learn some 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 Latent Semantic Analysis (LSA) (Deerwester *et al.* 1990), in order to extract latent semantic features for each low-level image feature, we can apply matrix factorization techniques to decompose $\mathbf{G}$ 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. Then $\mathbf{u}_i$ can be treated as a latent semantic representation of the i^{th} image low-level feature, and $\mathbf{v}_{1j}$ can be treated as a latent semantic representation of j^{th} tag. However, the matrix $\mathbf{G}$ may be very sparse, resulting in the decomposition on $\mathbf{G}$ may not be precise.

Recall that we have another relational matrix $\mathbf{F} \in \mathbb{R}^{k \times h}$ between documents and tags. We can also decompose it as

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

where $\mathbf{W} \in \mathbb{R}^{k \times g}$, $\mathbf{V}_2 \in \mathbb{R}^{h \times g}$. Then $\mathbf{w}_i$ can be treated as a latent semantic representation of document $\mathbf{d}_i$, and $\mathbf{v}_{2j}$ can be treated as a latent semantic representation of the j^{th} tag. Since the matrix $\mathbf{F}$ is relatively dense compared to $\mathbf{G}$, the decomposition on $\mathbf{F}$ may be more precise. Therefore, our motivation is to use the results of the decomposition on $\mathbf{F}$ to help the decomposition on $\mathbf{G}$ to learn a more precise $\mathbf{U}$. Note that if we can decompose $\mathbf{G}$ and $\mathbf{F}$ perfectly, then we may get $\mathbf{V}_1 = \mathbf{V}_2$ as 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 first proposed by Singh and Gordon (2008). It has been shown that when relational matrices are sparse, decomposing them simultaneously can get 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. Thus one common technique to solve this kind of optimization problem is to fix two matrices and optimize the left one iteratively until the results are convergent. The detailed is shown in Algorithm 1. The empirical study of the convergence of the algorithm is presented in the experimental section.

Constructing New Representation

In the previous section, we have described how to learn a semantic view $\mathbf{U}$ for each low-level image feature. In this section, we show how to map the target images $\mathbf{X}$ to the semantic feature representation for image classification. We first transform 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 train standard classifiers on $\{\widetilde{\mathbf{x}}_i, y_i\}$'s to make predictions on the testing images $\widetilde{\mathbf{X}}^*$, on which we apply the same feature representation construction.

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 $\lambda$, 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

Dataset and Processing

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

Caltech-256 image dataset consists of 256 categories of images, with each category usually containing hundreds of images. We randomly select 19 categories from the 256 categories, and build $\binom{19}{2} = 171$ pairs of binary classification tasks. The selected 19 categories and the corresponding number of images in each category are summarized as follows: *tennis-racket* (298), *american-flag* (299), *school-bus* (361), *cake* (96), *cd* (300), *chessboard* (299), *greyhound* (299), *fried-egg* (300), *dog* (359), *lighthouse* (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. We will use other data sources to test the method in the future.

In our experiments, we use the 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 testing data as the evaluation criterion, which is defined as follows,

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

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

For each binary classification task, there are hundreds of images. We randomly select 5 of them as training instances and the rest as testing instances. We repeat this for 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 compared 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.

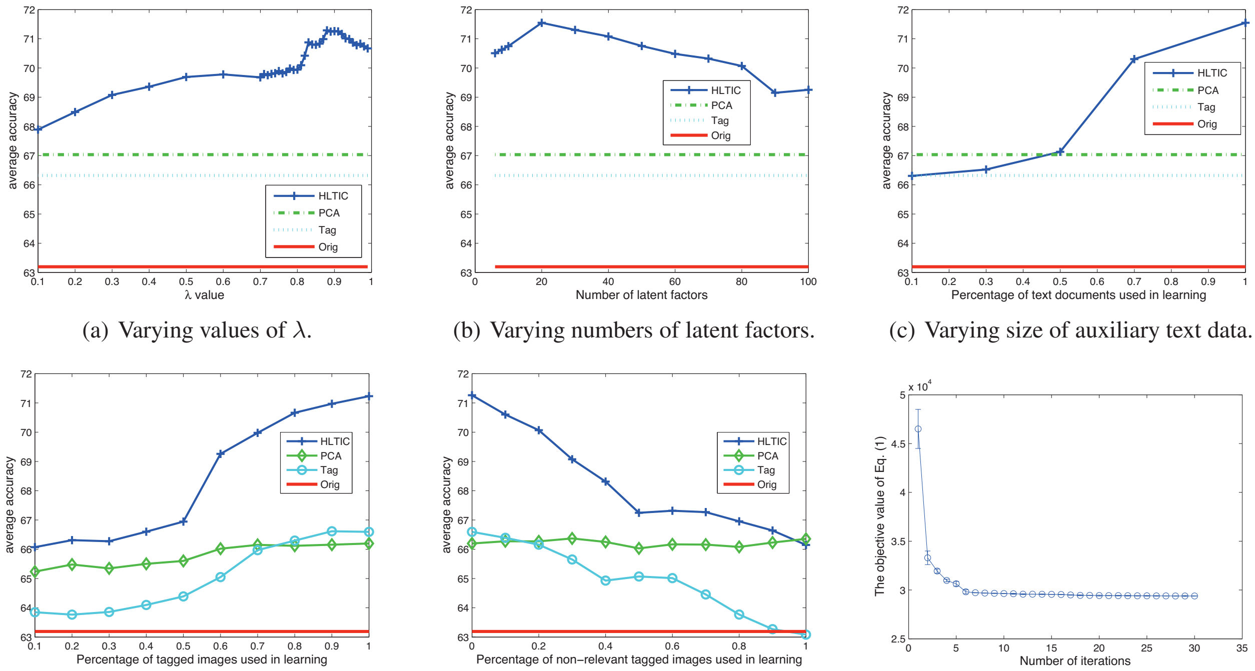
Experimental Results

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 of improvement can be negative if **HTLIC** performs worse than **Orig**. The last row in the table shows the average results

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

Table 2: Comparison results with labeled training instances.

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 size of auxiliary text data.

(d) Varying size of annotated image data.

(e) Varying size of relevant annotated image data.

(f) Varying numbers of iteration.

Figure 3: Experiments on parameter sensitivity.

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 quantity of the auxiliary text data to the overall performance of **HTLIC** in image classification. In this experiment, we denote "standard" or 1 for short, to be the whole document set crawled from the Web, and denote 0.1 to be the 10% documents sampled from the whole document set. The experimental results are shown in Figure 3(c). As we can see that when the size of the auxiliary document set increases, the performance of **HTLIC** increases as well. 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, resulting in the decomposition on $\mathbf{G}$ being more precise.

We also vary the size of the annotated images for each task. As shown in Figure 3(d), varying auxiliary image size affect the results for all the methods using auxiliary images. **HTLIC** and **Tag** have a clear curve of improving when there are 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). We share the same consensus that finding the correlation between images and text is critical to the understanding of images. However, our method is novel in that we use text and images from totally different sources and also the aim is different. Our work is also related to works on tagged images, e.g. (Wu *et al.* 2011).

Conclusions

In this paper, we propose the heterogeneous transfer learning method for image classification. We show that the performance of image classification can be improved by utilizing textual information. To bridge text documents and images, we use tagged images and create a semantic view for each target image by using collective matrix factorization technique, which effectively incorporates information in the auxiliary text into the tagging matrix. The experimental results also show our method outperforms other baselines when the labeled data in the target domain are short in supply.

Acknowledgement

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