Hybrid Heterogeneous Transfer Learning through Deep Learning

Joey Tianyi Zhou[†], Sinno Jialin Pan[‡], Ivor W. Tsang[§], and Yan Yan[♯]
[†]Nanyang Technological University, Singapore
[‡]Institute for Infocomm Research, Singapore
[§]University of Technology, Sydney, Australia
[♯]The University of Queensland, Australia
[†]tzhou1@ntu.edu.sg, [‡]jspan@i2r.a-star.edu.sg, [§]ivor.tsang@gmail.com, [♯]yanyan.tju@gmail.com

Abstract

Most previous heterogeneous transfer learning methods learn a cross-domain feature mapping between heterogeneous feature spaces based on a few cross-domain instance-correspondences, and these corresponding instances are assumed to be representative in the source and target domains respectively. However, in many real-world scenarios, this assumption may not hold. As a result, the constructed feature mapping may not be precise due to the bias issue of the correspondences in the target or (and) source domain(s). In this case, a classifier trained on the labeled transformed-source-domain data may not be useful for the target domain. In this paper, we present a new transfer learning framework called *Hybrid Heterogeneous Transfer Learning* (HHTL), which allows the corresponding instances across domains to be biased in either the source or target domain. Specifically, we propose a deep learning approach to learn a feature mapping between cross-domain heterogeneous features as well as a better feature representation for mapped data to reduce the bias issue caused by the cross-domain correspondences. Extensive experiments on several multilingual sentiment classification tasks verify the effectiveness of our proposed approach compared with some baseline methods.

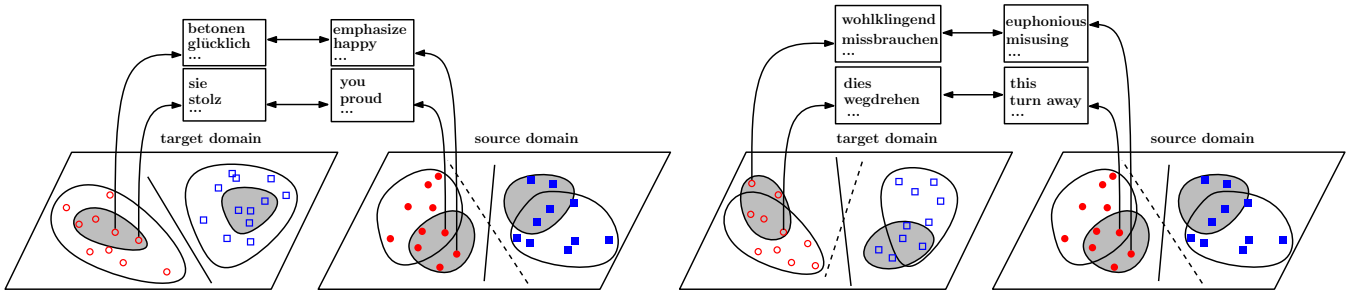
Introduction

Transfer learning or domain adaptation is a new machine learning paradigm, which aims to transfer knowledge extracted from an auxiliary domain, i.e., a source domain, where sufficient labeled data are available, to solve learning problems in a new domain, i.e., a target domain, with little or no additional human supervision (Pan and Yang 2010). Recently, more and more attention has been shifted from transferring knowledge across homogeneous domains to transferring knowledge across heterogeneous domains where the source and target domains may have heterogeneous types of features (Yang et al. 2009).

Different from homogeneous transfer learning, which assumes that the source and target domain data are represented in the same feature space of the same dimensionality (Blitzer, McDonald, and Pereira 2006; Pan, Kwok, and Yang 2008), and thus the domain difference is only caused by bias in feature or data distributions, heterogeneous transfer learning allows the source and target domains to be represented in different feature spaces. There are many real-world applications where heterogeneous transfer learning is crucial. For instance, many Natural Language Processing (NLP) tasks, such as named entity recognition, coreference resolution, etc., highly rely on a lot of annotated corpora and linguistic or semantic knowledge bases to build precise classifiers. For English, annotated corpora and knowledge bases are widely available, while for other languages, such as Thai, Vietnamese, etc., there are few resources. In this case, heterogeneous transfer learning is desirable to transfer knowledge extracted from rich English resources to solve NLP tasks in other languages whose resources are poor.

Most existing approaches to heterogeneous transfer learning aim to learn a feature mapping across heterogeneous feature spaces based on some cross-domain correspondences constructed either by labels (Kulis, Saenko, and Darrell 2011) or *a translator* (Dai et al. 2008). With the learned feature mapping, instances can be mapped from the target domain to the source domain or the other way round. In this way, source domain labeled data can be used to learn an accurate classifier for the target domain. There is a common assumption behind these methods that the selection of the instance-correspondences to learn the feature mapping is *unbiased*. In other words, the cross-domain corresponding instances are assumed to be representative in the source and target domains respectively. However, in many real-world scenarios, this assumption may not hold, which means that the selected corresponding instances may be biased and able to represent neither the source nor target domain data.

Taking cross-language document classification as a motivating example, which is illustrated in Figure 1(a), the objective is to learn a text classifier on documents in one language (e.g., German) only with a set of annotated English documents. To apply heterogeneous transfer learning methods to solve this task, one can simply construct German-English document-correspondences by translating some German documents into English by Google translator. However, the *wordbook* of the translated English documents may be quite different from that of the original English documents. For instance, the German word "betonen" is translated into the English word "emphasize" by Google translator. However, in an original English document, its corre-



(a) Case I: target corresponding instances are not biased

(b) Case II: target corresponding instances are biased

Figure 1: Hybrid Heterogeneous Transfer Learning.

sponding word is "highlight" or "stress". This is referred to as the "bias" issue in word-distribution between the translated documents and the original ones in the target domain. In this case, a feature mapping learned on such correspondences may not be effective.

As a second example as shown in Figure 1(b), we consider a multilingual sentiment classification task, where our task is to automatically classify overall sentiment polarities of songs reviews in German given labeled books reviews in English, and some unlabeled pairs of songs reviews in German and their English translations. Though one can make use of the German-English review-correspondences to learn a feature mapping from English to German, the labeled books reviews in English after transformation may not be useful to learn an accurate sentiment classifier for songs reviews in German. The reason is that, opinion and topic words used on different types of products can be very different (Li et al. 2012). In the context of this example, the "bias" between the transformed source domain data and the target domain data is caused by the difference of product domains.

Motivated by the above observations, we propose a new heterogeneous transfer learning framework called "hybrid heterogeneous transfer learning" (HHTL) to ensure knowledge transfer across heterogeneous domains to be effective even though the cross-domain correspondences are biased. Specifically, the HHTL framework consists of two main components: 1) learning a heterogeneous feature map between the source domain labeled data and the target unlabeled domain, and 2) discovering a latent representation to reduce the distribution bias between the transformed target domain unlabeled data and source domain labeled data. Thereafter, standard classification methods can be applied on the source domain labeled data with the latent representation to build a target classifier effectively. We present a deep learning method to simultaneously learn a feature transformation from the target domain to the source domain, and two mappings from lower-level features to higher-level features in the source and target domains respectively.

Note that the proposed HHTL framework is different from multi-view learning (Blum and Mitchell 1998), where fully correspondences between two views of data are required, and corresponding labels of the correspondences are assumed to be available in general. In HHTL, no label information of the cross-domain instance-correspondences is required. Moreover, in HHTL, labeled data is only assumed to be available in the source domain, while the goal is to learn a classifier to be used in the target domain.

Related Work

Homogeneous transfer learning aims to improve models' generalization ability across different domains of the same feature space. Learning a good feature representation for different domain data is crucial to homogeneous transfer learning. For instance, Pan et al. (2011) proposed a dimensionality reduction method, namely transfer component Analysis (TCA), to learn a low-dimensional space where the distance in distributions between a source domain and a target domain can be reduced. Raina et al. (2007) proposed a self-taught learning framework based on sparse coding to learn high-level features for transfer learning. Recently, deep learning techniques have been proposed for transfer learning. Glorot, Bordes, and Bengio (2011) applied stack denoised autoencoder (SDA) to learn hidden feature representations for cross-domain sentiment classification. In a follow-up work, Chen et al. proposed a variation of SDA for transfer learning, namely marginalized SDA (mSDA), which has been shown to be more effective and efficient.

Heterogeneous transfer learning aims to transfer knowledge across different feature spaces. Most methods for heterogeneous transfer learning aim to learn a common feature representation based on some correspondences between domains such that both source and target domain data can be represented by homogeneous features. Specifically, one can learn two different feature mappings to transform the source domain and target domain data to a latent feature space respectively (Shi et al. 2010; Prettenhofer and Stein 2010; Wang and Mahadevan 2011; Duan, Xu, and Tsang 2012). Alternatively, one can learn an asymmetric transformation to map data from one domain to another domain (Kulis, Saenko, and Darrell 2011; Zhou et al. 2014). Moreover, there are many methods (Dai et al. 2008; Wang and Mahadevan 2009) proposed to learn the mappings by incorporating instance correspondences between domains. Deep learning techniques have also been proposed to heterogeneous transfer learning (Socher et al. 2013), where knowledge is transferred from text to image based on a lot of text-image correspondences through deep learning. Our proposed framework can be considered as a more general case of heterogeneous transfer learning, where the bias of the correspondences between the source and target domains is considered.

A Deep Learning Approach

Problem Formulation

Given a set of target domain unlabeled data $\mathbf{D}_T = \{\mathbf{x}_{T_i}\}_{i=1}^{n_1}$, a set of source domain labeled data $\mathbf{D}_S = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_2}$, and an additional set of pairs of source and target domain unlabeled data, namely parallel data, $\mathbf{D}_C = \{(\mathbf{x}_{S_i}^{(c)}, \mathbf{x}_{T_i}^{(c)})\}_{i=1}^{n_c}$, where $\mathbf{x}_{S_i}$ or $\mathbf{x}_{S_i}^{(c)}$ is in $\mathbb{R}^{d_S \times 1}$, and $\mathbf{x}_{T_i}$ or $\mathbf{x}_{T_i}^{(c)}$ is in $\mathbb{R}^{d_T \times 1}$. Our objective is to learn weight matrices $\mathbf{W}_S$ and $\mathbf{W}_T$ to project the source domain labeled data and the target domain unlabeled data to hidden representations, $\mathbf{W}_S\mathbf{X}_S$ and $\mathbf{W}_T\mathbf{X}_T$, respectively, and a feature mapping $\mathbf{G}$ to map data from the target domain feature space to the source domain feature space,[1] such that the difference between the original source domain data and the transferred target domain data is small. With the learned weight matrices $\mathbf{W}_S$, $\mathbf{W}_T$ and the feature mapping $\mathbf{G}$, one can train a classifier f from $\{(\mathbf{W}_S\mathbf{X}_{S_i}, y_{S_i})\}_{i=1}^{n_1}$, and make prediction on a target domain unlabeled data $\mathbf{x}_T^*$ by applying $f(\mathbf{G}(\mathbf{W}_T\mathbf{X}_T^*))$.

As discussed in the two examples of cross language/lingual text mining, on one hand, the selection of corresponding instances in the source domain is always biased in general. This is because the corresponding instances in the source domain are not randomly selected based on the source domain data distribution but based on the selection of the corresponding instances in the target domain. On the other hand, the selection of the corresponding instances in the target domain can be either: 1) unbiased, e.g., in the example of cross language document classification, one can apply Google translator on a set of randomly selected documents in German to construct correspondences; or 2) bias, e.g., in the example of cross lingual sentiment classification, a set of reviews in German to be used to construct correspondences are about books while our target is to learn a sentiment classifier on music reviews. In the following sections, we will address these two problems through the proposed hybrid heterogeneous transfer learning framework.

High-Level Homogeneous Feature Learning

Inspired by the motivations and promising results of self-taught learning and deep learning in domain adaptation (Raina et al. 2007; Glorot, Bordes, and Bengio 2011; Chen et al. 2012), to address the bias issue in either the source domain only or both the source and target domains caused by instance shift or feature mismatch, we propose to apply Stacked Denoised Autoencoder (SDA) (Vincent et al. 2008) on the source domain data, i.e., the source domain labeled data and the source-domain corresponding unlabeled data[2], and the target domain data, i.e., the target domain unlabeled data and the target-domain corresponding unlabeled data, to learn high-level representations. Specifically, SDA firstly randomly sets some values of source domain features to be 0, which is referred to as a "corruption" of source domain data. In total, one can obtain m different corruptions. After that SDA tries to learn high-level features by reconstructing these m corruptions. For example, German word "betonen" is translated to"emphasize" by using Google Translator. However in human writings, one may use the words "proud" and "like" instead. SDA aims to reconstruct the machine translated word "emphasize" by using the words "proud" and "like". Therefore, the learned high-level features have capability to reduce data or feature bias.

In particular, we adopt a recently proposed method named Marginalized Stacked Denoised Autoencoder (mSDA) for high-level feature learning on homogeneous features. mSDA is an extension of SDA, which simplifies the reconstruction from two-level encoder and decoder to a single mapping. The reasons why we use mSDA are two folds: 1) the effectiveness of mSDA has been shown in homogeneous domain adaptation problems (Chen et al. 2012), and 2) compared to the standard SDA method, mSDA has proven to be much more efficient. For simplicity in presentation, following the notations used in (Chen et al. 2012), we absorb a constant feature into the feature vector as $\mathbf{x}_S = [\mathbf{x}_S^\top \ 1]^\top$ or $\mathbf{x}_T = [\mathbf{x}_T^\top \ 1]^\top$, and incorporate a bias term $\mathbf{b}$ within the weight matrix as $\mathbf{W} = [\mathbf{W} \ \mathbf{b}]$. We further denote $\overline{\mathbf{X}}_S = [\mathbf{X}_S^\top \ \mathbf{X}_S^{c\top}]^\top$ the union source domain data, and $\overline{\mathbf{X}}_T = [\mathbf{X}_T^\top \ \mathbf{X}_T^{c\top}]^\top$ the union target domain data.

Firstly, for the source domain data, we apply mSDA on $\overline{\mathbf{X}}_S$ to learn a weight matrix $\mathbf{W}_S \in \mathbb{R}^{(d_S+1)\times(d_S+1)}$ by minimizing the squared reconstruction loss as follows,

$$\sum_{i=1}^{m} \left\| \overline{\mathbf{X}}_S - \mathbf{W}_S \overline{\mathbf{X}}_S^{(i)} \right\|_F^2, \tag{1}$$

where $\overline{\mathbf{X}}_S^{(i)}$ denotes the i-th corrupted version of $\overline{\mathbf{X}}_S$. The solution to (1) depends on how the original features are corrupted which can be explicitly expressed as follows,

$$\mathbf{W}_S = \mathbf{P}\mathbf{Q}^{-1} \ \text{ with } \ \mathbf{Q} = \widetilde{\mathbf{X}}_S\widetilde{\mathbf{X}}_S^\top \ \text{ and } \ \mathbf{P} = \widehat{\mathbf{X}}_S\widetilde{\mathbf{X}}_S^\top, \tag{2}$$

where $\widehat{\mathbf{X}}_S = [\overline{\mathbf{X}}_S \ \overline{\mathbf{X}}_S \ \cdots \ \overline{\mathbf{X}}_S]$ denotes the m-times repeated version of $\overline{\mathbf{X}}_S$, and $\widetilde{\mathbf{X}}_S$ is the corrupted version of $\widehat{\mathbf{X}}_S$. In general, to alleviate bias in estimation, a large number of m over the training data with random corruptions are required, which is computationally expensive. To address this issue, mSDA introduces a corruption probability p to model infinite corruptions, i.e., $m \longrightarrow \infty$. Define a feature vector $\mathbf{q} = [1-p, \cdots, 1-p, 1]^\top \in \mathbb{R}^{d_S+1}$, where $\mathbf{q}_i$ represents the probability of a feature indexed by i "surviving" after the corruption. Thus, we can obtain the expectation of (1), and its solution can be written analytically as

$$\mathbf{W}_S = \mathbb{E}[\mathbf{P}]\mathbb{E}[\mathbf{Q}]^{-1}, \tag{3}$$

where $\mathbb{E}[\mathbf{P}]_{ij} = \mathbf{S}_{ij}\mathbf{q}_j$, $\mathbf{S} = \overline{\mathbf{X}}_S\overline{\mathbf{X}}_S^\top$, and

$$\mathbb{E}[\mathbf{P}]_{ij} = \begin{cases} \mathbf{S}_{ij}\mathbf{q}_i\mathbf{q}_j, & \text{if } i \neq j, \\ \mathbf{S}_{ij}\mathbf{q}_i, & \text{otherwise.} \end{cases} \tag{4}$$

After $\mathbf{W}_S$ is learned, the nonlinearity of features is injected through the nonlinear encoder function $h(\cdot)$ that is learned

[1] Alternatively, one can learn a feature mapping $\mathbf{G}^\top$ to map data from the source domain to the target domain.

[2] In practice, if there is an additional set of unlabeled data in the source domain, one can use it as well for high-level feature learning. However, in this paper, for simplicity in description, we do not assume that additional source domain unlabeled data is available.

together with the reconstruction weights $\mathbf{W}_S$, mSDA applies a nonlinear squashing-function, e.g., the hyperbolic tangent function $\tanh(\cdot)$, on the outputs of mSDA, $\overline{\mathbf{H}}_S = \tanh(\mathbf{W}_S\overline{\mathbf{X}}_S)$, to generate nonlinear features.

Feature Learning for Source and Target Domains

The above process can be recursively done by replacing $\overline{\mathbf{X}}_S$ with $\overline{\mathbf{H}}_S$ to obtain a series of weight matrices $\{\mathbf{W}_S^k\}$'s for each layer of feature learning. Similarly, for the union target domain data $\overline{\mathbf{X}}_T$, we can recursively apply mSDA to learn a series of reconstruction weights $\{\mathbf{W}_T^k\}$'s and generate high-level nonlinear features for each layer of feature learning, similarly. Note that when there is no data or feature bias in the target domain, one can simply set $\mathbf{W}_T$ to be the identity matrix of the dimensionality $d_T + 1$, and replace $\tanh(\cdot)$ by the identical function, respectively.

Heterogeneous Feature Mapping

So far, in a specific layer k of feature learning, we have learned a pair of reconstruction weights $\mathbf{W}_{S,k}$ and $\mathbf{W}_{T,k}$, and higher-level feature representations $\overline{\mathbf{H}}_{S,k}$ and $\overline{\mathbf{H}}_{T,k}$ for the union source and target domain data respectively. By denoting $\overline{\mathbf{H}}_{S,k}^{(c)}$ and $\overline{\mathbf{H}}_{T,k}^{(c)}$ the higher-level feature representations of the cross-domain corresponding instances in the source and target domains respectively, we now introduce how to learn a feature mapping across heterogeneous features $\overline{\mathbf{H}}_{S,k}^{(c)}$ and $\overline{\mathbf{H}}_{T,k}^{(c)}$.

Specifically, in layer k, we aim to learn a feature transformation $\mathbf{G}_k \in \mathbb{R}^{(d_S+1)\times(d_T+1)}$, where a bias term is incorporated within the transformation, by minimizing the following objective,

$$\|\mathbf{H}_{S,k} - \mathbf{G}\mathbf{H}_{T,k}\|_F^2 + \lambda\|\mathbf{G}_k\|_F^2, \tag{5}$$

where $\lambda > 0$ is a parameter of the regularization term on $\mathbf{G}_k$, which controls the tradeoff between the alignment of heterogeneous features and the complexity of $\mathbf{G}_k$. It can be shown that the optimization problem (5) has a closed form solution which can be written as follows,

$$\mathbf{G}_k = (\mathbf{H}_{S,k}\mathbf{H}_{T,k}^\top)(\mathbf{H}_{T,k}\mathbf{H}_{T,k}^\top + \lambda\mathbf{I})^{-1}, \tag{6}$$

where $\mathbf{I}$ is the identity matrix of the dimensionality $d_T + 1$.

Prediction by Stacking Layers

By defining the total number of layers K of deep learning, one can recursively apply mSDAs and the heterogeneous feature mapping learning algorithm on the source and target domain data in each layer to generate different levels of features and feature transformations between heterogeneous features. After learning high-level features and feature mappings, for each source domain labeled instance $\mathbf{x}_{S_i}$, by denoting $\mathbf{h}_{S_i,k}$ its corresponding higher-level feature representation in the k-th layer, we can define a new feature vector $\mathbf{z}_{S_i}$ by augmenting its original features and high-level features of all layers as $\mathbf{z}_{S_i} = [\mathbf{h}_{S_i,1}^\top \cdots \mathbf{h}_{S_i,K}^\top]^\top$, where $\mathbf{h}_{S_i,1} = \mathbf{x}_{S_i}$. We then apply a standard classification algorithm on $\{\mathbf{z}_{S_i}, y_{S_i}\}$'s to train a target classifier f. For making a prediction on a target domain instance $\mathbf{x}_T^*$, we first generate its higher-level feature representations $\{\mathbf{h}_{T,k}^*\}_{k=1}^K$, where $\mathbf{h}_{T,1}^* = \mathbf{x}_T$, of each layer by using the weight matrices $\{\mathbf{W}_{T,k}\}_{k=1}^K$, where $\mathbf{W}_{T,1} = \mathbf{I}$, and do the feature augmentation, $\mathbf{z}_T = [(\mathbf{G}_1\mathbf{h}_{T,1}^*)^\top \cdots (\mathbf{G}_K\mathbf{h}_{T,K}^*)^\top]^\top$. Finally, we apply the learned classifier f on $\mathbf{z}_T$ to make prediction $f(\mathbf{z}_T)$. The reason that we augment different layers of features for both training and testing is because we aim to incorporate additional high-level features to alleviate the bias for both two domains without losing original feature information. The overall algorithm is summarized in Algorithm 1.

Algorithm 1 Hybrid Heterogeneous Transfer Learning.

Input: target domain unlabeled data $\mathbf{D}_T = \{\mathbf{x}_{T_i}\}_{i=1}^{n_1}$, source domain labeled data $\mathbf{D}_S = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_2}$, cross-domain parallel data, $\mathbf{D}_C = \{(\mathbf{x}_{S_i}^{(c)}, \mathbf{x}_{T_i}^{(c)})\}_{i=1}^{n_c}$, a feature corruption probability p in mSDA, a trade-off parameter λ, and the number of layers K.

Initializations: $\overline{\mathbf{X}}_S = [\mathbf{X}_S\ \mathbf{X}_S^{(c)}]$, $\overline{\mathbf{X}}_T = [\mathbf{X}_T\ \mathbf{X}_T^{(c)}]$, $\overline{\mathbf{H}}_{S,1} = \overline{\mathbf{X}}_S$, $\overline{\mathbf{H}}_{T,1} = \overline{\mathbf{X}}_T$, and learn $\mathbf{G}_1$ by solving

$$\min_{\mathbf{G}_1} \|\mathbf{H}_{S,1}^{(c)} - \mathbf{G}_1\mathbf{H}_{T,1}^{(c)}\|_F^2 + \lambda\|\mathbf{G}_1\|_F^2.$$

for $k = 2, ..., K$ **do**

1: Apply mSDA on $\overline{\mathbf{H}}_{S,k-1}$ and $\overline{\mathbf{H}}_{T,k-1}$:

$$\{\mathbf{W}_{S,k}, \overline{\mathbf{H}}_{S,k}\} = \text{mSDA}(\overline{\mathbf{H}}_{S,k-1}, p),$$
$$\{\mathbf{W}_{T,k}, \overline{\mathbf{H}}_{T,k}\} = \text{mSDA}(\overline{\mathbf{H}}_{T,k-1}, p).$$

Note: if there is no data or feature bias in the target domain, simply set $\mathbf{W}_{T,k} = \mathbf{I}$, and $\overline{\mathbf{H}}_{T,k} = \overline{\mathbf{X}}_T$.

2: Learn heterogeneous feature mapping $\mathbf{G}_k$:

$$\min_{\mathbf{G}_k} \|\mathbf{H}_{S,k}^{(c)} - \mathbf{G}_k\mathbf{H}_{T,k}^{(c)}\|_F^2 + \lambda\|\mathbf{G}_k\|_F^2.$$

end for

Do feature augmentation on source domain labeled data

$$\mathbf{Z}_S = [\mathbf{H}_{S,1}^\top\ \cdots\ \mathbf{H}_{S,K}^\top]^\top,$$

and train a classifier f with $\{\mathbf{Z}_S, \mathbf{Y}_S\}$.

Output: f, $\{\mathbf{G}_k\}_{k=1}^K$, $\{\mathbf{W}_{S,k}\}_{k=2}^K$, and $\{\mathbf{W}_{T,k}\}_{k=2}^K$.

Experiments

In experiments, we verify the proposed framework on several cross-language classification tasks in terms of classification accuracy, impact of layers and parameter sensitivity.

Experiment Setting

Dataset The cross-language sentiment dataset (Prettenhofer and Stein 2010) comprises of Amazon product reviews of three product categories: books, DVDs and music. These reviews are written in four languages: English (EN), German (GE), French (FR), and Japanese (JP). For each language, the reviews are split into a train file and a test file, including 2,000 reviews per categories. We use the English reviews in the train file as the source domain labeled data, non-English (each of the other 3 languages) reviews in a train file as target

domain unlabeled data. Moreover, we apply Google translator on the non-English reviews in a test file to construct cross-domain (English v.s. non-English) unlabeled parallel data. The performance of all methods are evaluated on the target domain unlabeled data.

Baselines In our heterogeneous transfer learning setting, no labeled data is provided for the target domain. Most HTL methods which require target domain labeled data cannot be used as baselines. Therefore, we only compare the proposed HHTL framework with the following baselines:

- **SVM-SC**: We first train a classifier on the source domain labeled data and then predict on the source domain parallel data. By using the correspondence, the predicted labels for source parallel data can be transferred into target parallel data. Next, we train a model on the target parallel data with predicted labels to make predictions on the target domain test data. We name this baseline as the SVM-Source-Correspondence transfer (SVM-SC).
- **CL-KCCA**: We apply Cross-Lingual Kernel Canonical Component Analysis (CL-KCCA) (Vinokourov, Shawe-Taylor, and Cristianini 2002) on the unlabeled paralle data to learn two projections for the source and target languages, and then train a monolingual classifier with the projected source domain labeled data.
- **HeMap**: We apply heterogeneous Spectral Mapping (HeMap) (Shi et al. 2010) to learn mappings to project two domain data onto a common feature subspace. However, HeMap does not take the instance correspondence information into consideration.
- **mSDA-CCA**: We apply Multimodal Deep Learning (Ngiam et al. 2011) to learn a shared feature representation for "multimodal" domains. In experiments, for the fair comparison, instead of using RBM for high-level feature learning, we adopt mSDA and conduct CCA on the correspondences between domains in same layers.

For all experiments, we employ the linear support vector machine (SVM) (Fan et al. 2008) with default parameter settings. We use the cross-validation to adjust the model parameters. Specifically, we choose λ from $\{0.01, 0.1, 1, 10, 100\}$ for HHTL, choose corruption probability p from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ for mSDA from, and fix the number of layers used in mSDA to be 3. We tune the parameter κ for CL-KCCA (see (5) in (Vinokourov, Shawe-Taylor, and Cristianini 2002)), mSDA-CCA, and the parameter β for HeMap (See (1) in (Shi et al. 2010)) from $\{0.01, 0.1, 1, 10, 100\}$. In this paper, we employ the deep learning network to learn better high-level feature representations. Due to the cost of memory and computation, we only study 3 layers. And due to the limit of space, we only report the results of 1 and 3 layers, denoted by HHTL(1) and HHTL(3), respectively. For the mSDA-CCA, we only report the results with a 3-layer structure.

Performance Comparison

We evaluate the performance of proposed methods under two learning settings: 1) learning with unbiased target correspondence instances, and 2) learning with biased target correspondence instances.

Table 1: Learning with unbiased target correspondence instances: comparison results in terms of testing accuracy (%).

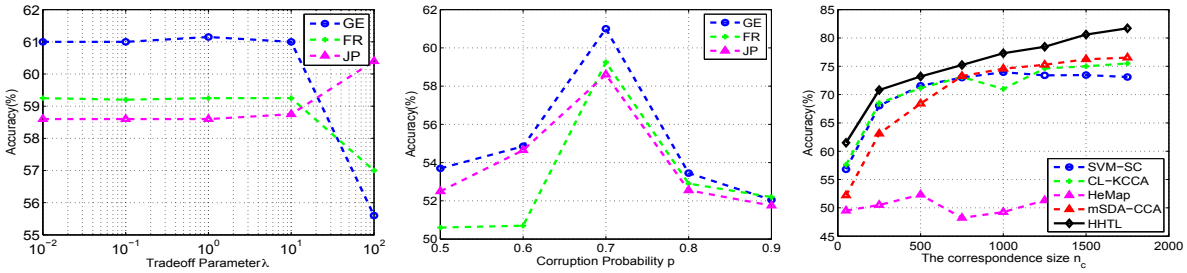
Target	SVM-SC	CL-KCCA	HeMap	mSDA-CCA	HHTL(1)	HHTL(3)
FR	73.10	75.50	50.23	74.89	74.01	**82.50**
GE	73.05	75.00	49.83	76.53	73.08	**82.70**
JP	65.50	66.82	51.30	68.19	66.12	**75.56**

Learning with unbiased target correspondence instances In this setting, we consider a data bias in the source domain. During the training process, all the original English reviews that consist of 3 categories are used as the source domain labeled data, and non-English reviews are considered as target domain data. We randomly choose 2,000 non-English reviews of all the categories, and translate them to English to form the parallel data. The remaining non-English reviews form the unlabeled target data. The averaged results in terms of accuracy on the reviews of each non-English language over 10 repetitions are reported in Table 1.

From Table 1, we can observe that the proposed HHTL method with 3 layers outperforms all the other baselines significantly. The reason is that HHTL benefits a lot from learning the high-level features which can alleviate the data bias between the translated and original reviews in the source domain. Besides, the performance of CL-KCCA and SVM-SC is much better than HeMap. The inferior performance of HeMap is caused by the fact that HeMap discards the valuable corresponding information in training. Moreover, mSDA-CCA performs slightly better than CL-KCCA, and much better than all the other baselines because of the powerful representations learned by the deep structure. However, it still performs worse than HHTL because the representations learned by mSDA-CCA are based on the biased correspondences. From the results, we can also see that the performance of all the methods in the Japanese domain is much worse than that in the other two language domains. It is due to the fact that Japanese is more different from English compared to German and French. French and German all come from the similar family of languages in the West. Thus they share many similar words with English, while Japanese belongs to the family of Eastern languages. Therefore, it is more difficult to transfer knowledge from the English domain to the Japanese domain.

Learning with biased target correspondence instances In this setting, we focus on cross-language cross-category learning between English and the other 3 languages. For the comprehensive comparisons, we construct 18 cross-language cross-category sentiment classification tasks as follows: EN-B-FR-D, EN-B-FR-M, EN-B-GE-D, EN-B-GE-M, EN-B-JP-D, EN-B-JP-M, EN-D-FR-B, EN-D-FR-M, EN-D-GE-B, EN-D-GE-M, EN-D-JP-B, EN-D-JP-M, EN-M-FR-B, EN-M-FR-D, EN-M-GE-B, EN-M-GE-D, EN-B-JP-B, EN-B-JP-D. For example, the task EN-B-FR-D uses all the Books reviews in French in the test file and its English translations as the parallel data, the DVD reviews in French as the target language test data, and original English Books reviews as the source domain labeled data.

The results are summarized in the Table 2. This setting is more challenging than the previous one due to larger data bias after feature transformation. Therefore, the performance of all target languages is dropped compared to that



(a) Accuracy v.s. λ (b) Accuracy v.s. p (c) Accuracy v.s. n_c on GE

Figure 2: Parameter Analysis.

Table 2: Learning with biased target correspondence instances: comparison results in terms of testing accuracy (%).

TASK	SVM-SC	CL-KCCA	HeMap	mSDA-CCA	HHTL(1)	HHTL(3)
EN-B-FR-D	73.25	50.00	49.45	72.96	72.65	**76.75**
EN-B-FR-M	62.40	47.20	50.70	64.29	63.30	**67.65**
EN-B-GE-D	72.55	77.50	66.12	**78.23**	70.95	75.10
EN-B-GE-M	58.95	50.00	48.35	62.52	59.60	**69.55**
EN-B-JP-D	69.50	69.46	49.55	71.95	69.45	**72.56**
EN-B-JP-M	53.15	55.23	52.05	57.17	53.56	**62.39**
EN-D-FR-B	71.35	47.10	48.35	72.32	71.25	**79.27**
EN-D-FR-M	71.65	47.10	49.90	68.87	71.70	**75.43**
EN-D-GE-B	76.40	54.87	50.80	75.32	74.00	**79.35**
EN-D-GE-M	69.05	54.87	50.55	75.69	74.05	**78.55**
EN-D-JP-B	70.00	61.16	50.12	**72.60**	65.55	68.12
EN-D-JP-M	59.20	66.89	49.80	57.75	66.02	**70.58**
EN-M-FR-B	**76.20**	73.54	50.55	74.23	73.80	75.84
EN-M-FR-D	73.65	47.10	50.87	72.76	74.15	**77.04**
EN-M-GE-B	74.50	**81.50**	49.45	76.82	75.45	78.30
EN-M-GE-D	74.60	80.20	50.20	72.28	74.45	**81.42**
EN-B-JP-B	67.85	63.75	48.85	68.83	65.35	**71.65**
EN-B-JP-D	69.35	60.46	48.80	71.06	68.55	**74.25**

shown in Table 1. Our proposed method HHTL still performs much more stable and better than the other 3 baselines except for a few tasks. This is because that our proposed method can largely reduce the data bias by generating the more powerful and higher-level features for both the source and target domains, which can lead to better cross-language mappings in each layer.

Impact of Layers in the HHTL Structure

As shown in Table 2, the more layers are used, the better performance is achieved. SVM-SC manifests comparable results with HHTL using 1 layer, where the learned high-level features are not sufficiently useful. However, with increasing number of layers, HHTL can enhance the cross-language classification performance by generating more useful and higher-level features that alleviate the data bias.

Parameter Sensitivity Study

The proposed method has two parameters to tune: 1) the tradeoff parameter λ, and 2) the feature corruption probability p. Besides these, the number of correspondences n_c from the parallel data may also affect the performance of cross-language classification. In this section, we conduct a series of experiments to study the sensitivity issues of the parameters, λ, p and n_c.

In the first experiment, we aim to analyze how performance changes with varying values of λ in the range of $[0.001, 0.01, 1, 10, 100]$ with $p = 0.7$ and $n_c = 50$. From Figure 2(a), we can observe that the performance is stable when λ is no more than 10. Besides, we set $\lambda = 0.01$, $n_c = 50$ and vary p in the range of $[0.5, 0.6, 0.7, 0.8, 0.9]$. The results of all the target languages are illustrated in Figure $2(b)$. We can see that corruption probability cannot be either too large or too small to achieve good performance. On one hand, if we corrupt many features, the original feature information will be discarded a lot, resulting in failure of discovering powerful hidden features in higher layers by reconstructions. On the other hand, if we only corrupt a few features, the high-level hidden features recovered from the original features always tends to be similar to the original ones. In this case, we cannot learn informative hidden features for knowledge transfer.

In the second experiment, we study the influence of the size of the unlabeled parallel data ($[50, 250, 500, 750, 1000, 1250, 1500, 1750]$) to the overall performance of HHTL. The results for German are reported in Figure 2(c).[3] All the methods which use the unlabeled parallel data consistently outperform HeMap, which discards the correspondence information. Nevertheless, HTLL performs the best and achieve more stable improvement with increasing size of parallel data. Different from HTLL, the accuracies of CL-KCCA and SVM-SC almost stop increasing when the size of parallel data is larger than 750. The reason is that the impact of domain distribution mismatch hinders the improvement of the overall performance even though with more correspondences. mSDA-CCA performs even worse than SVM-SC and CL-KCCA when the number of correspondences is smaller than 750. This is because that the multimodal deep learning method requires sufficient correspondence data between two modalities to learn reliable feature representations. Though mSDA-CCA outperforms other baselines when n_c is larger than 750, it still performs much worse than HHTL. All the results demonstrate the effectiveness and robustness of HHTL for cross-language sentiment classification.

Conclusions

In this paper, we have proposed a Hybrid Heterogeneous Transfer Learning (HHTL) framework to transfer knowledge across different feature spaces and simultaneously correct the data bias on the transformed feature space. Based on the framework, we proposed a deep learning approach. Extensive experiments demonstrated the superiority of the proposed method on several multilingual text mining tasks.

[3] Due to the limit of space, we only report the results on German. However, we observe similar results on the other languages.

Acknowledgments

This research was in part supported by the Australian Research Council Future Fellowship FT130100746.

References

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*, 120–128. ACL.

Blum, A., and Mitchell, T. M. 1998. Combining labeled and unlabeled sata with co-training. In *COLT*, 92–100.

Chen, M.; Xu, Z. E.; Weinberger, K. Q.; and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. In *ICML*.

Dai, W.; Chen, Y.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2008. Translated learning: Transfer learning across different feature spaces. In *NIPS*, 353–360.

Duan, L.; Xu, D.; and Tsang, I. W. 2012. Learning with augmented features for heterogeneous domain adaptation. In *ICML*.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* 9:1871–1874.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 513–520.

Kulis, B.; Saenko, K.; and Darrell, T. 2011. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 1785–1792.

Li, F.; Pan, S. J.; Jin, O.; Yang, Q.; and Zhu, X. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *ACL*, 410–419.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*, 689–696.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22(10):1345–1359.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22(2):199–210.

Pan, S. J.; Kwok, J. T.; and Yang, Q. 2008. Transfer learning via dimensionality reduction. In *AAAI*, 677–682.

Prettenhofer, P., and Stein, B. 2010. Cross-language text classification using structural correspondence learning. In *ACL*, 1118–1127.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. Y. 2007. Self-taught learning: transfer learning from unlabeled data. In *ICML*, 759–766.

Shi, X.; Liu, Q.; Fan, W.; Yu, P. S.; and Zhu, R. 2010. Transfer learning on heterogenous feature spaces via spectral transformation. In *ICDM*, 1049–1054.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*, 935–943.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*, 1096–1103.

Vinokourov, A.; Shawe-Taylor, J.; and Cristianini, N. 2002. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS*, 1473–1480.

Wang, C., and Mahadevan, S. 2009. A general framework for manifold alignment. In *AAAI Fall Symposium: Manifold Learning and Its Applications*.

Wang, C., and Mahadevan, S. 2011. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 1541–1546.

Yang, Q.; Chen, Y.; Xue, G.-R.; Dai, W.; and Yu, Y. 2009. Heterogeneous transfer learning for image clustering via the socialweb. In *ACL/IJCNLP*, 1–9.

Zhou, J. T.; Tsang, I. W.; Pan, S. J.; and Tan, M. 2014. Heterogeneous domain adaptation for multiple classes. In *AISTATS*, 1095–1103.