Proactive Transfer Learning for Heterogeneous Feature and Label Spaces

Seungwhan Moon ✉ and Jaime Carbonell

Language Technologies Institute,
School of Computer Science, Carnegie Mellon University,
5000 Forbes Ave, PA, 15213, U.S.A.
{seungwhm,jgc}@cs.cmu.edu

Abstract. We propose a framework for learning new target tasks by leveraging existing heterogeneous knowledge sources. Unlike the traditional transfer learning, we do not require explicit relations between source and target tasks, and instead let the learner actively *mine* transferable knowledge from a source dataset. To this end, we develop (1) a transfer learning method for source datasets with heterogeneous feature and label spaces, and (2) a proactive learning framework which progressively builds *bridges* between target and source domains in order to improve transfer accuracy. Experiments on a challenging transfer learning scenario (learning from *hetero-lingual* datasets with non-overlapping label spaces) show the efficacy of the proposed approach.

1 Introduction

The notion of enabling a machine to learn a new task by leveraging an auxiliary source of knowledge has long been the focus of transfer learning. While many different flavors of transfer learning approaches have been developed, most of these methods assume explicit relatedness between source and target tasks, such as the availability of source-target correspondent instances (*e.g.* multi-view / multimodal learning), or the class relations information for multiple datasets sharing the same feature space (*e.g.* zero-shot learning, domain adaptation), etc. These approaches have been effective in their respective scenarios, but very few limited studies have investigated learning from heterogeneous knowledge sources that lie in both different feature and label spaces. See Section 2 for the detailed literature review.

Given an unforeseen target task with limited label information, we seek to mine useful knowledge from a plethora of heterogeneous knowledge sources that have already been curated, albeit in different feature and label spaces. To address this challenging scenario we first need an algorithm to estimate *how* the source and the target datasets may be related. One common aspect of any dataset for a classification task is that each instance is eventually assigned to some abstract concept(s) represented by its category membership, which often has its own *name*. Inspired by the Deep Visual-Semantic Embedding (DeViSE) model [7] which assigns the unsupervised word embeddings to label terms, we propose to map heterogeneous source and target labels into the same word embedding space, from which we can obtain their semantic class relations. Using information from the class relations as an anchor, we first attempt to uncover a shared latent subspace where both source and target features can be mapped. Simultaneously, we learn a shared projection from this intermediate layer into the final embedded labels space, from which we can predict labels using the shared knowledge.

The quality of transfer essentially depends on how well we can uncover the *bridge* in the projected space where the two datasets are semantically linked. Intuitively, if the two datasets describe completely different concepts, very little information can be transferred from one to the other. We therefore also propose a proactive transfer learning framework which expands the labeled target data to actively *mine* transferable knowledge and to progressively improve the target task performance.

We evaluate the proposed combined approach on a unique learning problem of a *hetero-lingual* text classification task, where the objective is to classify a novel target text dataset given only a few labels along with a source dataset in a different language, describing different classes from the target categories. While this is a challenging task, the empirical results show that the proposed approach improves over the baselines.

The rest of the paper is organized as follows: we position our approach in relation to the previous work in Section 2, and formulate the heterogeneous transfer learning problem in Section 3. Section 4 describes in detail the proposed proactive transfer learning framework and presents the optimization problem. The empirical results are reported and analyzed in Section 5, and we give our concluding remarks and proposed future work in Section 6.

2 Related Work

Transfer learning with heterogeneous feature spaces :Multi-view representation learning aims at aggregating multiple heterogeneous "views" (feature sets) of an instance that describe the same concept to train a model. Most notably, [29] proposes Deep Canonically Correlated Autoencoders (DCCAE) which learn a representation that maximizes the mutual information between different views under an autoencoder regularization. While DCCAE is reported to be state of the art on multi-view representation learning using Canonical Correlation Analysis (CCA) [4], their approach (as well as other CCA-based methods) strictly require access to paired observations from two views belonging to the same class. [3] proposes translated learning which aims to learn a target task in the same label space as the source task, using source-correspondent instances such as image-text parallel captions as an anchor. [33] proposes Hybrid Heterogeneous Transfer Learning (HHTL) which extends the previous translated learning work with an added objective of learning an unbiased feature mapping through marginalized stacked denoising autoencoders (mSDA), given correspondent instances. [28] develops a similar approach in bilingual content classification tasks, and proposes to generate correspondent samples through an available machine translation system. [5] propose the Heterogeneous Feature Augmentation (HFA) method for a shared homogeneous binary classification task, which relaxes the previous limitations that require correspondent instances, and instead aims to discover a common subspace that can map two heterogeneous features. Our approach generalizes all the previous work by allowing for heterogeneous label spaces between source and target, thus not requiring explicit source-target correspondent instances or classes.

Transfer learning with a heterogeneous label space : Zero-shot learning aims at building a robust classifier for unseen novel classes in the target task, often by relaxing categorical label space into a distributed vector space via transferred knowledge. For instance, [19] uses image co-occurrence statistics to describe a novel image class category, while [30, 7, 27, 24, 8, 32, 15] embed labels into semantic word vector space according to their label terms, where textual embeddings are learned from auxiliary text documents in an unsupervised manner. More recently, [13] proposes to learn domain-adapted projections to the embedded label space. While these approaches are reported to improve robustness and generalization on novel target classes, they assume that source datasets are in the same feature space as the target dataset (*e.g.* image). We extend the previous research by adding the joint objective of uncovering relatedness among datasets with heterogeneous feature spaces, via anchoring the semantic relations between the source and the target label embeddings.

Domain Adaptation approaches aim to minimize the marginal distribution difference between source and target datasets, assuming their class conditional distribution remains the same for homogeneous feature and label spaces. This is typically implemented via instance re-weighting [11, 2, 14], subspace mapping [31], or via identification of transferable features [16]. [23] provide an exhaustive survey on other traditional transfer learning approaches.

Active learning provides an alternative solution to the label scarcity problem, which aims at reducing sample complexity by iteratively querying the most informative samples with the highest utility given the labeled sampled thus far [25, 21]. **Transfer active learning** approaches [6, 2, 12, 35, 26] aim to combine transfer learning with the active learning framework by conditioning transferred knowledge as priors for optimized selection of target instances. Specifically, [9] overcomes the common *cold-start* problem at the beginning phase of active learning with zero-shot class-relation priors. However, many of the previously proposed transfer active learning methods do not apply to our setting because they require source and target data to be in either homogeneous feature space or the same label space or both. Therefore, we propose a *proactive transfer learning* approach for heterogeneous source and target datasets, where the objective is to progressively find and query *bridge* instances that allow for more accurate transfer, given a sampling budget.

Our contributions are three-fold: we propose (1) a novel transfer learning method with both heterogeneous feature and label spaces, and (2) a proactive transfer learning approach for identifying and querying *bridge* instances between target and source tasks to improve transfer accuracy effectively. (3) We evaluate the proposed approach on a novel transfer learning problem, the *hetero-lingual* text classification task.

3 Problem Formulation

We formulate the proposed framework for learning a target multiclass classification task given a source dataset with heterogeneous feature and label spaces as follows: We first define a dataset for the target task $\mathbf{T} = \{\mathbf{X_T}, \mathbf{Y_T}, \mathbf{Z_T}\}$, with the target task features $\mathbf{X_T} = \{\mathbf{x_T^{(i)}}\}_{i=1}^{N_T}$ for $\mathbf{x_T} \in \mathbb{R}^{M_T}$, where N_T is the target sample size and M_T is the target feature dimension, the ground-truth labels $\mathbf{Z_T} = \{\mathbf{z_T^{(i)}}\}_{i=1}^{N_T}$, where $\mathbf{z_T} \in \mathcal{Z}_T$ for a categorical target label space $\mathcal{Z}_T$, and the corresponding high-dimensional label descriptors $\mathbf{Y_T} = \{\mathbf{y_T^{(i)}}\}_{i=1}^{N_T}$ for $\mathbf{y_T} \in \mathbb{R}^{M_E}$, where M_E is the dimension of the embedded labels, which can be obtained from *e.g.* unsupervised word embeddings, etc. We also denote L_T and UL_T as a set of indices of labeled and unlabeled target instances, respectively, where $|L_T| + |UL_T| = N_T$. For a novel target task, we assume that we are given zero or a very few labeled instances, thus $|L_T| = 0$ or $|L_T| \ll N_T$. Similarly, we define a heterogeneous source dataset $\mathbf{S} = \{\mathbf{X_S}, \mathbf{Y_S}, \mathbf{Z_S}\}$, with $\mathbf{X_S} = \{\mathbf{x_S^{(i)}}\}_{i=1}^{N_S}$ for $\mathbf{x_S} \in \mathbb{R}^{M_S}$, $\mathbf{Z_S} = \{\mathbf{z_S^{(i)}}\}_{i=1}^{N_S}$ for $\mathbf{z_S} \in \mathcal{Z}_S$, $\mathbf{Y_S} = \{\mathbf{y_S^{(i)}}\}_{i=1}^{N_S}$ for $\mathbf{y_S} \in \mathbb{R}^{M_E}$, and L_S, accordingly. For the source dataset we assume $|L_S| = N_S$. Note that in general, we assume $M_T \neq M_S$ (heterogeneous feature space) and $\mathcal{Z}_T \neq \mathcal{Z}_S$ (heterogeneous label space).

Our goal is then to build a robust classifier $\mathbf{f} : \mathcal{X}_T \rightarrow \mathcal{Z}_T$ for the target task, trained with $\{\mathbf{x_T^{(i)}}, \mathbf{y_T^{(i)}}, \mathbf{z_T^{(i)}}\}_{i \in L_T}$ as well as transferred knowledge from $\{\mathbf{x_S^{(i)}}, \mathbf{y_S^{(i)}}, \mathbf{z_S^{(i)}}\}_{i \in L_S}$.

4 Proposed Approach

Our approach aims to leverage a source data that lies in different feature and label spaces from a target task. Transferring knowledge directly from heterogeneous spaces is intractable, and thus we begin by obtaining a unified vector representation for different source and target categories. Specifically, we utilize a skip-gram based language model that learns semantically meaningful vector representations of words, and map our categorical source and target labels into the word embedding space (Section 4.1). In parallel, we learn compact representations for the source and the target features that encode abstract information of the raw features (Section 4.2), which allows for more tractable transfer through affine projections. Once the label terms for the source and the target datasets are anchored in the word embedding space, we first learn projections into a new latent common feature space from the source and the target feature spaces ($\mathbf{W_S}$ and $\mathbf{W_T}$), respectively, from which $\mathbf{W_f}$ maps the joint features into the embedded label space (Section 4.3). Lastly, we actively query and expand the labeled

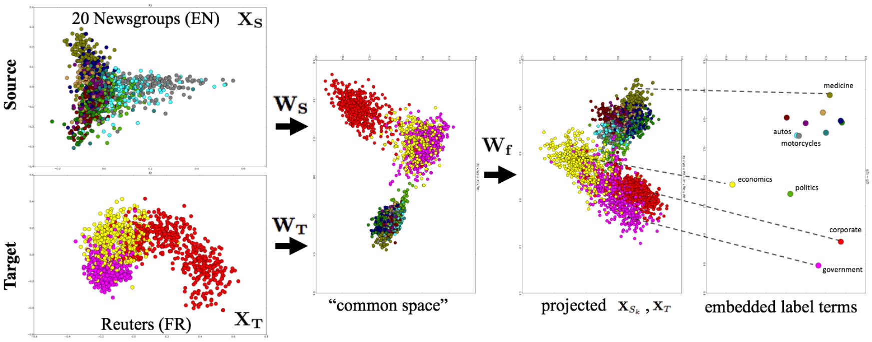


Fig. 1. An illustration of the proposed approach. Source (20 Newsgroups: English) and target (Reuters Multilingual: French) datasets lie in different feature spaces ($\mathbf{x_S} \in \mathbb{R}^{M_S}$, $\mathbf{x_T} \in \mathbb{R}^{M_T}$), and describe different categories ($\mathcal{Z}_S \neq \mathcal{Z}_T$). First, categorical labels are embedded into the dense continuous vector space (*e.g.* via text embeddings learned from unsupervised documents.) The objective is then to learn $\mathbf{W_f}$, $\mathbf{W_S}$, and $\mathbf{W_T}$ jointly such that $\mathbf{W_S}$ and $\mathbf{W_T}$ map the source and target data to the latent common feature space, from which $\mathbf{W_f}$ can project to the same space as the embedded label space. Note that the shared projection $\mathbf{W_f}$ is learned from both the source and the target datasets, thus we can more robustly predict a label for a projected instance by finding its nearest label term projection.

set L_T to jointly improve the joint classifier $\mathbf{W_f}$ and the transfer accuracy (Section 4.4). Figure 1 shows the illustration of the proposed approach, visualized with the real datasets (20 Newsgroups and Reuters Multilingual Datasets).

4.1 Language Model Label Embeddings

The skip-gram based language model [20] has proven effective in encoding semantic information of words, which can be trained from unsupervised text. We use the obtained label term embeddings as *anchors* for source and target datasets, and drive the target model to learn indirectly from source instances that belong to semantically similar categories. In this work, we use 300-D word embeddings trained from the `Google News` dataset[1] (about 100 billion words).

4.2 Unsupervised Representation Learning for Features

In order to project source and target feature spaces into the joint latent space effectively, as a pre-processing step we first obtain abstract and compact representations of raw features to allow for more tractable transformation. Unlike the similar zero-shot learning approaches [7], we do not use the embeddings obtained from a fully supervised network (e.g. the activation embeddings at the

[1] word2vec: `https://code.google.com/archive/p/word2vec/`

top of the trained visual model), because we assume the target task is scarce in labels. For our experiments with text features, we use the latent semantic analysis (LSA) method [10] to transform the raw tf-idf features into a 200-D low-rank approximation.

4.3 Transfer Learning for Heterogeneous Feature and Label Spaces

We define $\mathbf{W_S}$ and $\mathbf{W_T}$ to denote the sets of learnable parameters that project source and target features into a latent joint space, where the mappings can be learned with deep neural networks, kernel machines, etc. For simplicity, we treat $\mathbf{W_S}$ and $\mathbf{W_T}$ as linear transformation layers, thus $\mathbf{W_S} \in \mathbb{R}^{M_S \times M_C}$ and $\mathbf{W_T} \in \mathbb{R}^{M_T \times M_C}$ for projection into the M_C-dimension common space. Similarly, we define $\mathbf{W_f} \in \mathbb{R}^{M_C \times M_E}$ which maps from the common feature space into the embedded label space.

To learn these parameters simultaneously, we solve the following joint optimization problem with hinge rank losses (similar to [7]) for both source and target.

$$\min_{\mathbf{W_f}, \mathbf{W_S}, \mathbf{W_T}} \; \frac{1}{|L_S|} \sum_{i=1}^{|L_S|} l(\mathbf{S^{(i)}}) + \frac{1}{|L_T|} \sum_{j=1}^{|L_T|} l(\mathbf{T^{(j)}}) \tag{1}$$

where

$$l(\mathbf{S^{(i)}}) = \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_S^{(i)}}} \max[0, \epsilon - \mathbf{x_S^{(i)}} \mathbf{W_S} \mathbf{W_f} \mathbf{y_S^{T(i)}} + \mathbf{x_S^{(i)}} \mathbf{W_S} \mathbf{W_f} \tilde{\mathbf{y}}^{\mathbf{T}}]$$

$$l(\mathbf{T^{(j)}}) = \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_T^{(j)}}} \max[0, \epsilon - \mathbf{x_T^{(j)}} \mathbf{W_T} \mathbf{W_f} \tilde{\mathbf{y}}_{\mathbf{T}}^{\mathbf{T(j)}} + \mathbf{x_T^{(j)}} \mathbf{W_T} \mathbf{W_f} \tilde{\mathbf{y}}^{\mathbf{T}}]$$

where $l(\cdot)$ is a per-instance hinge loss, $\tilde{\mathbf{y}}$ refers to the embeddings of other label terms in the source and the target label space except the ground truth label of the instance, and ϵ is a fixed margin. We use $\epsilon = 0.1$ for all of our experiments.

In essence, we train the weight parameters to produce a higher dot product similarity between the projected source or target instance and the word embedding representation of its correct label than between the projected instance and other incorrect label term embeddings. The intuition of the model is that the learned $\mathbf{W_f}$ is a shared and more generalized linear transformation capable of mapping the joint intermediate subspace into the embedded label space.

We solve Eq.1 efficiently with stochastic gradient descent (SGD), where the gradient is estimated from a small minibatch of samples.

Once $\mathbf{W_S}$, $\mathbf{W_T}$, and $\mathbf{W_f}$ are learned, at test time we build a label-producing nearest neighbor (NN) classifier for the target task as follows:

$$\text{NN}(\mathbf{x_T}) = \operatorname*{argmax}_{\mathbf{z} \in \mathcal{Z}_T} \mathbf{x_T} \mathbf{W_T} \mathbf{W_f} \mathbf{y_z^T} \tag{2}$$

where $\mathbf{y_z}$ maps a categorical label term $\mathbf{z}$ into its word embeddings space. Similarly, we can build a NN classifier for the source task as well, using the projection $\mathbf{W_S W_f}$.

4.4 Proactive Transfer Learning

The quality of the learned parameters for the target task $\mathbf{W_T}$ and $\mathbf{W_f}$ depends on the available labeled target training samples (L_T). As such, we propose to expand L_T by querying a near-optimal subset of the unlabeled pool UL_T, which once labeled will improve the performance of the transfer accuracy and ultimately the target task, assuming the availability of unlabeled data and (limited) annotators. In particular, we relax this problem with a greedy pool-based active learning framework, where we iteratively select a small subset of unlabeled samples that maximizes the expected utility to the target model:

$$\hat{\mathbf{x_T}} = \underset{\mathbf{x_T} \in \{\mathbf{x_T^{(i)}}\}_{i \in UL_T}}{\operatorname{argmax}} \quad U(\mathbf{x_T}) \tag{3}$$

where $U(\mathbf{x_T})$ is a utility function that measures the value of a sample $\mathbf{x_T}$ defined by a choice of the query sampling objective. In traditional active learning, the uncertainty-based sampling [25, 17] and the density-weighted sampling strategies [22, 34] are often used for the utility function $U(\mathbf{x_T})$ in the target domain only. However, the previous approaches in active learning disregard the knowledge that we have in the source domain, thus being prone to query samples of which the information can be potentially redundant to the transferable knowledge. In addition, these approaches only aim at improving the target classification performance, whereas querying *bridge* instances to maximally improve the transfer accuracy instead can be more effective by allowing more information to be transferred in bulk from the source domain. Therefore, we propose the following two proactive transfer learning objectives for sampling in the target domain that utilize the source knowledge in various ways:

Maximal Marginal Distribution Overlap (MD): We hypothesize that the overlapping projected region is where the heterogeneous source and target data are semantically related, thus a good candidate for a *bridge* that maximizes the information transferable from the source data. We therefore propose to select unlabeled target samples ($\mathbf{x_T}$) in regions where the marginal distributions of projected source and target samples have the highest overlap:

$$U_{\mathrm{MD}}(\mathbf{x_T}) = \min\left(\hat{P_{\mathbf{T}}}(\mathbf{x_T}|\mathbf{W_T}, \mathbf{W_f}), \hat{P_{\mathbf{S}}}(\mathbf{x_T}|\mathbf{W_S}\mathbf{W_f})\right) \tag{4}$$

where $\hat{P_{\mathbf{T}}}$ and $\hat{P_{\mathbf{S}}}$ are the estimated marginal probability of the projected target and source instances, respectively. Specifically, we estimate each density with the non-parametric kernel method:

$$\begin{aligned}
\hat{P_{\mathbf{T}}}(\mathbf{x_T}|\mathbf{W_T}, \mathbf{W_f}) &= \frac{1}{N_T}\sum_{i=1}^{N_T} K_h(\mathbf{x_T}\mathbf{W_T}\mathbf{W_f} - \mathbf{x_T^{(i)}}\mathbf{W_T}\mathbf{W_f}) \\
\hat{P_{\mathbf{S}}}(\mathbf{x_T}|\mathbf{W_S}, \mathbf{W_f}) &= \frac{1}{N_S}\sum_{j=1}^{N_S} K_h(\mathbf{x_T}\mathbf{W_T}\mathbf{W_f} - \mathbf{x_S^{(j)}}\mathbf{W_S}\mathbf{W_f})
\end{aligned} \tag{5}$$

Algorithm 1 Proactive Transfer Learning

```
Input: source data $\mathbf{S}$, target data $\mathbf{T}$, active learning policy $U(\cdot)$, budget $B$, query
size per iteration $Q$
Randomly initialize $\mathbf{W_f}$, $\mathbf{W_T}$, $\mathbf{W_S}$
for $iter = 1$ to $B$ do
  1. Learn $\mathbf{W_f}$, $\mathbf{W_T}$, $\mathbf{W_S}$ by solving
     $$\min_{\mathbf{W_f},\mathbf{W_S},\mathbf{W_T}} \frac{1}{|L_S|}\sum_{i=1}^{|L_S|} l(\mathbf{S^{(i)}}) + \frac{1}{|L_T|}\sum_{j=1}^{|L_T|} l(\mathbf{T^{(j)}})$$
  2. Query $Q$ new samples
  for $q = 1$ to $Q$ do
    $\hat{\mathbf{i}} = \underset{\mathbf{i}\in UL_T}{\text{argmax}}\, U(\mathbf{x_T^{(i)}})$
    $UL_T := UL_T \backslash \{\hat{\mathbf{i}}\}, L_T := L_T \cup \{\hat{\mathbf{i}}\}$
  end for
end for
Output: $\mathbf{W_f}$, $\mathbf{W_T}$, $\mathbf{W_S}$
```

where K_h is a scaled Gaussian kernel with a smoothing bandwidth h. Solving $\max_{\mathbf{x_T}} \min(\hat{P}_T(\mathbf{x_T}), \hat{P}_S(\mathbf{x_T}))$ finds such instance $\mathbf{x_T}$ whose projection lies in the highest density overlap between source and target instances.

Maximum Projection Entropy (PE) aims at selecting an unlabeled target sample that has the maximum entropy of dot product similarities between a *projected* instance and its possible label embeddings:

$$U_{\text{PE}}(\mathbf{x_T}) = -\sum_{\mathbf{z}\in\mathcal{Z}_T} \log(\mathbf{x_T W_T W_f y_z^T})\mathbf{x_T W_T W_f y_z^T} \tag{6}$$

The projection entropy utilizes the information transferred from the source domain (via $\mathbf{W_f}$), thus avoiding information redundancy between source and target. After samples are queried via the maximum projection entropy method and added to the labeled target data pool, we re-train the weights such that projections of the target samples have less uncertainty in label assignment.

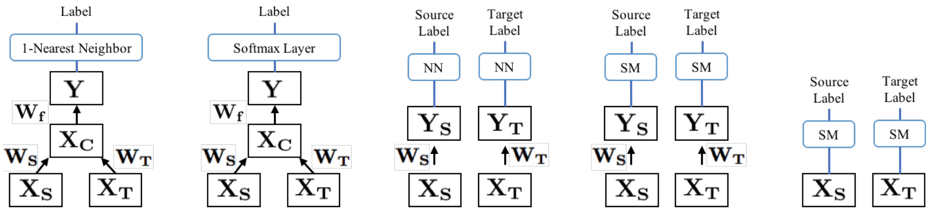
To reduce the active learning training time at each iteration, we query a small fixed number of samples ($= Q$) that have the highest utilities. Once the samples are annotated, we re-train the model with Eq.1, and select the next batch of samples to query with Eq.3. The overall process is summarized in Algorithm 1.

5 Empirical Evaluation

We evaluate the proposed approach on a hetero-lingual text classification task (Section 5.2) with the baselines described in Section 5.1.

5.1 Baselines

In our experiments we use a source dataset within heterogeneous feature and label spaces from a target dataset. Most of the previous transfer learning ap-



(a) W_Wf:**NN** (b) W_Wf:**SM** (c) W_:**NN** (d) W_:**SM** (e) -:**SM**

Fig. 2. The proposed method (a) and the baseline networks (b-e). At test time, the nearest neighbor-based models (a,c) return the nearest label in the embedding space ($\mathcal{Y}$) to the projection of a test sample, whereas the n-way softmax layer (SM) classifiers (b,d,e) are trained to produce categorical labels from their respective final projection. We use the notation **W_** to refer to **Wt** and **Ws**, as they share the same architecture.

proaches that allow only one of input or output spaces to be heterogeneous thus cannot be used as baselines (see Section 2 for the detailed comparison). We therefore compare the proposed heterogeneous transfer approach with the following baseline networks (illustrated in Figure 2):

- **W_Wf:NN** (**proposed approach**; heterogeneous transfer learning network): We learn the projections $\mathbf{W_S}$, $\mathbf{W_T}$, and $\mathbf{W_f}$ by solving the joint optimization problem in Eq.1. At test time, we use the 1-nearest neighbor classifier (NN) defined in Eq.2 and look for a category embedding that is closest to the projected source ($\mathbf{x_S W_S W_f}$) or target instance ($\mathbf{x_T W_T W_f}$) at the final layer. We use the notation **W_** to denote a placeholder for a source model (**WsWf:NN**) and a target model (**WtWf:NN**), as they share the same architecture.
- **W_Wf:SM**: We train the weights in the same way (Eq.1), and we add a softmax layer (SM) at the top projection layer (word embedding space) in replacement of the NN classifier.
- **W_:NN** ([7]; zero-shot learning networks with distributed word embeddings): We learn the projections $\mathbf{W_S} \in \mathbb{R}^{M_S \times M_E}$ and $\mathbf{W_T} \in \mathbb{R}^{M_T \times M_E}$ by solving two separate optimization problems for source and target networks respectively:

$$\min_{\mathbf{W_S}} \frac{1}{|L_S|} \sum_{i=1}^{|L_S|} l(\mathbf{S^{(i)}}), \quad \min_{\mathbf{W_T}} \frac{1}{|L_T|} \sum_{j=1}^{|L_T|} l(\mathbf{T^{(j)}}) \tag{7}$$

where the loss functions are defined in a similar way as in Eq.1:

$$\begin{aligned} l(\mathbf{S^{(i)}}) &= \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_S^{(i)}}} \max[0, \epsilon - \mathbf{x_S^{(i)} W_S y_S^{T(i)}} + \mathbf{x_S^{(i)} W_S \tilde{y}^T}] \\ l(\mathbf{T^{(j)}}) &= \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_T^{(j)}}} \max[0, \epsilon - \mathbf{x_T^{(j)} W_T y_T^{T(j)}} + \mathbf{x_T^{(j)} W_T \tilde{y}^T}] \end{aligned} \tag{8}$$

Table 1. Overview of datasets. $|\mathcal{Z}|$: the number of categories.

Dataset	$\|\mathcal{Z}\|$	Label Terms (*e.g.*)
20 Newsgroups (`20NEWS`)	20	'politics', 'religion', 'electronics', 'motorcycles', 'baseball', 'sale', $\cdots$
Reuters Multilingual (`FR`,`SP`,`GR`,`IT`)	6	'corporate', 'finance', 'economics' 'performance', 'government', 'equity'
Reuters R8 (`R8`)	8	'acquisition', 'interest', 'money' 'crude', 'trade', 'grain', $\cdots$

At test time, we use the NN classifier with projected source ($\mathbf{x_S W_S}$) and target ($\mathbf{x_T W_T}$) instances. The target task thus does not use the transferred information from the source task, but only uses the semantic word embeddings transferred from a separate unannotated corpus. This baseline can be regarded as an application of DeViSE [7] on non-image classification tasks.

- **W_:SM**: We train the weights with Eq.7, and we add a softmax layer.
- **-:SM**: We train two separate networks with logistic regression softmax layers for source and target tasks with $\mathbf{X_S}$ and $\mathbf{X_T}$, respectively.

5.2 Application: Hetero-lingual Text Classification

We apply the proposed approach to learn a target text classification task given a source text dataset with both a heterogeneous feature space (*e.g.* a different language) and a label space (*e.g.* describing different categories).

The datasets we use are summarized in Table 1. Note that the 20 Newsgroups[2] (English: 18,846 documents), the Reuters Multilingual [1] (French: 26,648, Spanish: 12,342, German: 24,039, Italian:12,342 documents), the R8 of RCV-1[3] (English: 7,674 documents) datasets describe different categories with varying degrees of relatedness. The original categories of some of the datasets were not in the format compatible to our word embeddings dictionary. We manually replaced those label terms to the semantically close words that exist in the dictionary (*e.g.* `sci.med` → 'medicine', etc.).

Task 1: Transfer Learning for Scarce Target

Setup: We assume a scenario where only a small fraction of the target samples are labeled ($\%_{L_T}$ = 0.1% or 1% depending on the size of the dataset) whereas the source dataset is fully labeled, and create various heterogeneous source-target pairs from the datasets summarized in Table 1. Table 2 reports

[2] `http://qwone.com/j̃ason/20Newsgroups/`

[3] `http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html`

Table 2. Hetero-lingual text classification test accuracy (%) on (1) the target task and (2) the source task, given a fully labeled source dataset and a partially labeled target dataset, averaged over 10-fold runs ($M_C = 320$). $\%_{L_T}$: the percentage of target samples labeled. The baselines are described in Figure 2.

Datasets			(1) Test: **Target** (%)					(2) Test: **Source** (%)				
Source	Target	$\%_{L_T}$	$\mathbf{W_T W_f}$**:NN**	$\mathbf{W_T W_f}$:SM	$\mathbf{W_T}$:NN	$\mathbf{W_T}$:SM	-:SM	$\mathbf{W_S W_f}$**:NN**	$\mathbf{W_S W_f}$:SM	$\mathbf{W_S}$:NN	$\mathbf{W_S}$:SM	-:SM
20NEWS	FR	0.1	**57.7**	46.1	55.7	44.4	39.4	**78.2**	77.1	78.0	77.3	77.6
	SP		**52.1***	43.0	46.6	42.7	43.8	77.8	77.3			
	GR		**56.2***	44.2	51.1	41.0	37.7	**78.5**	77.3			
	IT		**47.3**	39.8	46.2	35.2	31.8	77.2	77.4			
R8	FR	0.1	**56.5**	42.1	55.6	44.4	39.4	97.0	96.9	**97.2**	96.6	96.7
	SP		**50.6***	43.5	46.6	42.7	43.8	**97.2**	96.8			
	GR		**57.8***	45.1	51.1	41.0	37.7	97.0	96.8			
	IT		**49.7**	32.7	46.2	35.2	31.8	96.9	96.9			
FR	20NEWS	1	**44.7**	35.2	**44.4**	35.7	27.5	86.1	**86.0**	85.9	**86.0**	**86.0**
SP			44.2	36.0				**88.3**	88.1	88.2	88.2	88.1
GR			43.3	35.5				83.4	83.3	**83.5**	83.2	**83.5**
IT			**44.9**	34.1				**85.5**	85.3	85.3	85.1	85.1
FR	R8	0.1	61.8	52.1	**62.8**	52.3	48.1	**86.0**	**86.0**	85.9	**86.0**	**86.0**
SP			**67.3***	52.3				**88.3**	88.1	88.2	88.2	88.1
GR			**64.1**	50.9				83.3	83.1	**83.5**	83.2	**83.5**
IT			62.0	54.7				**85.4**	85.2	85.3	85.1	85.1

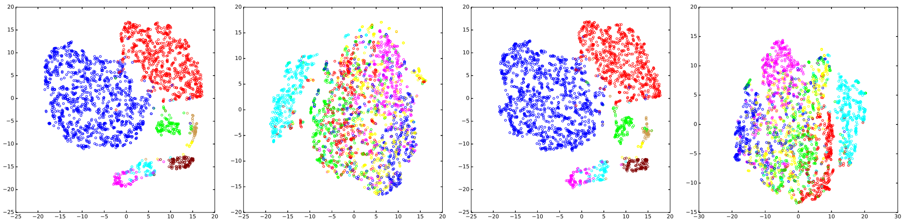
the text classification results for both source and target tasks in this experimental setting. The results are averaged over 10-fold runs, and for each fold we randomly select $\%_{L_T}$ of the target train instances to be labeled as indicated in Table 2. Bold denotes the best performing model for each test, and * denotes the statistically significant improvement ($p < 0.05$) over other methods.

Main results: Table 2 shows that the proposed approach (**WtWf:NN**) improves upon the baselines on several source-target pairs on the target classification task. Specifically, **WtWf:NN** shows statistically significant improvement over the single-modal baseline (**Wt:NN**) on the source-target pairs 20NEWS→SP, 20NEWS→GR, R8→SP, R8→GR, and SP→R8. The performance boost demonstrates that the transferred knowledge from a source dataset (in the form of $\mathbf{W_f}$) does improve the projection pathway from the target feature space to the embedded label space. Note that the transfer learning (**WtWf:NN**) from Reuters Multilingual datasets (FR, SP, GR, IT) to 20 Newsgroups (20NEWS) dataset specifically does not improve over the single-modal baseline (**Wt:NN**). The 20 Newsgroups dataset is in general harder to discriminate and spans over a larger label space than the Reuters Multilingual datasets, and thus this result indicates that the heterogeneous transfer is not as reliable if the target label space is more densely distributed than the source label space.

We observe that the nearest neighbor (**NN**) classifiers outperform the softmax (**SM**) classifiers in general. This is because the objectives in Eq.1 aim at learning a mapping such that each instance is mapped close to its respective label term embedding (in terms of dot product similarity), thus making the

Table 3. Label terms (word embeddings) cosine similarities summary for heterogeneous dataset pairs.

Datasets	Cosine Similarity		
	max	min	avg
`20NEWS` ↔ `FR,SP,GR,IT`	0.460	-0.085	0.090
`R8` ↔ `FR,SP,GR,IT`	0.342	-0.039	0.114



(a) Source, **Ws** (b) Target, **Wt** (c) Source, **WsWf** (d) Target, **WtWf**

Fig. 3. t-SNE visualization of the projected source (R8) and target (GR) instances, where (a), (b) are learned without the transferred knowledge (**W_:NN**), and (c), (d) use the transferred knowledge (**W_Wf:NN**).

nearest neighbor-finding approach a natural choice. The networks with a softmax layer perform poorly on our target classification task, possibly due to the small number of categorical training labels, making the task very challenging.

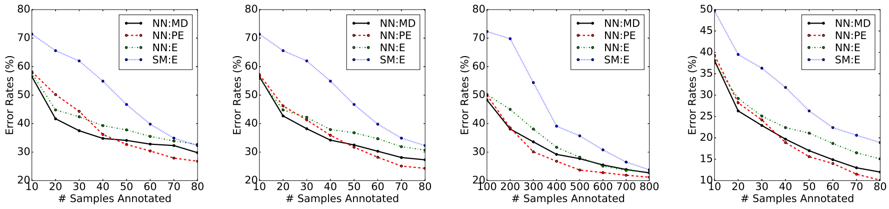
We also present the summary of cosine similarities in the embedded label space between the source and the target label terms in Table 3, which approximates the inherent *distance* between the source and the target tasks. While the R8 dataset tends to be more semantically related with the Reuters Multilingual datasets than the 20 Newsgroups dataset on average, we only observe marginal difference in their knowledge transfer performance, given the same respective source or target dataset.

Note also that both **WtWf:SM** and **Wt:SM** significantly outperform **-:SM**, a single softmax layer that does not use the auxiliary class relations information learned from word embeddings. This result demonstrates that the projection of samples into the embedded label space improves the discriminative quality of feature representation.

We observe that for a small portion of the target dataset neither helps nor hurts the source classification task, showing no statistically significant difference between the proposed approach (**WsWf:NN**) and other baselines. The learned $\mathbf{W_f}$ can thus be considered as a robust projection that maps the intermediate common subspace instances into the embedded label space which can describe both the source and the target categories.

Table 4. Comparison of performance (**WtWf:NN**) with varying intermediate embedding dimensions, averaged over 10-fold runs.

Datasets		Test Accuracy (%) vs. M_C					
S	T	20	40	80	160	320	640
20NEWS	FR	54.6	56.8	55.3	56.4	**57.7**	57.1
R8	FR	55.9	54.3	55.1	**57.0**	56.5	56.7



(a) 20NEWS → FR (b) R8 → FR (c) FR → 20NEWS (d) FR → R8

Fig. 4. Proactive transfer learning results. X-axis: the number of queried samples, Y-axis: error rate.

Feature visualization: To visualize the projection quality of the proposed approach, we plot the t-SNE embeddings [18] of the source and the target instances (R8→GR; $\%_{L_T} = 0.1$), projected with **W_Wf:NN** and **W_:NN**, respectively (Figure 3). We make the following observations: (1) The target instances are generally better discriminated with the projection learned from **WtWf:NN** which transfers knowledge from the source dataset, than the one learned from **Wt:NN**. (2) The projection quality of the source samples remains mostly the same. Both of these observations accord with the results in Table 2.

Sensitivity to the embedding dimension: Table 4 compares the performance of the proposed approach (**WtWf:NN**) with varying embedding dimensions (M_C) at the intermediate layer. We do not observe statistically significant improvement for any particular dimension, and thus we simply choose the embedding dimension that yields the highest average value on the two dataset pairs ($M_C = 320$) for all of the experiments.

Task 2: Proactive Transfer Learning

We consider a proactive transfer learning scenario, where we expand the labeled target set by querying an oracle given a fixed budget. We compare the proposed proactive transfer learning strategies (Section 4.4) against the conventional uncertainty-based sampling methods.

Setup: We choose 4 source-target dataset pairs to study: (a) 20NEWS→FR, (b) R8→FR, (c) FR→20NEWS, and (d) FR→R8. The lines **NN:MD** (maximal marginal distribution overlap; solid black) and **NN:PE** (maximum projection entropy; dashed red) refer to the proposed proactive learning strategies in Section 4.4, respectively, where the weights are learned with **WtWf:NN**. The baseline active learning strategies **NN:E** (entropy; dashdot green) and **SM:E** (entropy; dotted blue) select target samples that have the maximum class-posterior entropy given the original target input features only, which quantifies the uncertainty of samples in multiclass classification. The uncertainty-based sampling strategies are widely used in conventional active learning [25, 17], however these strategies do not utilize any information from the source domain. Once the samples are queried, **NN:E** learns the classifier **WtWf:NN**, whereas **SM:E** learns a 1-layer softmax classifier.

Main results: Figure 4 shows the target task performance improvement over iterations with various active learning strategies. We observe that both of the proposed active learning strategies (**NN:MD**, **NN:PE**) outperform the baselines on all of the source-target dataset pairs. Specifically, **NN:PE** outperforms **NN:E** on most of the cases, which demonstrates that reducing entropy in the projected space is significantly more effective than reducing class-posterior entropy given the original features. Because we re-train the joint network after each query batch, avoiding information redundancy between source and target while reducing target entropy is critical. Note that **NN:MD** outperforms **NN:PE** generally at the beginning, while the performance of **NN:PE** improves faster as it gets more samples annotated. This result indicates that selecting samples with the maximal source and target density overlap (**MD**) helps in building a *bridge* for transfer of knowledge initially, while this information may eventually get redundant, thus the decreased efficacy. Note also that the all of the projection-based methods (**NN:MD**, **NN:PE**, **NN:E**) significantly outperform **SM:E**, which measures the entropy and learns the classifier at the original feature space. This result demonstrates that the learned projections $\mathbf{W_T W_f}$ effectively encode input target features, from which we can build a robust classifier efficiently even with a small number of labeled instances.

6 Conclusions

We summarize our contributions as follows: We address a unique challenge of mining and leveraging transferable knowledge in the heterogenous case, where labeled source data differs from target data in both feature and label spaces. To this end, (1) we propose a novel framework for heterogeneous transfer learning to discover the latent subspace to map the source into the target space, from which it simultaneously learns a shared final projection to the embedded label space. (2) In addition, we propose a proactive transfer learning framework which expands the labeled target data with the objective of actively improving transfer accuracy and thus enhancing the target task performance. (3) An extensive empirical evaluation on the hetero-lingual text classification task demonstrates the efficacy of each part of the proposed approach.

Future Work: While the empirical evaluation was conducted on the text domain, our formulation does not restrict the input domain to be textual. We thus believe the approach can be applied broadly, and as future work, we plan to investigate the transferability of knowledge with diverse heterogeneous settings, such as image-aided text classification tasks, etc., given suitable source and target data. In addition, extending the proposed approach for learning selectively from *multiple* heterogeneous source datasets also remains as a challenge.

References

1. Amini, M., Usunier, N., Goutte, C.: Learning from multiple partially observed views-an application to multilingual text categorization. In: NIPS. pp. 28–36 (2009)
2. Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., Ye, J.: Joint transfer and batch-mode active learning. In: ICML. pp. 253–261 (2013)
3. Dai, W., Chen, Y., Xue, G.R., Yang, Q., Yu, Y.: Translated learning: Transfer learning across different feature spaces. In: NIPS. pp. 353–360 (2008)
4. Dhillon, P., Foster, D.P., Ungar, L.H.: Multi-view learning of word embeddings via cca. In: NIPS. pp. 199–207 (2011)
5. Duan, L., Xu, D., Tsang, I.: Learning with augmented features for heterogeneous domain adaptation. ICML (2012)
6. Fang, M., Yin, J., Tao, D.: Active learning for crowdsourcing using knowledge transfer. In: AAAI (2014)
7. Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NIPS (2013)
8. Fu, Z., Xiang, T., Kodirov, E., Gong, S.: Zero-shot object recognition by semantic manifold distance. In: CVPR. pp. 2635–2644 (2015)
9. Gavves, E., Mensink, T.E.J., Tommasi, T., Snoek, C.G.M., Tuytelaars, T.: Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In: ICCV (2015)
10. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review 53(2), 217–288 (2011)
11. Huang, J., Gretton, A., Borgwardt, K.M., Schölkopf, B., Smola, A.J.: Correcting sample selection bias by unlabeled data. In: NIPS (2007)
12. Kale, D., Liu, Y.: Accelerating active learning with transfer learning. In: ICDM. pp. 1085–1090 (2013)
13. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Unsupervised domain adaptation for zero-shot learning. In: ICCV (2015)
14. Kshirsagar, M., Carbonell, J., Klein-Seetharaman, J.: Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks (2013)
15. Li, X., Guo, Y., Schuurmans, D.: Semi-supervised zero-shot classification with label representation learning. In: ICCV. pp. 4211–4219 (2015)
16. Long, M., Wang, J.: Learning transferable features with deep adaptation networks. ICML (2015)
17. Loy, C., Hospedales, T., Xiang, T., Gong, S.: Stream-based joint exploration-exploitation active learning. In: CVPR. pp. 1560–1567 (June 2012)
18. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(2579-2605), 85 (2008)
19. Mensink, T., Gavves, E., Snoek, C.G.: Costa: Co-occurrence statistics for zero-shot classification. In: CVPR. pp. 2441–2448 (2014)

20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ICLR (2013)
21. Moon, S., Carbonell, J.: Proactive learning with multiple class-sensitive labelers. International Conference on Data Science and Advanced Analytics (DSAA) (2014)
22. Nguyen, H., Smeulders, A.: Active learning using pre-clustering. International Conference on Machine Learning (ICML) (2004)
23. Pan, S.J., Yang, Q.: A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions on 22(10), 1345–1359 (2010)
24. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: NIPS. pp. 46–54 (2013)
25. Settles, B., Craven, M.: Training text classifiers by uncertainty sampling. EMNLP pp. 1069 – 1078 (2008)
26. Shi, X., Fan, W., Ren, J.: Actively transfer domain knowledge. In: KDD, pp. 342–357. Springer (2008)
27. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.Y.: Zero Shot Learning Through Cross-Modal Transfer. In: NIPS (2013)
28. Sun, Q., Amin, M., Yan, B., Martell, C., Markman, V., Bhasin, A., Ye, J.: Transfer learning for bilingual content classification. In: KDD. pp. 2147–2156 (2015)
29. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. ICML (2015)
30. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: IJCAI'11 (2011)
31. Xiao, M., Guo, Y.: Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation. In: ECMLPKDD, pp. 525–540. Springer (2015)
32. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. In: ICCV (2015)
33. Zhou, J.T., Pan, S.J., Tsang, I.W., Yan, Y.: Hybrid heterogeneous transfer learning through deep learning. AAAI (2014)
34. Zhu, J., Wang, H., Tsou, B., Ma, M.: Active Learning With Sampling by Uncertainty and Density for Data Annotations. IEEE Transactions on Audio, Speech, and Language Processing 18 (2010)
35. Zhu, Z., Zhu, X., Ye, Y., Guo, Y.F., Xue, X.: Transfer active learning. In: CIKM. pp. 2169–2172 (2011)