Completely Heterogeneous Transfer Learning with Attention - What And What Not To Transfer

Seungwhan Moon, Jaime Carbonell
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
[seungwhm | jgc]@cs.cmu.edu

Abstract

We study a transfer learning framework where source and target datasets are heterogeneous in both feature and label spaces. Specifically, we do not assume explicit relations between source and target tasks *a priori*, and thus it is crucial to determine *what and what not to transfer* from source knowledge. Towards this goal, we define a new heterogeneous transfer learning approach that (1) selects and attends to an optimized subset of source samples to transfer knowledge from, and (2) builds a unified transfer network that learns from both source and target knowledge. This method, termed "Attentional Heterogeneous Transfer", along with a newly proposed unsupervised transfer loss, improve upon the previous state-of-the-art approaches on extensive simulations as well as a challenging hetero-lingual text classification task.

1 Introduction

Humans learn from heterogeneous knowledge sources and modalities, and given a novel task humans are able to make inferences by leveraging the combined knowledge base. Inspired by this observation, recent work [Moon and Carbonell, 2016] investigates a completely heterogeneous transfer learning (CHTL) scenario, where source and target tasks are heterogeneous in both feature and label spaces (*e.g.* document classification tasks in different languages and with different categories). In their work, CHTL is formulated as a subspace learning problem in which heterogeneous source and target knowledge are combined in a common latent space by the learned projection. To ground heterogeneous source and target label terms into a common distributed label space, they use word embeddings obtained from a language model.

However, most of the previous approaches on transfer learning do not take into account different instance-level heterogeneity within a source dataset, often leading to undesirable *negative transfer*. Specifically, CHTL can suffer from brute-force merge of heterogeneous sources because it does not assume explicit relations between source and target knowledge in both instance and dataset-level.

To this end, we propose a new transfer method called "Attentional Heterogeneous Transfer", with the aim of determining *what to transfer and what not to transfer* from heterogeneous source knowledge. The proposed joint optimization problem learns the parameters for transfer network as well as an optimized subset of source dataset, ignoring unnecessary or confounding source instances that exhibit a negative impact in learning the target task.

In addition, we propose a new joint unsupervised optimization for heterogeneous transfer network which leverages both unlabeled source and target data, leading to enhanced discriminative power in both tasks. Unsupervised training also allows for more tractable learning of deep transfer networks, whereas the previous literature was confined to linear transfer models due to a small number of labeled target data.

Note that CHTL tackles a broader range of problems than prior transfer learning approaches in that they often require parallel datasets with source-target correspondent instances (*e.g.* Hybrid Heterogeneous Transfer Learning (HHTL) [Zhou *et al.*, 2014] or CCA-based methods for a multi-view learning problem [Wang *et al.*, 2015]), and that they require either homogeneous feature spaces [Kodirov *et al.*, 2015; Long and Wang, 2015] or label spaces [Dai *et al.*, 2008; Duan *et al.*, 2012; Sun *et al.*, 2015]. We provide a comprehensive list of related work in the later section.

Our contributions are three-fold: we propose (1) a novel transfer learning algorithm that attends selectively to a subset of samples from a heterogeneous source to allow for a more tractable and accurate knowledge transfer, and (2) an unsupervised transfer with denoising auto-encoder loss unique to the heterogeneous transfer network, allowing for training deeper layers. (3) We show the efficacy of the proposed approaches on extensive simulation studies as well as a novel real-world transfer learning task.

2 Background: Completely Heterogeneous Transfer Learning (CHTL)

We begin by describing the completely heterogeneous transfer learning (CHTL) setting, where the target multiclass classification task is learned from both a target dataset and a source dataset with heterogeneous feature and label spaces. Figure 1 illustrates the overall pipeline.

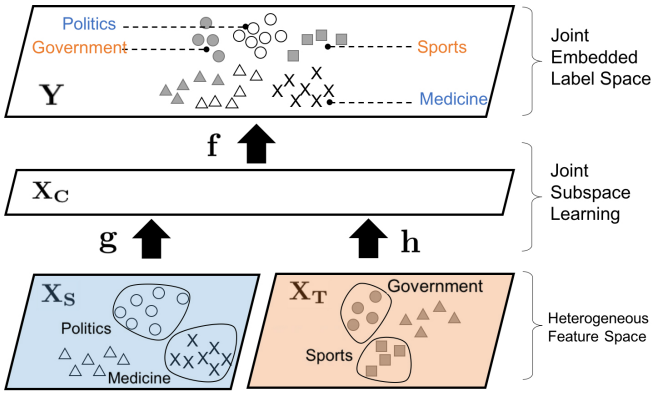


Figure 1: Completely Heterogeneous Transfer Learning (CHTL). Source and target lie in heterogeneous feature spaces ($\mathbf{x_S} \in \mathbb{R}^{M_S}$, $\mathbf{x_T} \in \mathbb{R}^{M_T}$), and describe heterogeneous labels ($\mathcal{Z}_S \neq \mathcal{Z}_T$). Heterogeneous source and target labels are first embedded into the joint label space via *e.g.* word embeddings from language models. CHTL learns projections $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$ simultaneously such that the shared projection $\mathbf{f}$ is trained with both source and target, thus leveraging knowledge from source in prediction of target tasks.

2.1 Notations

Let the target task $\mathbf{T} = \{\mathbf{X_T}, \mathbf{Y_T}, \mathbf{Z_T}\}$ be defined with the target samples $\mathbf{X_T} = \{\mathbf{x_T^{(i)}}\}_{i=1}^{N_T}$ 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 corresponding ground-truth labels $\mathbf{Z_T} = \{\mathbf{z_T^{(i)}}\}_{i=1}^{N_T}$, where $\mathbf{z_T} \in \mathcal{Z}_T$ for the categorical target label space $\mathcal{Z}_T$. and the parallel high-dimensional label representation $\mathbf{Y_T} = \{\mathbf{y_T^{(i)}}\}_{i=1}^{N_T}$ for $\mathbf{y_T} \in \mathbb{R}^{M_E}$, where M_E is the dimension of the embedded labels. Let L_T and UL_T be a set of indices of labeled and unlabeled target instances, respectively, for $|L_T| + |UL_T| = N_T$. Only a few labels are available for a novel target task, thus $|L_T| \ll N_T$. Similarly, define the 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 for $|L_S| = N_S$ (fully labeled source dataset), accordingly. The CHTL settings allow for $M_T \neq M_S$ (heterogeneous feature space) and $\mathcal{Z}_T \neq \mathcal{Z}_S$ (heterogeneous label space). CHTL aims at building a robust classifier for the target task ($\mathcal{X}_T \rightarrow \mathcal{Z}_T$), trained with $\{\mathbf{x_T^{(i)}}, \mathbf{y_T^{(i)}}, \mathbf{z_T^{(i)}}\}_{i \in L_T}$ as well as transferred knowledge from $\{\mathbf{x_S^{(i)}}, \mathbf{y_S^{(i)}}, \mathbf{z_S^{(i)}}\}_{i \in L_S}$.

2.2 Distributed Representation for Label Embeddings

In order to relax heterogeneity between source and target label spaces, it is important to obtain a common distributed label space where all of the source and target class categories can be mapped into.

In cases where source and target class categories are represented with label terms ("names"), we can effectively encode semantic information of words in distributed representations using (1) the skip-gram based language model [Mikolov *et al.*, 2013] trained from unsupervised text, or (2) the entity embeddings induced from a knowledge graph [Bordes *et al.*, 2013; Wang *et al.*, 2014; Nickel *et al.*, 2015] with WordNet [Miller, 1995]. The obtained label term embeddings $\mathbf{Y_S}$ and $\mathbf{Y_T}$ can be used as *anchors* for source and target, allowing for the target model to transfer knowledge from source instances with semantically similar categories.

2.3 Transfer Network

CHTL [Moon and Carbonell, 2016] builds a transfer network with three main transformation layers: $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$. $\mathbf{g} : \mathbb{R}^{M_S} \rightarrow \mathbb{R}^{M_C}$ and $\mathbf{h} : \mathbb{R}^{M_T} \rightarrow \mathbb{R}^{M_C}$ first project M_S-dimensional source features and M_T-dimensional target features into a M_C-dimensional joint latent space via linear transformation, respectively. Once source and target samples are projected onto the common latent space, the transfer network maps the projected source and target samples via a shared transformation $\mathbf{f} : \mathbb{R}^{M_C} \rightarrow \mathbb{R}^{M_E}$ onto the embedded label space. $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$ are learned simultaneously by solving the joint optimization objective with hinge rank losses for both source and target. While [Moon and Carbonell, 2016] only considers linear transformation layers, we provide a more generalized objective form where $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$ denote mappings implemented with DNNs.

$$\min_{\mathbf{W_f}, \mathbf{W_g}, \mathbf{W_h}} \mathcal{L}_{\text{HR}}(\mathbf{S}; \mathbf{W_g}, \mathbf{W_f}) + \mathcal{L}_{\text{HR}}(\mathbf{T}; \mathbf{W_h}, \mathbf{W_f}) + \mathcal{R}(\mathbf{W})$$

where

$$\mathcal{L}_{\text{HR}}(\mathbf{S}) = \frac{1}{|L_S|} \sum_{i=1}^{|L_S|} \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_S^{(i)}}} \max[0, \epsilon - \mathbf{f}(\mathbf{g}(\mathbf{x_S^{(i)}})) \cdot (\mathbf{y_S^{(i)}} - \tilde{\mathbf{y}})^\top]$$

$$\mathcal{L}_{\text{HR}}(\mathbf{T}) = \frac{1}{|L_T|} \sum_{j=1}^{|L_T|} \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_T^{(j)}}} \max[0, \epsilon - \mathbf{f}(\mathbf{h}(\mathbf{x_T^{(i)}})) \cdot (\mathbf{y_T^{(j)}} - \tilde{\mathbf{y}})^\top]$$

$$\mathcal{R}(\mathbf{W}) = \lambda_{\mathbf{f}} \|\mathbf{W_f}\|^2 + \lambda_{\mathbf{g}} \|\mathbf{W_g}\|^2 + \lambda_{\mathbf{h}} \|\mathbf{W_h}\|^2 \quad (1)$$

where $\mathcal{L}_{\text{HR}}(\cdot)$ is the hinge rank loss for source and target, $\mathbf{W} = \{\mathbf{W_f}, \mathbf{W_g}, \mathbf{W_h}\}$ are the learnable parameters for $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{h}$ respectively, $\tilde{\mathbf{y}}$ refers to the embeddings of other label terms in the source and the target label space except the ground truth label of the instance, ϵ is a fixed margin which we set as 0.1, $\mathcal{R}(\mathbf{W})$ is a weight decay regularization term, and $\lambda_{\mathbf{f}}, \lambda_{\mathbf{g}}, \lambda_{\mathbf{h}} \geq 0$ are regularization constants.

Intuitively, the weight parameters are trained 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. Note that $\mathbf{f}$ is trained and shared by both source and target samples, thus capable of leveraging knowledge learned from a source dataset for a target task. At test time, the following label-producing nearest neighbor (1-NN) classifier is used for the target task:

$$\text{1-NN}(\mathbf{x_T}) = \underset{\mathbf{z} \in \mathcal{Z}_T}{\text{argmax}} \, \mathbf{f}(\mathbf{h}(\mathbf{x_T})) \cdot \mathbf{y_z}^\top \quad (2)$$

where $\mathbf{y_z}$ maps a categorical label term $\mathbf{z}$ into its word embeddings space. A 1-NN classifier for the source task can be defined similarly, using the projection $\mathbf{f}(\mathbf{g}(\cdot))$ instead of $\mathbf{f}(\mathbf{h}(\cdot))$.

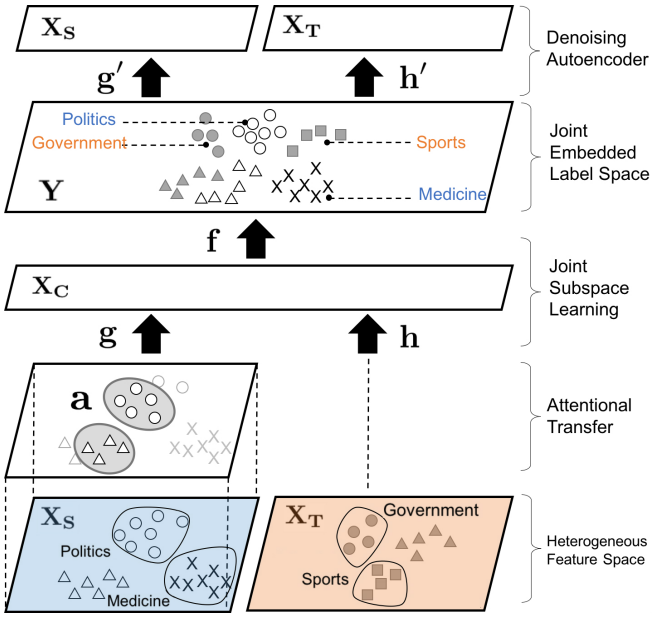


Figure 2: An illustration of CHTL with the proposed approach. The attention mechanism $\mathbf{a}$ filters and suppresses irrelevant source samples, and the denoising auto-encoders $\mathbf{g}'$ and $\mathbf{h}'$ improve robustness with unsupervised training.

3 Proposed Approaches

Figure 2 illustrates the proposed approaches.

3.1 Attentional Transfer - What And What Not To Transfer

While CHTL does not assume any explicit relations between source and target tasks, we speculate that there are certain instances within the source task that are more likely to be *transferable* than other samples. Inspired by successes of attention mechanism from recent literature [Xu *et al.*, 2015; Chan *et al.*, 2015], we propose an approach that selectively transfers useful knowledge by focusing only on a subset of source knowledge while avoiding others that may have a harmful impact on target learning. Specifically, the attention mechanism learns a set of parameters that specify a weight vector over a discrete subset of data, determining its relative importance or relevance in transfer. To enhance computational tractability we first pre-cluster the source dataset into K number of clusters $\mathbf{S_1}, \cdots, \mathbf{S_K}$, and formulate the following joint optimization problem that learns the parameters for the transfer network as well as a weight vector $\{\alpha_k\}_{k=1..K}$:

$$\min_{\mathbf{a},\mathbf{W_f},\mathbf{W_g},\mathbf{W_h}} \mu \sum_{k=1}^{K} \frac{\alpha_k}{|L_{S_k}|} \cdot \mathcal{L}_{\text{HR:K}}(\mathbf{S_k}) + \mathcal{L}_{\text{HR}}(\mathbf{T}) + \mathcal{R}(\mathbf{W})$$

where

$$\alpha_k = \frac{\exp(\mathbf{a}_k)}{\sum_{k=1}^{K} \exp(\mathbf{a}_k)}, \quad 0 < \alpha_k < 1$$

$$\mathcal{L}_{\text{HR:K}}(\mathbf{S_k}) = \sum_{i \in L_{S_k}} \sum_{\tilde{\mathbf{y}} \neq \mathbf{y_S^{(i)}}} \max[0, \epsilon - \mathbf{f}(\mathbf{g}(\mathbf{x_S^{(i)}})) \cdot (\mathbf{y_S^{(i)}} - \tilde{\mathbf{y}})^\top] \tag{3}$$

where $\mathbf{a}$ is a learnable parameter that determines the weight for each cluster, $\mathcal{L}_{\text{HR:K}}(\mathbf{S_k})$ is a cluster-level hinge loss for source, L_{S_k} is a set of source indices that belong to a cluster $\mathbf{S_k}$, and μ is a hyperparameter that penalizes $\mathbf{a}$ and $\mathbf{f}$ for simply optimizing for the source task only. Note that $\mathbf{f}$ is shared by both source and target networks, and thus the choice of $\mathbf{a}$ affects both $\mathbf{g}$ and $\mathbf{h}$. Essentially, the attention mechanism works as a regularization over source, suppressing the loss values for non-attended samples in knowledge transfer. In our experiments we use K-means clustering algorithm.

Optimization: We solve Eq.3 with a two-step alternating descent optimization. The first step involves optimizing for the source network parameters $\mathbf{W_g}, \mathbf{a}, \mathbf{W_f}$ while the rest are fixed, and the second step optimizes for the target network parameters $\mathbf{W_h}, \mathbf{W_f}$ while others are fixed.

3.2 Unsupervised Transfer Learning with Denoising Auto-encoder

We formulate unsupervised transfer learning with the CHTL architecture for added robustness, which is especially beneficial when labeled target data is scarce. Specifically, we add denoising auto-encoders where the pathway for predictions, $\mathbf{f}$, is shared and trained by both source and target through the joint subspace, thus benefiting from unlabelled source and target data. Finally, we formulate the CHTL learning problem with both supervised and unsupervised losses as follows:

$$\min_{\mathbf{a},\mathbf{W}} \mu \sum_{k=1}^{K} \frac{\alpha_k}{|L_{S_k}|} \cdot \mathcal{L}_{\text{HR:K}}(\mathbf{S_k}) + \mathcal{L}_{\text{HR}}(\mathbf{T}) + \mathcal{L}_{\text{AE}}(\mathbf{S}, \mathbf{T}; \mathbf{W})$$

where

$$\mathcal{L}_{\text{AE}}(\mathbf{S}, \mathbf{T}; \mathbf{W}) = \frac{1}{|UL_S|} \sum_{i=1}^{|UL_S|} \|\mathbf{g}'(\mathbf{f}(\mathbf{g}(\mathbf{x_S^{(i)}}))) - \mathbf{x_S^{(i)}}\|^2 + \frac{1}{|UL_T|} \sum_{j=1}^{|UL_T|} \|\mathbf{h}'(\mathbf{f}(\mathbf{h}(\mathbf{x_T^{(j)}}))) - \mathbf{x_T^{(j)}}\|^2 \tag{4}$$

where $\mathcal{L}_{\text{AE}}$ is the denoising auto-encoder loss for both source and target data (unlabelled), $\mathbf{g}'$ and $\mathbf{h}'$ reconstruct input source and target respectively, and the learnable weight parameters are defined as $\mathbf{W} = \{\mathbf{W_f}, \mathbf{W_g}, \mathbf{W_h}, \mathbf{W'_g}, \mathbf{W'_h}\}$.

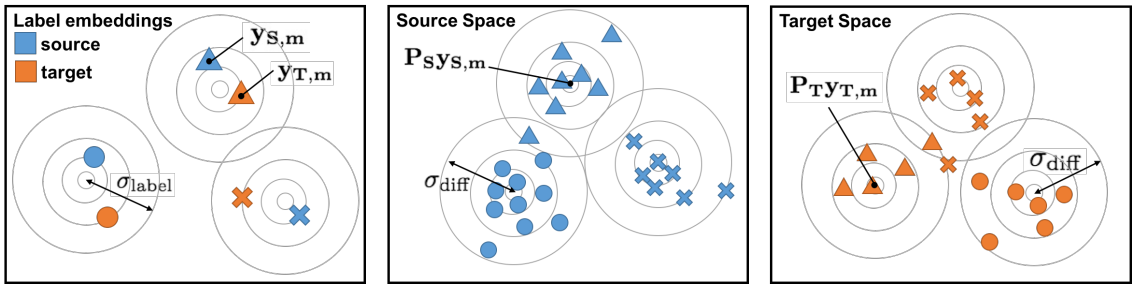
4 Empirical Evaluation

We validate the effectiveness of the proposed approaches via extensive simulations as well as a real-world application.

4.1 Baselines

Note that very few previous studies have addressed the transfer learning settings where both feature and label spaces are heterogeneous. The following baselines are considered.

- `CHTL:ATT+AE` (**proposed approach**; completely heterogeneous transfer learning (CHTL) network with attention and auto-encoder loss): the model is trained with the joint optimization problem in Eq.4.
- `CHTL:ATT` (CHTL with attention only): the model is trained with Eq.3. We evaluate this baseline to isolate the effectiveness of the attention mechanism.



(a) (b) (c)

Figure 3: Dataset generation process. (a) Draw a pair of source and target label embeddings $(\mathbf{y_{S,m}}, \mathbf{y_{T,m}})$ from each of M Gaussian distributions, all with $\sigma = \sigma_{\text{label}}$ (source-target label heterogeneity). For a random projection $\mathbf{P_S}, \mathbf{P_T}$, (b) Draw synthetic source samples from new Gaussian distributions with $\mathcal{N}(\mathbf{P_S y_{S,m}}, \sigma_{\text{diff}})$, $\forall m \in \{1, \cdots, M\}$. (c) Draw synthetic target samples from $\mathcal{N}(\mathbf{P_T y_{T,m}}, \sigma_{\text{diff}})$, $\forall m$. The resulting source and target datasets have heterogeneous label spaces (each class randomly drawn from a Gaussian with σ_{label}), as well as heterogeneous feature spaces ($\mathbf{P_S} \neq \mathbf{P_T}$).

- `CHTL` (CHTL without attention or auto-encoder; [Moon and Carbonell, 2016]): the model is trained with Eq.1.
- `ZSL` (Zero-shot learning networks with word embeddings; [Frome *et al.*, 2013]): the model is trained for target dataset only with label embeddings $\mathbf{Y_T}$ obtained from a language model. The model thus leverages knowledge from unsupervised text corpus, and is reported to be robust for low-resourced classification tasks. We solve the following optimization problem:

$$\min_{\mathbf{W_T}} \frac{1}{|L_T|} \sum_{j=1}^{|L_T|} l(\mathbf{T^{(j)}}) \tag{5}$$

where the loss function is defined as follows:

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

- `ZSL:AE` (ZSL with autoencoder loss): we add the autoencoder loss to the objective to Eq.5.
- `MLP` (A feedforward multi-layer perceptron): the model is trained for a target dataset only with categorical labels.

For each of the CHTL variations, we vary the number of fully connected (FC) layers (*e.g.* `1fc`,`2fc`,$\cdots$) as well as the label embedding methods as described in Section 2.2 (word embeddings (`W2V`), knowledge graph-induced embeddings (`G2V`), and random embeddings (`RAND`) as a reference).

4.2 Synthetic Datasets

We generate multiple pairs of source and target synthetic datasets and evaluate the performance with average classification accuracies on target tasks. Specifically, we aim to analyze the performance of the proposed approaches with varying source-target heterogeneity at varying task difficulty.

Datasets generation process is described in Figure 3. We generate synthetic source and target datasets each with M different classes, $\mathbf{S} = \{\mathbf{X_S}, \mathbf{Y_S}\}$, and $\mathbf{T} = \{\mathbf{X_T}, \mathbf{Y_T}\}$, such that their embedded label space are heterogeneous with a controllable hyperparameter σ_{label}. We first generate M isotropic Gaussian distributions $\mathcal{N}(\mu_m, \sigma_{\text{label}})$ for $m \in \{1, \cdots, M\}$. From each distribution we draw a pair of source and target label embeddings $\mathbf{y_{S,m}}, \mathbf{y_{T,m}} \in \mathbb{R}^{M_E}$. Intuitively, source and target datasets are more heterogeneous with a higher σ_{label}, as the drawn pair of source and target embeddings is farther apart from each other. We then generate source and target samples each with a random projection $\mathbf{P_S} \in \mathbb{R}^{M_S \times M_E}, \mathbf{P_T} \in \mathbb{R}^{M_T \times M_E}$ as follows:

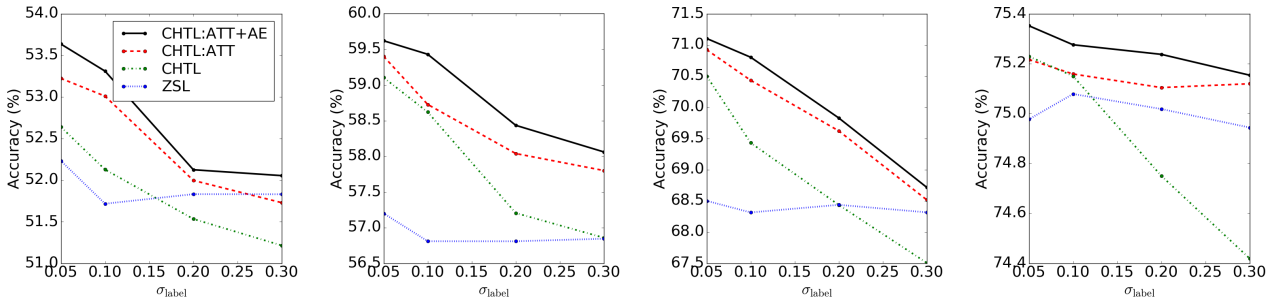
$$\mathbf{X_{S,m}} \sim \mathcal{N}(\mathbf{P_S y_{S,m}}, \sigma_{\text{diff}}),\ \mathbf{X_S} = \{\mathbf{X_{S,m}}\}_{1 \leq m \leq M}$$
$$\mathbf{X_{T,m}} \sim \mathcal{N}(\mathbf{P_T y_{T,m}}, \sigma_{\text{diff}}),\ \mathbf{X_T} = \{\mathbf{X_{T,m}}\}_{1 \leq m \leq M}$$

where σ_{diff} affects the label distribution classification difficulty. We denote $\%_{L_T}$ as the percentage of target samples labeled, and assume that only a small fraction of target samples is labeled ($\%_{L_T} \ll 1$).

For the following experiments, we set $N_S = N_T = 4000$ (number of samples), $M = 4$ (number of source and target dataset classes), $M_S = M_T = 20$ (original feature dimension), $M_E = 15$ (embedded label space dimension), $K = 12$ (number of attention clusters), $\sigma_{\text{diff}} = 0.5$, $\sigma_{\text{label}} \in \{0.05, 0.1, 0.2, 0.3\}$, and $\%_{L_T} \in \{0.005, 0.01, 0.02, 0.05\}$. We repeat the dataset generation process 10 times for each parameter set. We obtain 5-fold results for each dataset generation, and report the overall average accuracy in Figure 4.

Sensitivity to source-target heterogeneity: each subfigure in Figure 4 shows the performance of the baselines with varying σ_{label} (source-target heterogeneity). In general, `CHTL` baselines outperforms `ZSL`, but the performance degrades as heterogeneity increases. However, the attention mechanism (`CHTL:ATT`) is generally effective with higher source-target heterogeneity, suppressing the performance drop. Note that the performance improves in most cases when the attention mechanism is combined with the auto-encoder loss (`+AE`).

Sensitivity to target label Scarcity: we evaluate the tolerance of the algorithm at varying target task difficulty, measured with varying percentage of target labels given. When a small number of labels are given (Figure 4(a)), the improvement due to `CHTL` algorithms is weak, indicating that `CHTL` requires a sufficient number of target labels to build proper



(a) $\%_{L_T}$= 0.5% (b) $\%_{L_T}$= 1% (c) $\%_{L_T}$= 2% (d) $\%_{L_T}$= 5%

Figure 4: Simulation results with varying source-target heterogeneity (X-axis: σ_{label}, Y-axis: accuracy) at different $\%_L$. Baselines: CHTL:ATT+LD (black solid; proposed approach), CHTL:ATT (red dashes), CHTL (green dash-dots), ZSL (blue dots).

anchors with source knowledge. Note also that while the performance gain of CHTL algorithms begins to degrade as the target task approaches the saturation error rate (Figure 4(d)), the attention mechanism (CHTL:ATT) is more robust to this degradation and avoids negative transfer.

4.3 Hetero-lingual Text Classification

We apply the proposed methods on a hetero-lingual text classification task, where the objective is to learn a target task given a source data with heterogeneous feature space (different language) and heterogeneous labels (different categories).

Datasets: we use the RCV-1 dataset (English: 804,414 document; 116 classes) [Lewis *et al.*, 2004], the 20 Newsgroups[1] (English: 18,846 documents; 20 classes), the Reuters Multilingual [Amini *et al.*, 2009] (French (FR): 26,648, Spanish (SP): 12,342, German (GR): 24,039, Italian (IT): 12,342 documents; 6 classes), and the R8 [2] (English: 7,674 documents; 8 classes) datasets.

Main results (Table 1): all of the CHTL variations outperform the ZSL and MLP baselines, which indicates that knowledge from heterogeneous source domain does benefit target task. In addition, the proposed approach (CHTL:2fc+ATT+AE) outperforms other baselines in most of the cases, showing that the attention mechanism ($K = 40$) as well as the denoising autoencoder loss improve the transfer performance ($M_C = 320$, $M_E = 300$, label: word embeddings). While having two fully connected layers (CHTL:2fc) does not necessarily help CHTL performance by itself due to a small number of labels available for target data, it ultimately performs better when combined with the auto-encoder loss (CHTL:2fc+ATT+AE). Note that while both ZSL and MLP do not utilize source knowledge, ZSL with word embeddings shows a huge improvement over MLP, showing that ZSL is robust to low-resourced classification tasks. ZSL benefits from autoencoder loss as well, but the improvement is not as significant as in CHTL. Most of the results parallel the simulation results with the synthetic datasets, auguring well for the generality of our proposed approach.

[1] http://qwone.com/~jason/20Newsgroups/

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

Table 1: **Hetero-lingual text classification** test accuracy (%) on the target task, given a fully labeled source dataset and a partially labeled target dataset ($\%_{L_T} = 0.1$), averaged over 10-fold runs. Label embeddings with W2V.

Datasets		Target Task Accuracy (%)							
S	T	MLP	ZSL	(:AE)	CHTL	(:ATT	+AE)	(:2fc	**+ATT+AE)**
RCV1	FR	39.4	55.7	56.5	57.5	58.9	58.9	58.7	**59.0**
	SP	43.8	46.6	50.7	52.3	53.4	53.5	52.8	**54.2**
	GR	37.7	51.1	52.0	56.4	57.3	58.0	57.3	**58.4**
	IT	31.8	46.2	46.9	49.1	50.6	**51.2**	49.5	51.0
20 NEWS	FR	39.4	55.7	56.5	57.7	58.2	58.4	57.0	**58.6**
	SP	43.8	46.6	50.7	52.1	52.8	52.3	52.3	**53.1**
	GR	37.7	51.1	52.0	56.2	56.9	**57.5**	55.9	57.0
	IT	31.8	46.2	46.9	47.3	48.0	**48.1**	47.3	47.7
R8	FR	39.4	55.7	56.5	56.5	56.4	57.2	55.9	**57.7**
	SP	43.8	46.6	50.7	50.6	51.3	**51.8**	50.8	51.2
	GR	37.7	51.1	52.0	57.8	56.5	56.4	57.0	**58.0**
	IT	31.8	46.2	46.9	49.7	50.4	**50.5**	49.4	**50.5**
FR	R8	48.1	62.8	**63.5**	61.8	62.6	62.8	61.5	62.3
SP					67.3	66.7	67.1	67.4	**67.7**
GR					64.1	65.1	**65.5**	64.4	65.3
IT					62.0	63.4	**64.1**	61.6	63.0

Table 2: **CHTL with attention** test accuracy (%) on the target task, at varying K (number of clusters for attention), averaged over 10-fold runs. $\%_{L_T} = 0.1$, Method: CHTL:ATT.

Datasets		Accuracy (%)			
S	T	$K = 10$	$K = 20$	$K = 40$	$K = 80$
RCV1	FR	57.9	58.1	58.9	58.5
20NEWS	FR	57.7	58.0	58.2	58.3
R8	FR	57.0	57.3	56.4	56.6

Sensitivity to attention size K (Table 2): intuitively, $K \approx N_S$ leads to a potentially intractable training while $K \approx 1$ limits the ability to attend to subsets of source dataset, and thus an optimal value of K may exist. We set $K = 40$ for all experiments, which yields the highest average accuracy.

Visualization of attention: Figure 5 illustrates the effectiveness of the attention mechanism with an exemplary transfer learning task (source: R8, target: GR, method: CHTL:ATT, $K = 40$, $\%_{L_T} = 0.1$). The source instances that overlap with some of the target instances in the label space (near source label terms 'interest' and 'trade' and target label term 'finance') are given the most attention, which thus serve

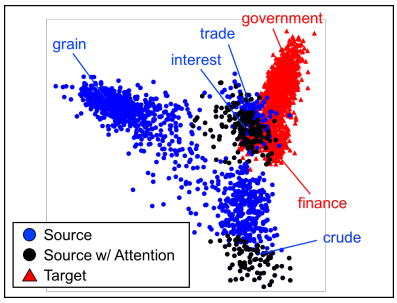


Figure 5: **Visualization of attention** (source: R8, target: GR). Shown in the figure is the 2-D PCA representation of source instances (blue circles), source instances with attention: top 5 source clusters with the highest weights (black circles), and target instances (red triangles) projected in the embedded label space ($\mathbb{R}^{M_E}$). Mostly the source instances that overlap with the target instances in the embedded label space are given attention during training.

Table 3: **CHTL with varying label embedding methods** (W2V: word embeddings, G2V: knowledge graph embeddings, Rand: random vector embeddings): test accuracy (%) on the target task averaged over 10-fold runs. $\%_{L_T} = 0.1$. Method: CHTL:2fc+ATT+AE.

Datasets		Accuracy (%)		
S	T	W2V	G2V	Rand
RCV1	FR	59.0	59.4	48.7
20NEWS	FR	58.6	58.9	51.8
R8	FR	57.7	57.0	52.1

as an *anchor* for knowledge transfer. Some of the source instances that are far from other target instances (near source label term 'crude') are also given high attention, which may be chosen to reduce the source task loss which is averaged over the attended instances. It can be seen that other heterogeneous source instances that may yield negative impact to knowledge transfer are effectively suppressed.

Choice of label embedding methods (Table 3): While W2V and G2V embeddings result in comparable performance with no significant difference, Rand embeddings perform much poorly. This shows that the quality of label embeddings is crucial in transfer of knowledge through CHTL.

5 Related Work

Attention-based learning: The proposed approach is largely inspired by the attention mechanism widely adapted in the recent deep neural network literature for various applications [Xu *et al.*, 2015; Sukhbaatar *et al.*, 2015]. The typical approaches learn parameters for recurrent neural networks (*e.g.* LSTM) which during the decoding step determines a weight over annotation vectors, or a relative importance vector over discrete subsets of input. The attention mechanism can be seen as a regularization preventing overfitting during training, and in our case avoiding negative transfer.

Limited studies have investigated **negative transfer**, most of which propose to prevent negative effects of transfer by measuring dataset- or task-level relatedness via parameter comparison in Bayesian models [Rosenstein *et al.*, 2005]. Our approach practically avoids instance-level negative transfer, by determining *which* knowledge within a source dataset to suppress or attend in learning of a transfer network.

Transfer learning with a heterogeneous label space: Zero-shot learning approaches train a model with distributed vector labels transferred from other domains, thus are more robust for unseen categories. Transfer sources include image co-occurrence statistics for image classification [Mensink *et al.*, 2014], text embeddings learned from auxiliary text documents [Weston *et al.*, 2011; Frome *et al.*, 2013; Socher *et al.*, 2013; Hendricks *et al.*, 2016], or other class-independent similarity functions [Zhang and Saligrama, 2015].

Transfer learning with heterogeneous feature spaces: Multi-view representation learning approaches aim at learning from heterogeneous "views" (feature sets) of multi-modal parallel datasets. The previous literature in this line of work include Canonical Correlation Analysis (CCA) based methods [Dhillon *et al.*, 2011] with an autoencoder regularization in deep nets [Wang *et al.*, 2015], translated learning [Dai *et al.*, 2008], Hybrid Heterogeneous Transfer Learning (HHTL) [Zhou *et al.*, 2014], [Gupta and Ratinov, 2008], etc., all of which require source-target correspondent parallel instances. When parallel datasets are not given initially, [Zhou *et al.*, 2016] propose an active learning scheme for iteratively finding optimal correspondences, or for text domain [Sun *et al.*, 2015] propose to generate correspondent samples through a machine translation system despite noise from imperfect translation. The Heterogeneous Feature Augmentation (HFA) method [Duan *et al.*, 2012] relaxes this limitation for a shared homogeneous binary classification task.

Domain adaptation with homogeneous feature and label spaces often assumes a homogeneous class conditional distribution between source and target, and aims to minimize the difference in their marginal distribution. The previous approaches include distribution analysis and instance reweighting or re-scaling [Huang *et al.*, 2007], subspace mapping [Xiao and Guo, 2015], basis vector identification via sparse coding [Kodirov *et al.*, 2015], or via layerwise deep adaptation [Long and Wang, 2015].

CHTL differs from the above transfer learning or domain adaptation approaches in that CHTL allows for arbitrarily heterogeneous feature and label spaces, and that it does not require instance-level correspondent datasets.

6 Conclusions

We propose a new method for completely heterogeneous transfer learning which uses the attention mechanism to determine instance-level transferability of source knowledge, as well as an unsupervised transfer loss which leads to more robust projections with deeper transfer networks. We provide both quantitative and qualitative analysis through comprehensive simulation studies as well as applications on real-world datasets. Results on synthetic datasets with varying heterogeneity and task difficulty provide new insights on the conditions and parameters in which CHTL can succeed. The proposed approach is general and thus can be applied in other domains, as indicated by the domain-free simulation results.

References

[Amini *et al.*, 2009] Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *NIPS*, pages 28–36, 2009.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[Chan *et al.*, 2015] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

[Dai *et al.*, 2008] 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.

[Dhillon *et al.*, 2011] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, pages 199–207, 2011.

[Duan *et al.*, 2012] Lixin Duan, Dong Xu, and Ivor Tsang. Learning with augmented features for heterogeneous domain adaptation. *ICML*, 2012.

[Frome *et al.*, 2013] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[Gupta and Ratinov, 2008] Rakesh Gupta and Lev-Arie Ratinov. Text categorization with knowledge transfer from heterogeneous data sources. In *AAAI*, pages 842–847, 2008.

[Hendricks *et al.*, 2016] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *CVPR*, 2016.

[Huang *et al.*, 2007] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

[Kodirov *et al.*, 2015] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015.

[Lewis *et al.*, 2004] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[Long and Wang, 2015] Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. *ICML*, 2015.

[Mensink *et al.*, 2014] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Moon and Carbonell, 2016] Seungwhan Moon and Jaime Carbonell. Proactive transfer learning for heterogeneous feature and label spaces. *ECML-PKDD*, 2016.

[Nickel *et al.*, 2015] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. *arXiv preprint arXiv:1510.04935*, 2015.

[Rosenstein *et al.*, 2005] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*, volume 2, page 7, 2005.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero Shot Learning Through Cross-Modal Transfer. In *NIPS*. 2013.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

[Sun *et al.*, 2015] Qian Sun, Mohammad Amin, Baoshi Yan, Craig Martell, Vita Markman, Anmol Bhasin, and Jieping Ye. Transfer learning for bilingual content classification. In *KDD*, pages 2147–2156, 2015.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119. Citeseer, 2014.

[Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. *ICML*, 2015.

[Weston *et al.*, 2011] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI'11*, 2011.

[Xiao and Guo, 2015] Min Xiao and Yuhong Guo. Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation. In *ECMLPKDD*. 2015.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.

[Zhang and Saligrama, 2015] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[Zhou *et al.*, 2014] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W. Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. *AAAI*, 2014.

[Zhou *et al.*, 2016] Joey Zhou, Sinno Pan, Ivor Tsang, and Shen-Shyang Ho. Transfer learning for cross-language text categorization through active correspondences construction. *AAAI*, 2016.