Deep Asymmetric Transfer Network for Unbalanced Domain Adaptation

Daixin Wang, Peng Cui, Wenwu Zhu
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
dxwang0826@gmail.com, cuip@mail.tsinghua.edu.cn, wwzhu@tsinghua.edu.cn

Abstract

Recently, domain adaptation based on deep models has been a promising way to deal with the domains with scarce labeled data, which is a critical problem for deep learning models. Domain adaptation propagates the knowledge from a source domain with rich information to the target domain. In reality, the source and target domains are mostly unbalanced in that the source domain is more resource-rich and thus has more reliable knowledge than the target domain. However, existing deep domain adaptation approaches often pre-assume the source and target domains balanced and equally, leading to a medium solution between the source and target domains, which is not optimal for the unbalanced domain adaptation.

In this paper, we propose a novel Deep Asymmetric Transfer Network (**DATN**) to address the problem of unbalanced domain adaptation. Specifically, our model will learn a transfer function from the target domain to the source domain and meanwhile adapting the source domain classifier with more discriminative power to the target domain. By doing this, the deep model is able to adaptively put more emphasis on the resource-rich source domain. To alleviate the scarcity problem of supervised data, we further propose an unsupervised transfer method to propagate the knowledge from a lot of unsupervised data by minimizing the distribution discrepancy over the unlabeled data of two domains. The experiments on two real-world datasets demonstrate that **DATN** attains a substantial gain over state-of-the-art methods.

Introduction

Nowadays, deep learning models have been successfully applied to many applications (Krizhevsky and Hinton 2011; Mikolov et al. 2010; Wang, Cui, and Zhu 2016). However, the performance for deep models relies heavily on the volume of training data, especially labeled data. When the labeled data are insufficient, the performance would be largely degraded.

To solve the problem, domain adaptation based on deep models attracts much interest in both academia and industry. These methods resort to an auxiliary domain, i.e. the source domain, and transfer its knowledge to the target domain. The key problem is how to bridge the source and the target domain to transfer the knowledge. Most of the existing methods either map them to a new common space (Hubert Tsai, Yeh, and Frank Wang 2016) or minimize the discrepancy between the latent representations of different domains to correlate the two domains (Zhuang et al. 2015; Shu et al. 2015). When these methods transfer the knowledge, they all pre-assume that the importance of the source and target domain data is equivalent. Thus, they always find medium solutions between the source and target domains. However, in most real-world cases, the source and target domains are unbalanced in that the source domain is often more resource-rich and has more reliable knowledge than the target domain. For example, we often use textual data to help image applications and it is much easier to classify texts than images because textual data have a smaller semantic gap (Shu et al. 2015; Zhang et al. 2013). Another typical example is that the English corpora is often used as the source domain to improve the task in minor languages, because we have much more annotated English documents to extract more reliable knowledge than those of minor languages (Zhou et al. 2014). But how to take the domain imbalance into domain adaptation is still an unsolved problem.

There are three challenges for unbalanced domain adaptation: (1) *Domain Heterogeneity*: Data from different domains have different statistical properties and distributions (Srivastava and Salakhutdinov 2012), resulting in the domain discrepancy, thus posing a great challenge for representation space alignment and knowledge transfer. (2) *Unbalanced Knowledge Transfer*. The knowledge in the source and target domains is unbalanced. We need to discriminate the two domains and transfer the knowledge in an asymmetric way to get an optimal solution, which makes existing transfer methods inadequate. (3) *Data Scarcity*. Most existing transfer methods only utilize labeled data to perform knowledge propagation. However, the labeled data are often scarce, which leads to the transfer model sensitive to the noises. How to incorporate more information to alleviate the scarcity is also critical for domain adaptation.

To address the above challenges, we propose a novel Deep Asymmetric Transfer Network (**DATN**), to perform unbalanced domain adaptation. Our model contains two pathways of deep models for source and target domains respectively, which are able to project the data into the high-level space, i.e. the semantic space, to bridge the *domain heterogeneity* (Srivastava and Salakhutdinov 2012). Considering that the knowledge in source and target domains is *unbalanced*,

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

we propose an asymmetric transfer model. Specifically, we learn a feature transfer in the high-level space from the target domain to the source domain to align their representation spaces. Then, we adapt the source domain classifier with more discriminative power to the transformed target domain space. These two asymmetric transfer processes of supervised information make the model more focus on the source domain richer knowledge. Furthermore, to solve the *scarcity* problem of the supervised data, we also utilize unsupervised data to perform transfer. We conduct the experiments on a heterogeneous and a homogeneous dataset. The experimental results demonstrate our proposed method achieves substantial gains compared to the existing approaches.

Related Work

Domain adaptation, also known as transfer learning (Pan and Yang 2010) aims at propagating the knowledge in the source domain to the target domain. Most of existing methods (Oquab et al. 2014; Glorot, Bordes, and Bengio 2011; Long and Wang 2015; Yosinski et al. 2014) focus on homogeneous domain adaptation, which assumes that data of the source and target domains lie in the same domain, such as the images in NUS-WIDE and ImageNet. For this branch, methods often use a share-parameter model for the two domains. Some other works work on heterogeneous domain adaptation, which assumes that the data of source and target domains lie in different domains and different feature spaces. For both homogeneous and heterogeneous domain adaptation, their key bottleneck is to alleviate the domain discrepancy to perform knowledge transfer. Most of existing methods (Zhu et al. 2011; Qi, Aggarwal, and Huang 2011; Shi et al. 2009; Dai et al. 2009) adopt shallow models attempting to explicitly reduce the discrepancy. However, the transferability of shallow models will be greatly limited due to the task-specific variability (Long and Wang 2015), thereby these models cannot achieve satisfied performance.

Recently deep neural networks have been demonstrated to be able to discover invariant factors underlying different datasets which are transferrable between different domains and tasks (Yosinski et al. 2014). Therefore, some recent works start to apply deep neural networks to bridge the source and target domains to perform domain adaptation (Zhuang et al. 2015; Shu et al. 2015; Hubert Tsai, Yeh, and Frank Wang 2016; Zhou et al. 2014). However, these methods assume that the knowledge in two domains are balanced and thus just find a medium solution between the source and target domains, which is not optimal as we have explained before. Although (Zhou et al. 2014) attempts to rely more on the source domain data, it assumes that all the knowledge in the source domain can be totally transferred, which is also detrimental to the transfer performance. Furthermore, most of these methods only use the labeled data to perform transfer, resulting in a non-robust solution when the labeled data is scarce. In this paper, we propose a deep transfer model, which is able to extract the knowledge from both the labeled and unlabeled data and adaptively transfer knowledge from the resource-rich source domain to the target domain in an asymmetric way.

The Methodology

Problem Statement and Notations

Given a set of target domain unlabeled data $X_T^{uL} = \{\mathbf{x}_{T_i}^{uL}\}_{i=1}^{n_T^{uL}}$ and labeled data $\{X_T^L, \mathbf{y}_T\} = \{(\mathbf{x}_{T_i}^L, y_{T_i})\}_{i=1}^{n_T^L}$, similarly a set of source domain unlabeled data $X_S^{uL} = \{\mathbf{x}_{S_i}^{uL}\}_{i=1}^{n_S^{uL}}$ and labeled data $\{X_S^L, \mathbf{y}_S\} = \{(\mathbf{x}_{S_i}^L, y_{S_i})\}_{i=1}^{n_S^L}$, and additionally a set of paired data across two domains $\{X_S^c, X_T^c\} = \{(\mathbf{x}_{S_i}^c, \mathbf{x}_{T_i}^c)\}_{i=1}^{n_c}$, where $\mathbf{x}_{S_i}^{uL}$, $\mathbf{x}_{S_i}^L$ and $\mathbf{x}_{S_i}^c$ are in $\mathbb{R}^{1\times d_S}$, $\mathbf{x}_{T_i}^{uL}$, $\mathbf{x}_{T_i}^L$ and $\mathbf{x}_{T_i}^c$ are in $\mathbb{R}^{1\times d_T}$ and y_{S_i}, $y_{T_i} \in \{1, ..., k\}$, our problem is to transfer the knowledge mined from the source domain data to the target domain.

To solve the problem, we propose a novel Deep Asymmetric Transfer Network (**DATN**) to perform unbalanced domain adaptation, whose framework is shown in Figure 1. It contains two components: the intra-domain representation learning (in Figure 1(a)) to initialize the deep models, and unbalanced domain adaptation (in Figure 1(b)) to propagate the knowledge. Some other notations regarding the model are listed in Table 1. Note that for simplicity, the superscript of the top layer's parameters is omitted. For example, W_S, W_T and $\mathbf{z}_{*_i}$ all denote the notations in the top layer of the deep model. In addition, the subscript and superscript of $\mathbf{z}^{(l)}$ are corresponded with its input. For example, $\mathbf{z}_{S_i}^{uL(l)}$ is the hidden representations corresponded with the input $\mathbf{x}_{S_i}^{uL}$.

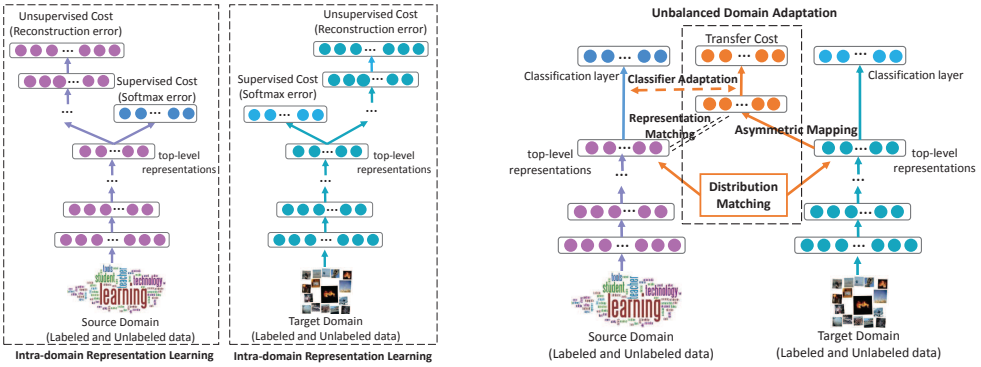
Table 1: Terms and Notations

Symbol	Definition
$*$	$* \in \{S, T\}$ represents the source or target domain
m_*	the number of hidden layers
k	the number of categories
d	the dimensionality of the top-layer representations
X_*	$X_* = X_*^{uL} \cup X_*^L \cup X_*^c = \{\mathbf{x}_{*_i}\}_{i=1}^{n_* = n_*^{uL} + n_*^L + n_c}$
$\hat{\mathbf{x}}_{*_i}$	The reconstruction of the input $\mathbf{x}_{*_i}$
$Z_*^{(l)}$	$Z_*^{(l)} = \{\mathbf{z}_{*_i}^{(l)}\}_{i=1}^{n_*}$, the l-th layer hidden representations
$W_*^{(l)}, \mathbf{b}_*^{(l)}$	the l-th layer's weight matrix or biases in encoder
$\hat{W}_*^{(l)}, \hat{\mathbf{b}}_*^{(l)}$	the l-th layer's weight matrix or biases in decoder
ϑ_*	the softmax parameters
θ_*	$\{W_*^{(l)}\}_{l=1}^{m_*} \cup \{\mathbf{b}_*^{(l)}\}_{l=1}^{m_*} \cup \{\vartheta_*\}$

Intra-domain High-level Representation Learning

We propose a semi-supervised deep autoencoder for each domain to utilize both of its labeled and unlabeled data to perform intra-domain representation learning. It acts as an initialization for the deep models, which facilitates the following unbalanced domain adaptation because it maps the data into the high-level space, where cross-domain data are easy to be correlated and transferred (Ngiam et al. 2011).

The proposed model, shown in Figure 1(a), consists of a deep autoencoder to utilize unsupervised information and a softmax layer to incorporate supervised information. For each layer, we use Relu (Nair and Hinton 2010), i.e. $g(x) = \max(0, x)$ as the non-linear function, because it can prevent gradient vanishing. Additionally, given the input, how to obtain the representations of each layer is the same as



(a) Intra-domain Representation Learning (b) Unbalanced Domain Adaptation

Figure 1: The framework of our proposed model **DATN**

many deep-model based papers do, like (Wang et al. 2015a; 2015b), which we omit in this paper due to the limit of space.

Deep autoencoder is an unsupervised model, which aims at minimizing the reconstruction error for the input samples and is able to capture the data manifolds smoothly (Salakhutdinov and Hinton 2009). Its loss function is:

$$\mathcal{L}_{*,recon} = \sum_{i=1}^{n_*} \|\hat{\mathbf{x}}_{*_i} - \mathbf{x}_{*_i}\|_2^2 \tag{1}$$

To further utilize the supervised information, we add a softmax layer on the deep autoencoder. The loss function can be formulated as follows:

$$\mathcal{L}_{*,soft} = -\frac{1}{n_*^L} \sum_{i=1}^{n_*^L} \sum_{j=1}^{k} 1\{y_{*_i} = j\} \log \frac{e^{\mathbf{z}_{*_i}^L \cdot \vartheta_{*_j}}}{\sum_{l=1}^{k} e^{\mathbf{z}_{*_i}^L \cdot \vartheta_{*_l}}} \tag{2}$$

where $1(\cdot)$ is the indicator function.

Then we combine Eq. 1 and Eq. 2 to form the loss function for intra-domain representation learning.

$$\mathcal{J}_*^{intra} = \mathcal{L}_{*,recon} + \mathcal{L}_{*,soft} + \mathcal{L}_{reg} \tag{3}$$

By minimizing Eq. 3 for each domain respectively, we can initialize the deep models for each domain. However, it cannot get satisfied performance on the target domain due to the lack of data. Then we propose unbalanced domain adaptation method in the next section to do knowledge transfer.

Unbalanced Domain Adaptation

To transfer the knowledge from the source domain to the target domain, two questions remain to be answered: where and how to transfer?

For the first question, (Srivastava and Salakhutdinov 2012) suggests that, cross-domain data have more explicit relationships in the high-level space, where representations of cross-domain data contain much semantic information. Therefore, we process all the transfers in the top-level space.

As we discussed before, in most cases source domain data have much richer and more reliable knowledge than target domain data. Then how to transfer the knowledge when the two domains are unbalanced is the other question. First, we propose an asymmetric transfer strategy which is able to emphasize on the richer knowledge within the source domain supervised data and adaptively transfer the knowledge to the target domain. Second, to alleviate the scarcity problem of supervised data, we utilize unsupervised data in both domains to perform unsupervised transfer by using distribution matching. The framework is shown in Figure 1(b).

Asymmetric Transfer To transfer the knowledge across domains, we first need to bridge the source and the target domain. Most methods seek to find a medium solution to symmetrically bridge the data in the two domains. These methods do not make good use of the richer and more reliable knowledge in the source domain. To solve the problem, we propose an asymmetric transfer process to bridge them. The process contains two parts, one to transfer the knowledge existing in the representation spaces and the other to further transfer the knowledge in the classifier.

In the first part, the key idea is that for a data pair, its high-level representations in the target domain can be transformed to approximate its high-level representations in the source domain. Accordingly, we map the data of the target domain to the source domain through a mapping function G and minimize the mis-alignment error. The loss function is shown as follows:

$$\mathcal{L}_{pair} = \|Z_S^c - Z_T^c \cdot G\|_F^2 + \lambda' \|G\|_F^2$$

where $\lambda' > 0$ is a parameter to balance the mis-alignment loss and the regularization penalty.

By doing the aforementioned asymmetric mapping, the high-level representation spaces of the source and target domains are aligned. Then we further adapt the source domain classifier to classify the target domain data through the learned mapping function G. The reason why we adapt the classifier of the source domain to the target domain is that the richer knowledge in the source domain will lead to a more discriminative source domain classifier. Additionally, we use the target domain labeled data to refine the adapted classifier. The objective function can be shown as follows:

$$\mathcal{L}_{trans} = -\frac{1}{n_T^L} \sum_{i=1}^{n_T^L} \sum_{j=1}^{k} 1\{y_{T_i} = j\} \log \frac{e^{\mathbf{z}_{T_i}^L \cdot G \cdot \vartheta_{S_j}}}{\sum_{l=1}^{k} e^{\mathbf{z}_{T_i}^L \cdot G \cdot \vartheta_{S_l}}}$$

where $1(\cdot)$ is the indicator function.

To summarize, our proposed asymmetric transfer method focuses more on the resource-rich source domain, which has richer and more reliable knowledge. It is able to transfer more informative source domain knowledge existing in both the representation space and classifier parameters to help the target domain. Furthermore, the proposed classifier adaptation method can prevent transferring unhelpful source domain knowledge to the target domain.

Unsupervised Transfer To make the model robust to the noises (Zhu 2005), we additionally utilize unlabeled data to perform transfer. However, unlike supervised data which have explicit information to bridge the two domains such as the labels or corresponded pairs, it is challenging to bridge the unlabeled data between different domains and thus is difficult to perform transferring.

To address the challenge, we use the distributions over the high-level representations as the bridge for cross-domain unlabeled data. Given enough unlabeled data, we assume that the marginal distributions over the high-level representations across two domains should be similar. Specifically, like many other papers do (Long et al. 2014), we use the Maximum Mean Discrepancy (MMD) (Sejdinovic et al. 2013) as the distance measure to compare two distributions. Then the loss functions for unsupervised transfer is introduced as follows, which measures the distribution discrepancy:

$$\mathcal{L}_{unsup} = MMD(Z_S, Z_T) = \| \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbf{z}_{S_i} - \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbf{z}_{T_i} \|_2^2$$

Final Objective and Optimization

Based on the above analysis, we derive the following objective function for our transfer model:

$$\mathcal{J}^{cross} = \mathcal{L}_{pair} + \alpha \mathcal{L}_{trans} + \beta \mathcal{L}_{unsup} + \mathcal{L}_{reg} \tag{4}$$

where $\mathcal{L}_{reg}$ is the regularization term defined as:

$$\mathcal{L}_{reg} = \lambda \sum_{*\in\{S,T\}} \sum_{l=1}^{m_*} (\|W_*^{(l)}\|_F^2 + \|b_*^{(l)}\|_2^2)$$

To train the model, we first optimize Eq. 3 to do intra-domain representation learning, which acts as an initialization. Then we do fine-tuning to perform the cross-domain knowledge transfer.

In fine-tuning, our goal is to minimize $\mathcal{J}^{cross}$ to get optimized parameters of the deep model θ_S, θ_T and additionally the cross-domain mapping function G. It is difficult to simultaneously derive the optimal θ_S, θ_T and G while minimizing $\mathcal{J}^{cross}$. To optimize them, we adopt Block Coordinate Descent (BCD) (Sontag, Globerson, and Jaakkola 2011), which iteratively optimizes the parameters of the two deep models θ_S, θ_T and G in a recurrent process.

In detail, to learn G, we fix the parameters of the deep models θ_S and θ_T and then get Z_T^c and Z_S^c. When Z_T^c and Z_S^c are obtained, from Eq. 4, the gradient of $\mathcal{L}_{trans}$ with respect to G is:

$$\frac{\partial \mathcal{L}_{trans}}{\partial G} = -2 {Z_T^c}^T \cdot (Z_S^c - Z_T^c \cdot G) + 2\lambda' G$$

Then by setting $\partial \mathcal{L}_{trans} / \partial G$ to zero, we can derive the the closed form of G as follows:

$$G = ({Z_T^c}^T \cdot Z_T^c + \lambda' I)^{-1} \cdot {Z_T^c}^T \cdot Z_S^c$$

where I is the identity matrix with the dimensionality of d.

To learn θ_S and θ_T, we fix G and use the back-propagation to update the parameters from the top layers down through the whole deep model. The full algorithm is shown in Alg. 1.

Algorithm 1 Training Algorithm for the *DATN*

Require: X_S, X_T
Ensure: Optimized parameters: $\tilde{\theta}_S$, $\tilde{\theta}_T$ and $\tilde{G}$
1: Randomly initialize parameters θ_S and θ_T.
2: *// Perform intra-domain representation learning*
3: **repeat**
4: Get $\mathcal{J}_*^{intra}(X_*; \theta_*)$ based on Eq. 3, $* \in \{S, T\}$.
5: $\theta_* = \theta_* - \mu \cdot \partial \mathcal{J}_*^{intra}(X_*; \theta_*) / \partial \theta_*$
6: **until** converge
7: $t = 0, \theta_S^{(1)} = \theta_S, \theta_T^{(1)} = \theta_T$
8: *// Perform Unbalanced Domain Adaptation*
9: **repeat**
10: $t = t + 1$
11: Get $G^{(t)}$ by using X_S, X_T, $\theta_S^{(t)}$ and $\theta_T^{(t)}$.
12: Get $\mathcal{J}^{cross}(X_S, X_T; \theta_S^{(t)}, \theta_T^{(t)}, G^{(t)})$ from Eq. 4.
13: $\theta_*^{(t+1)} = \theta_*^{(t)} - \mu \cdot \partial \mathcal{J}^{cross} / \partial \theta_*^{(t)}, * \in \{S, T\}$
14: **until** converge
15: $\tilde{\theta}_S = \theta_S^{(t+2)}, \tilde{\theta}_T = \theta_T^{(t+2)}, \tilde{G} = G^{(t+1)}$

Discussion

In this paper, we use deep autoencoder as the basic block to achieve the transfer. There are other kinds of deep models, such as the CNN (Krizhevsky, Sutskever, and Hinton 2012) and LSTM (Mikolov et al. 2010). Similarly, **DATN** can also be applied to other deep feed-forward models. For example, we can achieve the proposed asymmetric transfer and unsupervised transfer approach in the top layer of the CNN or LSTM. Since our main focus is to introduce the asymmetric transfer model to do unbalanced domain adaptation, we will omit the discussion about different deep architectures.

Then we introduce the complexity of the transfer method. In each iteration of BCD for fine-tuning, the complexity for optimizing θ_S and θ_T is $O(n_S + n_T)$. For the optimization of G, although it includes the inversion of a matrix, the matrix G's dimension is only $d \times d$. Therefore, the complexity for this part is $O(d^3 + d^2 n_c)$. Since d is a constant and often small, the overall complexity for cross-domain transfer is also linear to the number of the overall samples. Therefore, our transfer method is scalable for real applications.

Experiments

Datasets

In our experiments, we use two real-world datasets, i.e. NUS-WIDE and AMAZON REVIEWS.

NUS-WIDE (Chua et al. 2009) is a public web image dataset, which consists of 269,648 images from Flickr. These images are surrounded by a total of 5,018 unique

tags. We use tags as the source domain to help image classification. In our experiment, we use the data of the 10 categories, i.e. *birds*, *buildings*, *car*, *cat*, *dog*, *fish*, *horses*, *flowers*, *mountain* and *plane*. These categories can provide enough data to support our experiment and are also used by (Shu et al. 2015; Qi, Aggarwal, and Huang 2011). For textual information, we use the 1000 most frequent tags and thus texts are represented by 1000-dimensional tag occurrence vectors. For image features, we use both the provided 500-dimensional SIFT (Lowe) features and the 4096-dimensional CNN features generated by the VGGnet (Simonyan and Zisserman 2014) pretrained on the ImageNet.

AMAZON REVIEWS (Prettenhofer and Stein 2010) is a cross-language dataset, which contains the Amazon reviews of the products of three categories: books, DVDs and music on four languages: English (EN), German (GE), French (FR) and Japanese (JP). We use the English reviews as the source domain and each of the other three languages as the target domain to classify the reviews' category. Additionally, the Google translator is applied on part of the non-English reviews to construct domain-paired data. We use the 128-dimensional topic distributions obtained by Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) as the input features.

Some detailed statistics about the two datasets are shown in Table 2. There is no overlapped samples among $\mathbf{D}_*^{uL}$, $\mathbf{D}_*^{L}$, $\mathbf{D}_c$ and Test, $* \in \{S, T\}$. In these two datasets, the labeled data in the target domain are very limited, which is often the real case and a great challenge for many applications.

Table 2: The statistics of the datasets

Dataset	$\lvert\mathbf{D}_S^{uL}\rvert$	$\lvert\mathbf{D}_S^{L}\rvert$	$\lvert\mathbf{D}_T^{uL}\rvert$	$\lvert\mathbf{D}_T^{L}\rvert$	$\lvert\mathbf{D}_c\rvert$	Test
NUS-WIDE	15000	5000	15000	100	5000	1000
AMAZON	15000	3000	15000	50	2000	1000

Baseline and Evaluation Metrics

We use both non-transfer and transfer methods as the baselines. Non-transfer methods only utilize the target domain data to do classification. We use (1) **SVM** (Hearst et al. 1998): It is a well-known supervised shallow classification model and we use LibSVM (Chang and Lin 2011) as the implementation. (2) Deep Neural Network (**DNN**): It is a supervised deep model and uses softmax regression in the top layer to predict the labels. (3) Semi-Supervised Deep Autoencoder (**SSDAE**): Its loss function is the combination of the reconstruction error on the unlabeled data and softmax error on the labeled data.

For transfer methods, we use recently proposed heterogeneous domain adaptation methods as baselines. (1) **TTI** (Qi, Aggarwal, and Huang 2011): It is a shallow-structured transfer method. (2) **WSDTN** (Shu et al. 2015): It is a deep transfer method, which transfers the labeling information and seeks to find a medium solution between the source and target domain. (3) **HHTL** (Zhou et al. 2014): It is a deep transfer model, which assumes that all the knowledge in the source domain benefits the target domain. (4) **CDLS** (Hubert Tsai, Yeh, and Frank Wang 2016): It is a semi-supervised domain adaptation method which aims to symmetrically find cross-domain landmarks to help knowledge transfer. (5) **DATN_sup**: It is a simple version of **DATN**, which does not take unsupervised transfer into account.

Parameter Settings

For NUS-WIDE, inspired by (Wang et al. 2015a), we use a 5-layer deep model for image pathway, and a 4-layer for text pathway. For AMAZON REVIEWS, we all use 3-layer deep models for two domains. The number of neurons in each layer is summarized in Table 3. All the deep models use the same structure for a fair comparison.

The values of α and β are selected from $\{0, 0.5, 1, 2, 5, 10\}$. The regularization parameters of λ' and λ are set as 0.1 and 0.0001. The final values of all the parameters are determined by using 5-fold cross-validation on the training set. For NUS-WIDE, α is set as 2 and β is set as 1. For AMAZON REVIEWS, they are both set as 2. Our approach is implemented in Tensorflow. Throughout the experiments, the learning rate is set as 0.0001, the decay is set as 0.8 and the momentum is set as 0.8.

Table 3: Number of neurons of each layer of DATN

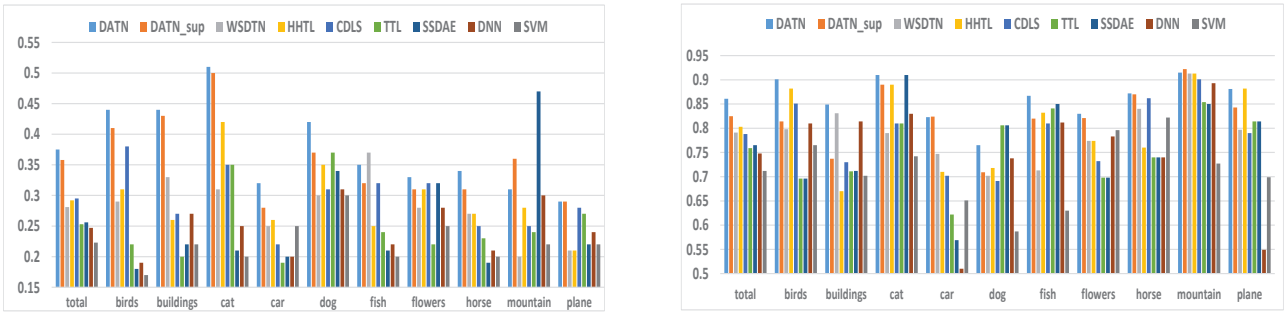
Dataset	Target Domain	Source Domain
NUS-WIDE (SIFT)	500-512-128-128-64	1000-512-128-64
NUS-WIDE (VGG16)	4092-1024-256-128-64	1000-512-128-64
AMAZON REVIEWS	128-100-100	128-100-100

Experimental Results

We first report the overall image classification accuracy and the accuracy over each category on NUS-WIDE using SIFT and VGG16 features. The results are reported in Figure 2.

From Figure 2, we can get the following observations:

- Regarding the overall accuracy, **DATN** achieves an at least 30% improvement when using SIFT features and 7% improvement when using VGG16 features over state-of-the-art methods. It demonstrates the effectiveness of our proposed **DATN** on heterogeneous domain adaptation.
- The result that **DATN_sup** outperforms **CDLS** and **WSDTN** demonstrates an advantage of the proposed asymmetric transfer model.
- The result that **DATN_sup** outperforms **HHTL** demonstrates that assuming the knowledge in the source domain is fully transferrable is detrimental to the performance on the target domain and our method of adaptively transferring source domain knowledge is essential.
- The result that **DATN** outperforms **DATN_sup** demonstrates the proposed unsupervised transfer is effective.
- The result that the performances of all the deep transfer learning methods are better than those of non-transfer methods demonstrates that deep domain adaptation is very essential for a domain with limited data.
- The result shows the performance using VGG16 features is much better than that using SIFT features. The reason is that VGG16 features is more powerful because they utilize much more side information on the ImageNet. Even if



(a) SIFT (b) VGG16

Figure 2: Image classification accuracy on NUS-WIDE using SIFT or VGG16 features.

in this case, the result that **DATN** can further improve the performance demonstrates the superiority of our method.

We also conduct the experiment on another application of transferring the knowledge mined from English corpora to classify the textual reviews on other languages like French, German and Japanese.

Table 4: Classification accuracy on AMAZON REVIEWS

Method	EN→ FR	EN→ GE	EN → JP
DATN	**0.737**	**0.729**	**0.755**
DATN_sup	0.724	0.718	0.734
WSDTN	0.675	0.665	0.678
HHTL	0.701	0.673	0.702
CDLS	0.692	0.672	0.681
TTI	0.572	0.563	0.568
SSDAE	0.628	0.602	0.623
DNN	0.601	0.588	0.600
SVM	0.544	0.558	0.571

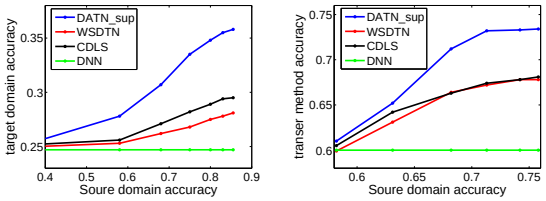
Table 4 shows that for the three target-domain languages, our method **DATN** can achieve better classification accuracy compared with baselines, which demonstrates the effectiveness of our method on the cross-language application.

Further Results and Discussions

Discussions about Asymmetric Transfer One assumption behind the unbalanced domain adaptation is that source domain data have richer and more reliable knowledge than those of target domain. Then one natural question is that how the quality of the source domain data affects the transfer performance on the target domain?

To conduct the experiment, we change the number of source domain labeled data to control the quality of the source domain. Then we report the classification accuracy on the target domain. We compare the performance of **DATN_sup** with the performance of **WSDTN**, **CDLS** and **DNN**. **WSDTN** and **CDLS** are symmetric-transfer based methods. **DNN** is non-transfer method. All methods only use the supervised data for training. The result is shown in Figure 3.

Figure 3 shows that when the source domain accuracy increases, the target domain accuracy of **DATN_sup** increases much faster than that of **WSDTN** and **CDLS**. The reason is that symmetric-transfer based methods, i.e. **WSDTN**



(a) NUS-WIDE (SIFT) (b) AMAZON (EN-JP)

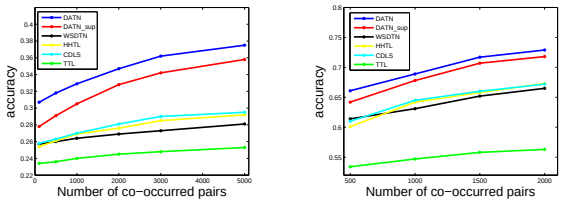
Figure 3: The target domain classification accuracy of **DATN_sup**, **WSDTN**, **CDLS** and **DNN** when changing the source domain accuracy performed on DNN.

and **CDLS**, find a medium solution between the source and the target domain, while our proposed asymmetric-transfer method will more focus on the knowledge in source domain data. Then for **DATN_sup**, as the accuracy of source domain increases, i.e. the knowledge in the source domain becomes more reliable, our proposed asymmetric transfer model is able to extract and transfer richer and more reliable source domain knowledge to the target domain, thus the performance on the target domain increases much faster. Furthermore, the result also shows that if the accuracy of the source domain is not that good, such as at 0.4 accuracy on NUS-WIDE, **DATN_sup** achieves 0.255 accuracy, which outperforms 0.247 accuracy of **DNN**, because **DATN_sup** transfers the knowledge of the source domain adaptively.

To summarize, when source domain data contain richer knowledge, our proposed asymmetric transfer model can achieve a significant improvement over symmetric transfer methods. When the knowledge is not that rich, the improvement also exists but not that significant.

Furthermore, we regard the domain-paired data as the bridge to connect the representation spaces between source and target domains, which facilitate the classifier adaptation. Therefore, we evaluate the effect of the number of the co-occured pairs on Figure 4.

From Figure 4, we find that the average accuracy of all the methods increases as the number of the co-occurred pairs increases, which demonstrates the importance of domain-paired data for domain adaptation. In addition, we observe that our proposed method is robust to the number of the

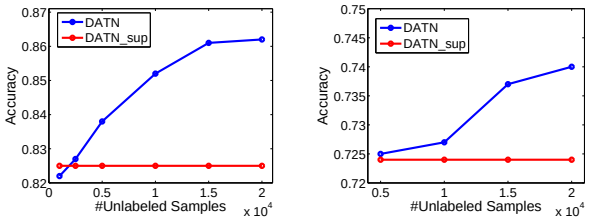


(a) NUS-WIDE (SIFT) (b) AMAZON (EN-GE)

Figure 4: The classification accuracy of transfer methods when changing the number of domain-paired data, i.e. $|\mathbf{D}_C|$.

paired data because when we only have 100 pairs in NUS-WIDE, **DATN** also significantly outperforms baselines.

Discussions about Unsupervised Transfer Then we evaluate how the number of unlabeled samples in the source domain affects the transfer accuracy in Figure 5.



(a) NUS-WIDE (VGG16) (b) AMAZON (EN-FR)

Figure 5: The classification accuracy of **DATN** on NUS-WIDE and AMAZON REVIEWS when the size of the unlabeled data in source domain , i.e. $|\mathbf{D}_S^{uL}|$, changes.

From Figure 5, we observe that when the size of the source domain unlabeled dataset is small, the performance of **DATN** is similar, or even worse than **DATN_sup**. The reason is that a small size of dataset cannot characterize the data distributions well and thus our method of matching the distributions may introduce noises to degrade the accuracy. When the size of unlabeled data continuously increases, the performance becomes better because more data can characterize more reliable distributions to be transferred.

Furthermore, we discard the asymmetric transfer part to see the result of the unsupervised transfer, denoted as **DATN_unsup**. The performance is shown in Table 5.

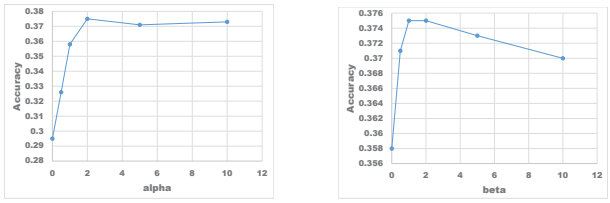
Table 5: Transfer Performance for **DATN**, **DATN_sup** and **DATN_unsup**

Dataset	**DATN**	**DATN_sup**	**DATN_unsup**
NUS-WIDE (SIFT)	**0.375**	0.358	0.285
NUS-WIDE (VGG)	**0.861**	0.825	0.785
Amazon (EN→ FR)	**0.737**	0.724	0.638
Amazon (EN→ GE)	**0.729**	0.718	0.642
Amazon (EN→ JP)	**0.755**	0.734	0.661

Table 5 shows that the result of **DATN_unsup** is far less than **DATN** and **DATN_sup**. It demonstrates that transferring supervision information, i.e. domain-paired data and classifier, is more informative for domain adaptation than transferring unsupervised information. Although unsupervised information is also helpful, it should be combined with supervised information, which is able to better bridge the source and target domains and then fascilitate the unsupervised transfer process.

Parameter Sensitivity

We investigate the parameter sensitivity of α and β defined in Eq. 4 on NUS-WIDE.



(a) The influence of α (b) The influence of β

Figure 6: The study of parameter sensitivity for DATN on NUS-WIDE (SIFT).

From the results, we can see that our model is not sensitive to the parameter settings in general. Specifically, comparing Figure 6(a) and Figure 6(b), we find that α has a greater influence on the performance than β, which implies that the proposed supervised asymmetric transfer model, especially the classifier adaptation method is more important than the unsupervised transfer, which is consistent with the conclusion we just got in the previous experiment.

Conclusion

In this paper, we propose a novel Deep Asymmetric Transfer Network (**DATN**) to perform unbalanced domain adaptation. Our model is able to propagate much richer and more robust knowledge in the source domain to the target domain in an asymmetric way. Furthermore, to make the model more robust, we do unsupervised transfer by distribution matching over high-level representations across domains. The experiments conducted on two real-world datasets demonstrate a significant improvement of **DATN** over baselines. We also find that the proposed supervised asymmetric transfer model, especially the classifier adaptation method, has larger effect than the unsupervised transfer on the classification accuracy. The future directions may focus on transferring the knowledge from more complex data, such as the natural languages, heterogeneous networks and so on.

Acknowledgement

This work was supported by National Program on Key Basic Research Project, No. 2015CB352300; National Natural Science Foundation of China Major Project No. U1611461; National Natural Science Foundation of China, No. 61772304, No. 61521002, No. 61531006. Thanks for the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, and the Young Elite Scientist Sponsorship Program by CAST.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2(3):27.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*.

Dai, W.; Chen, Y.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2009. Translated learning: Transfer learning across different feature spaces. In *Advances in neural information processing systems*, 353–360.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 513–520.

Hearst, M. A.; Dumais, S. T.; Osman, E.; Platt, J.; and Scholkopf, B. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* 13(4):18–28.

Hubert Tsai, Y.-H.; Yeh, Y.-R.; and Frank Wang, Y.-C. 2016. Learning cross-domain landmarks for heterogeneous domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5081–5090.

Krizhevsky, A., and Hinton, G. E. 2011. Using very deep autoencoders for content-based image retrieval. In *ESANN*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Long, M., and Wang, J. 2015. Learning transferable features with deep adaptation networks. *CoRR, abs/1502.02791* 1:2.

Long, M.; Wang, J.; Ding, G.; Pan, S. J.; and Philip, S. Y. 2014. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 26(5):1076–1089.

Lowe, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, 3.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10):1345–1359.

Prettenhofer, P., and Stein, B. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1118–1127. Association for Computational Linguistics.

Qi, G.-J.; Aggarwal, C.; and Huang, T. 2011. Towards semantic knowledge propagation from text corpus to web images. In *Proceedings of the 20th international conference on World wide web*, 297–306. ACM.

Salakhutdinov, R., and Hinton, G. 2009. Semantic hashing. *International Journal of Approximate Reasoning*.

Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; Fukumizu, K.; et al. 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics* 41(5):2263–2291.

Shi, Y.; Lan, Z.; Liu, W.; and Bi, W. 2009. Extending semi-supervised learning methods for inductive transfer learning. In *2009 Ninth IEEE International Conference on Data Mining*, 483–492. IEEE.

Shu, X.; Qi, G.-J.; Tang, J.; and Wang, J. 2015. Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In *Proceedings of the 23rd ACM international conference on Multimedia*, 35–44. ACM.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sontag, D.; Globerson, A.; and Jaakkola, T. 2011. Introduction to dual decomposition for inference. *Optimization for Machine Learning* 1(219-254):1.

Srivastava, N., and Salakhutdinov, R. 2012. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*.

Wang, D.; Cui, P.; Ou, M.; and Zhu, W. 2015a. Deep multimodal hashing with orthogonal regularization. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*.

Wang, D.; Cui, P.; Ou, M.; and Zhu, W. 2015b. Learning compact hash codes for multimodal representations using orthogonal deep structure. *IEEE Transactions on Multimedia* 17(9):1404–1416.

Wang, D.; Cui, P.; and Zhu, W. 2016. Structural deep network embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*.

Zhang, H.; Zha, Z.-J.; Yang, Y.; Yan, S.; Gao, Y.; and Chua, T.-S. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, 33–42. ACM.

Zhou, J. T.; Pan, S. J.; Tsang, I. W.; and Yan, Y. 2014. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, 2213–2220.

Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S. J.; Xue, G.-R.; Yu, Y.; and Yang, Q. 2011. Heterogeneous transfer learning for image classification. In *AAAI*.

Zhu, X. 2005. Semi-supervised learning literature survey.

Zhuang, F.; Cheng, X.; Luo, P.; Pan, S. J.; and He, Q. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In *Int. Joint Conf. Artif. Intell.*