# Transferable Attention for Domain Adaptation

**Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long,✉ Jianmin Wang**
School of Software, Tsinghua University, China
KLiss, MOE; BNRist; Research Center for Big Data, Tsinghua University, China
{wxm17,liliang17,ywr16}@mails.tsinghua.edu.cn {mingsheng,jimwang}@tsinghua.edu.cn

## Abstract

Recent work in domain adaptation bridges different domains by adversarially learning a domain-invariant representation that cannot be distinguished by a domain discriminator. Existing methods of adversarial domain adaptation mainly align the global images across the source and target domains. However, it is obvious that not all regions of an image are transferable, while forcefully aligning the untransferable regions may lead to negative transfer. Furthermore, some of the images are significantly dissimilar across domains, resulting in weak image-level transferability. To this end, we present Transferable Attention for Domain Adaptation (TADA), focusing our adaptation model on transferable regions or images. We implement two types of complementary transferable attention: transferable local attention generated by multiple region-level domain discriminators to highlight transferable regions, and transferable global attention generated by single image-level domain discriminator to highlight transferable images. Extensive experiments validate that our proposed models exceed state of the art results on standard domain adaptation datasets.

## Introduction

Deep learning has been a huge success in diverse applications across many fields, such as computer vision, robotic control and natural language processing with the help of large-scale labeled datasets. However, owing to a phenomenon known as *dataset bias* or *domain shift* (Torralba and Efros 2011), deep networks trained on one large labeled dataset do not generalize well to novel datasets and tasks. The typical solution of further fine-tuning pre-trained networks on task-specific datasets may be impractical because it is often prohibitively expensive to collect enough labeled data to properly fine-tune the considerable number of network parameters. This dilemma has motivated the research on domain adaptation, which aims to establish effective algorithms to reduce the labeling cost, typically by leveraging readily-available labeled data from a different but related domain (Quionero-Candela et al. 2009; Pan and Yang 2010).

Domain adaptation, a special scenario of transfer learning, aims to learn a discriminative model that reduces the dataset shift between the training and testing distributions. Previous domain adaptation methods either bridge the source domain and target domain by learning domain-invariant representations (Pan et al. 2011; Gong et al. 2012) or estimating instance importances (Huang et al. 2006; Gong, Grauman, and Sha 2013) using labeled source data and unlabeled target data. Recent researches have indicated that deep neural networks can learn more transferable representations (Oquab et al. 2014; Donahue et al. 2014; Yosinski et al. 2014), by disentangling explanatory factors of variations behind domains. The latest advances of deep domain adaptation for classification task extract domain-invariant representations by embedding domain adaptation modules in deep architectures, through minimizing the discrepancy between feature distributions (Tzeng et al. 2014; Long et al. 2015; 2016; 2017) or adversarially learning the feature representations to deceive some domain discriminator (Tzeng et al. 2015; Ganin and Lempitsky 2015; Ganin et al. 2016; 2016; Long et al. 2018).

Despite their efficacy in various tasks, existing adversarial domain adaptation methods mainly align the representations extracted from the entire images across domains, without considering the complex distributions of different regions. It is obvious that different regions of an image are not equally transferable. Some regions in the image, like background, may not contribute much to domain adaptation though it is possible to be aligned across domains in the feature space. Moreover, some images that are significantly dissimilar across domains in the feature space should not be forcefully aligned together across domains, otherwise it may be vulnerable to negative transfer of irrelevant knowledge. However, these problems were not considered by previous domain adaptations methods.

In this paper, we tackle the aforementioned challenges in a unified multi-adversarial learning framework while further exploring the attention mechanism. Recent advances in domain adaptation reveal that fine-grained alignment of features extracted from different domains (Pei et al. 2018) can yield better performance in many transfer learning tasks. Besides, the attention mechanism (Vaswani et al. 2017) is an effective method to focus on important regions of an image, with numerous successes in deep learning tasks such as classification, segmentation and detection. Therefore, this paper presents Transferable Attention for Domain Adaptation (TADA), which realizes two types of complementary

transferable attention: transferable *local* attention generated by multiple region-level domain discriminators to highlight transferable regions, and transferable *global* attention generated by single image-level domain discriminator to highlight transferable images. The transferable local attention considers the variability in different regions' transferability, which is implemented by a multi-adversarial network over the representations extracted from different regions. Furthermore, the transferable global attention takes the variability of different images' transferability into account, exploring the fact that the images more similar across domains will contribute more to the classification task. Comprehensive experiments demonstrate that our models achieve state of the art performance on standard domain adaptation datasets.

## Related Works

**Domain Adaptation** In order to mitigate the generalization bottleneck introduced from *domain shift*, many domain adaptation methods (Pan and Yang 2010; Quionero-Candela et al. 2009) have been proposed through the past decade. Early domain adaptation methods either bridge the source domain and the target domain by learning domain-invariant feature representations (Pan et al. 2011; Duan, Tsang, and Xu 2012) or estimating instance importances (Huang et al. 2006; Sugiyama, Krauledat, and Ller 2007) using labeled source data and unlabeled target data. Since deep networks prove to be able to learn more transferable representations (Yosinski et al. 2014), they have been widely adopted to generate domain-invariant representation for transfer learning (Glorot, Bordes, and Bengio 2011; Oquab et al. 2014), multi-modal and multi-task learning (Collobert et al. 2011; Ngiam et al. 2009), leading to significant performance gains against previous shallow transfer learning methods.

These deep domain adaptation methods gain huge improvement, however, deep representations can only reduce, but not remove, the cross-domain discrepancy according to some recent research (Glorot, Bordes, and Bengio 2011; Tzeng et al. 2014). To seamlessly integrate deep learning and domain adaptation, some methods (Tzeng et al. 2014; Long et al. 2015; 2016; 2017) add adaptation layers in deep convolutional networks to match the feature distributions of the source and the target domains, while others add a subnetwork as the domain discriminator to distinguish features extracted from different domains and train a deep classification model to confuse the domain discriminator at the same time (Ganin and Lempitsky 2015; Tzeng et al. 2015; 2017). To consider the complex multimode structures underlying the data distributions, Pei *et al.* (2018) utilizes multiple domain discriminators each associated with a class to enable fine-grained alignment of different data distributions.

**Adversarial Learning** Since Ganin *et al.* (2015) successfully achieved the adaptation behavior via a gradient reversal layer, adversarial adaptation methods have gained growing interest within the field of transfer learning. These methods utilize an adversarial objective loss function regarding to a domain discriminator to minimize the domain discrepancy and thus generate more transferable representations. Since the standard Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) may encounter the technical difficulty of mode collapse, some innovative ideas were addressed by recent proposals (Mirza and Osindero 2014; Che et al. 2016; Metz et al. 2016; Odena, Olah, and Shlens 2017). In particular, Generative Multi-Adversarial Network (GMAN) (Durugkar, Gemp, and Mahadevan 2016), a framework that extends GANs to multiple discriminators, explores several design perspectives with the discriminator role ranging from formidable adversary to forgiving teacher, which significantly eases model training and enhances distribution matching. Recently, increasing researches applied GANs to the domain transfer problem, for example, the CoGAN (Liu and Tuzel 2016) train two GANs to generate the source and target images respectively. In addition, the CyCADA (Hoffman et al. 2018) introduced the CycleGAN (Zhu et al. 2017) into the problem of semantic segmentation for domain adaptation and achieved impressive performances.

**Global and Local Attention** Recently, attention mechanisms have received remarkable advances in network architectures, even without the assistance of recurrence or convolutions entirely (Vaswani et al. 2017). These methods allow the networks to weight features at the pixel level. Considering different levels of attention, the global and local attention was further put forward. It is reported that global and local attention models have achieved promising performance in image caption (Li et al. 2017), image segmentation (Chen et al. 2016) and image classification (Wang et al. 2017). Additionally, researchers (Moon and Carbonell 2017) proposed Attentional Heterogeneous Transfer, a method with a newly-designed transfer loss to determine the transferable samples from the source domain to the target domain.

## Transferable Attention for Domain Adaptation

In this paper, we focus on the unsupervised domain adaptation problem, which constitutes a labeled source domain $\mathcal{D}_s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and an unlabeled target domain $\mathcal{D}_t = \{\boldsymbol{x}_j^t\}_{j=1}^{n_t}$, where $\boldsymbol{x}_i$ is an example and $y_i$ is the associated label. The goal of this paper is to build a deep network, which can be trained on the labeled source data and generalize well to the unlabeled target data. Note that the source and target domains follow different probability distributions. The discrepancy between these two distributions raises the key technical challenge of domain adaptation.

The main technique of exiting methods is to bridge different domains by closing the distribution discrepancy. To realize this, an intuitive idea is to formally define some statistical distance in the probabilitic metric space, while learning a new representation of the source and target data to minimize that distance (Long et al. 2015). A more sophisticated idea is inspired by the generative adversarial networks (GANs) (Goodfellow et al. 2014). Here a two-player minimax game is constructed, in which the first player is a domain discriminator to distinguish the source from the target, and the second player is a feature extractor trained adversarially to deceive the domain discriminator (Ganin et al. 2016). These two ideas have incubated the mainstream deep

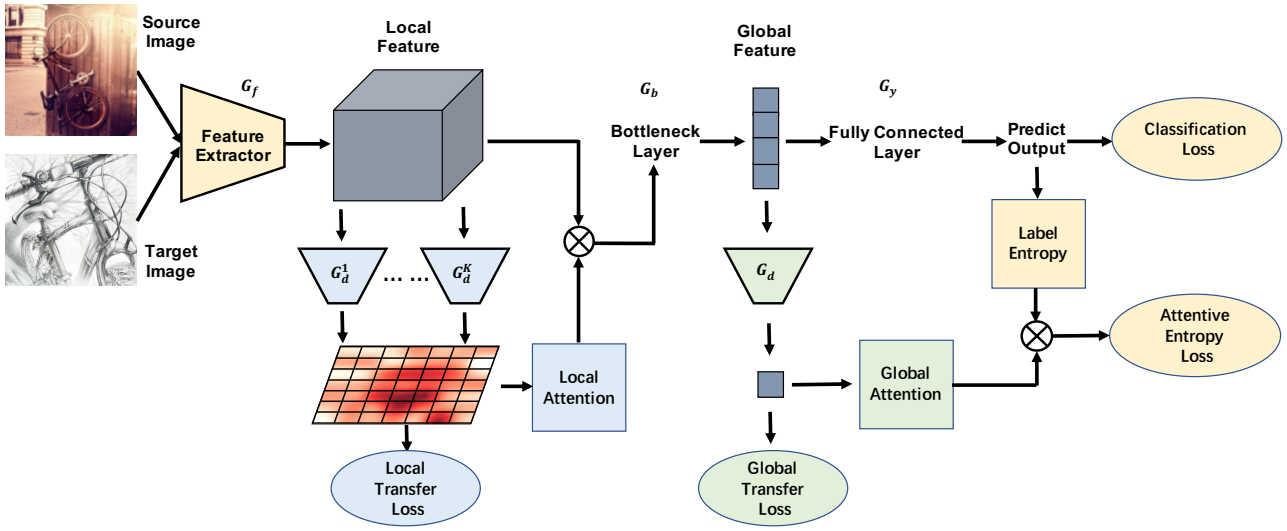



Figure 1: The architectures of Transferable Attention for Domain Adaptation (TADA), where multi-adversarial network (blue) is developed for local attention to highlight the representations of those regions with higher transferability, and global adversarial network (green) is utilized to enhance the prediction certainty of the images more similar in the feature space across domains.

domain adaptation methods in recent years.

While significant performance gains have been witnessed, there is a common intrinsic limitation of this line of work: transfer is in coarse-grain. In other words, each image is reasoned as a whole to be transferred or not, without exploiting its fine-grained structures, e.g. regions or patches. A more realistic consideration is that, the rich structures of an image should be further exploited to enable fined-grained transfer. Also we should be able to reason about which structure of an image is transferable and which is not.

This observation is motivated from human learning: when a person learns to transfer the knowledge behind an image, he will mainly attend to the structures analogous (in other words, transferable) to his target task of interest. Thus, the attention mechanism of humans is beyond just paying attention to objects—instead, they pay more attention to the objects useful for reasoning their particular target task. In this paper, we envision the idea of *transferable attention*, which is described as follows.

**Definition 1 (Transferable Attention)** *A transferable attention, in the context of image recognition, is the mechanism that a human not only pays attention to a source object, but also connects this attention to a target object of interest.*

More intuitively, the standard attention studied in machine learning (Vaswani et al. 2017) focuses on a specific object in an image, while the transferable attention studied in this paper concentrates on the similarity or distinction between two objects—to reason about whether they are transferable. An important fact is that: the attention may be multi-scale, that is, sometimes people observe all objects as a whole, and sometimes they attend to only a specific part of an object. This distinction also applies to the transferable attention. In this work, we explore two types of transferable attention: transferable *local* attention and transferable *global* attention.

## Transferable Local Attention

First, we introduce the transferable local attention, which focuses our domain adaptation model to those transferable regions. Lets recall the well-established domain adversarial neural network (DANN) (Ganin et al. 2016), in which a two-player minimax game is constructed. The domain discriminator $G_d$ serves as the first player and its parameters $\theta_d$ are learned by minimizing the loss $L_d$ of domain discriminator, while the feature extractor $G_f$ is the second player whose parameters $\theta_f$ are learned by maximizing the loss $L_d$ in order to confuse the domain discriminator $G_d$. The objective of DANN (Ganin et al. 2016) can be formulated as

$$\begin{aligned} C_0\left(\theta_f, \theta_y, \theta_d\right) = & \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s} L_y\left(G_y\left(G_f\left(\boldsymbol{x}_i\right)\right), y_i\right) \\ & - \frac{\lambda}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d\left(G_d\left(G_f\left(\boldsymbol{x}_i\right)\right), d_i\right), \end{aligned} \tag{1}$$

where $n = n_s + n_t$ and $\lambda$ is a hyper-parameter that trade-offs the domain adaptation loss $L_d$ with the classification loss $L_y$ corresponding to the source classifier $G_y$.

As mentioned earlier, not all regions of the image are equally transferable and some regions in the image are more transferable than the others. Therefore, to match the source and target domains over the structures underlying the different regions of an image, we split the domain discriminator $G_d$ in Equation (1) into $K$ region-wise domain discriminators $G_d^k$, $k = 1, 2, \ldots, K$, and each is responsible for matching the source and target domain data corresponding to region $k$, as shown in Figure 1. Applying this to all $K$ domain discriminators $G_d^k$, $k = 1, 2, \ldots, K$ yields

$$L_l = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d(G_d^k(\boldsymbol{f}_i^k), d_i), \tag{2}$$

where $G_f$ represents the feature extractor, $\boldsymbol{f}_i^k = G_f(\boldsymbol{x}_i))^k$ is the representation in region $k$, $d_i$ is the domain label of point $\boldsymbol{x}_i$, $L_d$ is the cross-entropy loss of the domain discriminator $G_d^k$. Moreover, $K$ is determined by the network architecture, and in particular, $K = W \times H$, where $W$ and $H$ are the width and height of the feature map in the last convolutional layer. For example, we adopt ResNet-50 as our network backbone, where the last convolutional layer is of dimension $7 \times 7 \times 2048$, and thus $K = 49$.

The output $\hat{d}_i^k = G_d^k(\boldsymbol{f}_i^k)$ of each domain discriminator $G_d^k$ is the probability of the region $k$ in image $i$ belonging to the source domain. When the probability approaches 1, it indicates that the region $k$ belongs to the source domain, and 0 represents that it belongs to the target domain. Note that our goal of transferable local attention is to attend to the objects of interest that are also transferable across domains. Therefore, in order to focus on more transferable regions, a larger local attention value should be generated over these regions. In information theory, the entropy functional is an uncertainty measure defined as $H(p) = -\sum_j p_j \cdot \log(p_j)$ which nicely meets our need to quantify the transferability. We thus utilize the entropy criterion to generate the local attention value for each region $k$ as

$$w_i^k = 1 - H(\hat{d}_i^k). \tag{3}$$

The benefit of this kind of local attention is to enable a fine-grained transfer path from the source to the target. However, wrong local attention may hurt the domain adaptation task to some degree, so we mitigate such negative effect by adding a residual connection following the idea of existing attention methods (Wang et al. 2017). Such a residual mechanism is more robust to wrong local attention. Therefore, the locally attended features $\boldsymbol{h}_i^k$ can be finally transformed as

$$\boldsymbol{h}_i^k = (1 + w_i^k) \cdot \boldsymbol{f}_i^k. \tag{4}$$

In this way, the regions whose representations are more transferable will be weighted by a larger attention value, thus focusing the domain adaptation model on those important regions. As the attention value for each region is generated according to its transferability, this kind of local attention is naturally transferable across domains.

## Transferable Global Attention

In this section, we further introduce the transferable global attention, which focuses our domain adaptation model to those transferable images. In the local attention module, we focus on the transferability of each region for a fine-grained transfer. Admittedly, owing to translations, rotations or other transformations in the images, it is possible that the domain discriminators might find fewer regions to align, however, these regions are still more transferable than the other hard-to-align regions. Therefore, it is necessary to develop a global adversarial module which can transfer knowledge under domain variations due to translations, rotations or other transformations. Similar to DANN (Ganin and Lempitsky 2015), our global adversarial learning module is also added to the features $G_b(\boldsymbol{h}_i)$ before the classifier $G_y$. We train the global discriminator with the following objective function:

$$L_g = \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d(G_d(G_b(\boldsymbol{h}_i), d_i)), \tag{5}$$

where $G_b$ represents the bottleneck layer parametrized by $\theta_b$ (Figure 1), $d_i$ is the domain label of point $\boldsymbol{x}_i$, and $L_d$ is the cross-entropy loss of the global domain discriminator $G_d$.

Inspired by the minimum entropy regularization for semi-supervised learning (Grandvalet and Bengio 2005), RTN (Long et al. 2016) exploits the entropy minimization principle for refining the classifier adaptation, which encourages the low-density separation between classes by minimizing the entropy of class-conditional distribution on target domain data $\mathcal{D}_t$. Obviously, minimizing entropy increases the confidence of the classifier predictions. However, not all images in the target domain are transferable, such as the images that are significantly dissimilar in the feature space across domains. Negative results may incur if the entropy of these images are forcefully minimized. Since these dissimilar images are easier to be mistakenly classified, it may be harmful to increase their confidence of the classifier predictions since increasing their certainty will confuse the classifier.

Similar to the local attention generation mechanism, it is also reasonable to utilize the global discriminator's output $\hat{d}_i = G_d(G_b(\boldsymbol{h}_i))$ to generate an attention value for each image's entropy loss, aiming to enhance the certainty of those images that are more similar across domains. The global attention value for each image is generated by the following equation:

$$m_i = 1 + H(\hat{d}_i), \tag{6}$$

where $\hat{d}_i$ is the output of the global domain discriminator indicating the corresponding image's transferability. The more transferable the corresponding image is, the larger the global attention value $m_i$ is. Embedding the attention value $m_i$ into the entropy loss, the attentive entropy can be formulated as

$$L_h = -\frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} \sum_{j=1}^{c} m_i \cdot \boldsymbol{p}_{i,j} \cdot \log(\boldsymbol{p}_{i,j}), \tag{7}$$

where $c$ is the number of classes, and $\boldsymbol{p}_{i,j}$ is the probability of predicting point $\boldsymbol{x}_i$ to class $j$. In this way, the global discriminator's output is cleverly applied to highlight the entropy of the images which transfer better in the feature space. By minimizing the attentive entropy penalty, these images' predictions will become certain and thus improve the classifier's performance. As the attention value for each image is generated according to its transferability, this kind of global attention is naturally transferable across domains.

## Transferable Attention for Domain Adaptation

By applying transferable *local* attention and transferable *global* attention modules, *negative* transfer for each region is alleviated and *positive* transfer for each image is enhanced. The local attention module based on multi-adversarial network on different regions enables a fine-grained transfer path from the source to the target domain, while the global attention module embedded on the global feature before the

classifier can transfer knowledge under domain variations due to translations, rotations or other transformations. What we still lack is a proper classification loss function that leads the classifier to generate correct predictions. Like most domain adaptation methods, this loss function can be formulated on the source domain labeled data $\mathcal{D}_s$ as

$$L_y = \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s} L_y(G_y(G_b(\boldsymbol{h}_i)), y_i), \tag{8}$$

where $L_y$ is the cross-entropy loss function and $G_y$ is the source classifier used for making final predictions.

We enable effective unsupervised domain adaptation by the Transferable Attention for Domain Adaptation (TADA), which jointly learns transferable features and adaptive classifiers by integrating deep feature learning, global domain adaptation, local domain adaptation and transferable attention mechanism in an end-to-end deep architecture. Finally, the proposed TADA model can be formulated as

$$\begin{aligned} C(\theta_f, \theta_b, \theta_y, \theta_d, \theta_d^k|_{k=1}^K) &= L_y + \gamma L_h - \lambda(L_g + L_l) \\ &= \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s} L_y(G_y(G_b(\boldsymbol{h}_i)), y_i) \\ &- \frac{\gamma}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}} \sum_{j=1}^{C} m_i \cdot \boldsymbol{p}_{i,j} \cdot \log(\boldsymbol{p}_{i,j}) \\ &- \frac{\lambda}{n} [ \sum_{\boldsymbol{x}_i \in \mathcal{D}} L_d(G_d(G_b(\boldsymbol{h}_i), d_i)) \\ &+ \frac{1}{K} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{D}} L_d(G_d^k((G_f(\boldsymbol{x}_i))^k), d_i)] \end{aligned} \tag{9}$$

where $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_t$ and $\gamma, \lambda$ are hyper-parameter that respectively trade-off the attentive entropy objective and domain adaptation objective with the classification objective in the unified optimization problem. The minimax optimization problem is to find the network parameters $\hat{\theta}_f$, $\hat{\theta}_y$, $\hat{\theta}_b$, $\hat{\theta}_d$ and $\hat{\theta}_d^k (k = 1, 2 \ldots, K)$ that jointly satisfy

$$\begin{aligned} (\hat{\theta}_f, \hat{\theta}_b, \hat{\theta}_y) &= \arg\min_{\theta_f, \theta_b, \theta_y} C\left(\theta_f, \theta_b, \theta_y, \theta_d, \theta_d^k|_{k=1}^K\right), \\ (\hat{\theta}_d, \hat{\theta}_d^1, ..., \hat{\theta}_d^K) &= \arg\max_{\theta_d, \theta_d^1, \ldots, \theta_d^K} C\left(\theta_f, \theta_b, \theta_y, \theta_d, \theta_d^k|_{k=1}^K\right). \end{aligned} \tag{10}$$

The overall system can be efficiently trained end-to-end by back-propagation, using the auto-differentiation technique.

## Experiments and Results

We evaluate the proposed Transferable Attention for Domain Adaptation (TADA) model with state of the art domain adaptation methods. Code and datasets will be available at github.com/thuml.

### Setup

**Office-31** (Saenko et al. 2010), a standard benchmark for visual domain adaptation, contains 4,652 images and 31 categories from three distinct domains: *Amazon* (**A**), which contains images downloaded from amazon.com, *Webcam* (**W**) and *DSLR* (**D**). We evaluate all methods across three transfer tasks **A** $\rightarrow$ **W**, **D** $\rightarrow$ **W** and **W** $\rightarrow$ **D**, which are widely used by previous deep transfer learning methods (Tzeng et al. 2014; Ganin and Lempitsky 2015), and another three transfer tasks **A** $\rightarrow$ **D**, **D** $\rightarrow$ **A** and **W** $\rightarrow$ **A** as used by (Tzeng et al. 2015; Long et al. 2015; 2016).

**Office-Home** (Venkateswara et al. 2017) is a more challenging dataset for domain adaptation evaluation. It consists of around 15,500 images in total from 65 categories of everyday objects in office and home settings. There are four significantly different domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**) and Real-World images (**Rw**). The images of these domains have substantially different appearances and backgrounds, and the number of categories is much larger than that of *Office-31*, making it more difficult to transfer across domains.

We follow the standard evaluation protocols for unsupervised domain adaptation (Long et al. 2015; Ganin and Lempitsky 2015). We set $\lambda = 1.0$ and $\gamma = 0.1$ throughout all experiments. Our methods were implemented based on the **PyTorch**, and ResNet-50 (He et al. 2016) models pretrained on the ImageNet dataset (Russakovsky et al. 2014). We fine-tune all convolutional and pooling layers and apply back-propagation to train the classifier layer and all domain discriminators. Whatever module trained from scratch, its learning rate was set to be 10 times that of the lower layers. We adopt mini-batch stochastic gradient descent (SGD) with momentum of 0.95 using the learning rate and progressive training strategies as in (Ganin and Lempitsky 2015).

### Results

The classification accuracies on the *Office-31* dataset for unsupervised domain adaptation based on ResNet-50 are shown in Table 1. For fair comparison, the results of all baselines are directly reported from their original papers wherever available. The TADA model significantly outperforms all comparison methods on most transfer tasks. It is remarkable that TADA promotes the classification accuracies substantially on hard transfer tasks, e.g. **D** $\rightarrow$ **A**, **W** $\rightarrow$ **A**, **A** $\rightarrow$ **W** and **A** $\rightarrow$ **D**, and produce comparable classification performance on easy transfer tasks, **D** $\rightarrow$ **W** and **W** $\rightarrow$ **D**. However, the results for TADA (local) on **D** $\rightarrow$ **A** and **W** $\rightarrow$ **A** are much lower than the existing approaches. As we know, the two domains *Webcam* (W) and *DSLR* (D) have only 795 and 498 images in total for 31 classes respectively. This is not sufficient for learning good local attention, which constitutes fine-grained regional information. Furthermore, it is possible that the domain discriminators might find fewer regions to align due to translations, rotations or other transformations. The global attention successfully remedies this weakness of the local attention, enabling TADA (global+local) to achieve strong results.

As reported in Table 2, the TADA approach overpasses the comparison methods on all transfer tasks on *Office-Home* and improve their accuracy with a larger rooms though the domains in this dataset are with more categories. Some transfer learning tasks are even improved by more than 10

Table 1: Accuracy (%) on *Office-31* for unsupervised domain adaption (ResNet)

| Method | A→W | D→W | W→D | A→D | D→A | W→A | Avg |
|---|---|---|---|---|---|---|---|
| ResNet-50 (He et al. 2016) | 68.4 ± 0.2 | 96.7 ± 0.1 | 99.3 ± 0.1 | 68.9 ± 0.2 | 62.5 ± 0.3 | 60.7 ± 0.3 | 76.1 |
| TCA (Pan et al. 2011) | 72.7 ± 0.0 | 96.7 ± 0.0 | 99.6 ± 0.0 | 74.1 ± 0.0 | 61.7 ± 0.0 | 60.9 ± 0.0 | 77.6 |
| GFK (Gong et al. 2012) | 72.8 ± 0.0 | 95.0 ± 0.0 | 98.2 ± 0.0 | 74.5 ± 0.0 | 63.4 ± 0.0 | 61.0 ± 0.0 | 77.5 |
| DAN (Long et al. 2015) | 80.5 ± 0.4 | 97.1 ± 0.2 | 99.6 ± 0.1 | 78.6 ± 0.2 | 63.6 ± 0.3 | 62.8 ± 0.2 | 80.4 |
| RTN (Long et al. 2016) | 84.5 ± 0.2 | 96.8 ± 0.1 | 99.4 ± 0.1 | 77.5 ± 0.3 | 66.2 ± 0.2 | 64.8 ± 0.3 | 81.6 |
| DANN (Ganin et al. 2016) | 82.0 ± 0.4 | 96.9 ± 0.2 | 99.1 ± 0.1 | 79.7 ± 0.4 | 68.2 ± 0.4 | 67.4 ± 0.5 | 82.2 |
| ADDA (Tzeng et al. 2017) | 86.2 ± 0.5 | 96.2 ± 0.3 | 98.4 ± 0.3 | 77.8 ± 0.3 | 69.5 ± 0.4 | 68.9 ± 0.5 | 82.9 |
| JAN (Long et al. 2017) | 85.4 ± 0.3 | 97.4 ± 0.2 | **99.8** ± 0.2 | 84.7 ± 0.3 | 68.6 ± 0.3 | 70.0 ± 0.4 | 84.3 |
| MADA (Pei et al. 2018) | 90.0 ± 0.1 | 97.4 ± 0.1 | 99.6 ± 0.1 | 87.8 ± 0.2 | 70.3 ± 0.3 | 66.4 ± 0.3 | 85.2 |
| SimNet (Pinheiro 2018) | 88.6 ± 0.5 | 98.2 ± 0.2 | 99.7 ± 0.2 | 85.3 ± 0.3 | **73.4** ± 0.8 | 71.6 ± 0.6 | 86.2 |
| GTA (Sankaranarayanan et al. 2018) | 89.5 ± 0.5 | 97.9 ± 0.3 | 99.8 ± 0.4 | 87.7 ± 0.5 | 72.8 ± 0.3 | 71.4 ± 0.4 | 86.5 |
| **TADA** (local) | 89.4 ± 0.4 | **98.7** ± 0.2 | 99.8 ± 0.2 | 87.2 ± 0.2 | 66.4 ± 0.2 | 65.3 ± 0.3 | 84.5 |
| **TADA** (global) | 92.9 ± 0.4 | 98.2 ± 0.2 | 99.8 ± 0.2 | 88.9 ± 0.2 | 69.6 ± 0.2 | 71.0 ± 0.3 | 86.7 |
| **TADA** (local+global) | **94.3** ± 0.3 | **98.7** ± 0.1 | **99.8** ± 0.2 | **91.6** ± 0.3 | 72.9 ± 0.2 | **73.0** ± 0.3 | **88.4** |

Table 2: Accuracy (%) on *Office-Home* for unsupervised domain adaption (ResNet)

| Method | Ar→Cl | Ar→Pr | Ar→Rw | Cl→Ar | Cl→Pr | Cl→Rw | Pr→Ar | Pr→Cl | Pr→Rw | Rw→Ar | Rw→Cl | Rw→Pr | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ResNet-50 (He et al. 2016) | 34.9 | 50.0 | 58.0 | 37.4 | 41.9 | 46.2 | 38.5 | 31.2 | 60.4 | 53.9 | 41.2 | 59.9 | 46.1 |
| DAN (Long et al. 2015) | 43.6 | 57.0 | 67.9 | 45.8 | 56.5 | 60.4 | 44.0 | 43.6 | 67.7 | 63.1 | 51.5 | 74.3 | 56.3 |
| DANN (Ganin et al. 2016) | 45.6 | 59.3 | 70.1 | 47.0 | 58.5 | 60.9 | 46.1 | 43.7 | 68.5 | 63.2 | 51.8 | 76.8 | 57.6 |
| JAN (Long et al. 2017) | 45.9 | 61.2 | 68.9 | 50.4 | 59.7 | 61.0 | 45.8 | 43.4 | 70.3 | 63.9 | 52.4 | 76.8 | 58.3 |
| **TADA** (local) | 47.3 | 69.1 | 75.2 | 56.9 | 66.4 | 69.1 | 55.9 | 46.9 | 75.7 | 68.2 | 56.2 | 80.4 | 63.9 |
| **TADA** (global) | 51.3 | 66.0 | 76.5 | 58.6 | 69.3 | 70.3 | 58.3 | 52.0 | 77.1 | 70.2 | 57.0 | 81.5 | 65.7 |
| **TADA** (local+global) | **53.1** | **72.3** | **77.2** | **59.1** | **71.2** | **72.1** | **59.7** | **53.1** | **78.4** | **72.4** | **60.0** | **82.9** | **67.6** |

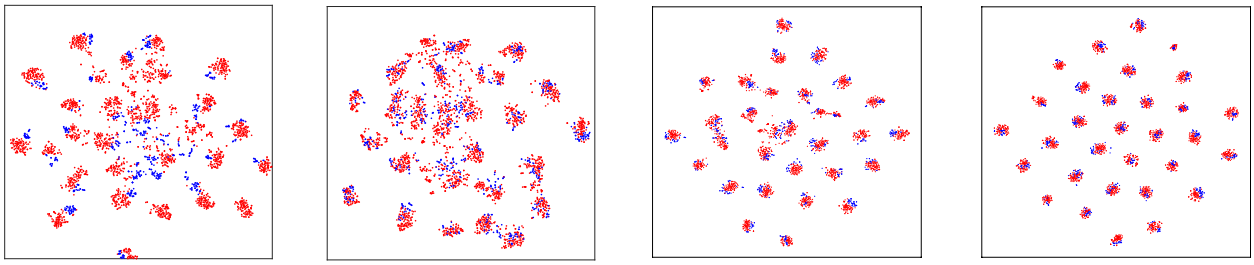
points compared with JAN (Long et al. 2017), such as **Pr** → **Cl**. It is encouraging that TADA yields larger improvements on such difficult transfer learning tasks, which suggests that TADA is able to learn more transferable representations for effective domain adaptation.

## Analysis

**Ablation Study** To tooth apart the separate contributions of the transferable local attention and the transferable global attention modules, we denote by **TADA (local)** the combination of classification, local attention and entropy modules, and by **TADA (global)** the combination of classification, global attention and attentive entropy modules. Experimental results reveal that both TADA (local) and TADA (global) gain significant improvements over baselines but TADA (global) works better on some difficult tasks than TADA (local). The reason is that TADA (global) can better transfer knowledge under domain variations due to translations, rotations or other transformations. What is more, TADA (local+global) also improves with a large room over either TADA (local) or TADA (global), revealing that TADA (local+global) is the most effective while TADA (local) and TADA (global) are well complementary to each other.

**Feature Visualization** To show the feature transferability, we visualize in Figures 2(a)–2(d) the network activations of the bottleneck layer from task **A** → **W** (31 classes) learned by ResNet, DANN, MADA and TADA respectively using t-SNE embeddings (Donahue et al. 2014). From left (ResNet) to right (TADA), the source and target domains are made more and more indistinguishable. In particular, the representations generated by TADA formed exactly 31 clusters with clear boundaries. TADA's better visualization result suggests that it is able to match the complex multimodal structures of the source and target data distributions both globally and locally, thus learning more transferable features for domain adaptation.

**Attention Map Visualization** To verify that the attention map can focus on the desirable regions (in particular, the foreground objects) in the image, we randomly sample some input images from the source domain (**Ar**) and target domain (**Rw**). As shown in Figure 3, different regions in the images have different corresponding attention masks in our network. The hotter the color, the larger the attention value. Take the image on the top-right as an example, the clock mask is highlighted with red color while the background mask diminishes in blue color though the background are complicated enough with person and other messy objects. Meanwhile, only the more transferable regions of the foreground were highlighted. The images in the bottom line indicate that only the flames of the candles are highlighted since they are more transferable than the body of candles. These results intuitively show that the transferable attention mechanism can generate meaningful regions for fine-grained transfer.



(a) ResNet (b) DANN (c) MADA (d) TADA

Figure 2: The t-SNE visualization of features learned by (a) ResNet, (b) DANN, (c) MADA, and (d) TADA (red: **A**; blue: **W**).

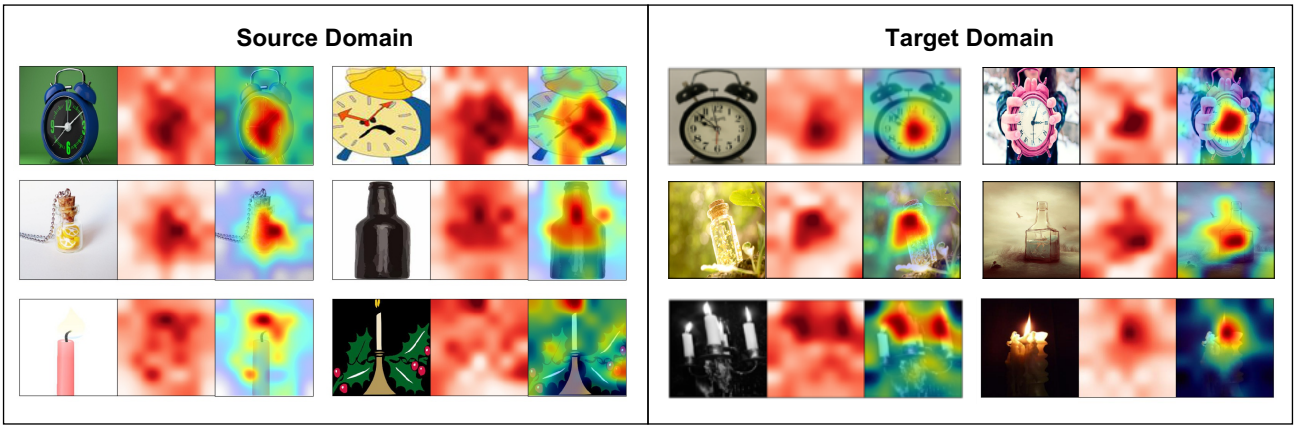



Figure 3: Attention visualization of the last convolutional layer of ResNet on *Office-Home*. The images on the left are randomly sampled from source domain (**Ar**) while the right from target domain (**Rw**). In each group of images, the original input images, the corresponding attentions and the attentions shown in the original input images are illustrated from left to right respectively.

## Conclusion

This paper presented Transferable Attention for Domain Adaptation (TADA), a novel multi-adversarial domain adaptation approach with both global and local attention mechanism. Unlike previous adversarial domain adaptation methods that only match the feature extracted from the entire images across domains, the proposed approach further exploits the complex multimodal structures by considering the transferability of different regions or images. Our approach studies two types of complementary transferable attention: local attention generated by multiple region-level domain discriminators to highlight transferable regions, and global attention generated by single image-level domain discriminator to highlight transferable images. Comprehensive experiments show that the proposed approach outperforms state of the art results on standard domain adaptation datasets.

## Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2016YFB1000701) and Natural Science Foundation of China (61772299, 61502265, 71690231).

## References

Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2016. Mode regularized generative adversarial networks. *CoRR* abs/1612.02136.

Chen, L.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 3640–3649.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. P. 2011. Natural language processing (almost) from scratch. *JMLR* 12:2493–2537.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*.

Duan, L.; Tsang, I. W.; and Xu, D. 2012. Domain transfer multiple kernel learning. *TPAMI* 34(3):465.

Durugkar, I.; Gemp, I.; and Mahadevan, S. 2016. Generative Multi-Adversarial Networks. *arXiv:1611.01673 [cs]*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML - Volume 37*, ICML'15, 1180–1189. JMLR.org.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016.

Domain-adversarial training of neural networks. *JMLR* 17:59:1–59:35.

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

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2066–2073.

Gong, B.; Grauman, K.; and Sha, F. 2013. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 222–230.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Grandvalet, Y., and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *NIPS*, 529–536.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 1994–2003.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; and Schölkopf, B. 2006. Correcting sample selection bias by unlabeled data. In *NIPS*, 601–608.

Li, L.; Tang, S.; Deng, L.; Zhang, Y.; and Tian, Q. 2017. Image caption with global-local attention. In *AAAI*, 4133–4139.

Liu, M., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*, 469–477.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. In *ICML*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 136–144.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*, 2208–2217.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *NIPS*.

Metz, L.; Poole, B.; Pfau, D.; and Sohl-Dickstein, J. 2016. Unrolled Generative Adversarial Networks. *arXiv:1611.02163*.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *CoRR* abs/1411.1784.

Moon, S., and Carbonell, J. G. 2017. Completely heterogeneous transfer learning with attention - what and what not to transfer. In *IJCAI*, 2508–2514.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2009. Multimodal deep learning. In *ICML*, 689–696.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, volume 70, 2642–2651.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 1717–1724.

Pan, S. J., and Yang, Q. 2010. A Survey on Transfer Learning. *TKDE* 22(10):1345–1359.

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

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-adversarial domain adaptation. In *AAAI*.

Pinheiro, P. O. 2018. Unsupervised domain adaptation with similarity learning. In *CVPR*.

Quionero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. *Dataset Shift in Machine Learning*. The MIT Press.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2014. ImageNet Large Scale Visual Recognition Challenge.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*.

Sugiyama, M.; Krauledat, M.; and Ller, K. R. 2007. Covariate shift adaptation by importance weighted cross validation. *JMLR* 8(1):985–1005.

Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*, 1521–1528.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *CoRR* abs/1412.3474.

Tzeng, E.; Hoffman, J.; Darrell, T.; and Saenko, K. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*, 2962–2971.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 6000–6010.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. *arXiv:1706.07522 [cs]*.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *CVPR*, 6450–6458.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *NIPS*, 3320–3328.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2242–2251.