Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers

Hong Liu 1 2 Mingsheng Long 1 3 Jianmin Wang 1 3 Michael I. Jordan 4

Abstract

Domain adaptation enables knowledge transfer from a labeled source domain to an unlabeled target domain. A mainstream approach is adversarial feature adaptation, which learns domain-invariant representations through aligning the feature distributions of both domains. However, a theoretical prerequisite of domain adaptation is the adaptability measured by the expected risk of an ideal joint hypothesis over the source and target domains. In this respect, adversarial feature adaptation may potentially deteriorate the adaptability, since it distorts the original feature distributions when suppressing domain-specific variations. To this end, we propose Transferable Adversarial Training (TAT) to enable the adaptation of deep classifiers. The approach generates transferable examples to fill in the gap between the source and target domains, and adversarially trains the deep classifiers to make consistent predictions over the transferable examples. Without learning domain-invariant representations at the expense of distorting the feature distributions, the adaptability in the theoretical learning bound is algorithmically guaranteed. A series of experiments validate that our approach advances the state of the arts on a variety of domain adaptation tasks in vision and NLP, including object recognition, learning from synthetic to real data, and sentiment classification.

1. Introduction

Transferring knowledge from a source domain with sufficient supervision to an unlabeled target domain is advantageous, since manual annotation for a new machine learning task is often prohibitive. However, deep neural networks are sensitive to cross-domain distribution shift. The same network can make spurious prediction on a target domain dissimilar to the source domain (Quionero-Candela et al., 2009). A notable example is that models trained on labeled synthetic data which come in abundance, may fail when generalizing to real-world unlabeled data (Liu et al., 2017; Hoffman et al., 2018).

Domain adaptation aims at learning an accurate classifier for such a scenario. Recent advances in deep neural networks have enhanced the transferability of feature representations (Yosinski et al., 2014) and the disentanglement of explanatory factors behind data (Bengio et al., 2013). Therefore, a reasonable approach to domain adaptation is harnessing the power of deep neural networks to extract domain-invariant feature representations. One possible way is to minimize some measure of distance between the source and target feature distributions such as maximum mean discrepancy (Long et al., 2015). On par with distance minimizing methods, adversarial domain adaptation incorporates adversarial learning as a two-player game similar to GANs (Goodfellow et al., 2014). In this paradigm, the base network is divided into a feature extractor and a classifier. A domain discriminator is induced to discriminate the source domain from the target domain, while the feature extractor learns domain-invariant representations to fool the domain discriminator (Ganin et al., 2016; Tzeng et al., 2017; Long et al., 2018).

However, this class of techniques face critical restrictions. Based on the domain adaptation theory (Ben-David et al., 2010), an essential prerequisite for domain adaptation is the
Motivation of our approach. (a) Learning from synthetic to real data on VisDA-2017 dataset (Peng et al., 2017). (b) The error of ideal joint hypothesis. (c) The distributions of the singular values of the deep classifier’s adapted weight matrices. Best viewed in color.
3. Transferable Adversarial Training

Consider the problem of unsupervised domain adaptation, with \(n_s\) i.i.d. observations \(\{x_s^{(i)}, y_s^{(i)}\}_{i=1}^{n_s}\) from a source domain of distribution \(P(x_s, y_s)\), and \(n_t\) i.i.d. observations \(\{x_t^{(i)}\}_{i=1}^{n_t}\) from a target domain of distribution \(Q(x_t, y_t)\). Note that the i.i.d. assumption is violated across domains as \(P \neq Q\). Our goal is to adapt a deep category classifier \(y = C(f)\) under the feature representation \(f = F(x)\), which guarantees lower generalization error on the target domain.

In this paper, we present a general approach to adapting deep classifiers across domains with guaranteed adaptability \(\lambda\). The approach, Transferable Adversarial Training (TAT), constitutes two alternating steps: adversarial generation of transferable examples and adversarial training with transferable examples, both without distorting feature distributions.

3.1. Adversarial Generation of Transferable Examples

Existing adversarial feature adaptation methods diminish domain-specific variations by learning domain-invariant representations. Denote by \(f = F(x)\) the feature extractor and by \(d = D(f)\) the domain discriminator. \(D\) and \(F\) form a two-player minimax game: \(D\) is trained to distinguish the source from the target while \(F\) is trained simultaneously to confuse \(D\). However, such a procedure may deteriorate the adaptability \(\lambda\). To guarantee adaptability, we propose to fix the feature representations and generate transferable examples to bridge domain gap. Concretely, we still train the domain discriminator \(D\) to distinguish the source domain from the target domain through the following loss function:

\[
\ell_d(\theta_D, f) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log[D(f_s^{(i)})] - \frac{1}{n_t} \sum_{i=1}^{n_t} \log[1 - D(f_t^{(i)})].
\]

(4)

The deep category classifier \(C\) is also trained to perform well on the source domain through the cross-entropy loss:

\[
\ell_c(\theta_C, f) = \frac{1}{n_s} \sum_{i=1}^{n_s} \ell_{ce}(C(f_s^{(i)}), y_s^{(i)}).
\]

(5)

Different from existing adversarial feature adaptation methods, we diminish the distributional variations by filling the gap between the source and target domains with transferable examples generated in a new adversarial training paradigm. To enable adversarial generation of transferable examples, we compute the gradients of the above loss functions \(\ell_d\) and \(\ell_c\) w.r.t. each example in terms of learned features \(f_s\) and \(f_t\). Note that we opt not to compute the gradients w.r.t. the raw images \(x_s\) and \(x_t\) as in standard adversarial training (Goodfellow et al., 2015). Further, the feature extractor \(F\)
is not updated using these gradients. As analyzed in the previous section, this guarantees good adaptability $\lambda$.

More formally, we propose to generate the transferable examples taking the philosophy of adversarial training. First, the transferable examples should effectively confuse the domain discriminator $D$, such that they can fill in the gap and bridge the source and target domains. Second, the transferable examples should be able to deceive the category classifier $C$, such that they can push the decision boundary away from data points. Hence, the transferable examples are generated adversarially through a joint loss of $\ell_c$ and $\ell_d$:

$$f_{s,k+1} \leftarrow f_{s,k} + \beta \nabla_f \ell_d(\theta_D, f_{s,k})$$

$$f_{s,k+1} \leftarrow f_{s,k} + \beta \nabla_f \ell_d(\theta_D, f_{s,k})$$

$$f_{t,k+1} \leftarrow f_{t,k} + \beta \nabla_f \ell_d(\theta_D, f_{t,k})$$

where $K$ is the number of iterations for generating each transferable example, and $k = 0, 1, \ldots, K - 1$ is the current iteration. Note that $f_{0,s} = f_{t}, f_{0,t} = f_{s}, f_{k,s} = f_{t,k}, f_{k,t} = f_{t,k}$. To generate examples that are sufficiently transferable, we need a sufficient number of iterations, typically $K = 10$. In addition, to avoid divergence of the generated examples, we control the $\ell_2$-distance between the generated examples and the original examples with hyper-parameters $\gamma$ and $\beta$.

### 3.2. Adversarial Training with Transferable Examples

We aim at enabling the robustness of the category classifier $C$ against domain distribution shift. To reach this goal, a reasonable way is to train the classifier to make accurate predictions for the transferable examples $f_{s,*}$ in the source domain. Furthermore, we require the classifier to make consistent predictions for the transferable examples $f_{t,*}$ and their original counterparts $f_t$ in the target domain. As analyzed empirically by Miyato et al. (2018), adversarial training improves local smoothness of the output distribution. Taking similar explanation, training the category classifier with transferable examples can be interpreted as improving the robustness of the classifier’s prior distribution against both adversarial perturbations and domain variations. Note that we have access to labels of the source domain but the target labels are absent. Thus, the loss function for adversarial training of the category classifier $C$ is formulated as follows,

$$\ell_{c, \text{adv}}(\theta_C, f_s) = \frac{1}{n_s} \sum_{i=1}^{n_s} \ell_{ce}(C(f_{s,*}^{(i)}), y_{s,*}^{(i)})$$

$$\ell_{d, \text{adv}}(\theta_D, f_s) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log[D(f_{s,*}^{(i)})]$$

$$\ell_{d, \text{adv}}(\theta_D, f_s) = -\frac{1}{n_t} \sum_{i=1}^{n_t} \log[1 - D(f_{t,*}^{(i)})].$$

Finally, we enable transferable adversarial training (TAT) by generating transferable examples and training classifiers on them. We jointly minimize error (4) and error (9) with
We have analyzed that the adversarial feature adaptation methods weaken the adaptability $\lambda$. TAT fixes the feature extractor $F$ throughout the training procedure to keep $\lambda$ unchanged, which is a complementary improvement to the previous methods. Further, by training the category classifier on the source domain, TAT minimizes error $\epsilon_s(C)$. Most importantly, our approach generates the transferable examples towards the opposite domains by confusing the domain discriminator. And training domain discriminator $D$ against these transferable examples will explicitly bound the $\mathcal{H} \Delta \mathcal{H}$-distance, as justified in Ganin & Lempitsky (2015).

Since the transferable examples will fill in the gap across domains (intuitively demonstrated in Figures 1 and 4), the $d_{\mathcal{H} \Delta \mathcal{H}}(X_s, X_t)$ term is further bounded in our approach. In summary, our approach conforms well with the domain adaptation theory, further yielding guaranteed adaptability.

4.2. Adversarial Training Theory

The proposed approach generates transferable examples, which are essentially the adversarial examples against both category classifier $C$ and domain discriminator $D$. This is in line with the adversarial training theory (Sinha et al., 2018).

**Theorem 2.** (Sinha et al., 2018) Assume $|\ell(\theta, x)| \leq M_\theta$ for all models $\theta \in \Theta$ and examples $x \in \mathcal{X}$. Then, for a fixed $t > 0$ and numerical constants $b_1, b_2 > 0$, with probability at least $1 - e^{-t}$, simultaneously for all $\theta \in \Theta$, $\rho \geq 0, \gamma \geq 0$,

$$\sup_{Q: W(P, Q) \leq \rho} \mathbb{E}_Q[\ell(\theta, X)] \leq \gamma \rho + \mathbb{E}_{P_n}[\phi_\gamma(\theta, X)] + O(1/\sqrt{n}, M_\theta, b_1, b_2),$$

where $W(P, Q)$ is the Wasserstein distance between $P$ and $Q$, $P_n$ is the empirical distribution of $P$, $O(\cdot)$ is a complexity function, and $\phi_\gamma(\theta, X)$ is the robust surrogate loss,

$$\phi_\gamma(\theta, X) = \sup_{x \in \mathcal{X}} \ell(\theta, x) - \gamma W(x_0, x).$$

Our approach trains the category classifier $C$ with the worst-case distributions within distance $\rho$ away from the source domain, which guarantees good performance if the target domain is in the range of distance $\rho$. By introducing hyper-parameters $\beta$ and $\gamma$ in Eq. (6)–(7), our approach executes the adversarial training procedure guaranteed by the theory.

5. Experiments

We evaluate TAT on five domain adaptation datasets. Codes and datasets are made available at github.com/thuml/Transferable-Adversarial-Training.

5.1. Experimental Setup

Office-31 (Saenko et al., 2010) is a standard dataset for visual domain adaptation. It contains 4,652 images across 31 categories from three domains: Amazon (A), Webcam (W), and DSLR (D). From this dataset, we construct 6 transfer tasks: $A \rightarrow W, D \rightarrow W, W \rightarrow D, A \rightarrow D, D \rightarrow A$, and $W \rightarrow A$.

ImageCLEF-DA (Long et al., 2017) is a dataset organized by selecting the 12 common classes shared by three public datasets (domains): Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). We evaluate all methods on 6 transfer tasks: $I \rightarrow P, P \rightarrow I, I \rightarrow C, C \rightarrow I, C \rightarrow P$, and $P \rightarrow C$.  

### Algorithm 1: Transferable Adversarial Training (TAT)

**Input:** original features $(f_s^{(i)}, y_s^{(i)})_{i=1}^n$, and $(f_t^{(i)})_{i=1}^n$.

**Output:** learned model parameters $\theta = (\theta_C, \theta_D)$.

Initialize $\theta = (\theta_C, \theta_D)$ randomly.

for $\text{iter} = 1$ to MaxIter do

Sample a mini-batch of $(\{f_s, y_s\})$ and $\{f_t\}$ uniformly from the training dataset in terms of original features.

for $k = 0$ to $K - 1$ do

$f_{s,k+1} \leftarrow f_s + \beta \nabla_{f_s} \ell_d(\theta_D, f_s) - \gamma \nabla_{\theta_C} \ell_2(f_s, f_o) + \beta \nabla_{\theta_C} \ell_1(\theta_C, f_s)$

$f_{t,k+1} \leftarrow f_t + \beta \nabla_{f_t} \ell_d(\theta_D, f_t) - \gamma \nabla_{\theta_C} \ell_2(f_t, f_o) + \beta \nabla_{\theta_C} \ell_1(\theta_C, f_t)$

end for

$\theta_C \leftarrow \theta_C - \alpha \nabla_{\theta_C} [\ell_C(\theta_C, f_s) + \ell_c,adv(\theta_C, f_s)]$

$\theta_D \leftarrow \theta_D - \alpha \nabla_{\theta_D} [\ell_D(\theta_D, f_s) + \ell_d,adv(\theta_D, f_s)]$

end for

respect to $D$, and error (5) and error (8) with respect to $C$. This leads to the optimization problem for the TAT approach:

$$\min_{\theta_D, \theta_C} \ell_d(\theta_D, f_s) + \ell_c(\theta_C, f_s) + \ell_d,adv(\theta_D, f_s) + \ell_c,adv(\theta_C, f_s).$$

We summarize the detailed training procedure in Algorithm 1. TAT runs over the feature-level examples $f$ and propagates only through the deep classifier $C$ (usually of no more than three layers), which is very computationally efficient (an order of magnitude faster than feature adaptation methods).
We investigate state-of-the-art domain adaptation methods: Adaptation (world images collected from MSCOCO (Lin et al., 2014). VisDA-2017 (Shu et al., 2018). For the sentiment classification tasks, we further compare with classic Marginalized Denoising Autoencoders (mSDA) (Chen et al., 2012) and the state of the art Transfer Denoising Autoencoders (TDA) (Long et al., 2016).

For image datasets, we use ResNet-50 (He et al., 2016) pretrained on ImageNet (Russakovsky et al., 2015) to extract original feature representations. We use Adam (Kingma & Ba, 2014) with initial learning rate $\eta_p = 10^{-4}$. We adopt the inverse-decay strategy of DANN (Ganin et al., 2016), where the learning rate changes by $\eta_p = \frac{\eta_0}{(1 + p)^\phi}$, $\phi = 0.75$, and $p$ is the progress ranging from 0 to 1. We use reverse validation for hyperparameter selection (Zhong et al., 2010). For image datasets, $\beta = 5$ and $\gamma = 1$. For sentiment datasets, we use Bag of Word (BoW) vectors and mSDA (Chen et al., 2012) representations as input.

For the proposed TAT approach, the category classifier is a two-layer fully connected network (2048 × 256 × #classes), and the domain discriminator consists of two fully connected layers with BatchNorm (Ioffe & Szegedy, 2015) and LeakyReLU non-linearity in the first layer. For the adversarial feature adaptation methods, we adopt gradient reversal layers (Ganin et al., 2016) for jointly training the domain discriminator with the category classifier.

### 5.2 Results

We report results of Office-31 based on ResNet-50 in Table 1. The proposed TAT outperforms all comparison methods on most tasks. We clearly observe that on $W \rightarrow A$ and $D \rightarrow A$ with relatively large domain shift and imbalanced domain scales, TAT exceeds all feature adaptation methods by large margins and even performs better than models incorporating complex generative architectures. This further testifies that transferable examples augment the original source domain and therefore mitigate the imbalance between domains.

On ImageCLEF-DA, TAT exceeds comparison methods in all tasks, but the boost is relatively minor, since domain scale is the same and the domain discrepancy is smaller.

When domain discrepancy is significant as on Office-Home,

---

**Table 1.** Classification accuracies (%) on Office-31 for unsupervised domain adaptation with ResNet-50.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>A→W</th>
<th>D→W</th>
<th>W→D</th>
<th>A→D</th>
<th>D→A</th>
<th>W→A</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50 (He et al., 2016)</td>
<td>68.4±0.2</td>
<td>96.7±0.1</td>
<td>99.3±0.1</td>
<td>68.9±0.2</td>
<td>62.5±0.3</td>
<td>60.7±0.3</td>
<td>76.1</td>
</tr>
<tr>
<td>DAN (Long et al., 2015)</td>
<td>80.5±0.4</td>
<td>97.1±0.2</td>
<td>99.6±0.1</td>
<td>78.6±0.2</td>
<td>63.6±0.3</td>
<td>62.8±0.2</td>
<td>80.4</td>
</tr>
<tr>
<td>DANN (Ganin et al., 2016)</td>
<td>82.6±0.4</td>
<td>96.9±0.2</td>
<td>99.3±0.2</td>
<td>81.5±0.4</td>
<td>68.4±0.5</td>
<td>67.5±0.5</td>
<td>82.7</td>
</tr>
<tr>
<td>ADDA (TZENG et al., 2017)</td>
<td>86.2±0.5</td>
<td>96.2±0.3</td>
<td>98.4±0.3</td>
<td>77.8±0.3</td>
<td>69.5±0.4</td>
<td>68.9±0.5</td>
<td>82.9</td>
</tr>
<tr>
<td>VADA (Shi et al., 2018)</td>
<td>86.5±0.5</td>
<td>98.2±0.4</td>
<td>99.7±0.2</td>
<td>86.7±0.4</td>
<td>70.1±0.4</td>
<td>70.5±0.4</td>
<td>85.4</td>
</tr>
<tr>
<td>GTA (Sankaranarayanan et al., 2018)</td>
<td>89.5±0.5</td>
<td>97.9±0.3</td>
<td>99.7±0.2</td>
<td>87.7±0.5</td>
<td>72.8±0.3</td>
<td>71.4±0.4</td>
<td>86.5</td>
</tr>
<tr>
<td>MCD (Saito et al., 2018)</td>
<td>88.6±0.2</td>
<td>98.5±0.1</td>
<td>100.0±0.0</td>
<td>92.2±0.2</td>
<td>69.5±0.1</td>
<td>69.7±0.3</td>
<td>86.5</td>
</tr>
<tr>
<td>CDAN (Long et al., 2018)</td>
<td>93.1±0.1</td>
<td>98.6±0.1</td>
<td>100.0±0.0</td>
<td>92.9±0.2</td>
<td>71.0±0.3</td>
<td>69.3±0.3</td>
<td>87.5</td>
</tr>
<tr>
<td>TAT</td>
<td>92.5±0.3</td>
<td>99.3±0.1</td>
<td>100.0±0.0</td>
<td>93.2±0.2</td>
<td>73.1±0.3</td>
<td>72.1±0.3</td>
<td>88.4</td>
</tr>
</tbody>
</table>

**Table 2.** Classification accuracies (%) on ImageCLEF for unsupervised domain adaptation with ResNet-50.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>I→P</th>
<th>P→I</th>
<th>I→C</th>
<th>C→I</th>
<th>C→P</th>
<th>P→C</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50 (He et al., 2016)</td>
<td>74.8±0.3</td>
<td>83.9±0.1</td>
<td>91.5±0.3</td>
<td>78.0±0.2</td>
<td>65.5±0.3</td>
<td>91.2±0.3</td>
<td>80.7</td>
</tr>
<tr>
<td>DAN (Long et al., 2015)</td>
<td>74.5±0.4</td>
<td>82.2±0.2</td>
<td>92.8±0.2</td>
<td>86.3±0.4</td>
<td>69.2±0.4</td>
<td>89.8±0.4</td>
<td>82.5</td>
</tr>
<tr>
<td>DANN (Ganin et al., 2016)</td>
<td>75.0±0.3</td>
<td>86.0±0.3</td>
<td>96.2±0.4</td>
<td>87.0±0.5</td>
<td>74.3±0.5</td>
<td>91.5±0.6</td>
<td>85.0</td>
</tr>
<tr>
<td>CDAN (Long et al., 2018)</td>
<td>76.7±0.3</td>
<td>90.6±0.3</td>
<td>97.0±0.4</td>
<td>90.5±0.4</td>
<td>74.5±0.3</td>
<td>93.5±0.4</td>
<td>87.1</td>
</tr>
<tr>
<td>TAT</td>
<td>78.8±0.2</td>
<td>92.0±0.2</td>
<td>97.5±0.3</td>
<td>92.0±0.3</td>
<td>78.2±0.4</td>
<td>94.7±0.4</td>
<td>88.9</td>
</tr>
</tbody>
</table>

Office-Home (Venkateswara et al., 2017) is a new dataset more difficult than Office-31. It has 15,500 images across 65 classes in office and home settings from four domains of large domain discrepancy: Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-World images (Rw).

VisDA-2017 (Peng et al., 2017) is the largest dataset to date for visual domain adaptation, providing with two distinct domains. Synthetic: renderings of 3D models from different angles and with different lightning conditions; Real: real-world images collected from MSCOCO (Lin et al., 2014).

Multi-Domain Sentiment (Blitzer et al., 2007) was widely adopted as the benchmark for domain adaptation in sentiment classification. It comprises of product reviews from amazon.com in four product domains: books (B), dvds (D), electronics (E), and kitchen appliances (K). Each review is assigned with a positive or negative polarity and is represented by Bag of Words (BoW) using term frequency.

We investigate state of the art domain adaptation methods: Deep Adaptation Network (DAN) (Long et al., 2015), Domain Adversarial Neural Network (DANN) (Ganin et al., 2016), Adversarial Discriminative Domain Adaptation (ADDA) (TZENG et al., 2017), Virtual Adversarial Domain Adaptation (VADA) (Shi et al., 2018), Generate to Adapt (GTA) (Sankaranarayanan et al., 2018), Maximum Classifier Discrepancy (MCD) (Saito et al., 2018), and Conditional Domain Adversarial Network (CDAN) (Long et al., 2018). For the sentiment classification tasks, we further compare with classic Marginalized Denoising Autoencoders (mSDA) (Chen et al., 2012) and the state of the art Transfer Denoising Autoencoders (TDA) (Long et al., 2016).
Table 3. Classification accuracies (%) on Office-Home for unsupervised domain adaptation (ResNet-50).

<table>
<thead>
<tr>
<th>METHOD</th>
<th>AR→CL</th>
<th>AR→PR</th>
<th>AR→RW</th>
<th>CL→AR</th>
<th>CL→PR</th>
<th>CL→RW</th>
<th>PR→AR</th>
<th>PR→CL</th>
<th>PR→RW</th>
<th>AR→RW</th>
<th>CL→PR</th>
<th>AVER.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESNET-50 (HE ET AL., 2016)</td>
<td>34.9</td>
<td>50.0</td>
<td>58.0</td>
<td>37.4</td>
<td>41.9</td>
<td>46.2</td>
<td>38.5</td>
<td>31.2</td>
<td>60.4</td>
<td>53.9</td>
<td>41.2</td>
<td>59.9</td>
</tr>
<tr>
<td>DAN (LONG ET AL., 2015)</td>
<td>43.6</td>
<td>57.0</td>
<td>67.9</td>
<td>45.8</td>
<td>56.5</td>
<td>60.4</td>
<td>44.0</td>
<td>43.6</td>
<td>67.7</td>
<td>63.1</td>
<td>51.5</td>
<td>74.3</td>
</tr>
<tr>
<td>DAN (GANIN ET AL., 2016)</td>
<td>45.6</td>
<td>59.3</td>
<td>70.1</td>
<td>47.0</td>
<td>58.5</td>
<td>60.9</td>
<td>46.1</td>
<td>43.7</td>
<td>68.5</td>
<td>63.2</td>
<td>51.8</td>
<td>76.8</td>
</tr>
<tr>
<td>CDAN (LONG ET AL., 2018)</td>
<td>49.0</td>
<td>69.3</td>
<td>74.5</td>
<td>54.4</td>
<td>66.0</td>
<td>68.4</td>
<td>55.6</td>
<td>48.3</td>
<td>75.9</td>
<td>68.4</td>
<td>55.4</td>
<td>80.5</td>
</tr>
<tr>
<td>TAT</td>
<td>51.6</td>
<td>69.5</td>
<td>75.4</td>
<td>59.4</td>
<td>69.5</td>
<td>68.6</td>
<td>59.5</td>
<td>50.5</td>
<td>76.8</td>
<td>70.9</td>
<td>56.6</td>
<td>81.6</td>
</tr>
</tbody>
</table>

Table 4. Classification accuracies (%) on Multi-Domain Sentiment Dataset for unsupervised domain adaptation.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>B→D</th>
<th>B→E</th>
<th>B→K</th>
<th>D→B</th>
<th>D→E</th>
<th>D→K</th>
<th>E→B</th>
<th>E→D</th>
<th>E→K</th>
<th>K→B</th>
<th>K→D</th>
<th>K→E</th>
<th>AVER.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (BoW) (CHANG &amp; LIN, 2011)</td>
<td>79.9</td>
<td>74.8</td>
<td>76.9</td>
<td>74.3</td>
<td>74.8</td>
<td>74.6</td>
<td>70.5</td>
<td>72.6</td>
<td>84.7</td>
<td>70.7</td>
<td>73.6</td>
<td>84.2</td>
<td>76.0</td>
</tr>
<tr>
<td>DANN (BoW) (GANIN ET AL., 2016)</td>
<td>78.4</td>
<td>73.3</td>
<td>77.9</td>
<td>72.3</td>
<td>75.4</td>
<td>78.3</td>
<td>71.3</td>
<td>73.8</td>
<td>85.4</td>
<td>70.9</td>
<td>74.0</td>
<td>84.3</td>
<td>76.3</td>
</tr>
<tr>
<td>mSDA (BoW) (CHEN ET AL., 2012)</td>
<td>83.5</td>
<td>74.5</td>
<td>84.6</td>
<td>83.6</td>
<td>79.5</td>
<td>87.1</td>
<td>78.8</td>
<td>80.9</td>
<td>85.3</td>
<td>80.2</td>
<td>82.3</td>
<td>86.9</td>
<td>82.3</td>
</tr>
<tr>
<td>TAT (BoW)</td>
<td>84.5</td>
<td>80.1</td>
<td>83.6</td>
<td>81.9</td>
<td>81.9</td>
<td>84.0</td>
<td>83.2</td>
<td>77.9</td>
<td>90.0</td>
<td>75.8</td>
<td>77.7</td>
<td>88.2</td>
<td>82.4</td>
</tr>
<tr>
<td>DANN (mSDA) (GANIN ET AL., 2016)</td>
<td>82.9</td>
<td>80.4</td>
<td>84.3</td>
<td>82.5</td>
<td>80.9</td>
<td>84.9</td>
<td>71.4</td>
<td>78.1</td>
<td>88.1</td>
<td>71.8</td>
<td>78.9</td>
<td>85.6</td>
<td>81.3</td>
</tr>
<tr>
<td>TDA (mSDA) (LONG ET AL., 2016)</td>
<td>84.1</td>
<td>85.0</td>
<td>87.5</td>
<td>84.9</td>
<td>85.7</td>
<td>88.6</td>
<td>82.0</td>
<td>82.7</td>
<td>87.7</td>
<td>81.5</td>
<td>83.3</td>
<td>86.8</td>
<td>81.5</td>
</tr>
<tr>
<td>TAT (mSDA)</td>
<td>86.8</td>
<td>85.9</td>
<td>88.6</td>
<td>86.4</td>
<td>86.4</td>
<td>89.4</td>
<td>83.7</td>
<td>83.5</td>
<td>90.4</td>
<td>81.4</td>
<td>84.7</td>
<td>89.2</td>
<td>86.3</td>
</tr>
</tbody>
</table>

Table 5. Classification accuracies (%) on VisDA-2017 (ResNet-50).

<table>
<thead>
<tr>
<th>METHOD</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESNET-50 (HE ET AL., 2016)</td>
<td>40.2</td>
</tr>
<tr>
<td>DANN (GANIN ET AL., 2016)</td>
<td>63.7</td>
</tr>
<tr>
<td>MCD (SAITO ET AL., 2018)</td>
<td>69.2</td>
</tr>
<tr>
<td>GTA (SANKARANARAYANAN ET AL., 2018)</td>
<td>69.5</td>
</tr>
<tr>
<td>CDAN (LONG ET AL., 2018)</td>
<td>70.0</td>
</tr>
<tr>
<td>TAT</td>
<td>71.9</td>
</tr>
</tbody>
</table>

We depict the decision boundary of TAT and compare it with the model trained solely on the source domain. We also show the distributions of the transferable examples.

As shown in Figure 4(a), the model trained on the source domain cannot accurately classify the target examples. In contrast, TAT’s decision boundary separates most target examples correctly (Figure 4(b)). In Figure 4(c), we illustrate the distribution of transferable examples. The generated examples are shown to fill in the gap between the source and target domains and thus reduce the domain discrepancy. By enforcing consistent predictions over transferable examples, we drive the decision boundary away from all examples.

Ablation study. We study the strategies for transferable examples generating and transferable adversarial training. The comparison between TAT and its variants are provided in Table 6. TAT (w\o c) and TAT (w\o d) refer to the proposed model without transferable examples generated from the gradient of category classifier $C$ and domain discriminator $D$, respectively. Results indicate that adversarial training towards the category classifier and the domain discriminator are both beneficial to bridging cross-domain discrepancy.

Cross-domain A-distance. As shown in the domain adaptation theory (1), two important factors bound the generalization error: $adaptability \lambda$ and $discrepancy \: d_{H,H'}$. Figure 2 shows that our approach yields the highest adaptability (that of ResNet-50). We further show in Table 7 the cross-domain A-distance (Ben-David et al., 2010), a proxy of $d_{H,H'}$. We compute the A-distance of TAT based on the transferable examples, which turns out to be the smallest in all methods.

6. Related Work

Domain Adaptation Domain adaptation generalizes a model under dataset shift (Pan & Yang, 2010). Moment matching minimizes the distance between feature statistics.
of the source and target. DAN and DDC minimize the maximum mean discrepancies (Long et al., 2015; Tzeng et al., 2014). Zellinger et al. (2017) proposed the central moment discrepancy as a discrepancy measure between distributions.

Table 7. Cross-domain A-distance of different approaches.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>D→W</th>
<th>W→A</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50 (He et al., 2016)</td>
<td>1.27</td>
<td>1.86</td>
</tr>
<tr>
<td>DANN (Ganin et al., 2016)</td>
<td>1.23</td>
<td>1.44</td>
</tr>
<tr>
<td>MCD (Saito et al., 2018)</td>
<td>1.22</td>
<td>1.60</td>
</tr>
<tr>
<td>TAT</td>
<td>1.06</td>
<td>1.04</td>
</tr>
</tbody>
</table>


In addition to enhancing the robustness of deep networks, adversarial training is also promising in a variety of machine learning problems. Miyato et al. (2018) incorporated virtual adversarial training (VAT) in semi-supervised context to smooth the output distributions as a regularization of deep networks. Virtual Adversarial Domain Adaptation (VADA) improves adversarial feature adaptation with VAT and harnesses the cluster assumption (Chapelle & Zien, 2005; Shu et al., 2018). Different from our method, it generates adversarial examples against only the classifier and still performs adversarial feature adaptation. Volpi et al. (2018) explored adversarial training in domain generalization scenarios.

7. Conclusion

We present a general approach, transferable adversarial training, to adapting deep classifiers. By deceiving both category classifier and domain discriminator, the approach generates transferable examples which bridge the gap across domains. Both high adaptability and small distribution discrepancy expected by the domain adaptation theory are achieved by the approach, as justified on both vision and NLP datasets.

8. Acknowledgements

This work was supported by the National Natural Science Foundation of China (61772299, 71690231, and 61672313).
References


Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers


