Characterizing and Avoiding Negative Transfer

Zirui Wang, Zihang Dai, Barnabás Póczos, Jaime Carbonell
Carnegie Mellon University
{ziruiw,dzihang,bapoczos,jgc}@cs.cmu.edu

Abstract

When labeled data is scarce for a specific target task, transfer learning often offers an effective solution by utilizing data from a related source task. However, when transferring knowledge from a less related source, it may inversely hurt the target performance, a phenomenon known as negative transfer. Despite its pervasiveness, negative transfer is usually described in an informal manner, lacking rigorous definition, careful analysis, or systematic treatment. This paper proposes a formal definition of negative transfer and analyzes three important aspects thereof. Stemming from this analysis, a novel technique is proposed to circumvent negative transfer by filtering out unrelated source data. Based on adversarial networks, the technique is highly generic and can be applied to a wide range of transfer learning algorithms. The proposed approach is evaluated on six state-of-the-art deep transfer methods via experiments on four benchmark datasets with varying levels of difficulty. Empirically, the proposed method consistently improves the performance of all baseline methods and largely avoids negative transfer, even when the source data is degenerate.

1. Introduction

The development of deep neural networks (DNNs) has improved the state-of-the-art performance on a wide range of machine learning problems and applications. However, DNNs often require a large amount of labeled data to train well-generalized models and as more classical methods, DNNs rely on the assumption that training data and test data are drawn from the same underlying distribution. In some cases, collecting large volumes of labeled training data is expensive or even prohibitive. Transfer learning [20] addresses this challenge of data scarcity by utilizing previously-labeled data from one or more source tasks. The hope is that this source domain is related to the target domain and thus transferring knowledge from the source can improve the performance within the target domain. This powerful paradigm has been studied under various settings [35] and has been proven effective in a wide range of applications [39, 16, 17].

However, the success of transfer learning is not always guaranteed. If the source and target domains are not sufficiently similar, transferring from such weakly related source may hinder the performance in the target, a phenomenon known as negative transfer. The notion of negative transfer has been well recognized within the transfer learning community [20, 35]. An early paper [24] has conducted empirical study on a simple binary classification problem to demonstrate the existence of negative transfer. Some more recent work [7, 10, 3] has also observed similar negative impact while performing transfer learning on more complex tasks under different settings.

Despite these empirical observations, little research work has been published to analyze or predict negative transfer, and the following questions still remain open: First, while the notion being quite intuitive, it is not clear how negative transfer should be defined exactly. For example, how should we measure it at test time? What type of baseline should we compare with? Second, it is also unknown what factors cause negative transfer, and how to exploit them to determine that negative transfer may occur. Although the divergence between the source and target domain is certainly crucial, we do know how large it must be for negative transfer to occur, nor if it is the only factor. Third and most importantly, given limited or no labeled target data, how to detect and/or avoid negative transfer.

In this work, we take a step towards addressing these questions. We first derive a formal definition of negative transfer that is general and tractable in practice. Here tractable means we can explicitly measure its effect given the testing data. This definition further reveals three underlying factors of negative transfer that give us insights on when it could occur. Motivated by these theoretical observations, we develop a novel and highly generic technique based on adversarial networks to combat negative transfer. In our approach, a discriminator estimating both marginal and joint distributions is used as a gate to filter potentially harmful source data by reducing the bias between source and target risks, which corresponds to the idea of importance reweighting [5, 38]. Our experiments
involving eight transfer learning methods and four benchmark datasets reveal the three factors of negative transfer. In addition, we apply our method to six state-of-the-art deep methods and compare their performance, demonstrating that our approach substantially improves the performance of all base methods under potential negative transfer conditions by largely avoiding negative transfer.

2. Related Work

Transfer learning [20, 36] uses knowledge learned in the source domain to assist training in the target domain. Early methods exploit conventional statistical techniques such as instance weighting [14] and feature mapping [19, 32]. Compared to these earlier approaches, deep transfer networks achieve better results in discovering domain invariant factors [37]. Some deep methods [16, 27] transfer via distribution (mis)match measurements such as Maximum Mean Discrepancy (MMD) [14]. More recent work [9, 29, 3, 26] exploit generative adversarial networks (GANs) [12] and add a subnetwork as a domain discriminator. These methods achieve state-of-the-art on computer vision tasks [26] and some natural language processing tasks [17]. However, none of these techniques are specifically designed to tackle the problem of negative transfer.

Negative transfer Early work that noted negative transfer [24] was targeted at simple classifiers such as hierarchical Naive Bayes. Later, similar negative effects have also been observed in various settings including multi-source transfer learning [7], imbalanced distributions [10] and partial transfer learning [3]. While the importance of detecting and avoiding negative transfer has raised increasing attention [35], the literature lacks in-depth analysis.

3. Rethink Negative Transfer

Notation. We will use $P_S(X, Y)$ and $P_T(X, Y)$, respectively, to denote the joint distribution in the source and the target domain, where $X$ is the input random variable and $Y$ the output. Following the convention, we assume having access to labeled source set $S = \{(x_i^s, y_i^s)\}_{i=1}^n$ sampled from the source joint $P_S(X, Y)$, a labeled target set $T_l = \{(x_j^l, y_j^l)\}_{j=1}^m$ drawn from the target joint $P_T(X, Y)$, and an unlabeled target set $T_u = \{x_k^u\}_{k=1}^n$ from the target marginal $P_T(X)$. For convenience, we define $T = (T_l, T_u)$.

Transfer Learning. Under the notation, transfer learning aims at designing an algorithm $A$, which takes both the source and target domain data $S, T$ as input, and outputs a better hypothesis (model) $h = A(S, T)$, compared to only using the target-domain data. For model comparison, we will adapt the standard expected risk, which is defined as

$$R_P_T(h) := \mathbb{E}_{x, y \sim P_T}[\ell(h(x), y)],$$

with $\ell$ being the specific task loss. To make the setting meaningful, it is often assumed that $n_s \gg n_t$.

Negative Transfer. The notion of negative transfer lacks a rigorous definition. A widely accepted description of negative transfer [20, 35] is stated as “transferring knowledge from the source can have a negative impact on the target learner”. While intuitive, this description conceals many critical factors underlying negative transfer, among which we stress the following three points:

1. Negative transfer should be defined w.r.t. the algorithm. Specifically, the informal description above does not specify what the negative impact is compared with. For example, it will be misleading to only compare with the best possible algorithm only using the target data, i.e., defining negative transfer as

$$R_{P_T}(A(S, T)) > \min_{A'} R_{P_T}(A'(\emptyset, T)),$$

because the increase in risk may not come from using the source-domain data, but the difference in algorithms. Therefore, to study negative transfer, one should focus on a specific algorithm at a time and compare its performance with and without the source-domain data. Hence, we define the negative transfer condition (NTC)\(^1\) for any algorithm $A$ as

$$R_{P_T}(A(S, T)) > R_{P_T}(A(\emptyset, T)).$$

For convenience, we also define the negative transfer gap (NTG) as a quantifiable measure of negative transfer:

$$R_{P_T}(A(S, T)) - R_{P_T}(A(\emptyset, T)),$$

and we say that negative transfer occurs if the negative transfer gap is positive and vice versa.

2. Divergence between the joint distributions is the root to negative transfer. As negative transfer is algorithm specific, it is natural to ask the question that whether there exists a transfer learning algorithm that can always improve the expected risk compared to its target-domain only baseline. It turned out this depends on the divergence between $P_S(X, Y)$ and $P_T(X, Y)$ [11]. As an extreme example, assume $P_S(X) = P_T(X)$ and $P_S(Y \mid x)$ is uniform for any $x$. In the case, there is no meaningful knowledge in $P_S(X, Y)$ at all. Hence, exploiting $S \sim P_S(X, Y)$ will almost surely harm the estimation of $P_T(Y \mid X)$, unless $P_T(Y \mid X)$ is uniform.

In practice, we usually deal with the case where there exists some “systematic similarity” between $P_S(X, Y)$ and $P_T(X, Y)$. Then, an ideal transfer would figure out and take advantage of the similar part, leading to improved performance. However, if an algorithm fails to discard the divergent part and instead rely on it, one can expect

\(^1\)More discussion in the supplementary.
negative transfer to happen. Thus, regardless of the algorithm choice, the distribution shift is the actual root to negative transfer.

3. Negative transfer largely depends on the size of the labeled target data. While the previous discussion focuses on the distribution level, an overlooked factor of negative transfer is the size of the labeled target data, which has a mixed effect.

On one hand, for the same algorithm and distribution divergence, NTC depends on how well the algorithm can do using target data alone, i.e. the RHS of Eq.(3). In zero-shot transfer learning\(^2\) [8, 21] where there is no labeled target data \((n_l = 0)\), only using unlabeled target data would result in a weak random model and thus NTC is unlikely to be satisfied. When labeled target data is available [24, 29, 17], a better target-only baseline can be obtained using semi-supervised learning methods and so negative transfer is relatively more likely to occur. At the other end of the spectrum, if there is an abundance of labeled target data, then transferring from an even slightly different source domain could hurt the generalization. Thus, this shows that negative transfer is relative.

On the other hand, the amount of labeled target data has a direct effect on the feasibility and reliability of discovering shared regularity between the joint distributions. As discussed above, the key component of a transfer learning algorithm is to discover the similarity between the source joint \(P_S(X, Y)\) and the target joint \(P_T(X, Y)\). When labeled target data is not available \((n_l = 0)\), one has to resort to the similarity between the marginals \(P_S(X)\) and \(P_T(X)\), which though has a theoretical limitation [2]. In contrast, if one has a considerable number of samples \((x_l, y_l) \sim P_T(X, Y)\) and \((x_u, y_u) \sim P_S(X, Y)\), the problem would be manageable. Therefore, an ideal transfer learning algorithm may be able to utilize labeled target data to mitigate the negative impact of unrelated source information.

With these points in mind, we next turn to the problem of how to avoid negative transfer in a systematic way.

4. Proposed Method

As discussed in Section 3, the key to achieving successful transfer and avoiding negative effects is to discover and exploit shared underlying structures between \(P_S(X, Y)\) and \(P_T(X, Y)\). In practice, there are many possible regularities one may take advantage of. To motivate our proposed method, we first review an important line of work and show how the observation in section 3 helps us to identify the limitation.

4.1. Domain Adversarial Network

As a notable example, a recent line of work [16, 8, 30] has successfully utilized a domain-invariant feature space assumption to achieve knowledge transfer. Specifically, it is assumed that there exists a feature space that is both shared by both source and target domains and discriminative enough for predicting the output. By learning a feature extractor \(F\) that can map both the source and target input to the same feature space, classifier learned on the source data can transfer to the target domain.

To find such a feature extractor, a representative solution is the Domain Adversarial Neural Network (DANN) [9], which exploits a generative adversarial network (GAN) framework to train the feature extractor \(F\) such that the feature distributions \(P(F(X_s))\) and \(P(F(X_T))\) cannot be distinguished by the discriminator \(D\). Based on the shared feature space, a simple classifier \(C\) is trained on both source and target data. Formally, the objective can be written as:

\[
\begin{align*}
\argmin_{F,C} \argmax_{D} & \mathcal{L}_{CLF}(F, C) - \mu \mathcal{L}_{ADV}(F, D), \\
\mathcal{L}_{CLF}(F, C) & = \mathbb{E}_{x_l,y_l \sim \mathcal{T}_l} [\ell_{CLF}(C(F(x_l)), y_l)] + \mathbb{E}_{x_u,y_u \sim \mathcal{S}} [\ell_{CLF}(C(F(x_u)), y_u)], \\
\mathcal{L}_{ADV}(F, D) & = \mathbb{E}_{x_u \sim \mathcal{T}_l}(x) [\log D(F(x_u))] + \mathbb{E}_{x_u \sim \mathcal{S}}(x) [\log (1 - D(F(x_u))].
\end{align*}
\]

Intuitively, \(\mathcal{L}_{CLF}\) is the supervised classification loss on both the target and source labeled data, \(\mathcal{L}_{ADV}\) is the standard GAN loss treating \(F(x_u)\) and \(F(x_s)\) as the true and fake features respectively, and \(\mu\) is a hyper-parameter balancing the two terms. For more details and theoretical analysis, we refer readers to the original work [8].

Now, notice that the DANN objective implicitly makes the following assumption: For any \(x_s \in \mathcal{X}_s\), there exists a \(x_t \in \mathcal{X}_t\) such that

\[
P_S(Y | x_s) = P_T(Y | x_t) = P(Y | F(x_s)) = P(Y | F(x_t)).
\]

In other words, it is assumed that every single source sample can provide meaningful knowledge for transfer learning. However, as we have discussed in Section 3, some source samples may not be able to provide any knowledge at all. Consider the case where there is a source input \(x_s \in \mathcal{X}_s\) such that \(P_S(Y | x_s) \neq P_T(Y | x_t)\) for any \(x_t\). Since \(P(F(X_s)) = P(F(X_t))\) as a result of the GAN objective, there exists a \(x' \in \mathcal{X}_t\) such that \(F(x') = F(x_s)\) and hence \(P(Y | F(x')) = P(Y | F(x_s))\). Then, if \(P(Y | F(x_s))\) is trained on the source data to match \(P_S(Y | x_s)\), it follows

\[
P(Y | F(x')) = P(Y | F(x_s)) = P(Y | x_s) \neq P(Y | x').
\]

As a result, relying on such “unrelated” source samples can hurt the performance, leading to negative transfer. Motivated by this limitation, we next present a simple yet ef-
Gradient Reversal Layer and location loss, appropriate weight, notice that the standard supervised learning eliminates the weakness, a natural idea is to reweight each assumption that all source samples are equally useful. To take both effective method to deal with harmful source samples in a systematic way.

4.2. Discriminator Gate

The limitation of DANN comes from the unnecessary assumption that all source samples are equally useful. To eliminate the weakness, a natural idea is to reweight each source sample in some proper manner. To derive an appropriate weight, notice that the standard supervised learning objective can be rewritten as

\[
\mathcal{L}^{\text{SUP}} = \mathbb{E}_{x, y \sim P_{T}(X, Y)} \left[ \mathcal{C}_{L}(C(F(x)), y) \right] \\
= \mathbb{E}_{x, y \sim P_{S}(X, Y)} \left[ \frac{P_{T}(x, y)}{P_{S}(x, y)} \mathcal{C}_{L}(C(F(x)), y) \right] \tag{8}
\]

where the density ratio naturally acts as an importance weight \cite{6, 38} for the source data. Hence, the problem reduces to the classic problem of density ratio estimation.

Here, we exploit a GAN discriminator to perform the density ratio estimation \cite{31}. Specifically, the discriminator takes both \(x\) and the paired \(y\) as input, and try to classify whether the pair is from the source domain (fake) or the target domain (true). At any point, the optimal discriminator is given by \(D(x, y) = \frac{P_{T}(x, y)}{P_{T}(x, y) + P_{S}(x, y)}\), implies

\[
P_{T}(x, y) = \frac{D(x, y)}{1 - D(x, y)}.
\]

In our implementation, to save model parameters, we reuse the feature extractor to obtain the feature of \(x\) and instantiate \(D(x, y) = D(F(x), y)\). With the weight ratio, we modify the classification objective \(6)\) in DANN as

\[
\mathcal{L}^{\text{CLF}}_{\text{CLF}}(C, F) = \mathbb{E}_{x_{i}, y_{i} \sim T_{L}} \left[ \mathcal{C}_{L}(C(F(x_{i})), y_{i}) \right] \\
+ \lambda \mathbb{E}_{x_{i}, y_{i} \sim S} \left[ \omega(x_{i}, y_{i}) \mathcal{C}_{L}(C(F(x_{i})), y_{i}) \right], \tag{9}
\]

where \(\omega(x_{i}, y_{i}) = \text{SG}(\frac{D(x_{i}, y_{i})}{1 - D(x_{i}, y_{i})})\)

where \(\text{SG}(\cdot)\) denotes stop gradient and \(\lambda\) is another hyperparameter introduce to scale the density ratio. As the density ratio acts like a gating function, we will refer to mechanism as discriminator gate.

On the other hand, we also augment the adversarial learning objective \(7)\) by incorporating terms for matching the joint distributions:

\[
\mathcal{L}^{\text{adv}}_{\text{ADV}}(F, D) = \mathbb{E}_{x_{i} \sim P_{T}(X)} [\log D(F(x_{i}), n i l)] \\
+ \mathbb{E}_{x_{i} \sim P_{S}(X)} [\log(1 - D(F(x_{i}), n i l))] \\
+ \mathbb{E}_{x_{i}, y_{i} \sim T_{L}} [\log D(F(x_{i}), y_{i})] \\
+ \mathbb{E}_{x_{i}, y_{i} \sim S} [\log(1 - D(F(x_{i}), y_{i}))], \tag{10}
\]

where \(\text{nil}\) denotes a dummy label which does not provide any label information and it is included to enable the discriminator \(D\) being used as both a marginal discriminator and a joint discriminator. As a benefit, the joint discriminator can utilize unlabeled target data since labeled data could be scarce. Similarly, under this objective, the feature network \(F\) will receive gradient from both the marginal discriminator and the joint discriminator. Theoretically speaking, the joint matching objective subsumes the the marginal matching objective, as matched joint distribution implied matched marginals. However, in practice, the labeled target data \(T_{L}\) is usually limited, making the joint matching objective itself insufficient. This particular design choice echoes our discussion about how the size of labeled target data can influence our algorithm design in Section 3.

Combining the gated classification objective \(9)\) and the augmented adversarial learning objective \(10)\), we arrive at our proposed approach to transfer learning

\[
\min_{F, C} \max_{D} \mathcal{L}^{\text{CLF}}_{\text{CLF}}(C, F) - \mu \mathcal{L}^{\text{adv}}_{\text{ADV}}(F, D). \tag{11}
\]

The overall architecture is illustrated in Figure 1. Finally, although the presentation of the proposed method is based on DANN, our method is highly general and can be applied directly to other adversarial transfer learning methods. In fact, we can even extend non-adversarial methods to achieve similar goals. In our experiments, we adapt six deep methods \cite{16, 27, 8, 30, 4, 26} of three different categories to demonstrate the effectiveness of our method.

5. Experiments

We conduct extensive experiments on four benchmark datasets to (1) analyze negative transfer and its three underlying aspects, and (2) evaluate our proposed discriminator gate on six state-of-the-art methods.

5.1. Datasets

We use four standard datasets with different levels of difficulties: (1) small domain shift: Digits dataset, (2) moderate domain shift: Office-31 dataset, and (3) large domain shift: Office-Home and VisDA datasets.
Digits contains three standard digit classification datasets: MNIST, USPS, SVHN. Each dataset contains large amount of images belonging to 10 classes (0-9). This dataset is relatively easy due to its simple data distribution and therefore we only consider a harder case: SVHN→MNIST. Specifically, SVHN [18] contains 73K images cropped from house numbers in Google Street View images while MNIST [15] consists of 70K handwritten digits captured under constrained conditions.

Office-31 [25] is the most widely used dataset for visual transfer learning. It contains 4,652 images of 31 categories from three domains: Amazon(A) which contains images from amazon.com, Webcam(W) and DSLR(D) which consist of images taken by web camera and SLR camera. We evaluate all methods across three tasks: W→D, A→D, and D→A. We select these three settings because the other three possible cases yield similar results.

Office-Home [33] is a more challenging dataset that consists of about 15,500 images of 65 categories that crawled through several search engines and online image directories. In particular, it contains four domains: Artistic images(Ar), Clip Art(Cl), Product images(Pr) and Real-World images(Rw). We want to test on more interesting and practical transfer learning tasks involving adaptation from synthetic to real-world and thus we consider three transfer tasks: Ar→Rw, Cl→Rw, and Pr→Rw. In addition, we choose to use the first 25 categories in alphabetic order to make our results more comparable to previous studies [4].

VisDA [22] is another challenging synthetic to real dataset. We use the training set as the synthetic source and the testing set as the real-world target (Synthetic→Real). Specifically, the training set contains 152K synthetic images generated by rendering 3D models and the testing set contains 72K real images from crops of Youtube Bounding Box dataset [23], both contain 12 categories.

5.2. Experimental Setup

To better study negative transfer effect and evaluate our approach, we need to control the three factors discussed in Section 3, namely algorithm factor, divergence factor and target factor. In our experiments, we adopt the following mechanism to control each of them.

Divergence factor: Since existing benchmark datasets usually contain domains that are similar to each other, we need to alter their distributions to better observe negative transfer effect. In our experiments, we introduce two perturbation rates, a, and b, to respectively control the marginal divergence and the conditional divergence between two domains. Specifically, for each source domain data we independently draw a Bernoulli variable of probability a, and if it returns one, we add a series of random noises to the input image such as random rotation, random salt&pepper noise, random flipping, etc (examples shown Figure 2). According to studies in [28, 1], such perturbation is enough to cause misclassification for neural networks and therefore is sufficient for our purpose. In addition, we draw a second independent Bernoulli variable of probability b, and assign a randomly picked label if it returns one.

![Example images before & after perturbation](image)

Target factor: Similar to previous works, we use all labeled source data for training. For the target data, we first split 50% as training set and the rest 50% for testing. In addition, we use all of target training data as unlabeled target data and use Ly percent of them as labeled target data. A symmetric study of source data can be found in [34].

Algorithm factor: To provide a more comprehensive study of negative transfer, we evaluate the performance of eight transfer learning methods of five categories: TCA [19], KMM [14], DAN [16], DCORAL [27], DANN a.k.a RevGrad [8], ADDA [29], PADA [4], GTA [26]. Specifically, (1) TCA is a conventional method based on MMD-regularized PCA, (2) KMM is a conventional sample reweighting method, (3) DAN and DCORAL are non-adversarial deep methods which use a distribution measurement as an extra loss, (4) DANN, ADDA and PADA use adversarial learning and directly train a discriminator, (5) GTA is a GAN based method that includes a generator to generate actual images in additional to the discriminator. We mainly follow the default settings and training procedures for model selection as explained in their respective papers. However, for fair comparison, we use the same feature extractor and classifier architecture for all deep methods. In particular, we use a modified LeNet as detailed in [26] for the Digits dataset. For other datasets, we fine-tune from the ResNet-50 pretrained on ImageNet with an added 256-dimension bottleneck layer between the res5c and fc layers. To compare the performance of our proposed approach, we adapt a gated version for each of the six deep methods (e.g. DANNgate is the gated DANN). Specifically, we extend DANN, ADDA and PADA straightforwardly as described in Section 4.2. For GTA, we extend the discriminator to take in class labels and output domain label predictions as gates. For DAN and DCORAL, we add an extra discriminator network to be used as gates but the general network is not trained adversarially. For hyper-parameters, we set = 1 and μ progressively increased from 0 to 1 in all our experiments. For each transfer task, we compare the average classification accuracy over five random repeats. To
Table 1. Classification accuracy (%) of DANN and DANN\textsubscript{gate} on tasks W→D and A→D. Perturbation rates are set equal, i.e. $\epsilon = \epsilon_x = \epsilon_y$. NTG\textsubscript{1} and NTG\textsubscript{2} are negative transfer gaps for DANN and DANN\textsubscript{gate}. $\Delta$ is the performance gain of DANN\textsubscript{gate} compared to DANN.

<table>
<thead>
<tr>
<th>W→D</th>
<th>A→D</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$</td>
<td>$\epsilon_x$</td>
<td>$\epsilon_y$</td>
</tr>
<tr>
<td>DANN</td>
<td>99.1±0.3</td>
<td>81.2±2.1</td>
</tr>
<tr>
<td>NTG\textsubscript{1}</td>
<td>-96.6</td>
<td>-80.5</td>
</tr>
<tr>
<td>DANN\textsubscript{gate}</td>
<td>98.9±0.6</td>
<td>83.3±2.1</td>
</tr>
<tr>
<td>NTG\textsubscript{2}</td>
<td>-96.3</td>
<td>-80.4</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>10.2</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Table 3. Classification accuracy (%) under the Covariate Shift assumption on task W→D. $\epsilon_y$ is fixed at 0. Negative transfer gap is shown in brackets.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\epsilon_x$</th>
<th>$L_{%}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DANN</td>
<td>81.3±(29.3)</td>
<td>83.8±(6.2)</td>
</tr>
<tr>
<td>DANN</td>
<td>83.5±(30.3)</td>
<td>86.1±(6.5)</td>
</tr>
<tr>
<td>GTA</td>
<td>85.5±(33.5)</td>
<td>88.1±(8.0)</td>
</tr>
</tbody>
</table>

5.3. Results and Analysis

5.3.1 Study of Negative Transfer

To reveal the three dependent factors, we study the effect of negative transfer under different methods with varying perturbation rates ($\epsilon_x$, $\epsilon_y$) and target labeled data ($L_{\%}$).

**Divergence factor.** The performance of DANN under different settings of $\epsilon$ and $L_{\%}$ on two tasks of Office-31 are shown in Table 1. We observe an increasing negative transfer gap as we increase the perturbation rate in all cases. In some cases such as $L_{\%} = 10\%$, we can even observe a change in the sign of NTG. For a more fine-grained study, we investigate a wider spectrum of distribution divergence by gradually increasing $\epsilon$ from 0.0 to 1.0 in Figure 3(a). Although DANN is better than DANN\textsubscript{gate} when $\epsilon$ is small, its performance degrades quickly as $\epsilon$ increases and drops below DANN\textsubscript{T}, indicating the occurrence of negative transfer.

On the other hand, by fixing $\epsilon_y = 0$ and using two domains $W$ and $D$ that are known to be particularly similar, we study negative transfer under the assumption of covariate shift in Table 3, and observe that negative transfer does *not* occur even with high $\epsilon_x$ and descent $L_{\%}$. These experimental results confirm that the distribution divergence is an important factor of negative transfer.

**Target factor.** Fixing a specific $\epsilon$, we observe that the negative transfer gap increases as $L_{\%}$ increases in Table 1. In the extreme case of unsupervised adaptation ($L_{\%} = 0\%$), NTG stays negative even if two domains are far apart ($\epsilon = 0.9$). In Figure 3(b), we fix $\epsilon = 0.2$ and plot the performance curve as $L_{\%}$ increases. We can see that while both DANN and DANN\textsubscript{T} perform better with more labeled target data, DANN is affected by the divergence factor and outperformed by DANN\textsubscript{T} when $L_{\%}$ becomes larger. This observation shows that negative transfer is relative and it depends on target labeled data.

**Algorithm factor.** In Table 2, we compare the results...
Table 2. Classification accuracy (%) of state-of-the-art methods on four benchmark datasets with negative transfer gap shown in brackets. Perturbation rates are fixed at $\epsilon_x = \epsilon_y = 0.7$. Target labeled ratio is set at $L_y = 10\%$ and we further enforce each task to use at most 3 labeled target samples per class.

<table>
<thead>
<tr>
<th>Method</th>
<th>SVHN→MNIST</th>
<th>W→D</th>
<th>A→D</th>
<th>D→A</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA[19]</td>
<td>58.7(18.2)</td>
<td>54.2(4.2)</td>
<td>11.4(20.5)</td>
<td>13.1(18.4)</td>
<td>-</td>
</tr>
<tr>
<td>KMM[4]</td>
<td>70.8(6.9)</td>
<td>58.7(4.5)</td>
<td>18.5(13.4)</td>
<td>17.7(13.4)</td>
<td>-</td>
</tr>
<tr>
<td>DAN[16]</td>
<td>78.5(4.4)</td>
<td>76.3(19.5)</td>
<td>35.0(1.3)</td>
<td>39.2(4.9)</td>
<td>32.0(8.0)</td>
</tr>
<tr>
<td>DAN_gate</td>
<td>82.2(8.1)</td>
<td>78.7(21.9)</td>
<td>60.4(6.7)</td>
<td>43.9(9.2)</td>
<td>37.6(7.8)</td>
</tr>
<tr>
<td>ΔDAN</td>
<td>$\uparrow$3.7</td>
<td>$\uparrow$2.4</td>
<td>$\uparrow$5.4</td>
<td>$\uparrow$4.7</td>
<td>$\uparrow$7.8</td>
</tr>
<tr>
<td>DCORAL[27]</td>
<td>75.1(1.2)</td>
<td>75.7(18.9)</td>
<td>53.8(0.4)</td>
<td>37.4(5.0)</td>
<td>34.9(12.1)</td>
</tr>
<tr>
<td>DCORAL_gate</td>
<td>81.7(7.0)</td>
<td>78.2(21.4)</td>
<td>59.0(5.6)</td>
<td>43.2(8.8)</td>
<td>38.0(10.6)</td>
</tr>
<tr>
<td>ΔDCORAL</td>
<td>$\uparrow$5.8</td>
<td>$\uparrow$2.5</td>
<td>$\uparrow$5.2</td>
<td>$\uparrow$5.8</td>
<td>$\uparrow$5.3</td>
</tr>
<tr>
<td>DANN[18]</td>
<td>68.3(7.7)</td>
<td>75.0(19.2)</td>
<td>51.0(2.3)</td>
<td>38.2(5.6)</td>
<td>29.9(6.0)</td>
</tr>
<tr>
<td>DANN_gate</td>
<td>78.1(2.1)</td>
<td>80.2(24.4)</td>
<td>61.8(8.5)</td>
<td>48.3(4.5)</td>
<td>45.0(4.6)</td>
</tr>
<tr>
<td>ΔDANN</td>
<td>$\uparrow$7.9</td>
<td>$\uparrow$5.2</td>
<td>$\uparrow$10.8</td>
<td>$\uparrow$10.1</td>
<td>$\uparrow$10.4</td>
</tr>
<tr>
<td>ADDA[30]</td>
<td>63.2(12.2)</td>
<td>74.5(18.1)</td>
<td>49.9(2.2)</td>
<td>38.3(5.1)</td>
<td>38.3(5.1)</td>
</tr>
<tr>
<td>ADDA_gate</td>
<td>79.4(4.0)</td>
<td>82.9(26.5)</td>
<td>64.2(12.1)</td>
<td>47.7(4.3)</td>
<td>47.7(4.3)</td>
</tr>
<tr>
<td>ΔADDA</td>
<td>$\uparrow$16.2</td>
<td>$\uparrow$7.4</td>
<td>$\uparrow$14.3</td>
<td>$\uparrow$9.4</td>
<td>$\uparrow$14.0</td>
</tr>
<tr>
<td>PAD[4]</td>
<td>69.7(5.5)</td>
<td>75.5(19.0)</td>
<td>50.2(1.9)</td>
<td>38.7(5.1)</td>
<td>38.7(5.1)</td>
</tr>
<tr>
<td>PAD_gate</td>
<td>81.8(5.6)</td>
<td>81.6(25.1)</td>
<td>62.1(10.0)</td>
<td>44.8(1.0)</td>
<td>44.8(1.0)</td>
</tr>
<tr>
<td>ΔPAD</td>
<td>$\uparrow$12.1</td>
<td>$\uparrow$5.9</td>
<td>$\uparrow$11.9</td>
<td>$\uparrow$6.1</td>
<td>$\uparrow$11.2</td>
</tr>
<tr>
<td>GTA[26]</td>
<td>81.2(6.8)</td>
<td>78.9(20.5)</td>
<td>58.4(7.2)</td>
<td>42.2(2.8)</td>
<td>42.2(2.8)</td>
</tr>
<tr>
<td>GTA_gate</td>
<td>83.4(4.9)</td>
<td>85.8(27.4)</td>
<td>66.7(15.5)</td>
<td>48.5(3.5)</td>
<td>48.5(3.5)</td>
</tr>
<tr>
<td>ΔGTA</td>
<td>$\uparrow$2.1</td>
<td>$\uparrow$6.9</td>
<td>$\uparrow$7.3</td>
<td>$\uparrow$6.3</td>
<td>$\uparrow$6.3</td>
</tr>
<tr>
<td>ΔAvg</td>
<td>$\uparrow$8.3</td>
<td>$\uparrow$5.2</td>
<td>$\uparrow$8.1</td>
<td>$\uparrow$7.1</td>
<td>$\uparrow$8.1</td>
</tr>
</tbody>
</table>

Table 4. Ablation Study on task A→D. DANN_gate-only applies only the discriminator gate while DANN_label-only uses label information without the gate. DANN_joint is a variant of DANN_gate where the feature network only matches the joint distribution (last two lines of Eq.10), DANN_marginal only matches the marginal distribution, and DANN_none matches none of them. DANN_oracle excludes perturbed source data via human oracle.

<table>
<thead>
<tr>
<th>Setting ($\epsilon_L, L_y$)</th>
<th>0.7, 30%</th>
<th>0.7, 10%</th>
<th>0.3, 30%</th>
<th>0.3, 10%</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>DANN</td>
<td>70.4</td>
<td>49.4</td>
<td>72.5</td>
<td>54.3</td>
<td>61.7</td>
</tr>
<tr>
<td>DANN_pr</td>
<td>79.5</td>
<td>50.7</td>
<td>80.3</td>
<td>50.1</td>
<td>65.2</td>
</tr>
<tr>
<td>DANN_oracle</td>
<td>81.6</td>
<td>58.5</td>
<td>89.1</td>
<td>85.4</td>
<td>78.7</td>
</tr>
<tr>
<td>DANN_gate-only</td>
<td>76.3</td>
<td>53.8</td>
<td>78.0</td>
<td>55.7</td>
<td>66.0</td>
</tr>
<tr>
<td>DANN_label-only</td>
<td>74.4</td>
<td>52.5</td>
<td>77.5</td>
<td>55.0</td>
<td>64.9</td>
</tr>
<tr>
<td>DANN_joint</td>
<td>82.3</td>
<td>57.6</td>
<td>83.1</td>
<td>59.4</td>
<td>70.6</td>
</tr>
<tr>
<td>DANN_marginal</td>
<td>80.6</td>
<td>56.5</td>
<td>81.5</td>
<td>58.6</td>
<td>69.3</td>
</tr>
<tr>
<td>DANN_none</td>
<td>79.6</td>
<td>52.4</td>
<td>79.7</td>
<td>57.5</td>
<td>67.3</td>
</tr>
<tr>
<td>DANN_gate</td>
<td>82.5</td>
<td>58.7</td>
<td>82.7</td>
<td>60.7</td>
<td>71.2</td>
</tr>
</tbody>
</table>

5.3.2 Evaluation of Discriminator Gate

We compare our gated models with their respective state-of-the-art methods on the benchmarks in Table 2. Even using limited amount of labeled target data, our proposed method consistently improves the performance for all deep methods on all tasks. More importantly, our method can largely eliminate the negative impact of less related source data and of all methods under a more practically interesting scenario of moderately different distributions and limited amount of labeled target data. We observe that some methods are more vulnerable to negative transfer than the other even using the same training data. For conventional methods, instance-reweighting method KMM achieves smaller NTG compared to feature selection method TCA, possibly because KMM can assign small weights to source instances with dissimilar input features. For deep methods, we find GTA to be the most robust method against negative transfer since it takes both label information and random noises as inputs to the generator network. More interestingly, we observe that methods based on distribution measurement such as MMD (e.g. DANN) achieve smaller NTG than methods based on adversarial networks (e.g. DANN), even though the later tends to perform better when distributions are similar. This is consistent with findings in previous works [3] and one possible explanation is that adversarial network’s better capability of matching source and target domains leads to more severe negative transfer. Similarly, ADDA has better matching power by using two separate feature extractors, but it results in larger NTG compared to DANN.

avoid negative transfer (e.g. DANN_gate achieves negative average NTG while DANN gets positive NTG). Specifically, our method achieves larger accuracy gains on harder tasks such as synthetic to real-world tasks in Office-Home and VisDA. This is mainly because source domains in these tasks tend to contain more unrelated samples. This finding is also consistent with results in Table 1 and Figure 3(a) where we can observe larger performance gains as perturbation rates increase. In the extreme case where the source domain is degenerate ($\epsilon = 1.0$ in Figure 3(a)), the gated model achieves comparable results to those of DANN_pr. On the other hand, the results of DANN and DANN_gate are similar when source domain is closely related to the target ($\epsilon = 0.0$ on task W→D in Table 1). This indicates that the discriminator gate can control the trade-off between maxi-
mal transfer and alleviating negative impact.

Ablation Study. We report the results of ablation study in Table 4 and analyze the effects of several components in our method subject to different settings of transfer tasks. First, both DANN\textsubscript{gate}-only and DANN\textsubscript{label}-only perform better than DANN but worse than DANN\textsubscript{gate}, showing that the discriminator gate and estimating joint distributions can both improve performance but their combination yields full performance benefit. Second, DANN\textsubscript{joint} obtains higher accuracy results than DANN\textsubscript{marginal} and DANN\textsubscript{none} since matching joint distributions is the key to avoid negative transfer when both marginal and conditional distributions shift. However, while DANN\textsubscript{joint} achieves comparable results as DANN\textsubscript{gate} when $L\% = 30\%$, it performs worse than DANN\textsubscript{gate} when $L\% = 10\%$. This shows that utilizing unlabeled target data to match marginal distributions can be beneficial when labeled target data is scarce. Lastly, it is inspiring to see DANN\textsubscript{gate} outperforms DANN\textsubscript{oracle} when perturbation rate is high. This is because less unperturbed source data are used for DANN\textsubscript{oracle} but DANN\textsubscript{gate} can utilize perturbed source data that contain related information. This further shows the effectiveness of our approach.

Feature Visualization. We visualize the t-SNE embeddings [6] of the bottleneck representations in Figure 4. The first column shows that, when perturbation rate is high, DANN cannot align the two domains well and it fails to discriminate both source and target classes as different classes are mixed together. The second column illustrates the discriminator gate can improve the alignment by assigning less weights to unrelated source data. For instance, we can see some source data from different classes mixed in the yellow cluster at the center right but they get assigned smaller weights. The third column shows the embeddings after we remove source data with small discriminator weights ($< 0.4$). We can observe that target data are much better clustered compared to that of DANN. These in-depth results demonstrate the efficacy of discriminator gate method.

Statistics of Instance Weights. We illustrate the discriminator output ($\frac{P_T(x_i, y_i) + P_S(x_i, y_i)}{P_T(x_i, y_i) + P_S(x_i, y_i)}$) for each source data in Figure 4(d). We can observe that DANN fails to discriminate unrelated source data as all weights concentrate around 0.5 in the middle. On the other hand, DANN\textsubscript{gate} assigns smaller weights to a large portion of source data (since perturbation rate is high) and thus filters out unrelated information. Figure 4(h) further shows that DANN assigns similar average weights for perturbed and unperturbed source data while DANN\textsubscript{gate} outputs much smaller values for perturbed data but higher ones for unperturbed data.

6. Conclusion

In this work, we analyze the problem of negative transfer and propose a novel discriminator gate technique to avoid it. We show that negative transfer directly relates to specific algorithms, domain divergence and target data. Experiments demonstrate these factors and the efficacy of our method. Our method consistently improves the performance of base methods and largely avoids negative transfer. Understanding negative transfer in more complex transfer tasks and settings should be addressed in a future research.
References


