Transferability vs. Discriminability: Batch Spectral Penalization for Adversarial Domain Adaptation

Xinyang Chen[1] Sinan Wang[1] Mingsheng Long[1] Jianmin Wang[1]

Abstract

Adversarial domain adaptation has made remarkable advances in learning transferable representations for knowledge transfer across domains. While adversarial learning strengthens the feature transferability which the community focuses on, its impact on the feature discriminability has not been fully explored. In this paper, a series of experiments based on spectral analysis of the feature representations have been conducted, revealing an unexpected deterioration of the discriminability while learning transferable features adversarially. Our key finding is that the eigenvectors with the largest singular values will dominate the feature transferability. As a consequence, the transferability is enhanced at the expense of over penalization of other eigenvectors that embody rich structures crucial for discriminability. Towards this problem, we present Batch Spectral Penalization (BSP), a general approach to penalizing the largest singular values so that other eigenvectors can be relatively strengthened to boost the feature discriminability. Experiments show that the approach significantly improves upon representative adversarial domain adaptation methods to yield state of the art results.

1. Introduction

Deep networks have achieved remarkable success in diverse machine learning areas. On the basis of large-scale labeled data, transferable features across multiple tasks and domains can be learned (Yosinski et al., 2014; Oquab et al., 2014; Donahue et al., 2014). Unfortunately, in many real-world applications, shortage of labeled data is not uncommon. To avoid both labeling efforts and overfitting issues, domain adaptation is extensively studied to overcome the domain shift or dataset bias such that a learner trained on other large-scale datasets can be leveraged (Torralba & Efros, 2011).

Domain adaptation tackles the problem of learning a model that reduces the dataset shift between training and testing distributions (Pan et al., 2010). Early domain adaptation methods in the shallow regime learn feature representations invariant across domains (Pan et al., 2011; Gong et al., 2012) or reweigh source instances based on their relevance to the target domain (Huang et al., 2007; Gong et al., 2013). Recent domain adaptation methods in the deep regime disentangle the explanatory factors of variations behind data and explore two strategies for aligning feature distributions across domains. The first strategy is moment matching, bridging the distributions by matching all their statistics (Long et al., 2015; Li et al., 2016; Long et al., 2017; Maria Carlucci et al., 2017). The second strategy is adversarial learning, where a domain discriminator is trained to distinguish the source from the target while feature representations are learned to confuse it simultaneously (Ganin & Lempitsky, 2015; Tzeng et al., 2015; Ganin et al., 2016; Tzeng et al., 2017; Luo et al., 2017; Long et al., 2018). These adversarial domain adaptation methods have yielded remarkable performance gains.

While it is widely believed by the the community that adversarial learning strengthens the feature transferability, which part of the features are transferable remains unclear. Further, how the feature discriminability will change in the process of learning transferable features has not been fully explored. From a spectral analysis viewpoint, we can decompose the feature representations into eigenvectors with importance quantified by the corresponding singular values. Intuitively, the feature transferability may mainly reside in the eigenvectors with top singular values, since the eigenvectors with low singular values may embody domain variations and should be discouraged. In contrast, the feature discriminability may depend on more eigenvectors, since the rich discriminative structures cannot be conveyed by only a few eigenvectors. As a result, there exists a contradiction when we try to make the representations both transferable and discriminative. As the target domain is fully unlabeled, this contradiction will mainly sacrifice the discriminability of target domain data.

In this paper, we try to address the dilemma of transferability against discriminability by understanding their behaviors in adversarial domain adaptation. We apply linear discriminant

[1] School of Software, BNRist, Research Center for Big Data, Tsinghua University. E-mail: chenxiny17@mails.tsinghua.edu.cn. Correspondence to: M. Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

analysis (**LDA**) (Fukunaga, 1990) to compute the largest ratio of the between-class variance and within-class variance in the projected space as a criterion of feature discriminability. We further train a joint classifier based on the feature representations on the source and target labeled data and use the error rate as another criterion of the discriminability. Both criteria indicate that the adversarial domain adaptation methods tend to enhance the transferability at the expense of deteriorating the discriminability. Based on this observation, we further introduce singular value decomposition (**SVD**) (Golub & Reinsch, 1970) to analyze the spectral properties of feature representations in batches. We confirm that the eigenvectors with the top singular values dominate the transferability of feature representations, while the eigenvectors with smaller singular values are overly penalized to convey deficient discriminability. Towards this problem, we present Batch Spectral Penalization (**BSP**), a general approach to penalizing the largest singular values so that the other eigenvectors can be relatively strengthened to boost the feature discriminability. Experimental results show that our method enables existing adversarial domain adaptation methods to learn both transferable and discriminative representations, yielding state of the art results on four benchmark datasets.

2. Transferability vs. Discriminability

In this paper, we study the unsupervised domain adaption problem, where we have access to labeled examples in the source domain while only unlabeled examples in the target domain. We are given n_s labeled examples from a source domain $\mathcal{S} = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}$ and n_t unlabeled examples from a target domain $\mathcal{T} = \{(\mathbf{x}_i^t)\}$, which are sampled from different distributions P and Q. Adversarial domain adaptation (Ganin et al., 2016; Long et al., 2018) has been arguably one of the most successful deep learning approaches to domain adaptation. By embedding adversarial learning modules into deep networks, we can learn transferable representations to suppress the distributional variations across domains.

There are two key criteria that characterize the goodness of feature representations to enable domain adaptation. One is **transferability**, which indicates the ability of feature representations to bridge the discrepancy across domains. With transferability, we can effectively transfer a learning model from the source domain to the target domain via the feature representations. The other is **discriminability**, which refers to the easiness of separating different categories by a supervised classifier trained over the feature representations. Adversarial domain adaptation methods make remarkable advances in enhancing the transferability of representations, however, the discriminability of the learned representations has only been attempted via minimizing the classification error on the source domain labeled data. In what follows, we will analyze that such a vanilla strategy is not good enough.

2.1. Adversarial Domain Adaptation

Adversarial domain adaptation methods, starting from Domain Adversarial Neural Network (**DANN**) (Ganin & Lempitsky, 2015), have become increasingly influential in domain adaptation. The fundamental idea behind is to learn transferable features that explicitly reduce the domain shift. Basically, these approaches constitute a feature extractor $\mathbf{f} = F(\mathbf{x})$ and a category classifier $\mathbf{y} = G(\mathbf{f})$, similar to standard supervised classification models. Additionally, a domain discriminator $\mathbf{d} = D(\mathbf{f})$ is added as the adversary to play a two-player minimax game against F. While the domain discriminator D is trained to distinguish the source domain from the target domain, the feature extractor F is trained to confuse the domain discriminator. As mathematically justified in (Ganin & Lempitsky, 2015), training the domain discriminator D to distinguish the source from the target is equivalent to maximizing some statistical distance $\mathsf{dist}_{P \leftrightarrow Q}(F, D)$ across the source and target distributions P and Q. By training F adversarially to deceive D, the feature representations are made *transferable* across domains. In addition, the feature extractor F and the source classifier G are learned simultaneously by minimizing the classification error $\mathcal{E}(F, G)$ on the source labeled data. This makes the feature representations *discriminative* across categories. Formally, adversarial domain adaptation is formulated as

$$\begin{aligned} &\min_{F,G} \ \mathcal{E}(F, G) + \delta \mathsf{dist}_{P \leftrightarrow Q}(F, D) \\ &\max_{D} \ \mathsf{dist}_{P \leftrightarrow Q}(F, D), \end{aligned} \tag{1}$$

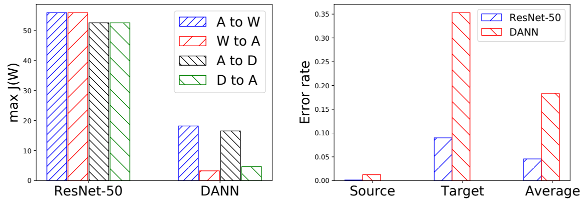
where δ is a hyperparameter to trade off the importance of transferability vs. discriminability in feature representation.

2.2. Discriminability of Feature Representations

In this paper, we rethink the adversarial domain adaptation by investigating both the transferability and discriminability of their learned feature representations. To investigate the discriminability, we need to define an evaluation criterion. Inspired by Linear Discriminant Analysis (**LDA**) (Fukunaga, 1990), we can calculate the between-class variance $\mathbf{S}_b$ and within-class variance $\mathbf{S}_w$ as follows:

$$\begin{aligned} \mathbf{S}_b &= \sum_{j=1}^{c} n_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathsf{T}} \\ \mathbf{S}_w &= \sum_{j=1}^{c} \sum_{\mathbf{f} \in \mathcal{F}_j} (\mathbf{f} - \boldsymbol{\mu}_j)(\mathbf{f} - \boldsymbol{\mu}_j)^{\mathsf{T}} \end{aligned} \tag{2}$$

where c is the number of classes, each class has n_j samples, $\mathbf{f} = F(\mathbf{x})$ is the deep feature extracted by F, $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}$ are the centers of the feature vectors in class j and in all classes respectively, and $\mathcal{F}_j$ is the set of all feature vectors in class j. $\mathbf{S}_b$ and $\mathbf{S}_w$ represent the total variance of the feature vectors across different classes and in the same class, respectively.



(a) $\max J(\mathbf{W})$ (b) Error Rate

Figure 1. Two experiments measuring discriminability of features: (a) $\max J(W)$; (b) Classification error rate on the representation.

We project the high-dimensional feature vectors into low-dimensional space and calculate the ratio of between-class variance to within-class variance. Intuitively, the representations with larger ratio imply stronger discriminability and vice versa. The discriminability criterion based on LDA is

$$\arg\max_{\mathbf{W}} J(\mathbf{W}) = \frac{\mathrm{tr}(\mathbf{W}^{\mathsf{T}}\mathbf{S}_{\mathrm{b}}\mathbf{W})}{\mathrm{tr}(\mathbf{W}^{\mathsf{T}}\mathbf{S}_{\mathrm{w}}\mathbf{W})} \tag{3}$$

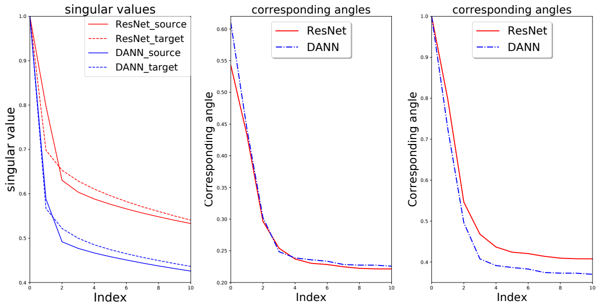
The optimal solution to the above optimization problem is the k (set to the num of classes) largest singular values of $\mathbf{S}_w^{-1}\mathbf{S}_b$, found by the Singular Value Decomposition (**SVD**):

$$\mathbf{S}_w^{-1}\mathbf{S}_b = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}. \tag{4}$$

More formally, the optimal solution is $\mathbf{W}^* = \mathbf{U}$.

In this study, we investigate the discriminability of feature representations of **DANN** (Ganin et al., 2016) (ResNet-50 backbone) and **ResNet-50** (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015). From Office-31, the most widely used dataset for visual domain adaptation, we choose 4 transfer tasks to compute the discriminability criterion $\max J(\mathbf{W})$: $\mathbf{A} \to \mathbf{W}$, $\mathbf{W} \to \mathbf{A}$, $\mathbf{A} \to \mathbf{D}$, $\mathbf{D} \to \mathbf{A}$, with results shown in Figure 1(a). It is very unexpected that the feature discriminability of DANN is lower than that of ResNet-50, implying that DANN's transferability is enhanced at the expense of worse discriminability.

We further confirm the above observation by another study. Motivated by the domain adaptation theory (Ben-David et al., 2010), we train a multilayer perceptrons (MLP) classifier G' with the representations learned by DANN (Ganin et al., 2016) (ResNet-50 backbone) and ResNet-50 (He et al., 2016), respectively. The MLP classifier is trained over all data with labels from both source and target domains, while the feature extractor F is fixed. Note that the target labels are only used for this pilot analysis. We compare the average error rates of the MLP classifier on both representations, with results shown in Figure 1(b). Again, we observe that the error rate on the representation of DANN is much higher than that of ResNet-50. Obviously, higher error rate implies weaker discriminability. This leads to worse generalization error bound as revealed by the domain adaptation theory (Ben-David et al., 2010).



(a) σ (b) $\cos(\psi)$ (c) $\cos(\psi)$

Figure 2. SVD analysis. With source and target feature matrices from different methods, we compute (a) the singular values (max-normalized); (b) squared root of the cosine values of corresponding angles (unnormalized); (c) squared root of the cosine values of corresponding angles (max-normalized). In the max-normalized version we scaled all singular values such that the largest one is 1.

With the above analyses, it is natural to raise the following questions: **(a) Why** features extracted by DANN present worse discriminability? **(b) How** can we enhance the transferability while guaranteeing acceptable discriminability?

2.3. Why Worse Discriminability?

To investigate the first question in-depth, we apply singular value decomposition (**SVD**) to compute respectively all singular values and eigenvectors of the source feature matrix $\mathbf{F}_s = [\mathbf{f}_1^s \dots \mathbf{f}_b^s]$ and target feature matrix $\mathbf{F}_t = [\mathbf{f}_1^t \dots \mathbf{f}_b^t]$:

$$\begin{aligned} \mathbf{F}_s &= \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^{\mathsf{T}} \\ \mathbf{F}_t &= \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^{\mathsf{T}} \end{aligned} \tag{5}$$

where b is the size of each batch used for mini-batch SGD training. Then we plot the singular values (max-normalized in each $\mathbf{\Sigma}$) of the feature matrix extracted by ResNet-50 (He et al., 2016) and DANN (Ganin et al., 2016) (ResNet-50 backbone). From Figure 2(a), we observe that the largest singular value of the DANN feature matrix is significantly larger than the other singular values, greatly weakening the informative signals of eigenvectors corresponding to smaller singular values. Such a sharp distribution of singular values intuitively imply deteriorated discriminability.

Next we investigate the transferability of each principal components in the feature matrix, i.e., eigenvectors in each direction. Shonkwiler (2009) introduced the **principal angles**, denoted by θ, to measure the similarity of two subspaces:

$$\begin{aligned} \cos(\theta_i) =& \frac{\langle \mathbf{a}_i, \mathbf{b}_i \rangle}{\|\mathbf{a}_i\| \, \|\mathbf{b}_i\|} \\ &= \max_{\mathbf{a},\mathbf{b}} \left\{ \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \, \|\mathbf{b}\|} : \mathbf{a} \perp \mathbf{a}_j, \mathbf{b} \perp \mathbf{b}_j, j = 1, ..., i-1 \right\} \end{aligned} \tag{6}$$

where $\mathbf{a}_i \in \mathbf{U}_s$ is an eigenvector of source feature matrix and $\mathbf{b}_i \in \mathbf{U}_t$ is an eigenvector of target feature matrix, and $\|\mathbf{a}_i\| = 1$, $\|\mathbf{b}_i\| = 1$.

However, the principal angles calculated by Mohammadi (2014) complete the pairing between the eigenvectors with the smallest angle, regardless of their corresponding singular values. In other words, all eigenvectors are treated equally when pairing. This is unreasonable because in recognition problems every feature may play a different role, and thus it does not make sense that eigenvectors with large singular values are easily transferred to the eigenvectors with small singular values simply because they have a small principal angle. More importantly, to understand the transferability of feature representations, it is necessary to explore the angles between the eigenvectors of the same singular value index in matrices $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$. In this paper we call it **corresponding angle**, denoted by ψ, and define it more formally as follows:

Definition 1 (Corresponding Angle) *The angle between two eigenvectors corresponding to the same singular value index, which are equally important in their feature matrices.*

The cosine value of the corresponding angle is calculated as

$$\cos(\psi_i) = \frac{\langle \mathbf{u}_{s,i}, \mathbf{u}_{t,i} \rangle}{\|\mathbf{u}_{s,i}\| \, \|\mathbf{u}_{s,i}\|}, \tag{7}$$

where $\mathbf{u}_{s,i}$ is the ith eigenvector in $\mathbf{U}_s$ with the ith largest singular value in the source feature matrix, and similarly for $\mathbf{u}_{t,i}$. Figures 2(b)–2(c) show the top-10 corresponding angles between the source and target eigenvectors $\mathbf{U}_s$ and $\mathbf{U}_t$. Results are averaged over batches. We observe that $\cos(\psi_1)$ is much larger than $\cos(\psi_i), i \geq 2$, which suggests that the eigenvector with the largest singular value dominates the transferability of feature representation. However, the decay trend in DANN features is even sharper than in ResNet-50 features, showing a severer dominance of the top eigenvectors for transferability in DANN.

The observations from the series of studies above suggest **why** the feature representations extracted by **DANN** present worse discriminability. That is, only the eigenvectors corresponding to largest singular values tend to carry transferable knowledge, while other eigenvectors may endow domain variations and thus be overly penalized. As a side effect, the crucial discriminative information conveyed in these eigenvectors will also be suppressed to weaken discriminability.

3. Approach

In this paper, we stress that transferability and discriminability are equally important for learning good representations in adversarial domain adaptation. While previous section focuses on uncovering the reasons on discriminability loss, this section focuses on **how** to enhance transferability while guaranteeing acceptable discriminability. The idea is two-fold. First, the eigenvectors corresponding to larger singular values should be fully leveraged to enhance transferability, as these eigenvectors generally suppress domain variations. Second, the eigenvectors corresponding to relatively smaller singular values should also be leveraged to convey richer structures, which are vitally important for discriminability.

3.1. Batch Spectral Penalization

The above analysis reveals two insights. The eigenvectors with large singular values are important for the entire feature matrix and its transferability. Although other eigenvectors, or dimensions, play a minor role for transferability, this does not mean that for discriminability they can be ignored. In contrast, to our knowledge, discriminative classifiers have necessary dependence on most dimensions. Therefore, it is necessary to suppress the dimension with top singular value to prevent it from standing out. In our approach, we first apply SVD to obtain the largest k singular values of source feature matrix $\mathbf{F}_s$ and target feature matrix $\mathbf{F}_t$ respectively. Then we propose Batch Spectral Penalization (**BSP**) as a regularization term over these largest k singular values:

$$L_{\text{bsp}}(F) = \sum_{i=1}^{k} (\sigma_{s,i}^2 + \sigma_{t,i}^2), \tag{8}$$

where $\sigma_{s,i}$ and $\sigma_{t,i}$ refer to the i-th largest singular values in $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ respectively. This BSP penalty can be applied directly among feature vectors within each batch.

Denote by $\mathbf{F} = [\mathbf{F}_s, \mathbf{F}_t]$ the concatenation of the source and target features in each batch. It is also intuitively possible to impose a BSP penalization over the singular values of $\mathbf{F}$. However, due to the domain difference, the eigenvectors of $\mathbf{F}$ may not be equally contributed by the source domain and target domain. To avoid potential distortion, it is necessary to penalize the singular values of the feature matrices from the source domain and target domain independently.

Computational complexity. Note that full SVD computing all the singular values on an $m \times n$ matrix would cost time $O(\min(m^2 n, mn^2))$. This is too expensive to be used in adversarial domain adaptation if computed over a nearly squared matrix, e.g. the weight parameters of deep networks. While the complexity of our batch spectral penalization is $O(b^2 d)$ where d is the dimension of features, since b is often small, the overall computational budget of BSP is nearly negligible in the mini-batch SGD training of deep networks.

3.2. Models with Batch Spectral Penalization

As stated before, the distributions of the singular values of the feature matrices $\mathbf{F}_s$ and $\mathbf{F}_t$ are pathological in DANN (Ganin et al., 2016). With batch spectral penalization (BSP)

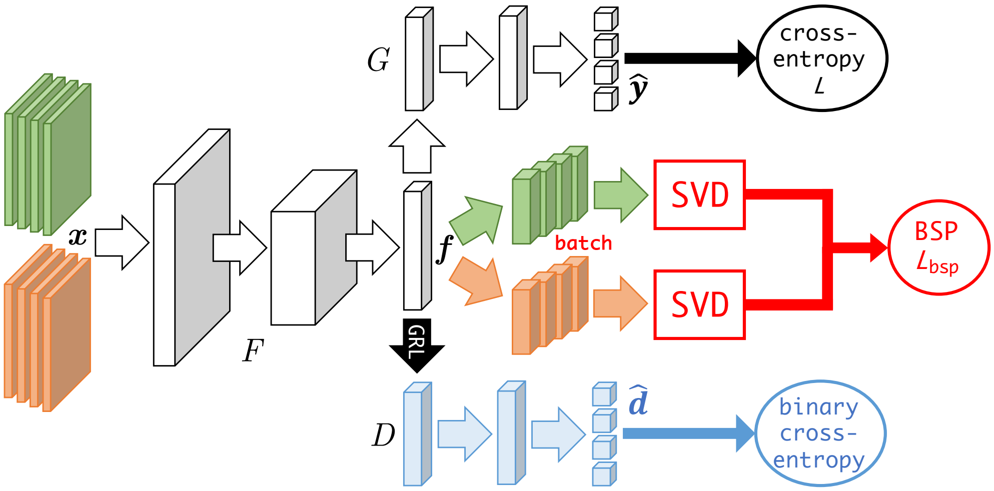


Figure 3. The architecture of **BSP+DANN** where BSP enhances discriminability while learning transferable features via domain adversarial network (DANN). BSP is a lightweight module readily pluggable into any deep domain adaptation networks, which is end-to-end trainable with the support of **differentiable SVD** in **PyTorch**. GRL denotes Gradient Reversal Layer widely used in adversarial domain adaptation.

applied to adversarial domain adaptation, the distributions of singular values are expected to be more balanced. The discriminability will be preserved as much as possible during adversarial learning, which is originally tailored to transferability. To learn representations with both transferability and discriminability, the minimax game of adversarial domain adaptation with batch spectral penalization is formulated as

$$\begin{aligned} &\min_{F,G} \ \mathcal{E}(F,G) + \delta \mathsf{dist}_{P\leftrightarrow Q}(F,D) + \beta L_{\text{bsp}}(F) \\ &\max_{D} \ \mathsf{dist}_{P\leftrightarrow Q}(F,D), \end{aligned} \tag{9}$$

where β is a hyperparameter for trading off BSP.

Next, we apply batch spectral penalization (BSP) to a vanilla and a state of the art adversarial domain adaptation models.

Domain Adversarial Neural Network (DANN). This vanilla model (Ganin et al., 2016) exactly fits into the adversarial domain adaptation framework in Equation (9), where $\mathcal{E}(F,G)$ uses cross-entropy loss to lower classification error on the source domain, and $\mathsf{dist}_{P\leftrightarrow Q}(F,D)$ uses negative cross-entropy loss to measure whether the feature representations $\mathbf{f}$ confuse the domain discriminator D successfully:

$$\begin{aligned} \mathcal{E}(F,G) &= \mathbb{E}_{(\mathbf{x}_i^s,\mathbf{y}_i^s)\sim P} L(G(\mathbf{x}_i^s),\mathbf{y}_i^s) \\ \mathsf{dist}_{P\leftrightarrow Q}(F,D) &= \mathbb{E}_{\mathbf{x}_i^s\sim P}\log[D(\mathbf{f}_i^s)] \\ &\quad + \mathbb{E}_{\mathbf{x}_i^t\sim Q}\log[1-D(\mathbf{f}_i^t)] \end{aligned} \tag{10}$$

where L is the cross-entropy loss of category classifier G. The architecture of **BSP+DANN** is shown in Figure 3. As shown, BSP can be easily plugged in as a penalization to the features extracted by any domain adaptation networks.

Conditional Domain Adversarial Network (CDAN). This state of the art model (Long et al., 2018) exploits the classifier predictions which are believed to convey rich discriminative information useful for conditional adversarial learning. Specifically, it conditions domain discriminator D on the classifier prediction $\mathbf{g}$ through the multilinear map:

$$T_{\otimes}(\mathbf{h}) = \mathbf{f} \otimes \mathbf{g}, \tag{11}$$

where $\mathbf{h} = [\mathbf{f},\mathbf{g}]$. Different from DANN, CDAN employs $\mathbf{h}$ instead of $\mathbf{f}$ as the input to the domain discriminator D:

$$\begin{aligned} \mathsf{dist}_{P\leftrightarrow Q}(F,D) &= \mathbb{E}_{\mathbf{x}_i^s\sim P}\log[D(\mathbf{h}_i^s)] \\ &\quad + \mathbb{E}_{\mathbf{x}_i^t\sim Q}\log[1-D(\mathbf{h}_i^t)]. \end{aligned} \tag{12}$$

And $\mathcal{E}(F,G)$ in CDAN is the same with that of DANN.

We use default $k=1$ due to the distribution of the singular values in domain adversarial networks, making batch spectral penalization suitable to be calculated more efficiently by the power method (Golub & Van der Vorst, 2001). However, since k may be different in other networks and SVD does not cost too much computation time in a feature matrix of a small batch, we still use SVD to compute BSP in this paper.

3.3. Theoretical Understanding

Ben-David et al. (2010) pioneered the domain adaptation theory that bounds the expected error $\mathcal{E}_{\mathcal{T}}(h)$ of a hypothesis h on the target domain by using three terms: (a) expected error of h on the source domain, $\mathcal{E}_{\mathcal{S}}(h)$; (b) $\mathcal{H}\Delta\mathcal{H}$-distance $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T})$, measuring domain shift as the discrepancy between the disagreement of two hypotheses $h,h' \in \mathcal{H}\Delta\mathcal{H}$; and (c) the error λ of the ideal joint hypothesis h^* on both source and target domains. The learning bound is

$$\mathcal{E}_{\mathcal{T}}(h) \leq \mathcal{E}_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda. \tag{13}$$

The error λ of the ideal joint hypothesis $h^* = \min_h \mathcal{E}_{\mathcal{S}}(h) + \mathcal{E}_{\mathcal{T}}(h)$ is

$$\lambda = \mathcal{E}_{\mathcal{S}}(h^*) + \mathcal{E}_{\mathcal{T}}(h^*). \tag{14}$$

Noteworthily, most adversarial domain adaptation methods treated the third term, λ, as a constant, which measures the discriminability of features especially in the target domain. From previous experimental analysis in Figure 1(b) we can find that the average error rate of the MLP classifier trained on the labeled data of source and target domains is half of λ. And features extracted by DANN (Ganin et al., 2016) incur higher error rate, implying worse discriminability than features extracted by pre-trained ResNet-50 (He et al., 2016).

To mitigate this pitfall of adversarial domain adaptation, the BSP approach penalizes the eigenvectors with largest singular values from standing out in the feature representations, such that the other eigenvectors with relatively smaller singular values (which also provide discriminative information) can be matched instead of being suppressed. This will essentially controls λ, yielding lower bound for the target error. We will extensively justify this in the empirical study.

4. Experiments

We embed the batch spectral penalization (**BSP**) into well-known adversarial domain adaptation methods and evaluate our method on several visual domain adaptation datasets. The code of BSP is available at `github.com/thuml/Batch-Spectral-Penalization`.

4.1. Setup

Office-31 (Saenko et al., 2010) is a vanilla dataset for visual domain adaptation with 4,652 images in 31 categories from three domains: *Amazon* (**A**), *Webcam* (**W**) and *DSLR* (**D**). We evaluate our methods on all six transfer tasks.

Office-Home (Venkateswara et al., 2017) is a more difficult dataset than Office-31, which consists of around 15,500 images from 65 classes in office and home settings, forming four extremely distinct domains: *Artistic images* (**Ar**), *Clip Art* (**Cl**), *Product images* (**Pr**), and *Real-World images* (**Rw**). We evaluate our methods on all twelve transfer tasks.

VisDA-2017 (Peng et al., 2017) is a challenging simulation-to-real dataset, with two domains: **Synthetic**, renderings of 3D models from different angles and with different lightning conditions; **Real**, real-world images. We evaluate our methods on the **Synthetic** → **Real** task.

Digits (Ganin et al., 2016). We use three digits datasets: **MNIST**, **USPS**, and **SVHN**. We build three transfer tasks: USPS to MNIST (**U** → **M**), MNIST to USPS (**M** → **U**), and SVHN to MNIST (**S** → **M**). We follow the experimental settings of (Hoffman et al., 2018).

We extend Domain Adversarial Neural Network (**DANN**) (Ganin et al., 2016), Adversarial Discriminative Domain Adaptation (**ADDA**) and Conditional Domain Adversarial Network (**CDAN**) (Long et al., 2018) by the proposed batch spectral penalization (**BSP**). We compare with 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), Joint Adaptation Network (**JAN**) (Long et al., 2017), Unsupervised Image-to-Image Translation (**UNIT**) (Zhu et al., 2017), Generate to Adapt (**GTA**) (Sankaranarayanan et al., 2018), Cycle-Consistent Adversarial Domain Adaptation (**CyCADA**) (Hoffman et al., 2018), Maximum Classifier Discrepancy (**MCD**) (Saito et al., 2018) and Conditional Domain Adversarial Network (**CDAN**) (Long et al., 2018).

We use **PyTorch** to implement our methods and fine-tune ResNet pre-trained on ImageNet (Russakovsky et al., 2015). Following the standard protocols for unsupervised domain adaptation (Long et al., 2018), all labeled source samples and unlabeled target samples participate in the training stage. We fix $\delta = 1$ and $\beta = 10^{-4}$ in all experiments. The learning rates of the layers trained from scratch are set to be 10 times those of fine-tuned layers. We adopt mini-batch SGD with momentum of 0.95 using the learning rate and progressive training strategies of DANN (Ganin & Lempitsky, 2015).

4.2. Results

The classification accuracies on Office-31 are shown in Table 1, with results of baselines directly reported from their original papers wherever available. Our method significantly improves the performance of domain adversarial networks and achieves state of the art results. There is an obvious boost in accuracies on relatively difficult tasks **D** → **A** and **W** → **A** where the source domain is quite small. As reported in Tables 2, 3 and 4, our method also boosts the accuracies of **DANN** and **CDAN**. As **CDAN** is a more advanced method, leading to relatively less room for improvement, our promotion on **DANN** is more obvious. This justifies the efficacy of BSP for improving the discriminability in the process of learning transferable features. A more inspiring result is that by training from scratch, **BSP** improves the performance of **ADDA** on the Digits dataset. This indicates that **BSP** is also helpful for the *asymmetric* adversarial adaptation methods.

4.3. Analyses

In this section, we employ the more difficult task **W** → **A** on Office-31 as the testbed. We analyze the features extracted by four networks: (a) **ResNet-50** pre-trained on ImageNet;

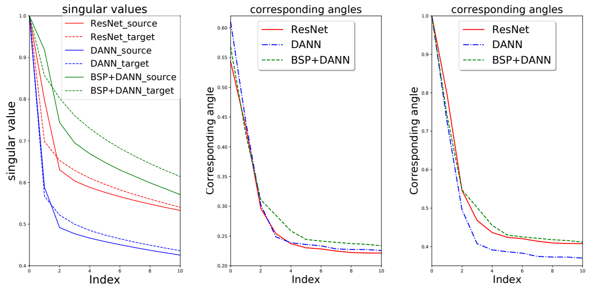
Table 1. Accuracy (%) on Office-31 for unsupervised domain adaptation (ResNet-50).

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
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
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
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
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
CDAN (Long et al., 2018)	93.1 ± 0.2	98.2 ± 0.2	**100.0** ± 0.0	89.8 ± 0.3	70.1 ± 0.4	68.0 ± 0.4	86.6
CDAN+E (Long et al., 2018)	**94.1** ± 0.1	**98.6** ± 0.1	**100.0** ± 0.0	92.9 ± 0.2	71.0 ± 0.3	69.3 ± 0.3	87.7
BSP+DANN (Proposed)	93.0 ± 0.2	98.0 ± 0.2	**100.0** ± 0.0	90.0 ± 0.4	71.9 ± 0.3	**73.0** ± 0.3	87.7
BSP+CDAN (Proposed)	93.3 ± 0.2	98.2 ± 0.2	**100.0** ± 0.0	**93.0** ± 0.2	**73.6** ± 0.3	72.6 ± 0.3	**88.5**

Table 2. Accuracy (%) on Office-Home for unsupervised domain adaptation (ResNet-50).

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
CDAN (Long et al., 2018)	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
CDAN+E (Long et al., 2018)	50.7	**70.6**	76.0	57.6	70.0	70.0	57.4	**50.9**	77.3	70.9	56.7	81.6	65.8
BSP+DANN (Proposed)	51.4	68.3	75.9	56.0	67.8	68.8	57.0	49.6	75.8	70.4	57.1	80.6	64.9
BSP+CDAN (Proposed)	**52.0**	68.6	**76.1**	**58.0**	**70.3**	**70.2**	**58.6**	50.2	**77.6**	**72.2**	**59.3**	**81.9**	**66.3**

(b) **ResNet-50** trained only with source samples; (c) **DANN**; (d) DANN with batch spectral penalization (**BSP+DANN**). In spectral analysis and corresponding angle analysis we use features extracted by (b)–(d), while in ideal joint hypothesis analysis we use features extracted by (a), (c) and (d). In addition, in each min-batch iteration, (c) uses 0.342s while (d) uses 0.359s (Titan V), which shows that SVD does not cost much computation for feature matrix of a small batch.

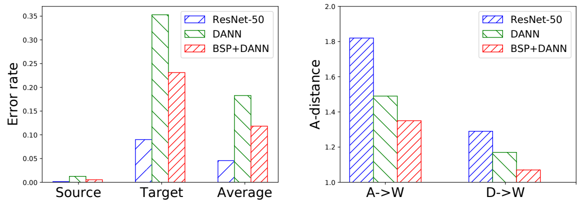


(a) σ (b) $\cos(\psi)$ (c) $\cos(\psi)$

Figure 4. SVD analysis. With source and target feature matrices from different methods, we compute (a) the singular values (max-normalized); (b) squared root of the cosine values of corresponding angles (unnormalized); (c) squared root of the cosine values of corresponding angles (max-normalized). In the max-normalized version we scaled all singular values such that the largest one is 1.

Spectral Analysis. Singular values of features extracted by **BSP+DANN** are shown in Figure 4(a). **BSP** successfully decreases the big difference between the largest and the rest. More dimensions pose positive influence on classification.

Corresponding Angle. Cosines of corresponding angles between the source and target domains are shown in Figures 4(b)–4(c). For **DANN**, the transferability of the eigenvector with the largest cosine far outweighs that of the rest. However, **BSP** gives consecutive eigenvectors a more prominent role during the transfer process. According to previous analysis, the feature matrix **BSP** constructs has a better-shaped distribution of singular values to enhance discriminability.



(a) Error Rate (b) A-distance

Figure 5. Discriminability and transferability of learned features: (a) Classification error rate on each representation; (b) A-distance.

Ideal Joint Hypothesis. We investigate the ideal joint hypothesis, which can be found by training an MLP classifier on all source and target data with labels. As analyzed before, it serves as a good indicator of discriminability. Moreover, this error rate is half of λ in Equation (14). The results are shown in Figure 5(a). As expected, while pre-trained ResNet has a lower λ than domain adversarial networks, **BSP** significantly enhances the discriminability of **DANN**.

Distribution Discrepancy. The A-distance (Ben-David et al., 2010) is a measure of domain discrepancy, defined as

Table 3. Accuracy (%) on VisDA-2017 for unsupervised domain adaptation (ResNet-101).

Method	plane	bcybl	bus	car	horse	knife	mcyle	person	plant	sktbrd	train	truck	mean
ResNet-101 (He et al., 2016)	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN (Long et al., 2015)	87.1	63.0	76.5	42.0	**90.3**	42.9	85.9	53.1	49.7	36.3	**85.8**	20.7	61.1
DANN (Ganin et al., 2016)	81.9	**77.7**	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD (Saito et al., 2018)	87.0	60.9	83.7	**64.0**	88.9	79.6	84.7	76.9	**88.6**	40.3	83.0	25.8	71.9
CDAN (Long et al., 2018)	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.7
BSP+DANN (Proposed)	92.2	72.5	**83.8**	47.5	87.0	54.0	86.8	72.4	80.6	66.9	84.5	37.1	72.1
BSP+CDAN (Proposed)	**92.4**	61.0	81.0	57.5	89.0	**80.6**	**90.1**	**77.0**	84.2	**77.9**	82.1	**38.4**	**75.9**

Table 4. Accuracy (%) on Digits for domain adaptation.

Method	M→U	U→M	S→M	Avg
DANN (Ganin et al., 2016)	90.4	94.7	84.2	89.8
ADDA (Tzeng et al., 2017)	89.4	90.1	86.3	88.6
UNIT (Zhu et al., 2017)	**96.0**	93.6	90.5	93.4
CyCADA (Hoffman et al., 2018)	95.6	96.5	90.4	94.2
CDAN (Long et al., 2018)	93.9	96.9	88.5	93.1
CDAN+E (Long et al., 2018)	95.6	98.0	89.2	94.3
BSP+DANN (Proposed)	94.5	97.7	89.4	93.9
BSP+ADDA (Proposed)	93.3	94.5	91.4	93.1
BSP+CDAN (Proposed)	95.0	**98.1**	**92.1**	**95.1**

$d_A = 2(1-2\epsilon)$, where ϵ is the error rate of a domain classifier trained to discriminate source domain and target domain. We employ the A-distance as a measure of transferability of feature representations. Results on tasks **A** →**W** & **W** →**D** are shown in Figure 4(b). The A-distance with features of **BSP+DANN** is smaller than that of **DANN**. This, combined with the experimental results above, reveals that **BSP** enhances not only discriminability but also transferability.

Table 5. Accuracy on VisDA-2017 for sensitivity of k.

k	1	2	4	8
Avg	72.1	72.3	71.9	71.5

Sensitivity Analysis. Sensitivity of a larger k in Equation (8) is conducted on the VisDA-2017 dataset, with results shown in Table 5. It is inspiring that results are better when $k = 2$, while performance drops a little with a larger k in BSP. When $k = 1$, once the largest singular value has been suppressed in an iteration, the second largest one may stand out and be suppressed in the next iteration. Thus $k = 1$ is good for most datasets. When k gets large, smaller singular values are also suppressed, which contradicts our analysis.

5. Related Work

The goal of domain adaptation (Candela et al., 2009; Pan et al., 2010) is to transfer knowledge from one domain to anther by reducing domain shift. Deep networks prove to be able to learn transferable representations (Yosinski et al., 2014). However, feature representations extracted by deep networks can only reduce, but not remove, the cross-domain discrepancy (Glorot et al., 2011). To mitigate the problem, some domain adaptation methods plug adaptation layers (Long et al., 2015; Li et al., 2016; Maria Carlucci et al., 2017; Long et al., 2017) or adaptive normalization layers (Li et al., 2016; Maria Carlucci et al., 2017) in deep networks to match the feature distributions of the source and target domains. In other methods, a subnetwork called domain discriminator, which is trained to distinguish source data from target data and the deep features are learned to confuse the discriminator in domain-adversarial training (Ganin et al., 2016; Tzeng et al., 2017; Luo et al., 2017; Long et al., 2018). The pixel-level adaptation methods provide alternative thinking for domain adaptation. Liu & Tuzel (2016) used two GANs to generate the source and target images respectively, and Gatys et al. (2016) introduced an additional reconstruction objective on the target domain for pixel-level adaptation. Chao et al. (2018) studied cross-dataset adaptation for Visual Question Answering. Tsai et al. (2018) imposed two independent domain discriminators on the feature and class layers for output space adaptation. Without domain discriminator, Saito et al. (2018) performed a minimax game between the feature generator and the two-branch classifiers to minimize the domain shift, while Kim et al. (2019) proposed a systematic and effective way to achieve hypothesis consistency through Gaussian processes.

All of the above methods enhance transferability of features in a variety of ways. In this paper, we try to improve domain adaptation methods based on in-depth analysis of discriminability, from a brand new perspective.

6. Conclusion

In this paper, we investigate how to learn feature representations that are both transferable and discriminative for deep domain adaptation. While previous works mainly focus on the transferability, we find that existing adversarial domain adaptation methods may inevitably deteriorate the discriminability. We thus propose a regularization approach based on spectral analysis of the feature representations. The approach is general and pluggable into any adversarial domain adaptation networks to yield significant performance gains.

Acknowledgements

We thank Yuchen Zhang at Tsinghua for insightful discussions. This work was supported by the National Key R&D Program of China (2017YFC1502003) and Natural Science Foundation of China (61772299, 71690231, and 61672313).

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Candela, J. Q., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. Dataset shift in machine learning, 2009.

Chao, W.-L., Hu, H., and Sha, F. Cross-dataset adaptation for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning(ICML)*, pp. 647–655, 2014.

Fukunaga, K. Introduction to statistical pattern recognition. 1990.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, pp. 1180–1189, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17(1): 2096–2030, 2016.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

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

Golub, G. H. and Reinsch, C. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14 (5):403–420, 1970.

Golub, G. H. and Van der Vorst, H. A. Eigenvalue computation in the 20th century. In *Numerical analysis: historical developments in the 20th century*, pp. 209–239. Elsevier, 2001.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2066–2073. IEEE, 2012.

Gong, B., Grauman, K., and Sha, F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 222–230, 2013.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 601–608, 2007.

Kim, M., Sahu, P., Gholami, B., and Pavlovic, V. Unsupervised visual domain adaptation: A deep max-margin gaussian process approach. *arXiv preprint arXiv:1902.08727*, 2019.

Li, Y., Wang, N., Shi, J., Liu, J., and Hou, X. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Liu, M.-Y. and Tuzel, O. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 469–477, 2016.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, pp. 97–105, 2015.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning (ICML)*, pp. 2208–2217. JMLR. org, 2017.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1647–1657, 2018.

Luo, Z., Zou, Y., Hoffman, J., and Fei-Fei, L. F. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 165–177, 2017.

Maria Carlucci, F., Porzi, L., Caputo, B., Ricci, E., and Rota Bulo, S. Autodial: Automatic domain alignment layers. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 5067–5075, 2017.

Mohammadi, B. Principal angles between subspaces and reduced order modelling accuracy in optimization. *Structural and Multidisciplinary Optimization*, 50(2):237–252, 2014.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–1724, 2014.

Pan, S. J., Yang, Q., et al. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010.

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

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, pp. 213–226. Springer, 2010.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3723–3732, 2018.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8503–8512, 2018.

Shonkwiler, C. Poincaré duality angles for riemannian manifolds with boundary. *arXiv preprint arXiv:0909.1967*, 2009.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528. IEEE, 2011.

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M. Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 4068–4076, 2015.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 4, 2017.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394. IEEE, 2017.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3320–3328, 2014.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.