Transferability vs. Discriminability: Batch Spectral Penalization for Adversarial Domain Adaptation (2019)

Xinyang Chen, Sinan Wang, Mingsheng Long, Jianmin Wang

Summary

Contributions

Domain adaptation (DA) tackles the problem of learning a model that reduces the dataset shift between training and testing distributions. One of the main strategies in DA is 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, yielding remarkable performance gains. While it is widely believed that adversarial learning strengthens the feature transferability, but which ones are made more transferable and how the feature discriminability will change in the process of learning transferable features? This paper tries to understand this trade-off between discriminability and transferability and understand their behaviors by studying the singular values of the learned features and by using linear discriminate analysis. And by analyzing the results, they propose Batch Spectral Penalization (BSP) as additional regularization to push the model to produce more discriminable features.

Method

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.

Discriminability of Feature Representations

The authors start by investigating the transferability and discriminability of the learned feature representations, to evaluate the discriminability, they propose different measures:

Linear Discriminant Analysis (LDA) to compare the between-class and within-class variances $\mathbf{S}_b$ and $\mathbf{S}_w$, with c classes and $\mathbf{f}$ as the output of the feature extractor $F(x)$, we compute them as follows:

$$\mathbf{S}_b = \sum_{j=1} n_j \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}\right) \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}\right)^\top$$

$$\mathbf{S}_w = \sum_{j=1}^{c} \sum_{\mathbf{f} \in \mathcal{F}_j} \left(\mathbf{f} - \boldsymbol{\mu}_j\right) \left(\mathbf{f} - \boldsymbol{\mu}_j\right)^\top$$

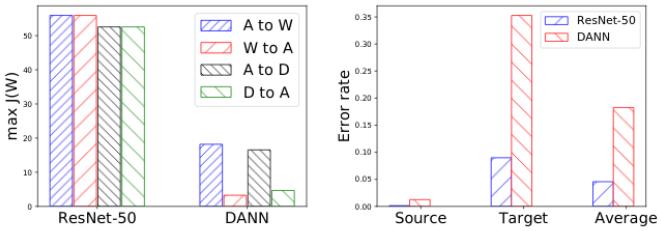
Using $\mathbf{S}_b$ and $\mathbf{S}_w$ we need to find the representations $\mathbf{W}$ with the largest ratio, implying strong discriminability:

$$\arg\max_{\mathbf{w}} J(\mathbf{W}) = \frac{\text{tr}\left(\mathbf{W}^\top \mathbf{S_b} \mathbf{W}\right)}{\text{tr}\left(\mathbf{W}^\top \mathbf{S_w} \mathbf{W}\right)}$$

with an optimal solution $\mathbf{W}^* = \mathbf{U}$ where $\mathbf{U}$ is found by the SVD of $\mathbf{S}_w^{-1}\mathbf{S}_b = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we can them compare $\max J(\mathbf{W})$ of both DANN and a resnet50.

Target error with an MLP Motivated by the domain adaptation theory, the authors train a multilayer perceptrons (MLP) classifier on top of the the representations learned by DANN and a Resnet50 with a fixed feature extractor F. In this case and only this case, the target labels are used.

The results of both measure are displayed in the figure bellow:



(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.

The results show that the feature discriminability of DANN is lower than that of ResNet-50, implying that DANNs transferability is enhanced at the expense of worse discriminability. Similar results for the target error, 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.

Why Worse Discriminability?

To see why we end-up with worst discriminability, we compute the SVD of the source and target features separately, $\mathbf{F}_s = [\mathbf{f}_1^* \dots \mathbf{f}_b^*]$ and $\mathbf{F}_t = [\mathbf{f}_1^t \dots \mathbf{f}_b^t]$, which are computed for a small batch of size b:

$$\mathbf{F}_s = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{V}_s^\top$$
$$\mathbf{F}_t = \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^\top$$

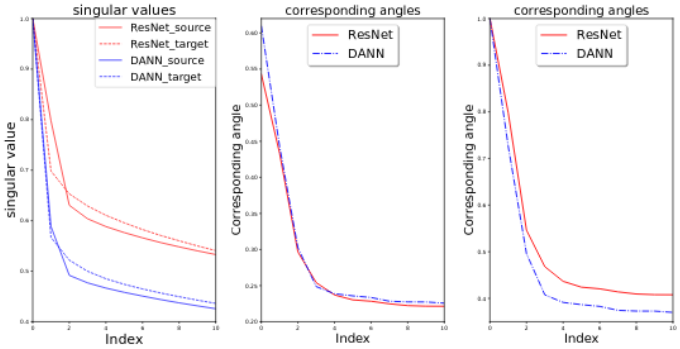
In addition to plotting the singular values σ in descending order, the authors also compute and plot the principal angles θ

$$\cos(\theta_i) = \frac{\langle \mathbf{a}_i, \mathbf{b}_i \rangle}{\|\mathbf{a}_i\| \, \|\mathbf{b}_i\|} = \max_{\mathbf{a},\mathbf{b}} \left\{ \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} : \mathbf{a} \perp \mathbf{a}_j, \mathbf{b} \perp \mathbf{b}_j, j = 1, \dots, i-1 \right\}$$

where $\mathbf{a}_i$ is an eigenvector of source feature matrix and $\mathbf{b}_i$ is an eigenvector of target feature matrix. But instead of finding the angles between any two eigen vector from source and target, the authors propose to compute it only over vector with the same position, and call this angle the Corresponding Angle:

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

The results are illustrated bellow:



(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.

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. For the Corresponding Angle, we observe that the first angle much larger that the rest, 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 eigen vectors for transferability in DANN.

BPS

As observed above, it is necessary to suppress the dimension with top singular valueto prevent it from standing out. To this end, the authors propose Batch Spectral Penalization as a regularization term over these largest k singular values:

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

Computed in pytorch as follows, where the target and source features are concatenated, and fed as input to the function:

```
def BSP(feature):
    feature_s = feature.narrow(0, 0, int(feature.size(0) / 2))
    feature_t = feature.narrow(0, int(feature.size(0) / 2),
                                  int(feature.size(0) / 2))
    _, s_s, _ = torch.svd(feature_s)
    _, s_t, _ = torch.svd(feature_t)
    sigma = torch.pow(s_s[0], 2) + torch.pow(s_t[0], 2)
    return sigma
```

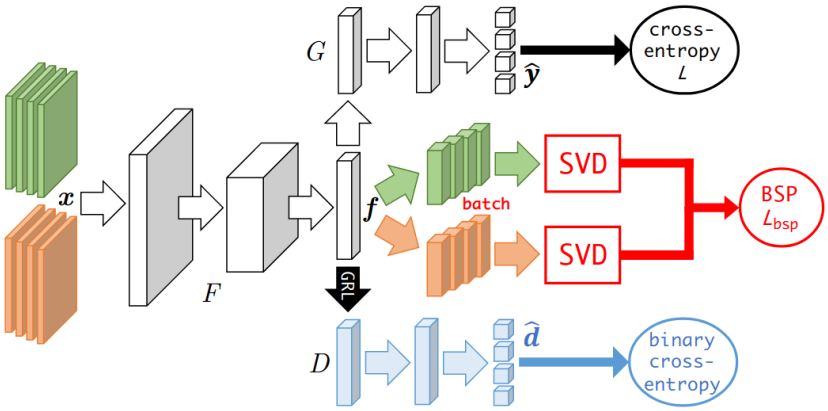


Figure 3. The architecture of **BSP+DANN** where BSP enhances discriminability while learning transferable features via domain adversarial network (DANN). BSP is a lightweight module readily pluggable into any deep domain adaptation networks, which is end-to-end trainable with the support of **differentiable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

Results

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

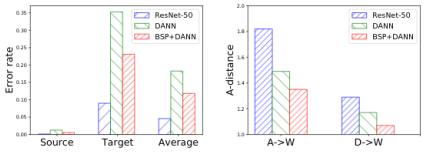
Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50 (He et al., 2016)	68.4 ± 0.2	96.7 ± 0.1	99.3 ± 0.1	68.9 ± 0.2	62.5 ± 0.3	60.7 ± 0.3	76.1
DAN (Long et al., 2015)	80.5 ± 0.4	97.1 ± 0.2	99.6 ± 0.1	78.6 ± 0.2	63.6 ± 0.3	62.8 ± 0.2	80.4
DANN (Ganin et al., 2016)	82.0 ± 0.4	96.9 ± 0.2	99.1 ± 0.1	79.7 ± 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**

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

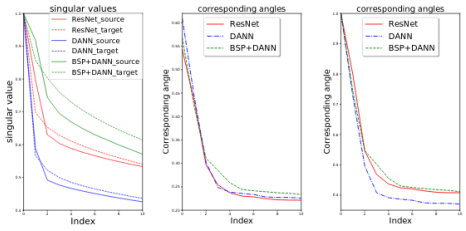
Method	plane	bcybl	bus	car	horse	knife	mcyle	person	plant	sktbrd	train	truck	mean
ResNet-101 (He et al., 2016)	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN (Long et al., 2015)	87.1	63.0	76.5	42.0	**90.3**	42.9	85.9	53.1	49.7	36.3	**85.8**	20.7	61.1
DANN (Ganin et al., 2016)	81.9	**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**



(a) Error Rate (b) A-distance

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

(a)



(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.

(b)