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

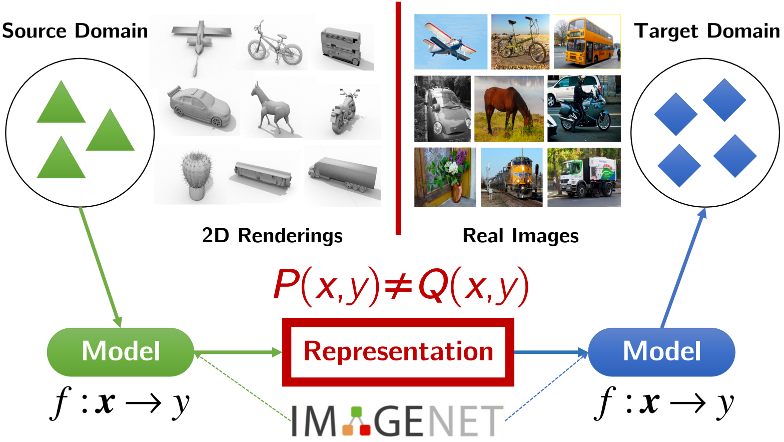
Xinyang Chen, Sinan Wang, Mingsheng Long, Jianmin Wang

School of Software
BNRist, Research Center for Big Data
Tsinghua University

International Conference on Machine Learning, 2019

Transfer Learning: Unsupervised Domain Adaptation

- Non-IID distributions $P \neq Q$
- Only unlabeled data in target domain



Adversarial Domain Adaption

Matching distributions across source and target domains s.t. $P \approx Q$

Adversarial adaptation: learning features indistinguishable across domains

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

We analysis features extracted by DANN[1] with a ResNet-50[2] pretrained on Imagenet

$$\begin{aligned} \mathcal{E}(F,G) &= \mathbb{E}_{(\mathbf{x}_i^s,\mathbf{y}_i^s)\sim P} L(G(F(\mathbf{x}_i^s)),\mathbf{y}_i^s) \\ \mathrm{dist}_{P \leftrightarrow Q}(F,D) &= \mathbb{E}_{\mathbf{x}_i^s \sim P} \log[D(\mathbf{f}_i^s)] \\ &+ \mathbb{E}_{\mathbf{x}_i^t \sim Q} \log[1 - D(\mathbf{f}_i^t)] \end{aligned} \tag{2}$$

[1] *Ganin et al. Unsupervised domain adaptation by backpropagation. ICML '15.*

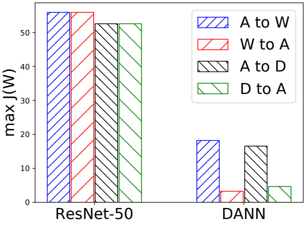
[2] *He et al. Deep residual learning for image recognition. CVPR '15.*

Discriminability of Feature Representations

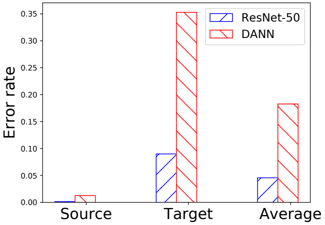
Two key criteria that characterize the goodness of feature representations

- Transferability: the ability of feature to bridge the discrepancy across domains
- Discriminability: the easiness of separating different categories by a supervised classifier trained over the feature representations

Discriminability of features extracted by DANN, worse discriminability is found:



(a) max $J(W)$

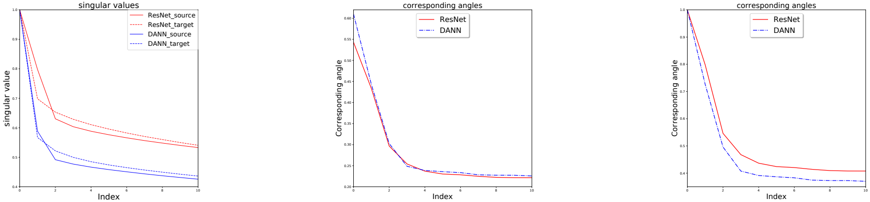


(b) Classification error rate

Figure: Analysis of discriminability of feature: (a) LDA, (b) MLP.

Why Discriminability Is Weakened?

- Corresponding Angles: corresponding angle is the angle between two eigenvectors corresponding to the same singular value index, which are equally important in their feature matrices.

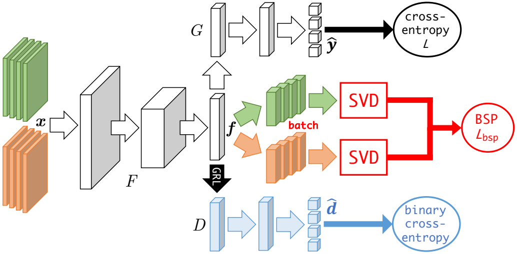


(a) σ (b) $\cos(\psi)$ (c) $\cos(\psi)$

Figure: SVD analysis. We compute (a) the singular values (normalized version); (b) corresponding angles (unnormalized version); (c) corresponding angles (normalized version). In normalized version we scale all values so that the largest value is 1.

- Only the eigenvectors with largest singular values tend to carry transferable knowledge

BSP: Batch Spectral Penalization



BSP combined with DANN to strengthen discriminability of feature

$$\min_{F,G} \mathcal{E}(F,G) + \gamma \mathrm{dis}_{P \leftrightarrow Q}(F,D) + \beta L_{\mathrm{bsp}}(F)$$
$$\max_{D} \ \mathrm{dis}_{P \leftrightarrow Q}(F,D), \tag{3}$$

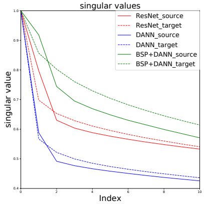
$$L_{\mathrm{bsp}}(F) = \sum_{i=1}^{k} (\sigma_{s,i}^2 + \sigma_{t,i}^2), \tag{4}$$

Results

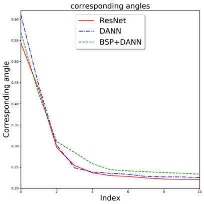
Table: Accuracy (%) on *Office-31* for unsupervised domain adaptation

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50	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	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	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	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	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	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	**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**

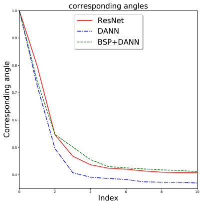
Analysis



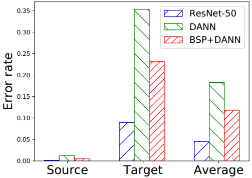
(a) σ



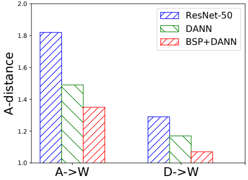
(b) $\cos(\psi)$



(c) $\cos(\psi)$



(d) Classification error



(e) A-distance

Thanks!

Poster: tonight at Pacific Ballroom #256