Cluster Alignment with a Teacher for Unsupervised Domain Adaptation

Zhijie Deng, Yucen Luo, Jun Zhu

Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Lab, THBI Lab, Tsinghua University

Tsinghua University

ICCV 2019 Seoul, Korea

Introduction

❑ **Deep learning** methods have shown promise in **unsupervised domain adaptation (UDA)**, which aims to leverage a labeled source domain to learn a classifier for the unlabeled target domain with a different distribution.

❑ **Marginal distribution alignment:**

1. Adversarial training [Tzeng et al., 2017; Ganin & Lempitsky, 2015]:

$$\min_{\theta}\max_{\phi} L_y(X_s, Y_s;\theta) + \alpha\big(E_{x\sim X_s}[\log c(f(x;\theta);\phi)] + E_{x\sim X_t}[\log(1-c(f(x;\theta);\phi))]\big)$$

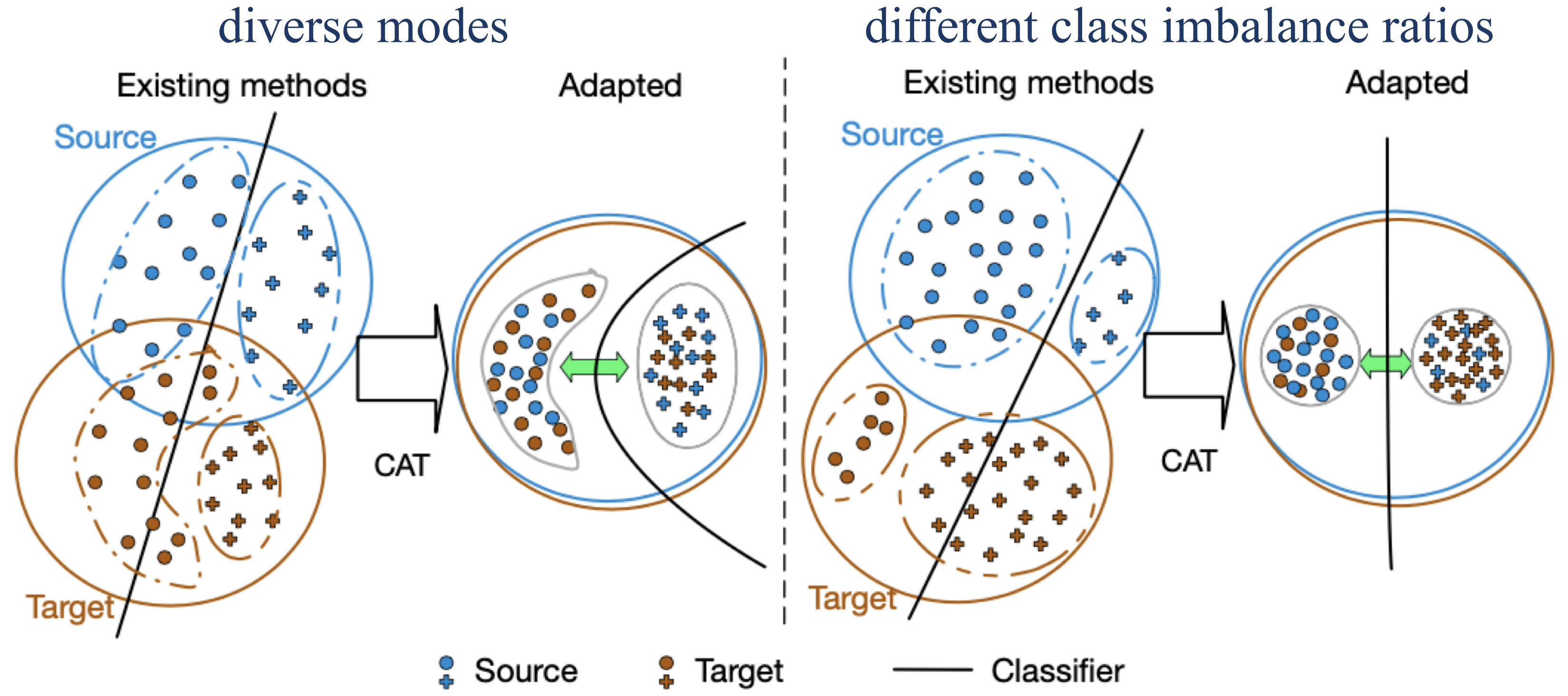
2. Kernelized training [Long et al., 2015]:

$$\min_{\theta} L_y(X_s, Y_s;\theta) + \alpha \mathrm{MMD}(f(X_s;\theta), f(X_t;\theta))$$

- Theoretical guarantee [Ben-David et al., 2010]: minimizing the divergence between the marginal distributions in the learned feature space is beneficial to reduce the classifier's error on target domain.

❑ **Observation: aligning the marginal is not enough in practice!**

diverse modes　　　different class imbalance ratios



- The classification data naturally presents a class-conditional multi-modal structure owing to the semantic similarity of samples from the same class.
- Existing methods aligning the marginal distributions while ignoring the class-conditional structures cannot perform well in challenging cases.

❑ **Motivation: incorporating the fine-grained class-conditional structure**

- Previous works (Shi & Sha, 2012; Pang et al., 2018) have validated that utilizing the class-conditional structure of data is beneficial in various tasks.
- In particular, matching the class-conditional structure in UDA enhances the discriminative power of the learned domain-invariant feature space and is compatible to the marginal distribution alignment methods.

Theoretical insight

$$\begin{aligned}\epsilon_t(h) &\le \epsilon_s(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(s,t) \\ &\quad + \min_{\hat{h}\in\mathcal{H}}(\epsilon_s(\hat{h}, l_s) + \epsilon_t(\hat{h}, l_t)) \\ &\le \epsilon_s(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(s,t) + \boxed{\epsilon_t(l_s, l_t)} \\ &\quad + \min_{\hat{h}\in\mathcal{H}}(\epsilon_s(\hat{h}, l_s) + \epsilon_t(\hat{h}, l_s))\end{aligned}$$

■ The target error of classifier h has this bound [Ben-David et al., 2010].

■ $\epsilon_t(l_s, l_t)$ could be large if the class-conditional structures, which determine the labeling functions, are not aligned, leading to unsatisfactory bound of $\epsilon_t(h)$.

Methodology

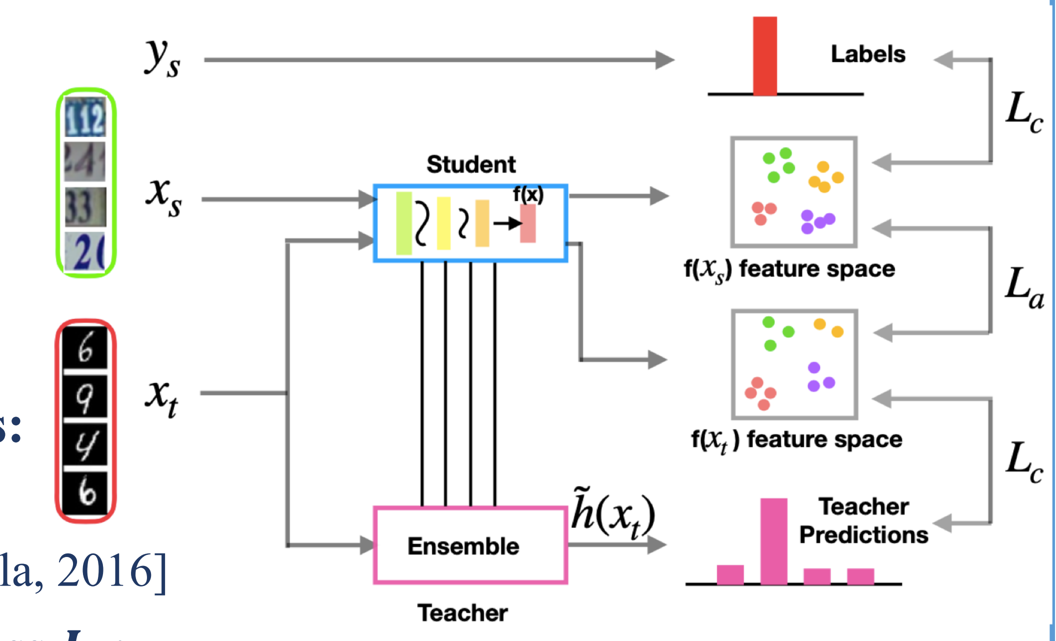
❑ **Overall objective:**

$$\min_{\theta} L_y + \alpha(L_c + L_a)$$

where α is a coefficient.

❑ **Build a teacher classifier to annotate the target samples:**

- Π model [Laine & Aila, 2016]
- Temporal ensemble [Laine & Aila, 2016]



❑ **Discriminative clustering loss L_c:**

- $L_c(X_s, X_t) = L_c(X_s) + L_c(X_t)$
- $L_c(X) = \frac{1}{|X|^2}\sum_{i,j}[\delta_{ij} d(f(x^i), f(x^j)) + (1-\delta_{ij})\max(0, m - d(f(x^i), f(x^j)))]$
- Concentrates features from the same class and separates features from different classes.

❑ **Cluster alignment loss L_a:**

- $L_a(X_s, X_t) = \frac{1}{K}\sum_{k=1}^{K} \| \lambda_{s,k} - \lambda_{t,k} \|_2^2$
- $\lambda_{s,k} = \frac{1}{|X_{s,k}|}\sum_{x_s^i\in X_{s,k}} f(x_s^i),\ \lambda_{t,k} = \frac{1}{|X_{t,k}|}\sum_{x_t^i\in X_{t,k}} f(x_t^i)$
- Works in a conditional feature matching way [Salimans et al., 2016], and can match the conditional distributions across domains theoretically.

❑ **Our method can be integrated into any marginal distribution alignment method when domains have analogous marginal distributions.**

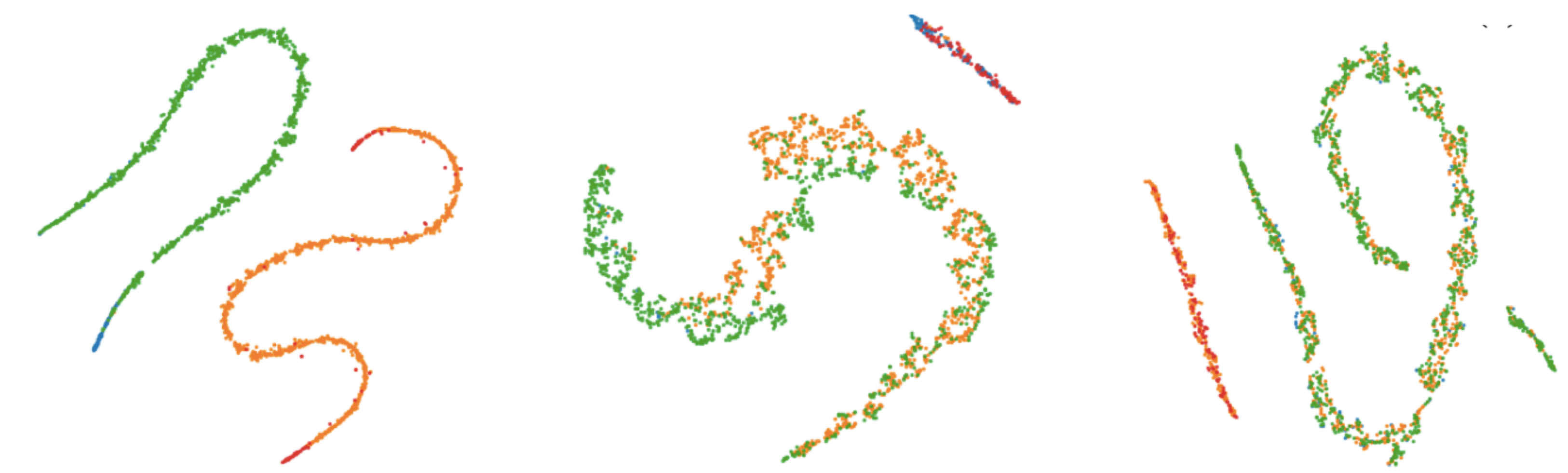
- RevGrad+CAT: $\min_{\theta}\max_{\phi} L_y(X_s, Y_s;\theta) + \alpha\big(E_{x\sim X_s}[\log c(f(x;\theta);\phi)] + E_{x\sim X_t}[\log(1-c(f(x;\theta);\phi))] + L_c + L_a\big)$

Experiments

❑ **Imbalanced *SVHN-MNIST-USPS* (synthetic task)**

➢ A challenging 2-class classification task: the source domains have 10 : 1 ratio of class imbalance while the target domains have 1 : 10.

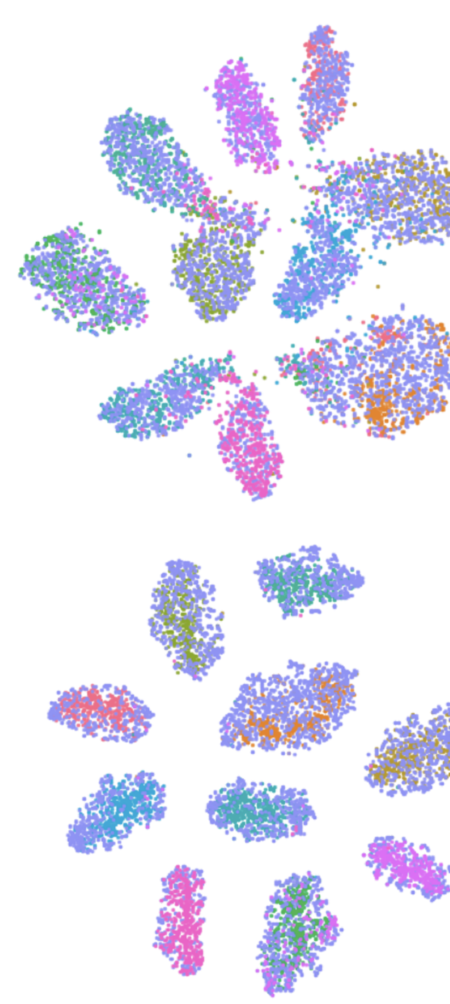
Method	*SVHN to MNIST*	*MNIST to USPS*	*USPS to MNIST*
RevGrad [7]	27.4 ± 6.3	26.7 ± 2.0	17.9 ± 1.4
MSTN [49]	25.8 ± 3.6	30.3 ± 1.0	29.4 ± 0.5
CAT	$\mathbf{100.0} \pm 0.05$	$\mathbf{100.0} \pm 0.0$	$\mathbf{99.9} \pm 0.2$



(a) CAT　　(b) RevGrad　　(c) MSTN

❑ ***SVHN-MNIST-USPS* task**

Method	*SVHN to MNIST*	*MNIST to USPS*	*USPS to MNIST*
Source Only	60.1 ± 1.1	75.2 ± 1.6	57.1 ± 1.7
DDC [45]	68.1 ± 0.3	79.1 ± 0.5	66.5 ± 3.3
CoGAN [20]	-	91.2 ± 0.8	89.1 ± 0.8
DRCN [8]	82.0 ± 0.1	91.8 ± 0.09	73.7 ± 0.04
ADDA [44]	76.0 ± 1.8	89.4 ± 0.2	90.1 ± 0.8
LEL [26]	81.0 ± 0.3	-	-
AssocDA [11]	97.6	-	-
MSTN [49]	91.7 ± 1.5	92.9 ± 1.1	-
CAT	98.1 ± 1.3	90.6 ± 2.3	80.9 ± 3.1
RevGrad [7]	73.9	77.1 ± 1.8	73.0 ± 2.0
RevGrad+CAT	98.0 ± 0.8	93.7 ± 1.1	95.7 ± 1.3
rRevGrad+CAT	$\mathbf{98.8} \pm 0.02$	94.0 ± 0.7	$\mathbf{96.0} \pm 0.9$
MCD [37]	96.2 ± 0.4	94.2 ± 0.7	94.1 ± 0.3
MCD+CAT	97.1 ± 0.2	$\mathbf{96.3} \pm 0.5$	95.2 ± 0.4
VADA [41]	94.5	-	-
VADA+CAT	95.2	-	-



(a) RevGrad

(b) rRevGrad +CAT

Conclusion

❑ **We propose CAT to exploit the class-conditional structures for effective adaptation in deep UDA.**

❑ **CAT is compatible to most existing UDA methods.**

❑ **CAT establishes new state-of-the-art baselines on a range of benchmarks.**

www.PosterPresentations.com