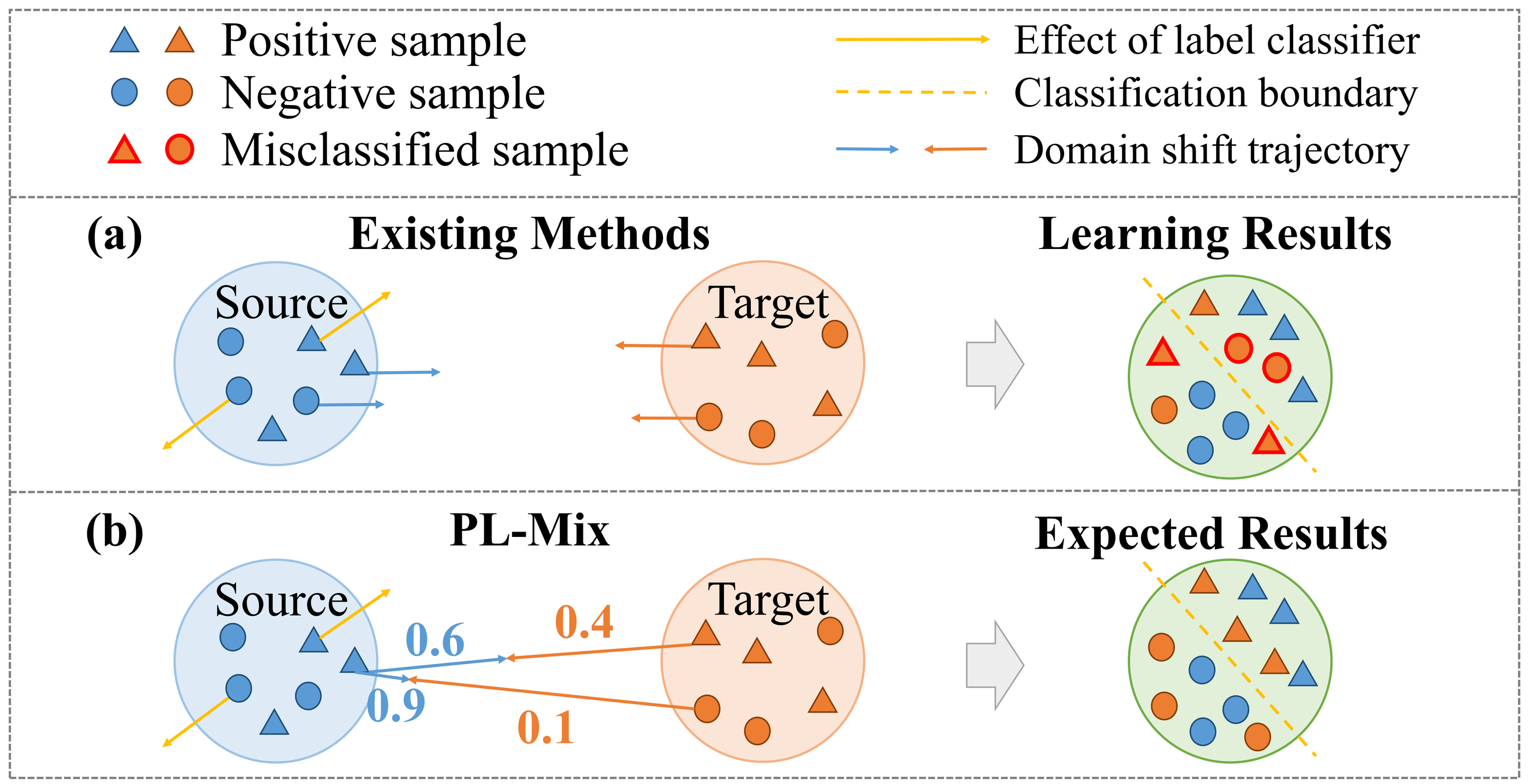
BDBC and SKLSDE
Beihang University, Beijing, China

On Unsupervised Domain Adaptation: Pseudo Label Guided Mixup for Adversarial Prompt Tuning

Fanshuang Kong[1], Richong Zhang[1,2], Ziqiao Wang[3], Yongyi Mao[3]
[1]SKLSDE, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]School of Electrical Engineering and Computer Science, University of Ottawa, Canada
{kongfs, zhangrc}@act.buaa.edu.cn, {zwang286, ymao}@uottawa.ca

GitHub Link
https://github.com/fskong/PL-Mix

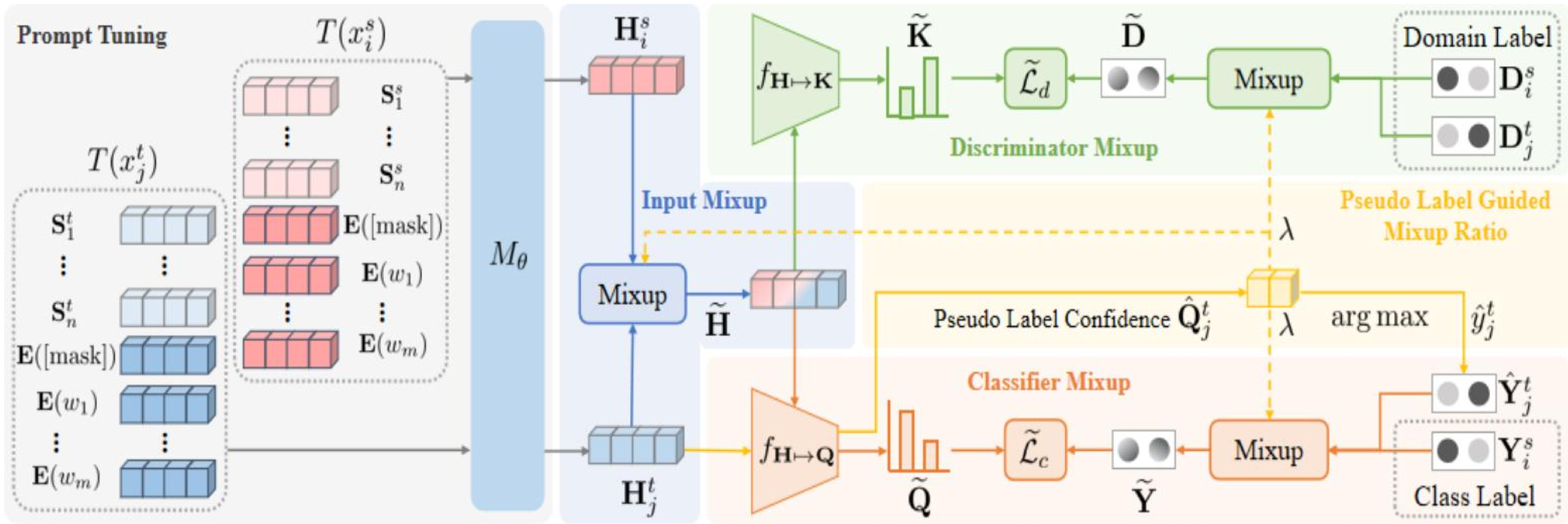
Motivation



No explicit mechanism that facilitates the positive (or negative) data of the source domain to be attracted towards the corresponding positive (or negative).

- The fusion of **pseudo labels** and **Mixup** creates intermediate synthetic data between source and target data of the same class, thereby **promoting the alignment between the two domains**.
- The source data should **move more** (Mixup ratio 0.6) toward the target data with **a same and highly confident label**.
- While the **movement should be limited** (Mixup ratio 0.9) for **a distinct and highly confident label**.

PL-Mix



☐ Confidence-dependent Mixup Ratio

Pseudo label confidence: $\hat{\mathbf{Q}}_j^{t[\hat{y}_j^t]}$

$$\alpha = \begin{cases} \hat{\mathbf{Q}}_j^{t[\hat{y}_j^t]}, & \text{if } \hat{y}_j^t = y_i^s \\ 1 - \hat{\mathbf{Q}}_j^{t[\hat{y}_j^t]}, & \text{otherwise} \end{cases}$$

Mixup ratio: $\lambda \sim Beta(\alpha, \alpha)$

$$\lambda := \begin{cases} 1, & \text{if } \hat{\mathbf{Q}}_j^{t[\hat{y}_j^t]} < \tau \\ \max(1-\lambda, \lambda), & \text{otherwise} \end{cases}$$

☐ Classifier Mixup

Pseudo class label Mix

$$\widetilde{\mathbf{H}} = \lambda \mathbf{H}_i^s + (1-\lambda)\mathbf{H}_j^t$$
$$\widetilde{\mathbf{Y}} = \lambda \mathbf{Y}_i^s + (1-\lambda)\hat{\mathbf{Y}}_j^t$$

Cross entropy loss of Classifier

$$\widetilde{\mathcal{L}}_c = -\sum_{i=1}^{N^s+N^t}\sum_{j=1}^{|\mathcal{Y}|}\widetilde{\mathbf{Y}}_i^{[j]}\log\widetilde{\mathbf{Q}}_i^{[j]}$$ **min**

☐ Discriminator Mixup

Domain label Mix

$$\widetilde{\mathbf{H}} = \lambda \mathbf{H}_i^s + (1-\lambda)\mathbf{H}_j^t$$
$$\widetilde{\mathbf{D}} = \lambda \mathbf{D}_i^s + (1-\lambda)\mathbf{D}_j^t$$

Cross entropy loss of Discriminator

max $$\widetilde{\mathcal{L}}_d = -\sum_{i=1}^{N^s+N^t}\sum_{j=1}^{|\mathcal{D}|}\widetilde{\mathbf{D}}_i^{[j]}\log\widetilde{\mathbf{K}}_i^{[j]}$$

PL-Mix Improves Generalization

Theorem 1 *Let the function space of F have the finite Natarajan dimension d_N. Assume that the loss function $\mathcal{L}_c(\cdot,\cdot;F)$ is R-subgaussian under P_{XY}^s. Then, for any F, there exists a constant $C > 0$ such that with probability $1-\delta$*

$$\mathcal{E}(F) \le C\sqrt{\frac{d_N \log|\mathcal{Y}| + \log\frac{1}{\delta}}{N_s}} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}\left(P_Z^t || P_Z^s\right)} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}\left(P_{Y|Z}^t || P_{\hat{Y}|Z}^t\right)} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}\left(P_{\hat{Y}|Z}^t || P_{Y|Z}^s\right)},$$

where $\mathrm{D}_{\mathrm{KL}}(\cdot||\cdot)$ denotes the KL divergence and $P_{\hat{Y}|Z}^t$ is the conditional pseudo label distribution of the target data.

☐ PL-Mix can improve the generalization guarantee

- Classifier Mixup Reduces the First Term
- Discriminator Mixup Reduces the Second Term
- Classifier Mixup Controls the Last Two Terms

Experiment

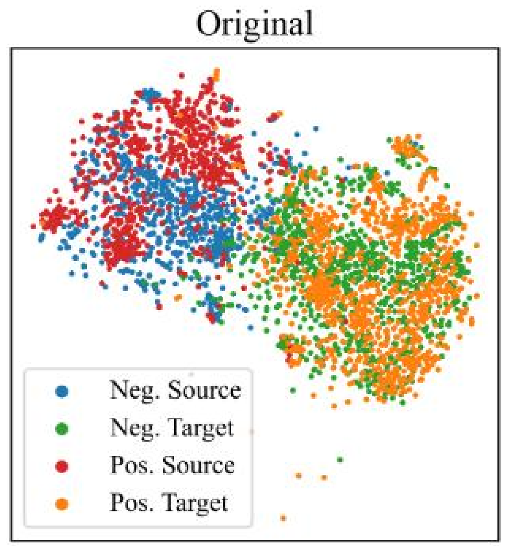
Bert-base-uncased

Model	B → D	B → E	B → K	D → B	D → E	D → K	E → B	E → D	E → K	K → B	K → D	K → E	Avg.
DANN	89.70	87.30	89.55	89.55	86.05	87.69	87.15	86.05	91.91	87.65	87.72	86.05	88.56
COBE	90.05	90.45	**92.90**	90.98	90.67	**92.00**	87.90	87.87	93.33	88.38	87.43	92.58	90.39
AdSPT	-	-	-	-	-	-	-	-	-	-	-	-	-
DANN	89.54	88.15	89.76	89.62	88.27	89.87	87.89	88.19	92.25	87.69	87.72	91.14	89.17
*COBE**	90.13	90.92	92.28	91.05	89.75	91.67	88.25	**88.88**	93.88	89.18	87.68	**92.87**	90.55
AdSPT	90.10	90.55	92.25	90.55	89.40	90.95	88.35	87.40	93.75	88.45	87.80	92.00	90.13
PL-Mix	**90.91**	**91.04**	91.82	**91.19**	**91.12**	91.84	**88.86**	88.56	**93.93**	**89.25**	**88.27**	92.77	**90.80**

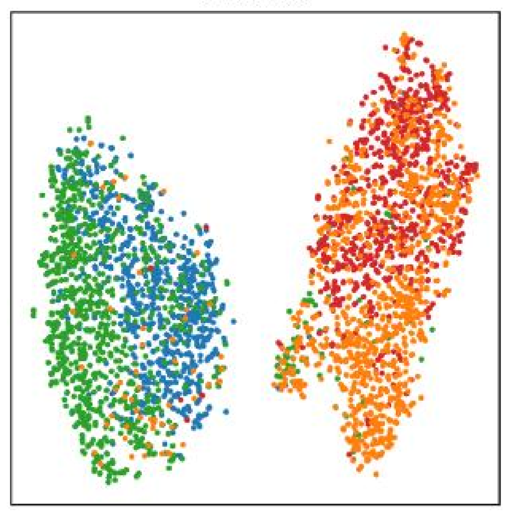
Roberta-base

Model	B → D	B → E	B → K	D → B	D → E	D → K	E → B	E → D	E → K	K → B	K → D	K → E	Avg.
DANN	-	-	-	-	-	-	-	-	-	-	-	-	-
COBE	-	-	-	-	-	-	-	-	-	-	-	-	-
AdSPT*	92.00	93.75	93.10	92.15	94.00	93.25	92.70	**93.15**	94.75	92.35	**92.55**	93.95	93.14
DANN	91.79	92.60	93.12	92.60	91.58	93.30	90.48	90.27	94.24	91.40	90.15	93.85	92.11
COBE	92.19	92.79	95.02	93.27	93.24	94.47	92.01	90.00	95.31	91.70	90.14	94.63	92.90
AdSPT	92.86	93.08	94.45	93.97	93.16	94.97	91.75	89.72	95.43	91.33	90.76	**94.70**	93.02
PL-Mix	**93.60**	**94.22**	**95.36**	**94.19**	**94.11**	**95.29**	**92.77**	92.02	**95.67**	**92.50**	91.71	94.65	**93.84**

Original



PL-Mix



☐ PL-Mix obtains convincing results

- PL-Mix outperforms SOTA models on Avg. both in Bert and Roberta
- PL-Mix aligns source data to target data according to their labels
- PL-Mix outperforms SOTA on multi-source setting (details in the paper)