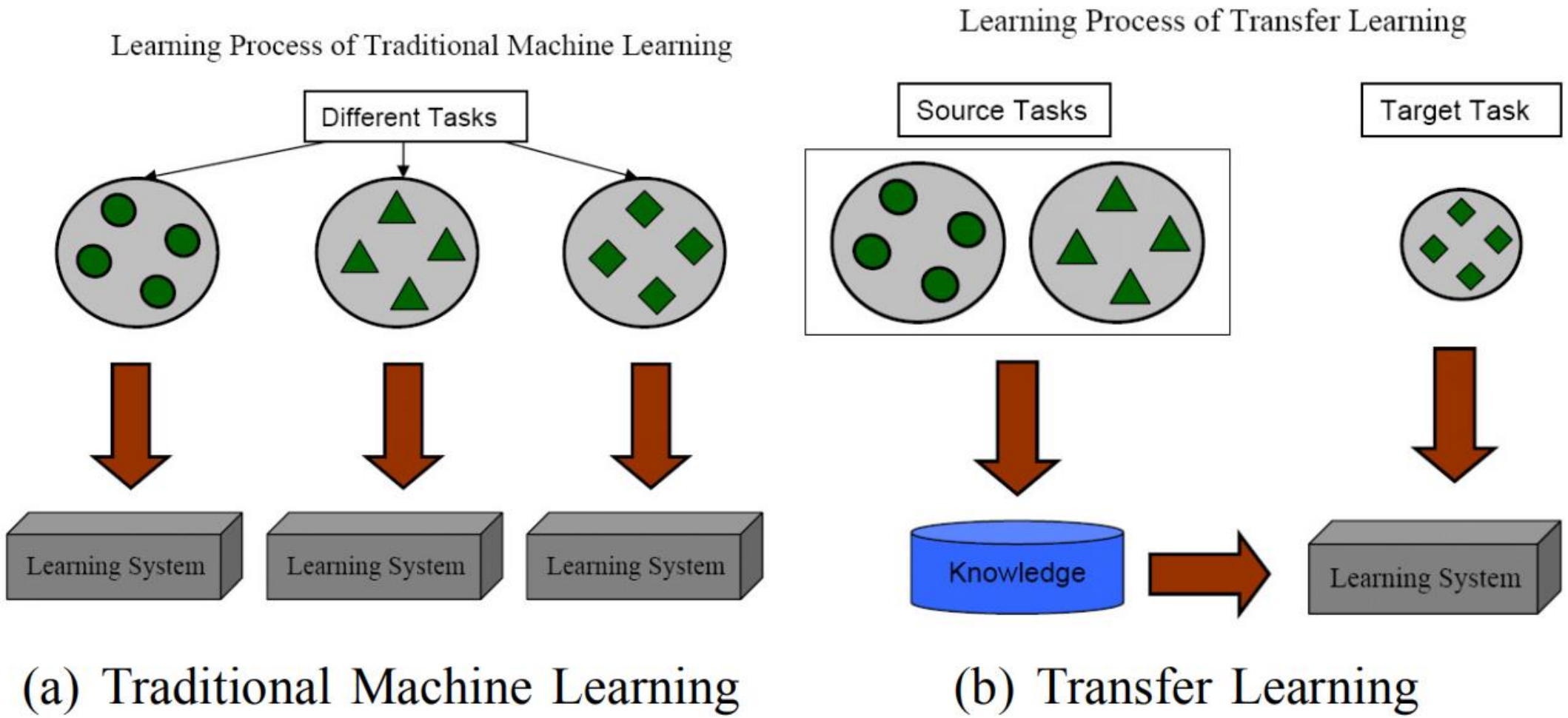
Characterizing and Avoiding Negative Transfer

Carnegie Mellon University

CVPR 2019

Introduction



(a) Traditional Machine Learning

(b) Transfer Learning

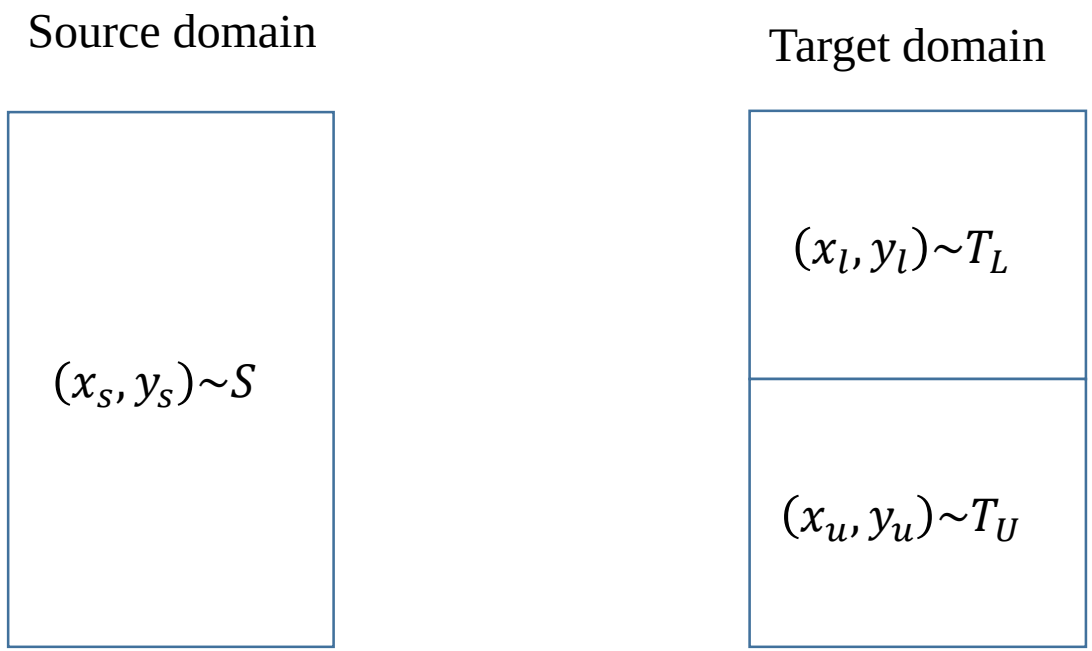
Introduction

Transfer learning uses knowledge learned in the source domain to assist training in the target domain.

Instance-transfer	To re-weight some labeled data in the source domain for use in the target domain.
Feature-representation-transfer	Find a “good” feature representation that reduces difference between the source and the target domains and the error of classification and regression models.

Negative transfer happens when the source domain data and task contribute to the **reduced performance** of learning in the target domain.

Domain Adversarial Neural Network



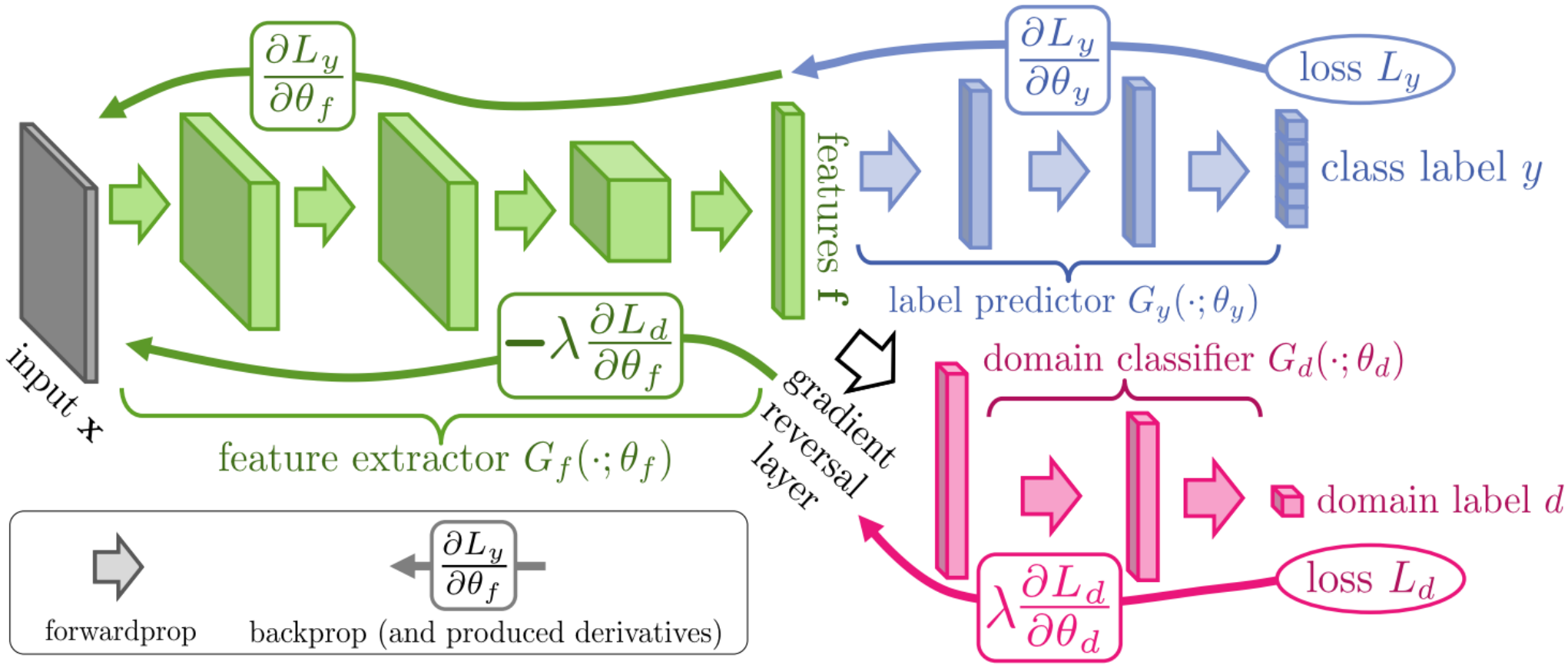
Domain Adversarial Neural Network

F	C	D
Feature extractor map both the source and target input to the same feature space	**Classifier** trained on both source and target data	**Discriminator** distinguish $P(F(X_S))$ and $P(F(X_T))$

$$\underset{F,C}{\operatorname{argmin}} \ \underset{D}{\operatorname{argmax}} \ \mathcal{L}_{\text{CLF}}(F,C) - \mu \mathcal{L}_{\text{ADV}}(F,D),$$

$$\begin{aligned}
\mathcal{L}_{\text{CLF}}(F,C) &= \mathbb{E}_{x_l,y_l \sim \mathcal{T}_L} \left[\ell_{\text{CLF}}(C(F(x_l)), y_l) \right] \\
&\quad + \mathbb{E}_{x_s,y_s \sim \mathcal{S}} \left[\ell_{\text{CLF}}(C(F(x_s)), y_s) \right], \\
\mathcal{L}_{\text{ADV}}(F,D) &= \mathbb{E}_{x_u \sim P_T(X)} \left[\log D(F(x_u)) \right] \\
&\quad + \mathbb{E}_{x_s \sim P_S(X)} \left[\log(1 - D(F(x_s))) \right].
\end{aligned}$$

Domain Adversarial Neural Network



Domain Adversarial Neural Network

Potential assumption

For any $x_s \in X_S$, there exists a $x_t \in X_T$ such that

$$P_S(Y|x_s) = P_T(Y|x_t) = P(Y|F(x_s))=P(Y|F(x_t)).$$

Not always be satisfied.

Discriminator Gate

Standard supervised learning objective

$$\begin{aligned}\mathcal{L}_{\text{SUP}} &= \mathbb{E}_{x,y\sim P_T(X,Y)}\left[\ell_{\text{CLF}}(C(F(x)),y)\right] \\ &= \mathbb{E}_{x,y\sim P_S(X,Y)}\left[\frac{P_T(x,y)}{P_S(x,y)}\ell_{\text{CLF}}(C(F(x)),y)\right]\end{aligned}$$

$$D(x,y) = \frac{P_T(x,y)}{P_T(x,y)+P_S(x,y)} \quad \frac{P_T(x,y)}{P_S(x,y)} = \frac{D(x,y)}{1-D(x,y)}.$$

Discriminator Gate

$$\mathcal{L}^{\text{gate}}_{\text{CLF}}(C, F) = \mathbb{E}_{x_l, y_l \sim \mathcal{T}_L} \left[\ell_{\text{CLF}}(C(F(x_l)), y_l) \right] + \lambda \mathbb{E}_{x_s, y_s \sim \mathcal{S}} \left[\omega(x_s, y_s) \ell_{\text{CLF}}(C(F(x_s)), y_s) \right]$$

$$\omega(x_s, y_s) = \mathbf{SG}\left(\frac{D(x_s, y_s)}{1 - D(x_s, y_s)} \right) \quad \text{(Stop gradient)}$$

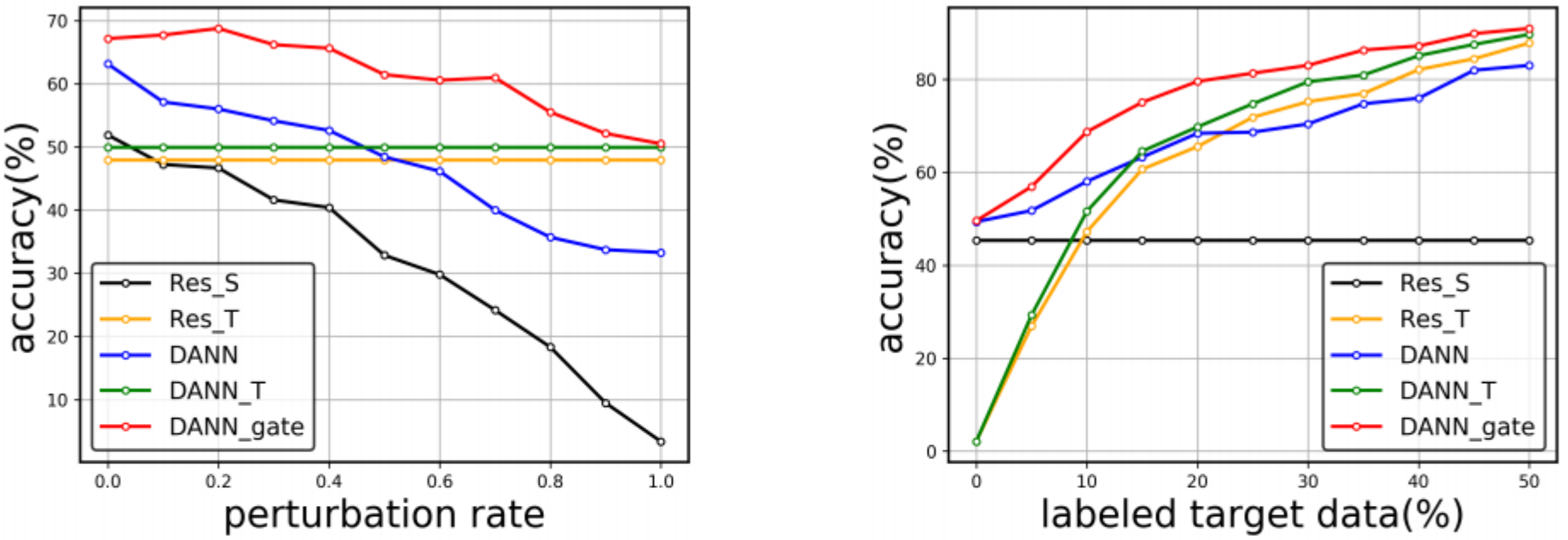
$$\begin{aligned} \mathcal{L}^{\text{aug}}_{\text{ADV}}(F, D) = {} & \mathbb{E}_{x_u \sim P_T(X)} \left[\log D(F(x_u), \texttt{nil}) \right] \\ & + \mathbb{E}_{x_s \sim P_S(X)} \left[\log(1 - D(F(x_s), \texttt{nil})) \right] \\ & + \mathbb{E}_{x_l, y_l \sim \mathcal{T}_L} \left[\log D(F(x_l), y_l) \right] \\ & + \mathbb{E}_{x_s, y_s \sim \mathcal{S}} \left[\log(1 - D(F(x_s), y_s)) \right], \end{aligned}$$

$$\underset{F,C}{\text{argmin}} \ \underset{D}{\text{argmax}} \ \mathcal{L}^{\text{gate}}_{\text{CLF}}(F, C) - \mu \mathcal{L}^{\text{aug}}_{\text{ADV}}(F, D).$$

Experiments

	W→D					A→D					
	ϵ =0.0	ϵ =0.3	ϵ =0.7	ϵ =0.9	Avg	ϵ =0.0	ϵ =0.3	ϵ =0.7	ϵ =0.9	Avg	$L_{\%}$
DANN	99.1±0.8	83.2±1.4	47.2±2.7	32.2±3.5	65.4	76.2±1.5	40.9±1.1	21.3±2.7	12.9±3.7	37.8	
NTG_1	-96.5	-80.3	-44.1	-28.3	-62.3	-73.7	-37.3	-17.2	-9.7	-34.5	
$DANN_{gate}$	98.9±0.6	83.3±2.1	48.4±2.5	32.1±3.1	65.7	76.0±1.2	41.0±1.6	21.5±3.1	13.2±2.4	37.9	0%
NTG_2	-96.3	-80.4	-45.3	-28.2	-62.6	-73.5	-37.4	-17.4	-10.0	-34.6	
Δ	↓0.2	↑0.1	↑1.2	↓0.1	↑0.3	↓0.2	↑0.1	↑0.2	↑0.3	↑0.1	
DANN	99.5±0.4	86.8±2.8	73.1±3.3	48.8±4.3	77.0	78.6±2.7	54.8±3.1	49.6±2.1	32.3±2.6	53.8	
NTG_1	-48.7	-37.8	-23.6	1.6	-27.1	-28.4	-4.4	1.2	18.4	-3.3	
$DANN_{gate}$	99.2±0.3	85.4±2.6	79.4±2.9	50.4±3.2	78.6	85.1±1.7	60.2±2.1	58.3±2.0	49.1±2.5	63.2	10%
NTG_2	-48.4	-36.4	-29.9	0.0	-28.7	-34.9	-9.8	-7.5	1.6	-12.7	
Δ	↓0.3	↓1.4	↑6.3	↑1.6	↑1.6	↑6.5	↑5.4	↑8.7	↑16.8	↑9.4	
DANN	99.6±0.2	89.7±1.6	78.4±2.5	70.5±4.3	84.6	80.2±2.0	73.3±2.2	70.2±3.3	51.3±4.3	68.8	
NTG_1	-18.5	-10.3	1.8	8.2	-4.7	-1.5	6.5	8.9	28.4	10.6	
$DANN_{gate}$	100.0±0.1	90.4±1.8	82.0±1.8	79.9±3.8	88.1	89.0±1.5	82.6±1.0	81.3±2.1	80.6±1.8	83.4	30%
NTG_2	-18.9	-11.0	-1.8	-1.2	-8.2	-10.3	-2.8	-2.2	-0.9	-4.1	
Δ	↑0.4	↑0.7	↑3.6	↑9.4	↑2.6	↑8.8	↑9.3	↑11.1	↑29.3	↑14.6	
DANN	100.0±0.0	92.2±1.7	85.8±2.3	78.2±4.8	89.1	84.5±1.9	77.6±3.8	70.6±4.9	65.4±6.3	74.5	
NTG_1	-11.7	-3.2	3.8	10.4	-0.2	4.6	12.1	18.8	23.2	14.7	
$DANN_{gate}$	100.0±0.0	93.3±1.7	91.2±1.5	89.5±3.4	92.5	93.2±1.3	91.4±1.2	90.2±2.0	89.8±1.9	91.2	50%
NTG_2	-11.7	-4.3	-1.6	-0.9	-4.6	-4.1	-1.7	-0.8	-1.2	-2.0	
Δ	→0.0	↑1.1	↑5.4	↑11.3	↑4.5	↑8.7	↑13.8	↑19.6	↑24.4	↑16.7	

Experiments



(a) $L_{\%}$ fixed at 20%

(b) ϵ fixed at 0.2

Figure 3. Incremental performance on task Pr→Rw. Res_S and Res_T are ResNet-50 baselines trained using only source data and only target data. Perturbation rates are set equal, i.e. $\epsilon = \epsilon_x = \epsilon_y$.

Experiments

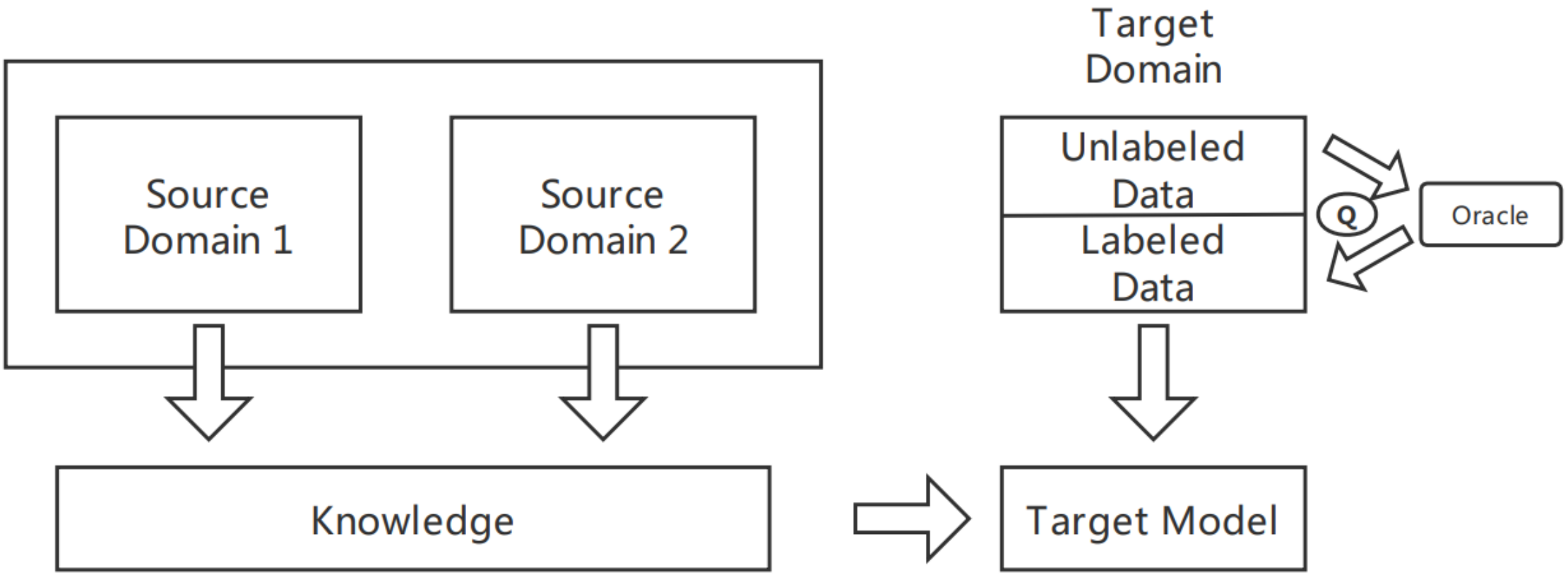
$$\epsilon = 0.7, L_{\%} = 10\%$$

Method	Digits	Office-31			Office-Home			VisDA	
	SVHN→MNIST	W→D	A→D	D→A	Ar→Rw	Cl→Rw	Pr→Rw	Synthetic→Real	Avg
TCA[19]	58.7(18.2)	54.2(-4.2)	11.4(20.5)	13.1(18.4)	-	-	-	-	34.4(13.2)
KMM[14]	70.9(6.0)	58.7(-8.5)	18.5(13.4)	17.7(13.8)	-	-	-	-	41.5(6.2)
DAN[16]	78.5(-4.4)	76.3(-19.5)	55.0(-1.3)	39.2(4.9)	43.2(3.8)	30.2(5.8)	47.2(4.0)	28.4(7.2)	49.8(0.1)
DAN_{gate}	82.2(-8.1)	78.7(-21.9)	60.4(-6.7)	43.9(0.2)	46.8(0.2)	38.0(-2.0)	50.4(0.8)	36.2(-0.6)	54.6(-4.7)
Δ_{DAN}	↑3.7	↑2.4	↑5.4	↑4.7	↑3.6	↑7.8	↑3.2	↑7.8	↑4.8
DCORAL[27]	75.2(-1.2)	75.7(-18.9)	53.8(-0.4)	37.4(5.0)	44.0(3.7)	32.4(4.1)	48.0(2.2)	30.5(5.7)	49.6(0.0)
$DCORAL_{gate}$	81.0(-7.0)	78.2(-21.4)	59.0(-5.6)	43.2(-0.8)	48.5(-0.8)	40.0(-3.5)	51.6(-1.4)	35.8(0.4)	54.7(-5.1)
Δ_{DCORAL}	↑5.8	↑2.5	↑5.2	↑5.8	↑4.5	↑7.6	↑3.6	↑5.3	↑5.1
DANN[8]	68.3(7.7)	75.0(-19.2)	51.0(2.3)	38.2(5.6)	42.8(4.2)	28.5(7.7)	42.0(10.0)	29.9(6.0)	47.0(3.0)
$DANN_{gate}$	78.1(-2.1)	80.2(-24.4)	61.8(-8.5)	48.3(-4.5)	51.2(-4.2)	43.8(-7.6)	55.2(-3.2)	40.5(-4.6)	57.4(-7.4)
Δ_{DANN}	↑9.8	↑5.2	↑10.8	↑10.1	↑9.4	↑14.7	↑13.2	↑10.6	↑10.4
ADDA[30]	63.2(12.2)	74.5(-18.1)	49.9(2.2)	38.3(5.1)	41.4(6.0)	25.2(13.5)	43.2(7.2)	28.0(7.3)	45.5(4.4)
$ADDA_{gate}$	79.4(-4.0)	82.9(-26.5)	64.2(-12.1)	47.7(-4.3)	52.2(-4.8)	48.0(-9.3)	58.2(-7.8)	43.0(-7.7)	59.5(-9.6)
Δ_{ADDA}	↑16.2	↑8.4	↑14.3	↑9.4	↑10.8	↑22.8	↑15.0	↑15.0	↑14.0
PADA[4]	69.7(6.5)	75.5(-19.0)	50.2(1.9)	38.7(5.1)	43.2(3.8)	30.1(5.5)	43.4(6.6)	32.2(5.5)	47.9(2.0)
$PADA_{gate}$	81.8(-5.6)	81.6(-25.1)	62.1(-10.0)	44.8(-1.0)	52.8(-5.8)	45.2(-9.6)	54.5(-4.5)	41.4(-5.7)	58.0(-8.1)
Δ_{PADA}	↑12.1	↑5.9	↑11.9	↑6.1	↑9.6	↑15.1	↑11.1	↑11.2	↑10.1
GTA[26]	81.2(-6.8)	78.9(-20.5)	58.4(-7.2)	42.2(2.8)	48.2(1.0)	33.1(5.1)	50.2(-0.1)	31.2(4.2)	52.9(-2.7)
GTA_{gate}	83.3(-8.9)	85.8(-27.4)	66.7(-15.5)	48.5(-3.5)	55.0(-5.8)	44.9(-6.7)	58.0(-7.7)	43.8(-8.4)	60.8(-10.6)
Δ_{GTA}	↑2.1	↑6.9	↑8.3	↑6.3	↑6.8	↑11.8	↑7.8	↑12.6	↑7.9
ΔAvg	↑8.3	↑5.2	↑8.1	↑7.1	↑7.5	↑13.3	↑8.9	↑10.4	

Conclusion

- Divergence between the joint distributions is the root to negative transfer.
- Negative transfer largely depends on the size of the labeled target data.

My work


Source Domain 1
Source Domain 2
Knowledge
Target Domain
Unlabeled Data
Labeled Data
Q
Oracle
Target Model

Linear Model Parameter Transfer

Assume that same kind of linear models learned from related tasks have a set of base models which can be used to represent these models.

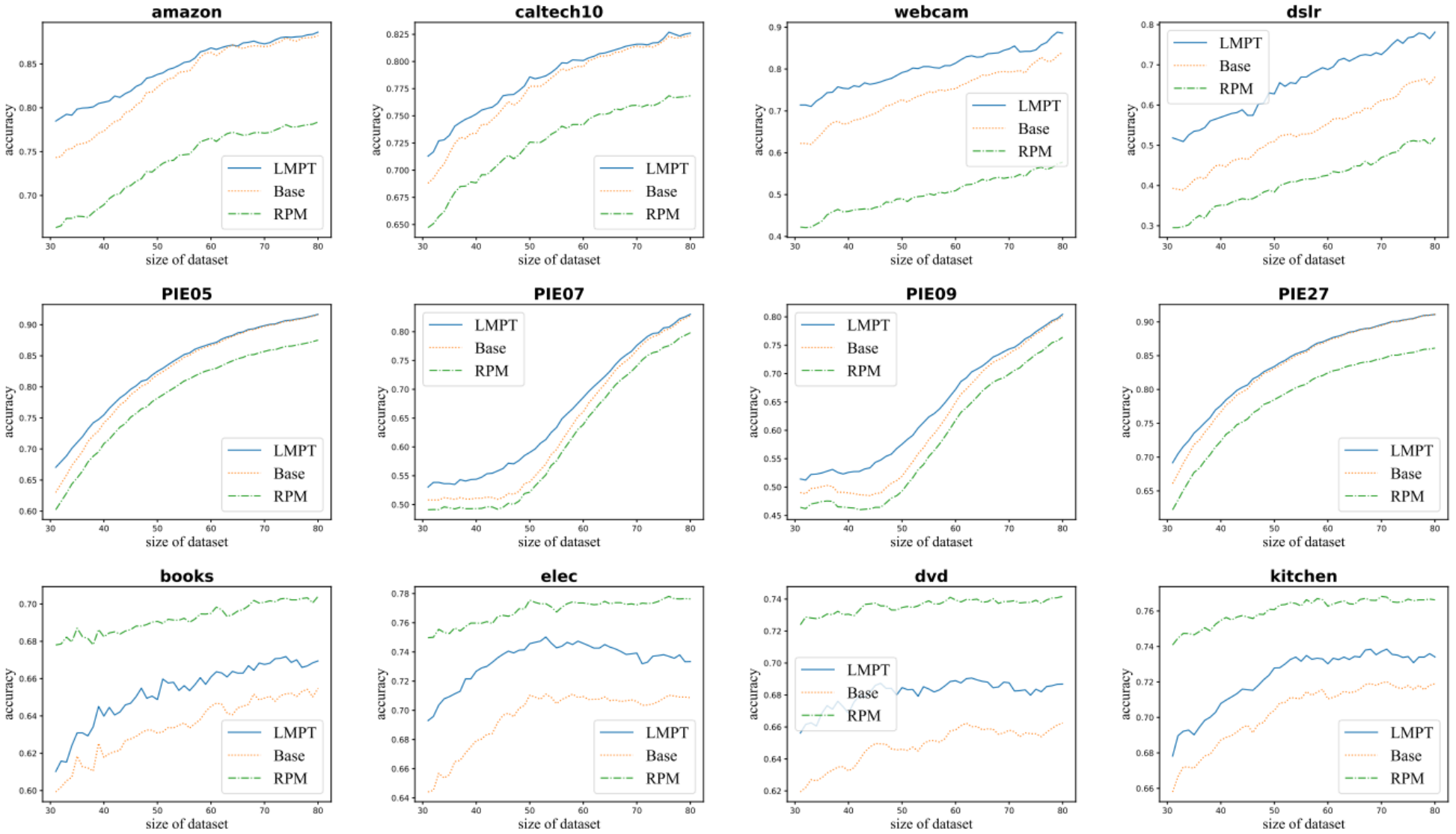
$$\min_{\mathbf{D},\mathbf{V}_S^i} \sum_{i=1}^{m} (\|\mathbf{W}_S^i - \mathbf{D}\mathbf{V}_S^i\|_F^2 + \eta\|\mathbf{V}_S^i\|_{2,1}) \qquad \mathbf{W}_S^i \in \mathbb{R}^{d\times c}$$

$$\min_{\mathbf{D},\mathbf{V}_S^i,\mathbf{V}_T,\mathbf{W}_T} \sum_{i=1}^{m} \left(\frac{1}{2}\|\mathbf{W}_S^i - \mathbf{D}\mathbf{V}_S^i\|_F^2 + \eta\|\mathbf{V}_S^i\|_{2,1} \right)$$

$$+\frac{\alpha}{2}\|\mathbf{W}_T - \mathbf{D}\mathbf{V}_T\|_F^2 + \eta\|\mathbf{V}_T\|_{2,1}$$

$$+\beta(\frac{1}{2}\|\mathbf{W}_T^T\mathbf{X}_L - \mathbf{Y}_L^T\|_F^2 + \|\mathbf{W}_T\|_F^2)$$

Experiment


amazon
caltech10
webcam
dslr
PIE05
PIE07
PIE09
PIE27
books
elec
dvd
kitchen
LMPT
Base
RPM
accuracy
size of dataset

Active Learning For Transfer

Assume that we have already well learnt from source and target domain, then what we lack is something **unique** from the target domain.

First, we train a base model $\boldsymbol{W}$ on $\boldsymbol{T_L}$ without transfer, which may not generalize well because of insufficient labeled data.
Then we train a transfer model $\boldsymbol{W_T}$ by any transfer method which we would like to use.
We tend to select instances with big difference between the two predictions to be queried.

Active Learning For Transfer

Reasons

- If the transfer make a positive impact on the prediction, which means that this kind of instances are not well learnt from the labeled data, we should query this instance to improve the prediction.
- If the transfer make a negative impact on the prediction, which means that the performance of prediction on this kind of instances is hindered by transfer, we should query this instance to add more supervised information of this kind to fix error.

If we directly choose the instance with biggest prediction difference, the noise or outlier is very possible to be queried.

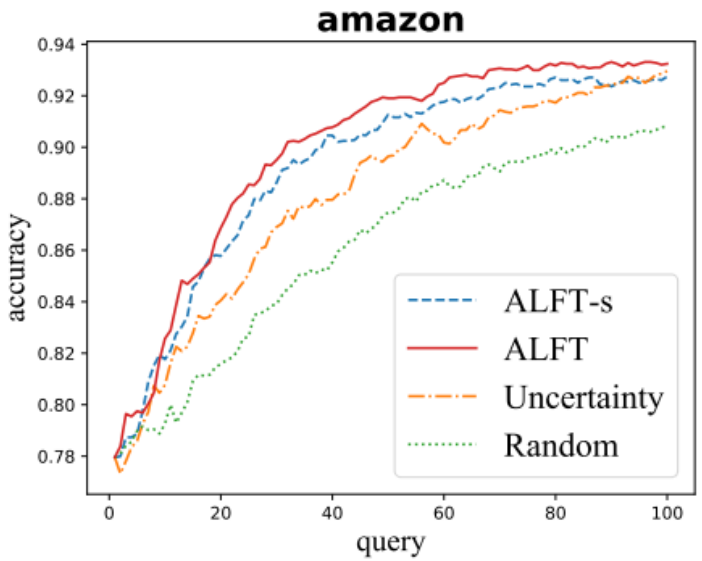
Active Learning For Transfer

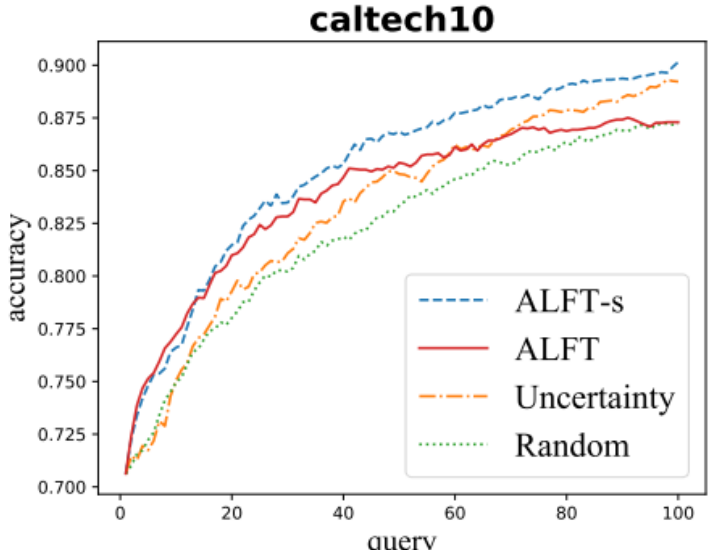
$$\min_{\boldsymbol{\alpha}} -\lambda \mathbf{d}^T \boldsymbol{\alpha} + MMD(\mathbf{P}_S, \mathbf{P}_L + \mathbf{P}_U \boldsymbol{\alpha})$$

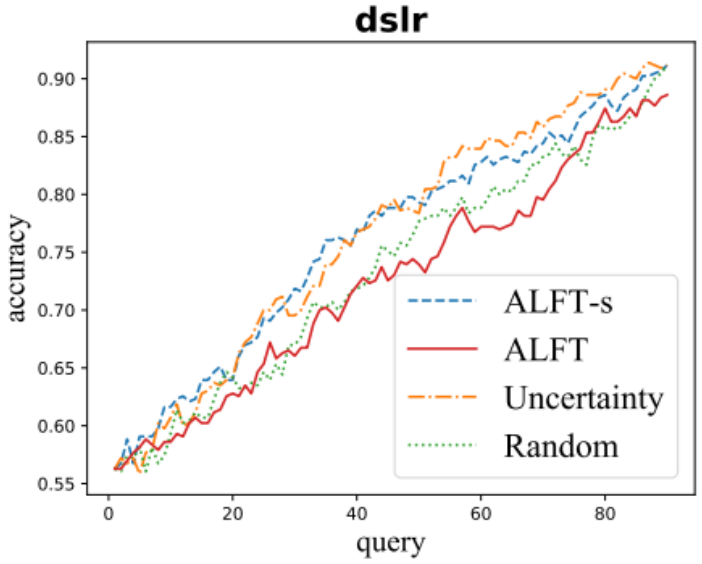
$$s.t. \qquad \alpha_i \in \{0,1\}, \sum_{i=1}^{n_u} \alpha_i = b$$

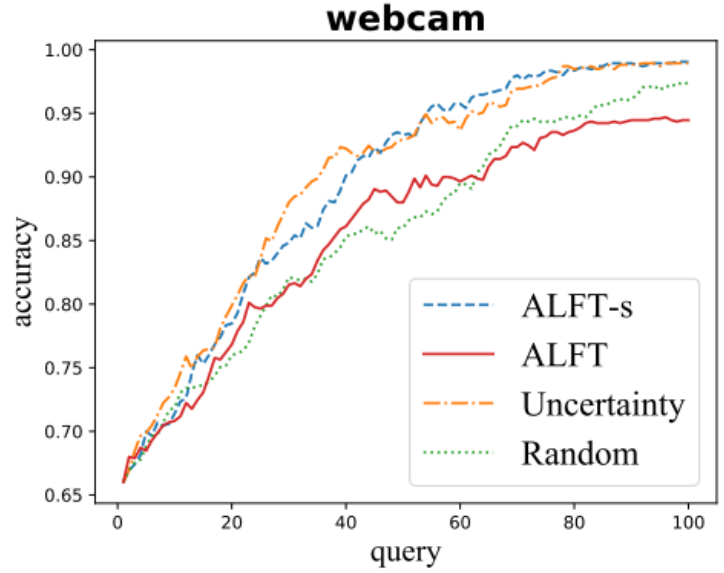
where $d^i = p_u^i - p_u^{i*}$, α_i means if or not query instance i and b is the batch-size of active learning.

Experiment

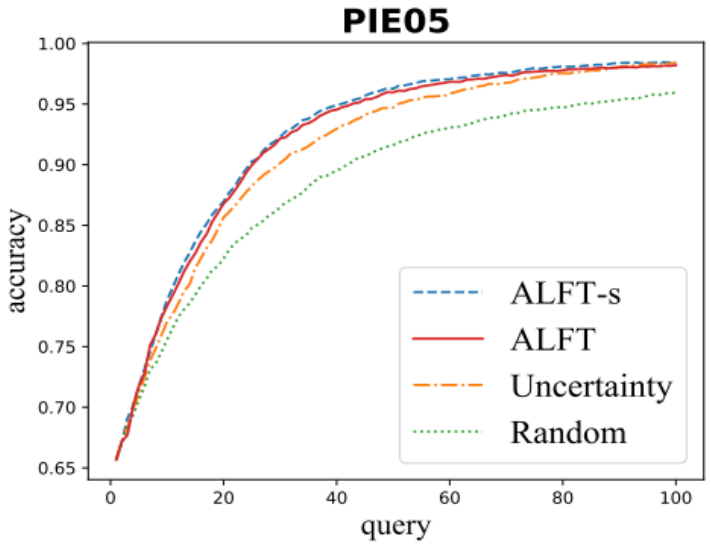

amazon
accuracy
query
ALFT-s
ALFT
Uncertainty
Random

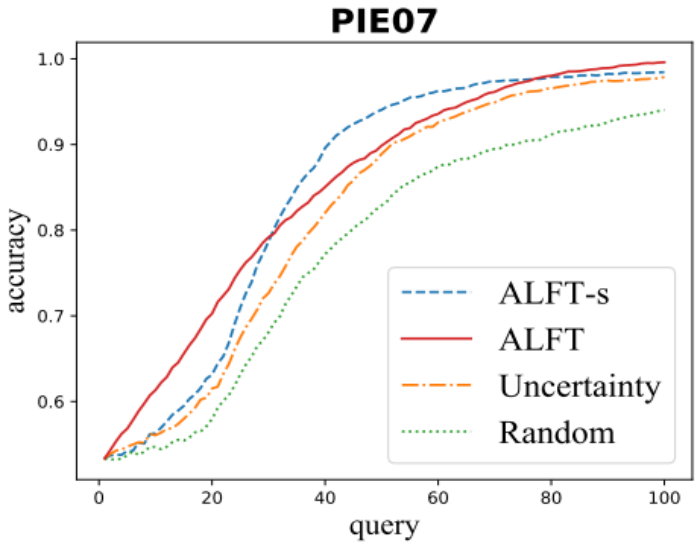

caltech10
accuracy
query
ALFT-s
ALFT
Uncertainty
Random

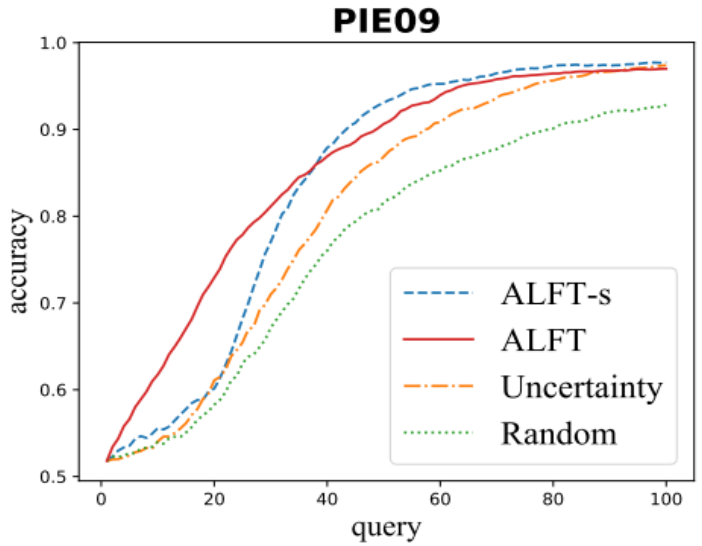

dslr
accuracy
query
ALFT-s
ALFT
Uncertainty
Random

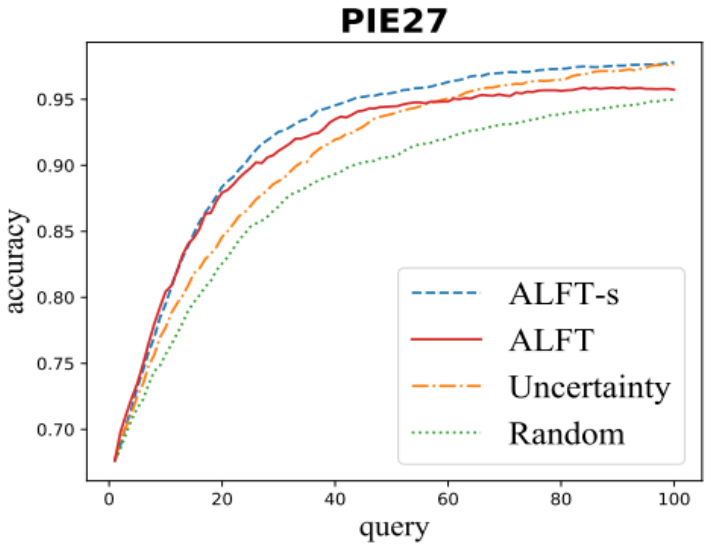

webcam
accuracy
query
ALFT-s
ALFT
Uncertainty
Random

Experiment


PIE05
accuracy
query
ALFT-s
ALFT
Uncertainty
Random


PIE07
accuracy
query
ALFT-s
ALFT
Uncertainty
Random


PIE09
accuracy
query
ALFT-s
ALFT
Uncertainty
Random


PIE27
accuracy
query
ALFT-s
ALFT
Uncertainty
Random