Few-Shot Adaptation of Pre-Trained Networks for Domain Shift

Paper ID: 2690

Wenyu Zhang[1], Li Shen[1], Wanyue Zhang[2], Chuan-Sheng Foo[1,3]

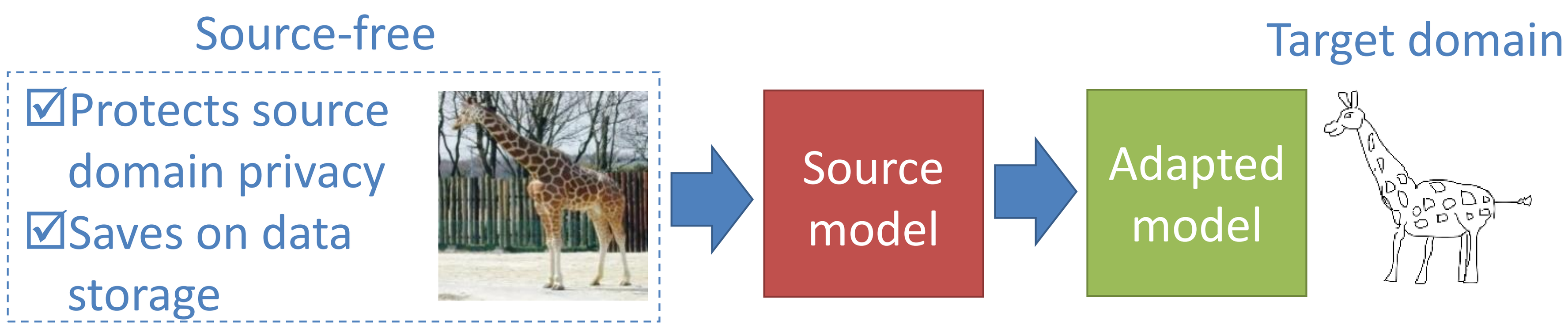
[1] Institute for Infocomm Research, [2] Max Planck Institute for Informatics, [3] Centre for Frontier AI Research

Background

Problem of Deep Learning

The success of deep learning in various applications is typically reliant on the assumption that train (source domain) and test (target domain) data distributions are the same. Many real-world environments are highly dynamic, hence we need timely adaptation to the target domain.

Source-Free Adaptation



Proposed Setup

Pre-Trained Model: $h_s: X_s \to Y_s$ trained on source domain s

Target Domain t: k–shot support set $L^{spt} \subset L^t$

- **Stronger constraints**: Labeled support set provides supervision
- **Practical**: Low requirement on target availability; No restriction on size and class composition of mini-batches during inference

Setup	Source data	Target data
Domain adaptation	L^s	U^t
k-shot domain adaptation	L^s	$L^{spt} \subset L^t$
Source-free domain adaptation	-	U^t
Test-time adaptation	-	Mini-batch U^t
Source-free k-shot adaptation	-	$L^{spt} \subset L^t$

(L^d and U^d denote labeled and unlabeled dataset for domain d.)

Proposed Method

Re-parameterizing BN operation

$$f(Z;\mu_t,\sigma_t,\gamma_t,\beta_t) = \left(\frac{Z-\mu_t}{\sigma_t}\right)\gamma_t + \beta_t = \left(\frac{Z-\mu_t}{\sigma_t}\right)\gamma_t\frac{\gamma_s}{\gamma_s} + \beta_t + \beta_s - \beta_s$$
$$= \left(\frac{Z-\tilde{\mu}_{opt}}{\tilde{\sigma}_{opt}}\right)\gamma_s + \beta_s = f(Z;\tilde{\mu}_{opt},\tilde{\sigma}_{opt},\gamma_s,\beta_s)$$

➔Optimal target domain BN layer can be obtained by optimizing only adaptation statistics $\tilde{\mu}_{opt}, \tilde{\sigma}_{opt}$ in pre-trained source model

➔We propose a low-dimensional approximation of $\tilde{\mu}_{opt}, \tilde{\sigma}_{opt}$

Linear Combination Coefficients for BN Statistics (LCCS)

☑Combination of BN statistics can be interpreted as *style mixing*

☑Models are adapted with *significantly fewer learnable parameters*

$$\tilde{\mu}_{opt} \approx \mu_{LCCS} = M\eta = \mu_s\eta_s + M_{spt}\eta_{spt}$$
$$\tilde{\sigma}_{opt} \approx \sigma_{LCCS} = \Sigma\rho = \sigma_s\rho_s + \Sigma_{spt}\rho_{spt}$$

$$f(Z;\mu_{LCCS},\sigma_{LCCS},\gamma_s,\beta_s) = \left(\frac{Z-\mu_{LCCS}}{\sigma_{LCCS}}\right)\gamma_s + \beta_s$$

- $\gamma_s, \beta_s \in \mathbb{R}^{C\times 1}$: Source domain BN parameters in pre-trained model
- $\mu_s, \sigma_s \in \mathbb{R}^{C\times 1}$: Source domain BN statistics in pre-trained model
- $M_{spt}, \Sigma_{spt} \in \mathbb{R}^{C\times n}$: n basis vectors of support set feature statistics
- $\eta_s, \rho_s \in \mathbb{R}$, $\eta_{spt}, \rho_{spt} \in \mathbb{R}^n$: Learnable LCCS parameters

Objective: minimize $L(\eta,\rho) = -\sum_{(x,y)\in L^{spt}} y \log h(x;\eta,\rho)$

Network	# params	# BN params	# LCCS params ($n=1$)
ResNet-18	12 million	9,600	80
ResNet-50	26 million	53,120	212
ResNet-101	45 million	105,344	416
DenseNet-121	29 million	83,648	484

Experiments: Image Classification

Evaluation is conducted on datasets covering style shift, synthetic-to-real shift and change in data collection environments.

Comparison with Test-Time Adaptation

Method	Test batch	PACS Bal.	PACS $\alpha=100$	VisDA Bal.	VisDA $\alpha=100$
Source model	any	83.1	76.9	64.7	54.7
+ Test-time BN	8	78.6	63.9	55.7	51.5
	32	83.3	66.8	60.7	55.2
	128	84.3	66.6	61.9	55.3
+ Tent	8	81.1	67.8	22.0	46.7
	32	86.1	69.4	59.0	57.5
	128	86.4	67.7	65.7	56.2
+ LCCS ($k=1$)	any	84.4	78.4	67.8	68.0
+ LCCS ($k=5$)	any	**87.1**	**81.9**	**76.0**	**77.8**

(α is the ratio of samples in the largest to smallest class.)

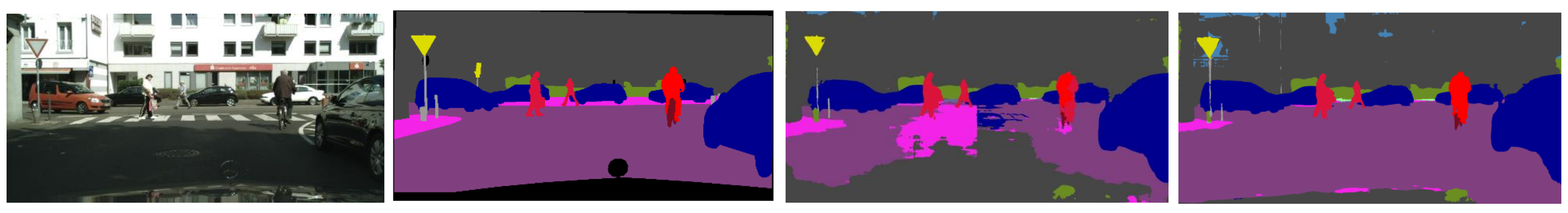
Comparison with Few-Shot Transfer Learning

Method	PACS $k=1$	PACS $k=5$	Camelyon17 $k=1$	Camelyon17 $k=5$	VisDA $k=1$	VisDA $k=5$
AdaBN	82.9	85.5	72.9	87.8	56.5	60.9
finetune BN	79.0	84.3	72.6	87.7	59.1	70.9
finetune classifier	82.5	83.7	70.5	70.4	67.6	69.7
finetune feat. extractor	83.6	86.0	79.3	86.5	67.3	68.4
L^2	**84.4**	85.8	**79.6**	88.2	66.0	66.4
L^2-SP	**84.4**	85.8	**79.6**	88.2	66.0	66.4
DELTA	**84.4**	85.8	**79.6**	88.2	65.9	66.5
Late Fusion	83.2	83.6	70.4	70.4	67.2	69.8
FLUTE	73.4	85.8	73.1	86.5	48.3	67.1
LCCS	**84.4**	**87.1**	76.6	**88.3**	**67.8**	**76.0**

Experiments: Semantic Segmentation

Evaluation by transferring from GTAV synthetic source domain to real or photo-realistic target domain with total 5 support samples.

Method	Cityscapes	BDD-100K	Mapillary	SYNTHIA
ERM	29.0	25.1	28.2	26.2
SW	29.9	27.5	29.7	27.6
IBN-Net	33.9	32.3	37.8	27.9
IterNorm	31.8	32.7	33.9	27.1
ISW	36.6	35.2	40.3	28.3
ISW + L^2	39.5	35.1	40.9	28.1
ISW + LCCS	**43.6**	**37.4**	**42.7**	**29.1**



Target image | Ground truth | ISW (SOTA) | ISW + LCCS

References

Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek. "Evaluating prediction-time batch normalization for robustness under covariate shift." *ArXiv* (2020).

S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge. "Improving robustness against common corruptions by covariate shift adaptation." *NeurIPS* (2020).

D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. "Tent: Fully test-time adaptation by entropy minimization." *ICLR* (2021).

S. Choi, S. Jung, H. Yun, J. Kim, S. Kim, and J. Choo. "RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening." *CVPR* (2021).

Contact: zhang_wenyu@i2r.a-star.edu.sg