Supplementary Material for "Deep Cocktail Network: Multi-source Unsupervised Domain Adaptation with Category Shift"

Ruijia Xu[1,†], Ziliang Chen[1,†], Wangmeng Zuo[2], Junjie Yan[3], Liang Lin[1,3*]
[1]Sun Yat-sen University [2]Harbin Institute of Technology [3]SenseTime Research
xurj3@mail2.sysu.edu.cn, c.ziliang@yahoo.com, wmzuo@hit.edu.cn,
yanjunjie@sensetime.com, linliang@ieee.org

Abstract

*This is the supplementary material for the paper entitled "**Deep Cocktail Network: Multi-source Unsupervised Domain Adaptation with Category Shift**". In the appendix.A, we introduce the architectures of deep cocktail network (DCTN) and the corresponding implementation information in the experiments. In the appendix.B, we present more detailed results about object and digit recognitions in the vanilla and category shift settings.*

1. Appendix.A

1.1. Architectures

In object recognition, deep domain adatation models [4] [5] mostly apply Alexnet [3] as their backbones. To achieve a fair comparison, we choose a DCTN architecture deriving from the Alexnet pipeline. As the Fig.1 illustrated, the representation module F is designed as a five-layer fully-convolutional network with three max-pooling operators, and the (multi-source) category classifier C is a three-layer fully-connected multi-task network. They are stacked into an exectly Alexnet-like pipeline to categorize images. For our multi-way adversarial adaptation, we adopt a CNN-based two-output classifier as two-source domain discriminator D.

Compared with natural images, digits contain less visual information. Hence we choose lighter structures as our representation module F, category classifier C and domain discriminator D for digit recognition, specifically as the Fig.2 reveals. We also combine F and C as the backbone of Source only and DAN to fairly compare our DCTN.

*Corresponding author: Liang Lin. † indicates equal contribution. This work was supported by National Science Foundation of China under Grant U1611461 and Special Funds for Guangdong College Students' Science and Technology Innovation (under Grant pdjhb0009).

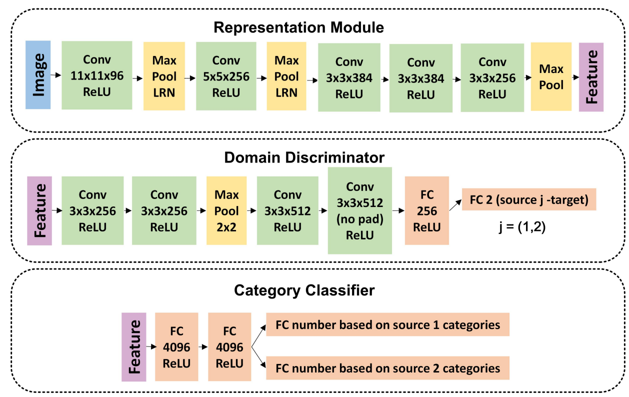

Figure 1. The representation module, domain discriminator and category classifier we used in the experiments about object recognition. (Best viewed in color)

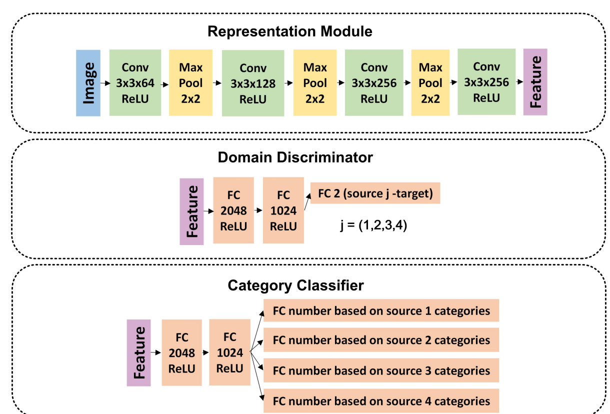

Figure 2. The representation module, domain discriminator and category classifier we used in the experiments about digit recognition. (Best viewed in color)

1.2. Stablizers

For the sake of legibility, we apply the sigmoid cross entropy loss to denote the multi-way adversary further inducing the perplexity score formula in our paper. In spite of a classical expression in many adversarial adaptations [1] [7], this loss function easily causes the gradient vanishing problem during the multi-way adversarial learning process. To overcome this, we utilize the least square measuring function [6] to execute the adversarial learning.

$$\mathcal{L}_{adv}^{(ls)}(F, D) = \frac{1}{N}\sum_{j}^{N}\mathbb{E}_{x\sim X_{s_j}}[(D_{s_j}(F(x)))^2] + \mathbb{E}_{x^t\sim X_t}[(1 - D_{s_j}(F(x^t)))^2] \tag{1}$$

Accordingly, the confusion loss has been revised as

$$\mathcal{L}_{cf}^{(ls)}(x; F, D_{s_j}) = (D_{s_j}(F(x)) - \frac{1}{2})^2 \tag{2}$$

then given a target instance x^t, we have a least square perplexity score

$$\mathcal{S}_{cf}^{(ls)}(x^t; F, D_{s_j}) = (D_{s_j}(F(x^t)))^2 + \alpha_{s_j} \tag{3}$$

. These revisions keep consistent in all our analysis in the paper, meanwhile stablize the multi-way adversarial learning in the DCTN.

No matter in training or testing, we require perplexity score weighting to predict the class of target instance. While in the adversarial learning process, the domain discriminator D must be gradually trained to accommodate the learning of representation module F. It means in the previous epoches, the perplexity score is not able to provide reliable probablistic relations between target and each source. This harms the pseudo-labeling scheme and further spoils the adversary at the next alternative step. Empirically we find that, such negative effect mostly comes from the unstable predictions toward single target instances. Hence we utilize the moving average to calculate the perpelxity score for each target instance.

$$\mathcal{S}_{cf}^{(ls)}(x_{N_T}^t; F, D_{s_j}) = \frac{1}{N_T}\sum_{i}^{N_T}(D_{s_j}(F(x_i^t)))^2 + \alpha_{s_j} \tag{4}$$

where N_T denotes how many times the target samples have been visited to train our model, and $x_{N_T}^t$ denotes the current target instance we are considering. We also apply the same strategy to obtain a stable consentration constant α_{s_j}.

$$\alpha_{s_j} = \frac{1}{N_T}\sum_{i}^{N_T}(D_{s_j}(1 - F(x_i^{s_j})))^2 \tag{5}$$

where $x_i^{s_j}$ denotes the source j instance come along with the coupled target instances in the adversarial learning.

1.3. Training

In object recognition, we initiate our DCTN by following the same way of DAN [4]. In terms of digit recognition, we perform DCTN learning from scratch. In order to execute online hard domain sample mining, we construct our mini-batch by sampling a equal number of images in each domain. For instance, let's consider a two-source domain adaptation with domain batch size 32. Then we have mini-batches with sizes as $96 = 32 \times (2+1)$ (2 and 1 denote two source domains and one target domain). In this situation, the length of epoch is decided by the size of the domain containing most instances. Finally, we adopt Adam [2] solver with $momentum = (0.9, 0.99)$ in all experiments to update our networks.

More hyper-parameter details have been further demonstrated in Table.1.

Table 1. The hyper-parameters setting in our experiment.

	Office-31	ImageCLEF-DA	Digit-five
domain batch size	32	32	128
pseudo threshold γ	0.9	0.98	0.9
learning rate	0.00001	0.000002	0.00001
image size	227×227	227×227	32×32

2. Appendix.B

In this section, we provide the extensive evaluation based on the main results in our paper. In the vanilla setting (Table 2-4), we append the single source transfer results and the average performances in all methods. In the category shift setting (Table 5-6), we offer the upperbound results (the corresponding accuracies in vanilla setting), then illustrate how to obtain the accuracy degradation and transfer gain in each result.

References

[1] J. Hoffman, E. Tzeng, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Domain Adaptation in Computer Vision Applications*, pages 173–187. Springer, 2017. 1

[2] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[4] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 1, 2

[5] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 1

[6] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. *arXiv preprint ArXiv:1611.04076*, 2016. 2

Table 2. Classification accuracy (%) on Office-31 benchmark for MDA in the vanilla setting.

Standards	Models	W→D	A→D	A→W	D→W	D→A	W→A	Avg
Single source	TCA	95.2	60.8	61.0	93.2	51.6	50.9	68.8
	GFK	95.0	60.6	60.4	95.6	52.4	48.1	68.7
	DDC	98.5	64.4	61.8	95.0	52.1	52.2	70.7
	DRCN	99.0	66.8	68.7	96.4	**56.0**	54.9	73.6
	WMMD	98.7	64.5	66.8	95.9	53.8	52.7	72.1
	RevGrad	99.2	72.3	73.0	96.4	53.4	51.2	74.3
	DAN	99.0	67.0	68.5	96.0	54.0	53.1	72.9
	RTN	**99.6**	71.0	73.3	96.8	50.5	51.0	73.7
		A,W→D		A,D→W		D,W→A		
Source combine	Source only	98.1		93.2		50.2		80.5
	DAN	98.8		95.2		53.4		82.5
Multi-source	Source only	98.2		92.7		51.6		80.8
	RDALR	31.2		36.9		20.9		29.7
	sFRAME	54.5		52.2		32.1		46.3
	SGF	39.0		52.0		28.0		39.7
	DCTN (ours)	**99.6**		**96.9**		54.9		**83.8**

Table 3. Classification accuracy (%) on ImageCLEF-DA benchmark for MDA in the vanilla setting.

Standards	Models	I→P	C→P	I→C	P→C	P→I	C→I	Avg
Single source	RevGrad	66.5	63.5	89.0	88.7	81.8	79.8	78.2
	DAN	67.3	61.6	87.7	88.4	80.5	76.0	76.9
	RTN	67.4	63.0	89.5	90.1	82.3	78.0	78.4
		I,C→P		I,P→C		P,C→I		
Source combine	Source only	68.3		88.0		81.2		79.2
	DAN	**68.8**		88.8		81.3		79.6
Multi-source	Source only	68.5		89.3		81.3		79.7
	DCTN (ours)	**68.8**		**90.0**		**83.5**		**80.8**

[7] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017. 1

Table 4. Classification accuracy (%) on Digits-five benchmark for MDA in the vanilla setting. Source only (ensemble) and DAN (ensemble) denote the target classifiers ensembled by their corresponding single source classifiers.

Standards	Models	mm → sv	mt → sv	sy→ sv	up → sv	mt → mm	sv → mm	sy → mm	up → mm	Avg
Single source	Source only	45.3	46.4	67.4	29.7	58.0	49.6	54.8	43.7	49.4
	DAN	43.2	42.2	67.1	38.5	53.5	51.8	58.8	40.5	49.5
		mm, mt, sy, up → sv				mt, sv, sy, up → mm				
Source combine	Source only	72.2				64.1				68.2
	DAN	71.0				66.6				68.8
Multi-source	Source only (ensemble)	64.6				60.7				62.7
	DAN (ensemble)	62.6				62.9				62.8
	DCTN (ours)	**77.5**				**70.9**				**74.2**

Table 5. Evaluations on Office-31 (A,D→ W) for MDA in the category shift setting.

Category Shift	Models	Accuracy	Upperbound	Accuracy Degrade	Transfer Gain
Overlap	Source only	84.4	92.7	84.4-92.7=-8.3	84.4-84.4=0
	DAN	87.8	94.2	87.8-94.2=-**6.4**	87.8-84.4=3.4
	DCTN(ours)	90.2	96.9	90.2-96.9=-6.7	90.2-84.4=**5.8**
Disjoint	Source only	78.1	92.7	78.1-92.7=-14.6	78.1-78.1=0
	DAN	75.5	94.2	75.5-94.2=-18.7	75.5-78.1=-2.6
	DCTN(ours)	82.9	96.9	82.9-96.9=**-14.0**	82.9-78.1=**4.8**

Table 6. Evaluations on ImageCLEF-DA (I,P→ C) for MDA in the category shift settings.

Category Shift	Models	Accuracy	Upperbound	Accuracy Degrade	Transfer Gain
Overlap	Source only	86.3	89.3	86.3-89.3=-3.0	86.3-86.3=0
	DAN	85.5	89.5	85.5-89.5=-4.0	85.5-86.3=-0.8
	DCTN(ours)	88.7	90.0	88.7-90.0=**-1.3**	88.7-86.3=**2.4**
Disjoint	Source only	81.5	89.3	81.5-89.3=**-7.8**	81.5-81.5=0
	DAN	71.0	89.5	71.0-89.5=-18.5	71.0-81.5=-10.5
	DCTN(ours)	82.0	90.0	82.0-90.0=-8.0	82.0-81.5=**0.5**