Clustering on Domain Adaptation

Pengcheng Xu

2020, July, 24

Paper list

- End-to-End Adversarial Attention Network for Multimodal Clustering
- Exploring Category-Agnostic Clusters for Open-Set Domain Adaptation
- Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering

End-to-End Adversarial-Attention Network for Multi-Modal Clustering

Runwu Zhou[1,2] Yi-Dong Shen [1*]
[1] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{*zhourw*, *ydshen*}@ios.ac.cn

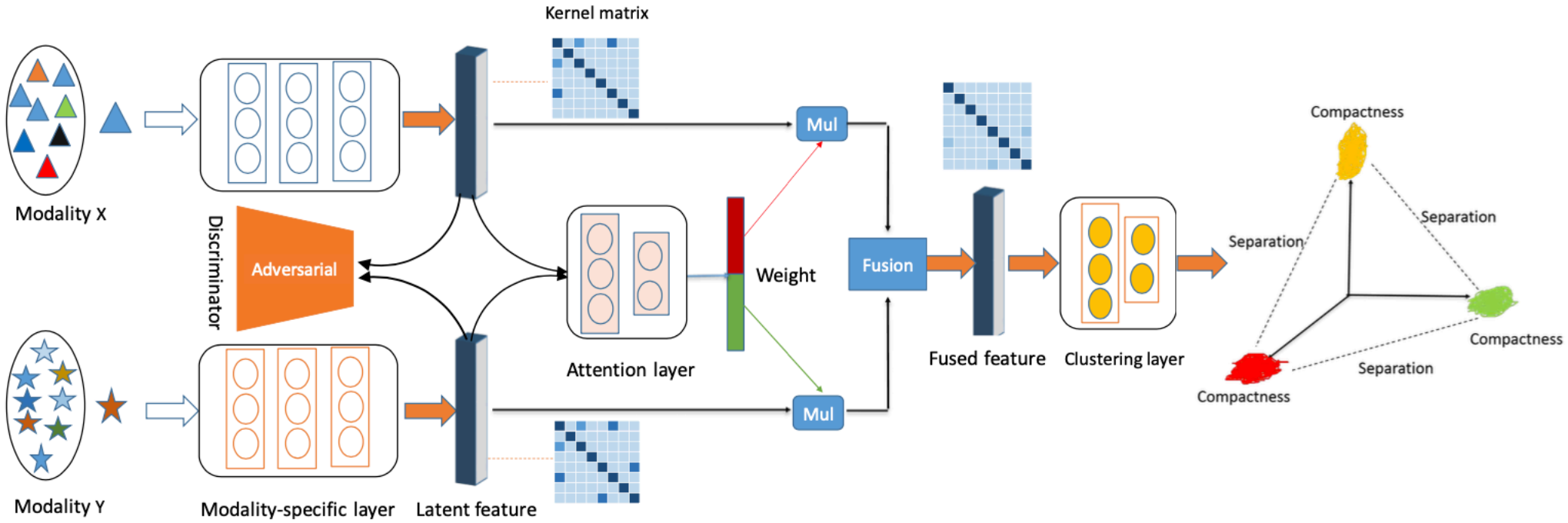
Motivation:

- Task: cluster a set of **n** data points consisting **V** modalities into **c** clusters. E.g. fuse the image info with its caption info.
- **Simultaneously** fuse and cluster data with multiple modalities
- End-to-End training from scratch with **batch-mode** and avoid auto encoder pre-training stage like the current methods.
- Explicitly encourage **separation** and **compactness** of clusters

Method

- Techniques: Adversarial Training, Attention and Divergence measures
- 1. Adversarial training to align the latent distribution of each modality
- 2. Attention to adaptively assign weight to each modality
- 3. Divergence loss to encourage separation and compactness.

Framework



Three parts: modality feature learning, modality fusion and cluster assignment

Modality Feature Learning

- Task: cluster a set of n data points consisting V modalities into c clusters.
- For **each modality**, extract features and measure the distance of features based on Gaussian metric.

$$\mathcal{D} = \{\mathbf{X}^1, ...\mathbf{X}^v, ..., \mathbf{X}^V\}$$

$$\mathbf{X}^v \in \mathcal{R}^{d_v \times n}$$

$$\mathbf{H}^v = E_v(\mathbf{X}^v; \theta_e^v).$$

$$\mathbf{K}_{ij}^v = exp(-||\mathbf{h}_i^v - \mathbf{h}_j^v||_2/2\sigma^2).$$

Modality Fusion

- Fuse diverse info of different modalities for comprehensive estimation
- Select a modality as **an anchor** and align distributions of the anchor between the rest modalities.
- V modalities, one anchor and V-1 discriminator between anchor and the rest.
- Input concatenated features h; output V dimension attention

$$\mathbf{h} = [\mathbf{h}^1, \mathbf{h}^2, ..., \mathbf{h}^V],$$
$$\mathbf{act} = \mathrm{FCs}(\mathbf{h}),$$
$$\mathbf{e} = \mathrm{Softmax}(\mathrm{sigmoid}(\mathbf{act})/\tau),$$
$$\mathbf{w} = \mathrm{Mean}(\mathbf{e}, \mathrm{dim}{=}0)$$
$$\mathbf{h}_f = \textstyle\sum_v \mathbf{w}_v \mathbf{h}^v.$$

$$\mathcal{L}_{adv} = \min_{\theta_e^v} \max_{\theta_d^v} \sum_{v=2}^{V} \mathbb{E}_{h^1 \sim p_1}[logD_v(h^1)] + \mathbb{E}_{h^v \sim p_v}[log(1 - D_v(h^v))]$$

$$\mathcal{L}_{att} = ||\mathbf{K}^f - \mathbf{K}^c||_F^2 \qquad \mathbf{K}^c = \textstyle\sum_v \mathbf{w}_v \mathbf{K}^v.$$

K^f is computed by the fused features

Clustering

- Fused features are used for clustering layers.
- Softmax outputs a soft cluster matrix represent assignment of data point **q** to cluster **l**

$$\mathbf{A} = [\alpha_{qi}], \text{ with elements } \alpha_{qi} \in (0, 1)$$

$$\mathcal{D}_{sc} = \frac{1}{k} \sum_{i=1}^{k-1} \sum_{j>i} \frac{\boldsymbol{\alpha}_i^T \mathbf{K} \boldsymbol{\alpha}_j}{\sqrt{\boldsymbol{\alpha}_i^T \mathbf{K} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j^T \mathbf{K} \boldsymbol{\alpha}_j}}$$

$$\mathcal{D}_{sim} = \frac{1}{k} \sum_{i=1}^{k-1} \sum_{j>i} \frac{\boldsymbol{\beta}_i^T \mathbf{K} \boldsymbol{\beta}_j}{\sqrt{\boldsymbol{\beta}_i^T \mathbf{K} \boldsymbol{\beta}_i \boldsymbol{\beta}_j^T \mathbf{K} \boldsymbol{\beta}_j}}$$

$$\mathcal{D}_{reg} = triu(\mathbf{A}^T \mathbf{A})$$

$$\mathcal{L}_c = \mathcal{D}_{sc} + \mathcal{D}_{sim} + \mathcal{D}_{reg}$$

The vectors $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, ..., \boldsymbol{\alpha_k}$ denote the columns of the hard cluster assignment matrix $\mathbf{A} \in \mathbb{R}^{n \times k}$. In our architecture, we relax the hard membership to soft one in order to preserve differentiability of the loss.

Experiment results

Dataset	NWC			RGB-D			VOC			CCV		
Metric	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
SC(1)	0.712	0.768	0.747	0.334	0.297	0.347	0.384	0.392	0.379	0.102	0.005	0.104
SC(2)	0.647	0.689	0.699	0.297	0.305	0.326	0.402	0.411	0.395	0.188	0.173	0.213
SC(3)	-	-	-	-	-	-	-	-	-	0.113	0.008	0.109
SC(con)	0.652	0.673	0.686	0.312	0.286	0.320	0.372	0.387	0.382	0.093	0.074	0.102
RMKMC	0.784	0.793	0.791	0.379	0.398	0.397	0.458	0.469	0.473	0.176	0.165	0.186
tRLMvc	0.873	0.849	0.869	0.445	0.439	0.460	0.534	0.547	0.556	0.212	0.226	0.231
CSMCS	0.824	0.813	0.829	0.392	0.414	0.426	0.488	0.496	0.517	0.194	0.186	0.198
WMSC	0.798	0.787	0.816	0.408	0.425	0.420	0.471	0.462	0.477	0.205	0.196	0.208
MCGC	0.853	0.862	0.876	0.438	0.447	0.453	0.527	0.546	0.539	0.224	0.216	0.240
DCCA	0.784	0.798	0.809	0.355	0.362	0.374	0.397	0.425	0.433	0.173	0.182	0.186
DMSC	0.877	0.864	0.876	0.419	0.426	0.433	0.541	0.538	0.566	0.183	0.194	0.196
DAMC	0.891	0.914	0.916	0.463	0.475	0.481	0.560	0.552	0.583	0.243	0.231	0.264
EAMC	**0.945**	**0.937**	**0.952**	**0.497**	**0.499**	**0.511**	**0.607**	**0.615**	**0.628**	**0.261**	**0.266**	**0.271**

Table 2. Clutering results on NWC, RGB-D, VOC and CCV datasets.

Exploring Category-Agnostic Clusters for Open-Set Domain Adaptation

Yingwei Pan[†], Ting Yao[†], Yehao Li[†], Chong-Wah Ngo[‡], and Tao Mei[†]
[†] JD AI Research, Beijing, China
[‡] City University of Hong Kong, Kowloon, Hong Kong
{panyw.ustc, tingyao.ustc, yehaoli.sysu}@gmail.com, cscwngo@cityu.edu.hk, tmei@jd.com

Motivation

- Open set domain adaptation
  - Source and Target do not have the same categories
  - Source: cat dog fish road laptop Target: cat road desk car sky

- Problem:
  - How to distinguish the unknown samples from known samples while classify known samples right?
  - How to do closed-set and open-set DA at the same time?

Motivation

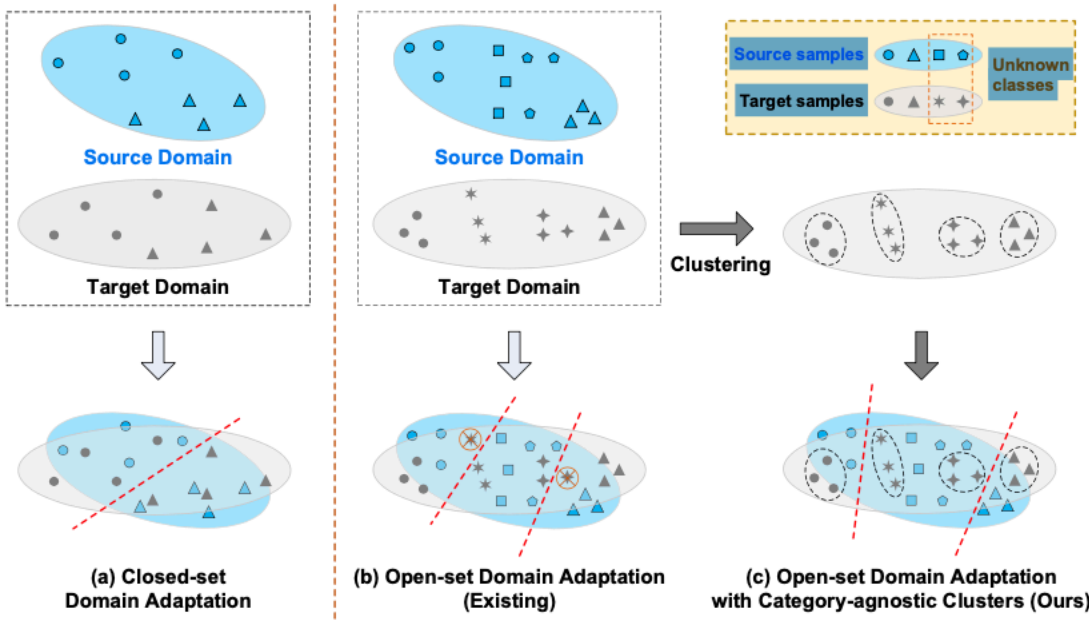
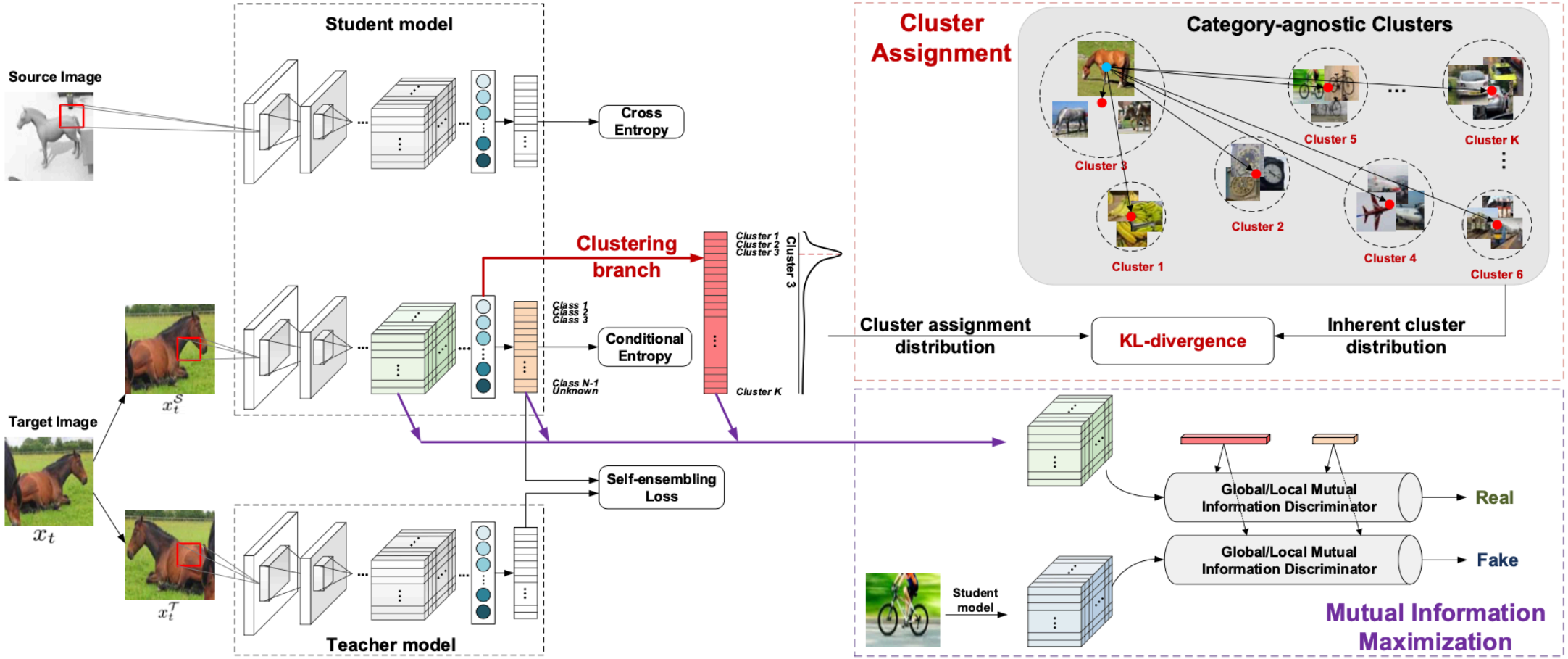


Figure 1. A comparison between (a) closed-set domain adaptation, (b) existing methods for open-set domain adaptation, and (c) our open-set domain adaptation with category-agnostic clusters.

Method

- 1. Mean-Teacher network for DA (self-ensembling loss)
- 2. Add a clustering branch to student branch
  - Enforce the target data maintain the inherent discriminative structure and easy to distinguish unknown samples (KL divergence)
- 3. Make feature suitable for downstream tasks
  - Mutual information max between input features, class probability and clustering probability.

Framework


Student model
Source Image
Cross Entropy
Clustering branch
Class 1
Class 2
Class 3
Conditional Entropy
Class N-1
Unknown
Cluster 1
Cluster 2
Cluster 3
Cluster K
Target Image
x_t
x_t^S
x_t^T
Self-ensembling Loss
Teacher model
Cluster Assignment
Category-agnostic Clusters
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster K
Cluster 6
Cluster assignment distribution
KL-divergence
Inherent cluster distribution
Global/Local Mutual Information Discriminator
Real
Global/Local Mutual Information Discriminator
Fake
Student model
Mutual Information Maximization

Experiment results

Table 1. Performance comparison with the state of arts on Office for open-set domain adaptation. ◇ indicates a different open-set setting without unknown source examples.

Method	A → D		A → W		D → A		D → W		W → A		W → D		Avg	
	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*
Source-only	67.1	67.0	64.6	63.8	61.9	60.7	90.6	92.3	60.2	59.7	96.7	98.7	73.5	73.7
RTN [17]	76.6	74.7	73.0	70.8	57.2	53.8	89.0	88.1	62.4	60.2	98.8	98.3	76.2	74.3
RevGrad [6]	78.3	77.3	75.9	73.8	57.6	54.1	89.8	88.9	64.0	61.8	98.7	98.0	77.4	75.7
AODA◇ [29]	76.6	76.4	74.9	74.3	62.5	62.3	94.4	94.6	81.4	81.2	96.8	96.9	81.1	80.9
ATI-λ [22]	79.8	79.2	77.6	76.5	71.3	70.0	93.5	93.2	76.7	76.5	98.3	99.2	82.9	82.4
FRODA [2]	**88.0**	-	78.7	-	76.5	-	**98.0**	-	73.7	-	94.6	-	84.9	-
SE-CC◇	80.6	84.0	82.4	84.2	83.2	**90.3**	92.9	96.6	82.7	85.9	96.8	99.1	86.4	90.0
SE-CC	85.3	**84.5**	**85.1**	**84.3**	**87.9**	89.5	97.7	**97.8**	**86.8**	**87.5**	**99.4**	**99.6**	**90.4**	**90.5**

Table 2. Performance comparison with the state of arts on VisDA for open-set adaptation (Known-to-Unknown Ratio = 1:10). ◇ indicates a different open-set setting without unknown source examples. † indicates the results are referred from the official leaderboard [1].

Method	aero	bike	bus	car	horse	knife	mbike	person	plant	skbrd	train	truck	unk	Knwn	Mean	Overall
Source-only	53.8	54.2	50.3	48.7	72.7	5.3	82.0	27.0	49.6	43.4	78.0	5.1	44.2	46.9	47.3	44.8
RevGrad [6]	33.0	57.3	44.1	33.9	72.1	46.9	82.2	26.8	36.8	50.4	89.4	9.8	47.8	48.6	48.5	47.8
RTN [17]	49.2	72.6	66.5	39.5	80.8	18.8	73.8	56.8	47.4	45.2	74.0	4.5	48.7	52.4	52.1	49.0
SE† [5]	**94.2**	74.1	**86.1**	68.1	**91.0**	26.1	**95.2**	46.0	**85.0**	40.4	79.2	11.0	51.0	66.4	65.2	52.7
AODA◇† [29]	80.2	63.1	59.1	63.1	83.2	12.1	89.1	5.0	61.0	14.0	79.2	0.0	69.0	50.8	52.2	67.6
ATI-λ [22]	85.7	74.9	60.3	49.9	80.0	19.3	88.8	40.8	54.0	**59.2**	66.4	**18.2**	59.5	58.1	58.2	59.3
SE-CC◇	82.1	80.7	59.7	50.0	80.6	36.7	83.1	56.2	56.6	21.9	57.7	4.0	70.6	55.8	56.9	69.2
SE-CC	**94.2**	**79.0**	83.4	**70.7**	**91.0**	**43.5**	89.3	**73.3**	69.4	58.8	**79.4**	12.8	**71.6**	**70.4**	**70.5**	**71.6**

Experiment results

Table 3. Performance comparison with the state of arts on VisDA dataset for closed-set domain adaptation.

Method	aero	bike	bus	car	horse	knife	mbike	person	plant	skbrd	train	truck	Mean
Source-only	67.1	51.4	50.8	64.5	83.4	13.0	89.9	34.4	78.8	47.0	88.1	2.0	55.9
RevGrad [6]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
RTN [17]	89.1	56.4	72.4	69.7	77.9	49.5	87.7	13.0	88.1	77.4	86.7	7.2	64.6
MCD [28]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
SimNet [24]	94.3	82.3	73.5	47.2	87.9	49.2	75.1	79.7	85.3	68.5	81.1	50.3	72.9
TPN [21]	93.7	85.1	69.2	**81.6**	93.5	61.9	89.3	81.4	93.5	81.6	84.5	49.9	80.4
SE [5]	96.2	**87.8**	**84.4**	66.5	**96.1**	96.1	90.5	81.5	**95.3**	91.5	87.5	51.6	85.4
SE-CC	**96.3**	86.5	82.4	81.3	**96.1**	**97.2**	**91.2**	**84.7**	94.4	**94.1**	**88.3**	**53.4**	**87.2**

Table 4. Performance comparison with the state of arts on Office dataset for closed-set domain adaptation.

Method	A → D	A → W	D → A	D → W	W → A	W → D	Avg
RTN [17]	77.5	84.5	66.2	96.8	64.8	99.4	81.6
RevGrad [6]	79.7	82.0	68.2	96.9	67.4	99.1	82.2
JAN [16]	85.1	86.0	69.2	96.7	70.7	99.7	84.6
SimNet [24]	85.3	88.6	73.4	98.2	71.8	99.7	86.2
GTA [30]	87.7	89.5	72.8	97.9	71.4	99.8	86.5
iCAN [36]	90.1	**92.5**	72.1	98.8	69.9	**100**	87.2
SE-CC	**91.4**	90.7	**74.0**	**99.0**	**72.9**	**100**	**88.0**

Table 5. Performance contribution of each design (i.e., Conditional Entropy (**CE**), KL-divergence Loss (**KL**), and Mutual Information Maximization (**MIM**)) in SE-CC on VisDA for open-set transfer.

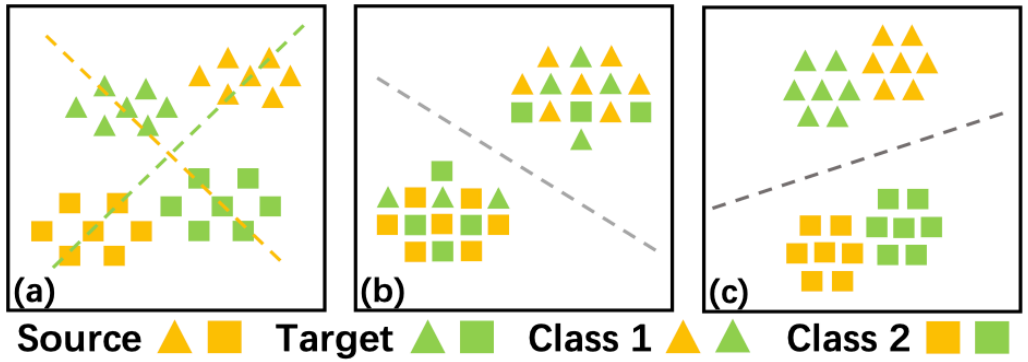
Method	CE	KL	MIM	Knwn	Mean	Overall
SE				66.4	65.2	52.7
+CE	✓			67.3	66.3	55.8
+KL	✓	✓		69.3	69.3	69.1
SE-CC	✓	✓	✓	**70.4**	**70.5**	**71.6**

Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering

Hui Tang, Ke Chen, and Kui Jia*
South China University of Technology
381 Wushan Road, Tianhe District, Guangzhou, Guangdong, China
eehuitang@mail.scut.edu.cn, {chenk,kuijia}@scut.edu.cn

Motivation

- 1. Explicit domain alignment damages the intrinsic target discrimination
  - 1. Source and Target data both have discriminative cluster structures.
  - 2. Clusters of the same class from different domains should be close.



- 2.Uncover the intrinsic discriminative structure of Target by clustering target being regularized by source

Method

- Deep discriminative clustering
  - Introduce an auxiliary counterpart Q as the label for classifier P
  - Minimize KL divergence of Q and P by alternating training Q and P
  - Do above for **feature space** and **label space**(after softmax)
  - Do above for **source** and **target**;
  - Target and source share network so source works as the regularization

Method

- Target label space
  - **P** is the probability from the classifier and **Q** is the auxiliary probability
  - KL divergence push P towards clustering assignment Q; Entropy balances cluster size and avoid degeneration.

$$\min_{\boldsymbol{Q}^t,\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}} \mathcal{L}^t_{f\circ\varphi} = \mathrm{KL}(\boldsymbol{Q}^t||\boldsymbol{P}^t) + \sum_{k=1}^{K} \varrho^t_k \log \varrho^t_k, \qquad \varrho^t_k = \tfrac{1}{n_t}\textstyle\sum_{i=1}^{n_t} q^t_{i,k}$$

$$\mathrm{KL}(\boldsymbol{Q}^t||\boldsymbol{P}^t) = \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{k=1}^{K} q^t_{i,k} \log \frac{q^t_{i,k}}{p^t_{i,k}}.$$

Method

- Target feature space
  - u_k is the learnable cluster center of **both source and target**

$$\widetilde{p}_{i,k}^{t} = \frac{\exp((1+||\boldsymbol{z}_i^t - \boldsymbol{\mu}_k||^2)^{-1})}{\sum_{k'=1}^{K}\exp((1+||\boldsymbol{z}_i^t - \boldsymbol{\mu}_{k'}||^2)^{-1})}.$$

$$\min_{\widetilde{\boldsymbol{Q}}^t,\boldsymbol{\theta},\{\boldsymbol{\mu}_k^t\}_{k=1}^K} \mathcal{L}_{\varphi}^{t} = \mathrm{KL}(\widetilde{\boldsymbol{Q}}^t||\widetilde{\boldsymbol{P}}^t) + \sum_{k=1}^{K}\widetilde{\varrho}_k^t \log \widetilde{\varrho}_k^t, \qquad \widetilde{\varrho}_k^t = \frac{1}{n_t}\sum_{i=1}^{n_t}\widetilde{q}_{i,k}^t.$$

$$\min_{\boldsymbol{Q}^t,\widetilde{\boldsymbol{Q}}^t,\{\boldsymbol{\theta},\boldsymbol{\vartheta}\},\{\boldsymbol{\mu}_k\}_{k=1}^K} \mathcal{L}_{\mathrm{SRDC}}^{t} = \mathcal{L}_{f\circ\varphi}^{t} + \mathcal{L}_{\varphi}^{t}.$$

Method

- Source label space and feature space
  - Change the corresponding Q with ground truth. Then KL divergence is basically cross-entropy

$$\min_{\boldsymbol{\theta},\boldsymbol{\vartheta}} \mathcal{L}^s_{f\circ\varphi} = -\frac{1}{n_s}\sum_{j=1}^{n_s}\sum_{k=1}^{K} \mathrm{I}[k = y^s_j] \log p^s_{j,k},$$

$$\min_{\boldsymbol{\theta},\{\boldsymbol{\mu}_k\}_{k=1}^K} \mathcal{L}^s_{\varphi} = -\frac{1}{n_s}\sum_{j=1}^{n_s}\sum_{k=1}^{K} \mathrm{I}[k = y^s_j] \log \widetilde{p}^s_{j,k},$$

$$\widetilde{p}^s_{j,k} = \frac{\exp((1+||\boldsymbol{z}^s_j - \boldsymbol{\mu}_k||^2)^{-1})}{\sum_{k'=1}^{K}\exp((1+||\boldsymbol{z}^s_j - \boldsymbol{\mu}_{k'}||^2)^{-1})}.$$

Method

- Enhancement via soft source selection
  - c^t is the cluster center of **target domain** and u_k are updated for every epoch

$$w^s(\boldsymbol{x}^s) = \frac{1}{2}\left(1 + \frac{\boldsymbol{c}_{y^s}^{t\top}\boldsymbol{x}^s}{||\boldsymbol{c}_{y^s}^t||\,||\boldsymbol{x}^s||}\right) \in [0,1].$$

$$\mathcal{L}^s_{f\circ\varphi(\cdot;\{w_j^s\}_{j=1}^{n_s})} = -\frac{1}{n_s}\sum_{j=1}^{n_s} w_j^s \sum_{k=1}^{K} \mathrm{I}[k = y_j^s] \log p_{j,k}^s, \quad (13)$$

$$\mathcal{L}^s_{\varphi(\cdot;\{w_j^s\}_{j=1}^{n_s})} = -\frac{1}{n_s}\sum_{j=1}^{n_s} w_j^s \sum_{k=1}^{K} \mathrm{I}[k = y_j^s] \log \widetilde{p}_{j,k}^s. \quad (14)$$

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
Source Model [21]	77.8±0.2	96.9±0.1	99.3±0.1	82.1±0.2	64.5±0.2	66.1±0.2	81.1
DAN [34]	81.3±0.3	97.2±0.0	99.8±0.0	83.1±0.2	66.3±0.0	66.3±0.1	82.3
DANN [16]	81.7±0.2	98.0±0.2	99.8±0.0	83.9±0.7	66.4±0.2	66.0±0.3	82.6
ADDA [53]	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
VADA [5				86.7±0.4	70.1±0.4	70.5±0.4	85.4
SimNet [43]				85.3±0.3	73.4±0.8	71.8±0.6	86.2
MSTN [59]	91.3	98.9	**100.0**	90.4	72.7	65.6	86.5
GTA [49]	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
MCD [48]	88.6±0.2	98.5±0.1	**100.0**±0.0	92.2±0.2	69.5±0.1	69.7±0.3	86.5
SAFN+ENT [60]	90.1±0.8	98.6±0.2	99.8±0.0	90.7±0.5	73.0±0.2	70.2±0.3	87.1
DAAA [28]	86.8±0.2	**99.3**±0.1	**100.0**±0.0	88.8±0.4	74.3±0.2	73.9±0.2	87.2
iCAN [63]	92.5	98.8	**100.0**	90.1	72.1	69.9	87.2
CDAN+E [35]	94.1±0.1	98.6±0.1	**100.0**±0.0	92.9±0.2	71.0±0.3	69.3±0.3	87.7
MSTN+DSBN [5]	92.7	99.0	**100.0**	92.2	71.7	74.4	88.3
TADA [56]	94.3±0.3	98.7±0.1	99.8±0.2	91.6±0.3	72.9±0.2	73.0±0.3	88.4
TAT [33]	92.5±0.3	**99.3**±0.1	**100.0**±0.0	93.2±0.2	73.1±0.3	72.1±0.3	88.4
SymNets [68]	90.8±0.1	98.8±0.3	**100.0**±0.0	93.9±0.5	74.6±0.6	72.5±0.5	88.4
BSP+CDAN [9]	93.3±0.2	98.2±0.2	**100.0**±0.0	93.0±0.2	73.6±0.3	72.6±0.3	88.5
MDD [66]	94.5±0.3	98.4±0.1	**100.0**±0.0	93.5±0.2	74.6±0.3	72.2±0.1	88.9
CAN [27]	94.5±0.3	99.1±0.2	99.8±0.2	95.0±0.3	**78.0**±0.3	77.0±0.3	90.6
SRDC	**95.7**±0.2	99.2±0.1	**100.0**±0.0	**95.8**±0.2	76.7±0.3	**77.1**±0.1	**90.8**

Table 3. Results (%) on Office-31 (ResNet-50).

Methods	I → P	P → I	I → C	C → I	C → P	P → C	Avg
Source Model [21]	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
DAN [34]	74.5±0.4	82.2±0.2	92.8±0.2	86.3±0.4	69.2±0.4	89.8±0.4	82.5
DANN [16]	75.0±0.6	86.0±0.3	96.2±0.4	87.0±0.5	74.3±0.5	91.5±0.6	85.0
JAN [37]	76.8±0.4	88.0±0.2	94.7±0.2	89.5±0.3	74.2±0.3	91.7±0.3	85.8
CDAN+E [35]	77.7±0.3	90.7±0.2	97.7±0.3	91.3±0.3	74.2±0.2	94.3±0.3	87.7
TAT [33]	78.8±0.2	92.0±0.2	97.5±0.3	92.0±0.3	78.2±0.4	94.7±0.4	88.9
SAFN+ENT [60]	79.3±0.1	93.3±0.4	96.3±0.4	91.7±0.0	77.6±0.1	95.3±0.1	88.9
SymNets [68]	80.2±0.3	93.6±0.2	97.0±0.3	93.4±0.3	78.7±0.3	96.4±0.1	89.9
SRDC	**80.8**±0.3	**94.7**±0.2	**97.8**±0.2	**94.1**±0.2	**80.0**±0.3	**97.7**±0.1	**90.9**

Table 4. Results (%) on ImageCLEF-DA (ResNet-50).

Methods	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
Source Model [21]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [34]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [16]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [37]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
SE [15]	48.8	61.8	72.8	54.1	63.2	65.1	50.6	49.2	72.3	66.1	55.9	78.7	61.5
DWT-MEC [45]	50.3	72.1	77.0	59.6	69.3	70.2	58.3	48.1	77.3	69.3	53.6	82.0	65.6
CDAN+E [35]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
TAT [33]	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
BSP+CDAN [9]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
SAFN [60]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
TADA [56]	53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	60.0	82.9	67.6
SymNets [68]	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	74.5	52.6	82.7	67.6
MDD [66]	**54.9**	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	**60.2**	82.3	68.1
SRDC	52.3	**76.3**	**81.0**	**69.5**	**76.2**	**78.0**	**68.7**	**53.8**	**81.7**	**76.3**	57.1	**85.0**	**71.3**

Table 5. Results (%) on Office-Home (ResNet-50).

Experiment results

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD*	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet*	73.58 (↓)	72.24 (↓)	69.21 (↓)	68.99 (↓)	71.06 (↓)	73.50 (↑)	71.02 (↓)
AT	74.08 (↓)	72.77 (↓)	70.55 (↓)	70.22 (↓)	72.31 (↓)	73.44 (↑)	71.43 (↓)
SP	73.83 (↓)	72.43 (↓)	69.67 (↓)	70.04 (↓)	72.69 (↓)	72.94 (↓)	72.68 (↓)
CC	73.56 (↓)	72.21 (↓)	69.63 (↓)	69.48 (↓)	71.48 (↓)	72.97 (↓)	70.71 (↓)
VID	74.11 (↓)	73.30 (↓)	70.38 (↓)	70.16 (↓)	72.61 (↓)	73.09 (↓)	71.23 (↓)
RKD	73.35 (↓)	72.22 (↓)	69.61 (↓)	69.25 (↓)	71.82 (↓)	71.90 (↓)	71.48 (↓)
PKT	74.54 (↓)	73.45 (↓)	70.34 (↓)	70.25 (↓)	72.61 (↓)	73.64 (↑)	72.88 (↓)
AB	72.50 (↓)	72.38 (↓)	69.47 (↓)	69.53 (↓)	70.98 (↓)	73.17 (↓)	70.94 (↓)
FT*	73.25 (↓)	71.59 (↓)	69.84 (↓)	70.22 (↓)	72.37 (↓)	72.86 (↓)	70.58 (↓)
FSP*	72.91 (↓)	n/a	69.95 (↓)	70.11 (↓)	71.89 (↓)	72.62 (↓)	70.23 (↓)
NST*	73.68 (↓)	72.24 (↓)	69.60 (↓)	69.53 (↓)	71.96 (↓)	73.30 (↓)	71.53 (↓)
CRD	**75.48** (↑)	**74.14** (↑)	**71.16** (↑)	**71.46** (↑)	**73.48** (↑)	**75.51** (↑)	**73.94** (↑)
CRD+KD	75.64 (↑)	74.38 (↑)	71.63 (↑)	71.56 (↑)	73.75 (↑)	75.46 (↑)	74.29 (↑)

Table 1: Test *accuracy* (%) of student networks on CIFAR100 of a number of distillation methods (ours is CRD); see Appendix for citations of other methods. ↑ denotes outperformance over KD and ↓ denotes underperformance. We note that CRD is the *only* method to always outperform KD (and also outperforms all other methods). We denote by * methods where we used our reimplementation based on the paper; for all other methods we used author-provided or author-verified code. Average over 5 runs.

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.6	64.6	70.36	70.5	71.82	70.5
KD*	67.37	67.35	73.81	74.07	74.45	74.83
FitNet*	64.14 (↓)	63.16 (↓)	70.69 (↓)	73.59 (↓)	73.54 (↓)	73.73 (↓)
AT	59.40 (↓)	58.58 (↓)	71.84 (↓)	71.73 (↓)	72.73 (↓)	73.32 (↓)
SP	66.30 (↓)	68.08 (↑)	73.34 (↓)	73.48 (↓)	74.56 (↑)	74.52 (↓)
CC	64.86 (↓)	65.43 (↓)	70.25 (↓)	71.14 (↓)	71.29 (↓)	71.38 (↓)
VID	65.56 (↓)	67.57 (↑)	70.30 (↓)	73.38 (↓)	73.40 (↓)	73.61 (↓)
RKD	64.52 (↓)	64.43 (↓)	71.50 (↓)	72.28 (↓)	73.21 (↓)	72.21 (↓)
PKT	67.13 (↓)	66.52 (↓)	73.01 (↓)	74.10 (↑)	74.69 (↑)	73.89 (↓)
AB	66.06 (↓)	67.20 (↓)	70.65 (↓)	73.55 (↓)	74.31 (↓)	73.34 (↓)
FT*	61.78 (↓)	60.99 (↓)	70.29 (↓)	71.75 (↓)	72.50 (↓)	72.03 (↓)
NST*	58.16 (↓)	64.96 (↓)	71.28 (↓)	74.12 (↑)	74.68 (↑)	74.89 (↑)
CRD	**69.73** (↑)	**69.11** (↑)	**74.30** (↑)	**75.11** (↑)	**75.65** (↑)	**76.05** (↑)
CRD+KD	69.94 (↑)	69.54 (↑)	74.58 (↑)	75.12 (↑)	76.05 (↑)	76.27 (↑)

Table 2: Top-1 test *accuracy* (%) of student networks on CIFAR100 of a number of distillation methods (ours is CRD) for transfer across very different teacher and student architectures. CRD outperforms KD and all other methods. Importantly, some methods that require very similar student and teacher architectures perform quite poorly. E.g. FSP (Yim et al., 2017) cannot even be applied; AT (Ba & Caruana, 2014) and FitNet (Zagoruyko & Komodakis, 2016a) perform very poorly etc. We denote by * methods where we used our reimplementation based on the paper; for all other methods we used author-provided or author-verified code. Average over 3 runs.

Q&A

Thanks For Listening