HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Gordon Center for Medical Imaging

Energy-constrained Self-training for Unsupervised Domain Adaptation

Xiaofeng Liu, Bo Hu, Xiongchang Liu, Jun Lu, Jane You, Lingsheng Kong

https://liu-xiaofeng.github.io/ liuxiaofengcmu@gmail.com

HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Gordon Center for Medical Imaging

Deep neural networks are usually data-starved and rely on the i.i.d assumption of training and testing.

Unsupervised Domain Adaptation

Self-training based UDA

Considering the noisy pseudo-label, previous work [12] proposes to construct a more conservative pseudo-label that smoothing the one-hot distribution or regularize it with the entropy.

The solution proposes in this paper is orthogonal with [12], which resorts to the additional supervision signal of EBM that independent of pseudo-label. Compared with the manually defined label smoothing in [12], the energy-constraint can adaptively regularize the training w.r.t. the input and the present network parameters.

HARVARD MEDICAL SCHOOL
MASSACHUSETTS GENERAL HOSPITAL
Gordon Center for Medical Imaging

optimizing the likelihood $\log p_{\mathbf{w}}(\mathbf{x},\mathbf{y}) = \boxed{\log p_{\mathbf{w}}(\mathbf{x})} + \boxed{\log p_{\mathbf{w}}(\mathbf{y}|\mathbf{x})}$ can be helpful for both the discrimination and generation task.

The additional optimization objective $\log p_{\mathbf{w}}(\mathbf{x})$ has been proven and evidenced that can improve the confidence calibration and robustness for conventional classification task [15].

Cross-entropy loss

Considering the target samples do not have ground truth label, the self-training methods [12] utilize the inaccurate pseudo label to calculate the cross-entropy loss. Therefore, optimizing $\log p_{\mathbf{w}}(\mathbf{x})$ can potentially be more helpful for UDA setting. Actually, $\log p_{\mathbf{w}}(\mathbf{x})$ is adaptive w.r.t. the input $\mathbf{x}$ and network parameter $\mathbf{w}$, and irrelevant to the inaccurate pseudo label, which can be an ideal regularizer of self-training based UDA.

HARVARD MEDICAL SCHOOL MASSACHUSETTS GENERAL HOSPITAL Gordon Center for Medical Imaging

However, how to modeling $\log p_{\mathbf{w}}(\mathrm{x})$ can be a challenging

Usually we rely on the sophisticate Markov Chain Monte Carlo sampler to train EBMs.

Considering $\frac{\partial \log p_{\mathbf{w}}(\mathrm{x})}{\partial \mathbf{w}}$ can be approximated with $-\frac{\partial E_{\mathbf{w}}(\mathrm{x})}{\partial \mathbf{w}}$ [15], it is possible to modeling the energy function $E_{\mathbf{w}}(\mathrm{x})$ instead of $\log p_{\mathbf{w}}(\mathrm{x})$. Following [15], we can define an EBM of the joint distribution $p_{\mathbf{w}}(\mathrm{x},\mathrm{y}) = \exp(f_{\mathbf{w}}(\mathrm{x})[k])/Z(\mathrm{w})$, by defining $E_{\mathbf{w}}(\mathrm{x},\mathrm{y}) = -f_{\mathbf{w}}(\mathrm{x})[k]$. By marginalizing out y, we have $p_{\mathbf{w}}(\mathrm{x}) = \frac{\sum_k \exp(f_{\mathbf{w}}(\mathrm{x})[k])}{Z(\mathbf{w})}$ [15]. Considering $p_{\mathbf{w}}(\mathrm{x}) = \exp(-E_{\mathbf{w}}(\mathrm{x}))/Z(\mathrm{w})$, the energy function of x can be

$$E_{\mathbf{w}}(\mathrm{x}) = -\log \sum_k \exp(f_{\mathbf{w}}(\mathrm{x})[k]) \tag{1}$$

In this setting, $p_{\mathbf{w}}(\mathrm{x}|\mathrm{y}) = \frac{p_{\mathbf{w}}(\mathrm{x},\mathrm{y})}{p_{\mathbf{w}}(\mathrm{x})} = \frac{\exp(f_{\mathbf{w}}(\mathrm{x})[k])/Z(\mathbf{w})}{\sum_k \exp(f_{\mathbf{w}}(\mathrm{x})[k])/Z(\mathbf{w})}$. The normalization constant $Z(\mathrm{w})$ will be canceled out and yielding the standard softmax function, which bridges the EMB and conventional classifiers.

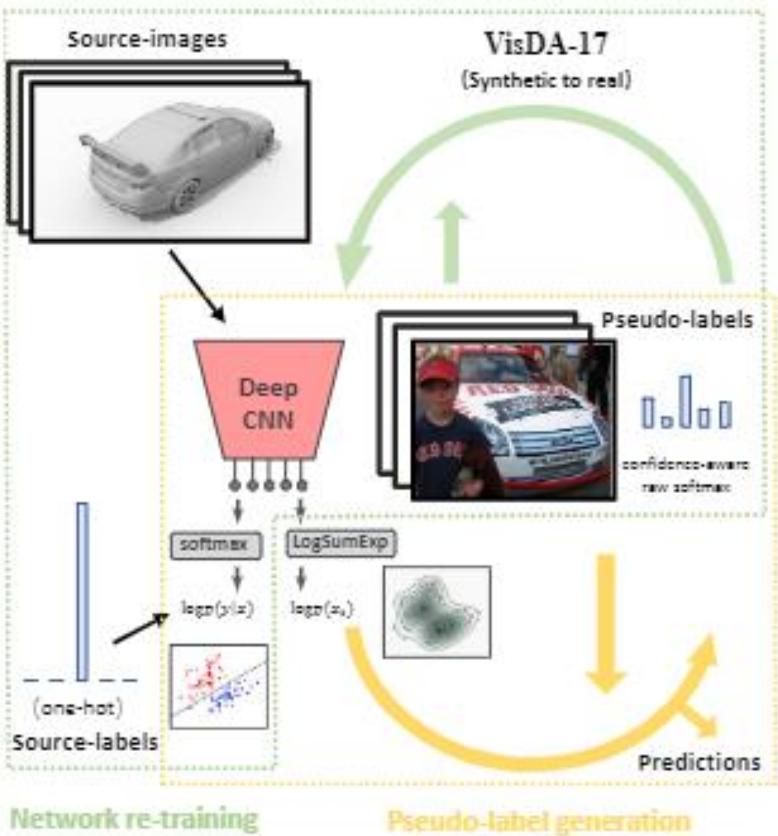

Fig. 1: The illustration of our Energy-constrained Self-training framework for UDA. Minimizing the pseudo label-irrelevant energy of $E_{\mathbf{w}}(\mathbf{x}_t)$ is introduced as additional objective for the target sample.

$$E_{\mathbf{w}}(\mathrm{x}) = -\log \sum_{k} \exp(f_{\mathbf{w}}(\mathrm{x})[k]) \quad (1)$$

In this setting, $p_{\mathbf{w}}(\mathrm{x}|\mathrm{y}) = \frac{p_{\mathbf{w}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{w}}(\mathbf{x})} = \frac{\exp(f_{\mathbf{w}}(\mathbf{x})[k])/Z(\mathbf{w})}{\sum_k \exp(f_{\mathbf{w}}(\mathbf{x})[k])/Z(\mathbf{w})}$. The normalization constant $Z(\mathrm{w})$ will be canceled out and yielding the standard softmax function, which bridges the EMB and conventional classifiers.

HARVARD MEDICAL SCHOOL
MASSACHUSETTS GENERAL HOSPITAL
Gordon Center for Medical Imaging

Following the formulation in our CRST [12], the self-training with EBM regularization (R-EBM) for target sample, i.e., $E_{\mathbf{w}}(\mathbf{x}_t)$, can be formulated as

$$\min_{\mathbf{w},\hat{\mathbf{Y}}_T} \mathcal{L}_{R-EBM}(\mathbf{w},\hat{\mathbf{Y}}) = -\sum_{s\in S}\sum_{k=1}^{K} y_s^{(k)} \log p_{\mathbf{w}}(k|\mathbf{x}_s)$$
$$-\sum_{t\in T}\{\sum_{k=1}^{K}[\hat{y}_t^{(k)} \log p_{\mathbf{w}}(k|\mathbf{x}_t) - \hat{y}_t^{(k)}\log \lambda_k] - \alpha E_{\mathbf{w}}(\mathbf{x}_t)\}$$
$$s.t.\ \hat{\mathbf{y}}_t \in \Delta^{K-1} \cup \{\mathbf{0}\}, \forall t \quad (2)$$

Step 1) Pseudo-label generation Fix w and solve:

$$\min_{\hat{\mathbf{Y}}_T} -\sum_{t\in T}\{\sum_{k=1}^{K}\hat{y}_t^{(k)}[\log p_{\mathbf{w}}(k|\mathbf{x}_t)-\log\lambda_k]-\alpha E_{\mathbf{w}}(\mathbf{x}_t)\}$$
$$s.t.\ \hat{y}_t\in\Delta^{K-1}\cup\{\mathbf{0}\},\forall t \tag{3}$$

For solving step **1)**, there is a global optimizer for arbitrary $\hat{\mathbf{y}}_t=(\hat{y}_t^{(1)},...,\hat{y}_t^{(K)})$ as [12]:

$$\hat{y}_t^{(k)*}=\begin{cases}1, \text{ if } k=\underset{k}{\text{argmax}}\ \dfrac{p_{\mathbf{w}}(k|\mathbf{x}_t)}{\lambda_k} \\ \qquad \text{and} \quad p_{\mathbf{w}}(k|\mathbf{x}_t)>\lambda_k \\ 0, \text{ otherwise}\end{cases} \tag{4}$$

Step 2) Network retraining Fix $\hat{\mathrm{Y}}_T$ and minimize

$$-\sum_{s \in S} \sum_{k=1}^{K} y_s^{(k)} \log p_{\mathbf{w}}(k|\mathbf{x}_s) - \sum_{t \in T} \sum_{k=1}^{K} \hat{y}_t^{(k)} \log p_{\mathbf{w}}(k|\mathbf{x}_t) \quad (5)$$

w.r.t. w. Carrying out step **1)** and **2)** for one time is defined as one round in self-training.

HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Gordon Center for Medical Imaging

Method	Base Net	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg.	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Source	DRN26	42.7	26.3	51.7	5.5	6.8	13.8	23.6	6.9	75.5	11.5	36.8	49.3	0.9	46.7	3.4	5.0	0.0	5.0	1.4	21.7
CyCADA [47]		79.1	33.1	77.9	23.4	17.3	32.1	33.3	31.8	81.5	26.7	69.0	62.8	14.7	74.5	20.9	25.6	6.9	18.8	20.4	39.5
Source	DRN105	36.4	14.2	67.4	16.4	12.0	20.1	8.7	0.7	69.8	13.3	56.9	37.0	0.4	53.6	10.6	3.2	0.2	0.9	0.0	22.2
MCD [42]		90.3	31.0	78.5	19.7	17.3	28.6	30.9	16.1	83.7	30.0	69.1	58.5	19.6	81.5	23.8	30.0	5.7	25.7	14.3	39.7
Source	PSPNet	69.9	22.3	75.6	15.8	20.1	18.8	28.2	17.1	75.6	8.00	73.5	55.0	2.9	66.9	**34.4**	30.8	0.0	18.4	0.0	33.3
DCAN [48]		85.0	30.8	81.3	25.8	21.2	22.2	25.4	26.6	83.4	36.7	76.2	58.9	24.9	80.7	29.5	42.9	2.50	26.9	11.6	41.7
Source	DeepLabv2	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
AdaptSegNet [49]		86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
AdvEnt [50]	DeepLabv2	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
Source	DeepLabv2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	29.2
FCAN [51]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.6
Source	DeepLabv2	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
DPR [52]		92.3	51.9	**82.1**	29.2	25.1	24.5	33.8	**33.0**	82.4	32.8	**82.2**	58.6	27.2	84.3	33.4	**46.3**	2.2	29.5	32.3	46.5
Source	DeepLabv2	73.8	16.0	66.3	12.8	22.3	29.0	30.3	10.2	77.7	19.0	50.8	55.2	20.4	73.6	28.3	25.6	0.1	27.5	12.1	34.2
PyCDA [53]		90.5	36.3	84.4	32.4	**28.7**	34.6	36.4	31.5	86.8	37.9	78.5	62.3	21.5	**85.6**	27.9	34.8	18.0	22.9	**49.3**	47.4
Source	DeepLabv2	71.3	19.2	69.1	18.4	10.0	35.7	27.3	6.8	79.6	24.8	72.1	57.6	19.5	55.5	15.5	15.1	11.7	21.1	12.0	33.8
CBST [31]		89.9	55.0	79.9	29.5	20.6	37.8	32.9	13.9	84.0	31.2	75.5	60.2	27.1	81.8	29.7	40.5	7.62	28.7	41.4	45.6
CBST+R_{EBM}		91.1	53.9	80.6	31.6	21.0	40.4	35.0	19.8	86.8	35.9	76.4	**63.3**	31.4	83.0	22.5	38.6	24.2	32.2	39.4	47.8
Source	DeepLabv2	71.3	19.2	69.1	18.4	10.0	35.7	27.3	6.8	79.6	24.8	72.1	57.6	19.5	55.5	15.5	15.1	11.7	21.1	12.0	33.8
CRST [12]		89.0	51.2	79.4	**31.7**	19.1	38.5	34.1	20.4	84.7	35.4	76.8	61.3	30.2	80.7	27.4	39.4	10.2	32.2	43.3	46.6
CRST+R_{EBM}		**92.5**	**56.6**	80.9	26.2	20.5	**40.5**	**35.3**	24.4	**86.9**	**37.3**	77.5	63.4	**30.5**	81.3	28.8	39.2	**24.6**	**33.5**	41.3	**48.5**

TABLE II: Experimental results for GTA5 to Cityscapes.

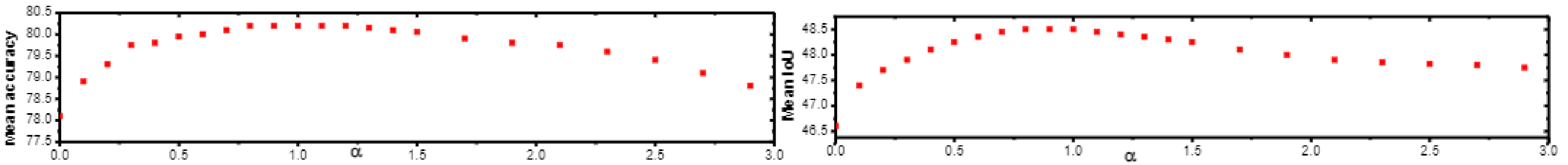

Fig. 2: Sensitive analysis of hyper-parameter α in VisDA17 (left) and CTA52Sityscapes (right) with CRST+R_{EBM}.

HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Gordon Center for Medical Imaging

Method	Aero	Bike	Bus	Car	Horse	Knife	Motor	Person	Plant	Skateboard	Train	Truck	Mean
Source-Res101 [38]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MMD [39]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN [40]	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
ENT [41]	80.3	75.5	75.8	48.3	77.9	27.3	69.7	40.2	46.5	46.6	79.3	16.0	57.0
MCD [42]	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
ADR [43]	87.8	79.5	**83.7**	65.3	**92.3**	61.8	88.9	73.2	87.8	60.0	**85.5**	32.3	74.8
DEV [44]	81.83	53.48	82.95	71.62	89.16	72.03	**89.36**	75.73	**97.02**	55.48	71.19	29.17	72.42
TDDA [38]	88.2	78.5	79.7	71.1	90.0	81.6	84.9	72.3	92.0	52.6	82.9	18.4	74.03
CBST [31]	87.1±1.2	79.5±2.3	58.3±2.6	50.4±3.9	82.8±2.1	73.7±7.2	80.9±2.6	71.8±3.1	81.6±3.2	88.4±3.3	75.2±1.2	68.4±3.4	74.8±0.5
CBST+R_{EBM}	87.9±1.6	79.6±1.5	68.5±1.3	68.6±1.9	83.2±1.2	78.4±1.9	83.5±1.5	72.2±1.5	82.2±1.6	84.3±1.5	80.9±1.4	67.5±1.3	77.0±0.6
CRST[12]	89.2±1.6	79.6±4.6	64.2±4.0	57.8±3.4	87.8±1.9	79.6±8.5	85.6±2.6	75.9±4.2	86.5±2.2	85.1±2.4	77.7±2.2	68.5±0.9	78.1±0.7
CRST+R_{EBM}	**90.3**±1.5	**82.6±1.2**	72.4±1.5	**71.7±1.8**	87.6±1.8	**81.8±1.9**	85.4±1.5	**80.8±1.5**	87.1±1.6	**89.9±1.5**	83.6±1.6	**71.5±1.3**	**80.2±0.5**
SimNet* [45]	**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
GTA* [46]	-	-	-	-	-	-	-	-	-	-	-	-	77.1
CRST+R_{EBM}*	93.2±1.3	**85.8±1.2**	73.7±1.0	**74.3±1.5**	89.5±0.6	**87.6±1.6**	88.2±1.6	**82.2±1.6**	90.9±1.3	**91.6±1.8**	85.1±1.4	**79.9±1.4**	**82.8±0.5**

TABLE I: Experimental results for VisDA17-val setting. We use ResNet101 as backbone except SimNet and GTA.*ResNet152 backbone.

Energy-constrained Self-training for Unsupervised Domain Adaptation

Xiaofeng Liu, Bo Hu, Xiongchang Liu, Jun Lu, Jane You, Lingsheng Kong

https://liu-xiaofeng.github.io/ liuxiaofengcmu@gmail.com