Augmented Self-Labeling for Source-Free Unsupervised Domain Adaptation

Hao Yan
Carleton University
haoyan6@cmail.carleton.ca

Yuhong Guo
Carleton University
Canada CIFAR AI Chair, Amii
yuhong.guo@carleton.ca

Chunsheng Yang
National Research Council Canada
chunsheng.yang@nrc-cnrc.gc.ca

Abstract

Unsupervised domain adaptation aims to learn a prediction model that generalizes well on a target domain given labeled source data and unlabeled target data. However, source data sometimes can be unavailable due to data privacy or decentralized learning architectures. In this paper, we address the source-free unsupervised domain adaptation problem where only the pretrained source model and unlabeled target data are given. To this end, we propose an Augmented Self-Labeling (ASL) method that jointly optimizes the prediction model and the pseudo-labels for the target data starting from the initial source model. It involves two alternating steps: augmented self-labeling improves pseudo-labels by solving an optimal transport problem with the Sinkhorn-Knopp algorithm, and model re-training trains the model with the supervision of improved pseudo-labels. We further introduce model regularization terms to improve the model re-training. Experiments show that our method achieves comparable or better results than the state-of-the-art methods on the standard benchmarks.

1 Introduction

Unsupervised domain adaptation tackles the setting where labeled source data and unlabeled target data are available when adapting to target domains. However, source domain data might be inaccessible in some privacy-sensitive applications. For example, federated learning collaboratively trains a model using decentralized data on mobile devices without fetching data into a centralized machine [2]. When adapting the model trained via federated learning, we have no access to the source data. This induces the source-free unsupervised domain adaptation setting, where only the trained source model and the unlabeled target data are given. Traditional unsupervised domain adaptation methods are not applicable to this setting because they usually seek to align distributions of source and target domains with data samples from both domains. A few methods that tackle this source-free unsupervised domain adaptation has recently been proposed in the literature. For example, SHOT [16] alternately refines the pseudo-labels of target data with a prototype classifier and fine-tunes the feature extractor together with a model regularization term to maximize the mutual information between model inputs and outputs. PPDA [11] assigns pseudo-labels based on a prototype classifier and a sample-level re-weighting scheme. 3C-GAN [15] and SDDA [12] utilize the conditional GAN to generate labeled target data through input-level adversarial training.

In this paper, we propose a new Augmented Self-Labeling (ASL) method for the source-free unsupervised domain adaptation problem. It involves two alternating steps: the augmented self-labeling

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

step aims to improve the pseudo-labels of target data and the model re-training step retrains the target model with the self-labeled target data. These two steps work together to gradually adapt the prediction model by incorporating the target data into training. In particular, we propose to ensemble the predicted probabilities corresponding to multiple randomly augmented versions of the same sample for self-labeling. We derive this problem into an instance of the optimal transport problem. In order to avoid the degenerate solution, we add the equipartition constraint on the labels and solve the problem efficiently via a fast version of the Sinkhorn-Knopp algorithm [5]. Moreover, we also introduce several model regularization terms to improve the model re-training. We apply the proposed ASL method to the source-free unsupervised domain adaptation tasks. Experiments show that our method can achieve comparable or even better results than the state-of-the-art methods on the standard benchmarks.

2 Augmented Self-Labeling (ASL) for Source-Free Domain Adaptation

This paper tackles the source-free unsupervised domain adaptation problem where only the pretrained source model and the unlabeled target data are available. Specifically, we assume the source prediction model $f_s : \mathcal{X}_s \to \mathcal{Y}_s$ trained in the source domain and the unlabeled target data $\{x_i\}_{i=1}^N$ are given. The goal is to produce a prediction model f that works well on the target data. We propose to adapt the source model into the target domain to induce a good target prediction model f by alternatively performing pseudo-label refinement on the target data through augmented self-labeling and conducting model retraining with the refined pseudo-labels.

2.1 Augmented Self-Labeling

We initialize the target model f with the parameters of the source model f_s. Given the unlabeled target data $\{x_i\}_{i=1}^N$, their pseudo-labels can be obtained by choosing the highly confident predictions based on the current model f, which are further used to fine-tune the model in an alternative manner.

Due to the existence of domain discrepancy, pseudo-labels for target data can be noisy which may lead to error accumulation in the target model. We thus propose an Augmented Self-Labeling method to optimize the target labels from the weighted average of multiple output predictions corresponding to samples with random data augmentations. Specifically, M different augmented version of samples $\{x_i^m\}_{m=1}^M$ can be obtained from the original sample x_i by independently applying random data augmentations M times, i.e.

$$\{x_i^1, x_i^2, \cdots, x_i^M\} = RandAugment(x_i), \tag{1}$$

where $RandAugment(\cdot)$ denotes a combination of multiple random data augmentations. The data augmentations we used include random resized crop, random auto-contrast and random color distortion [3]. In order to reduce the noise in the predicted probability, we take the ensemble of the $M+1$ probabilities corresponding to the M augmented version and the unaugmented version of each sample x_i to get the following average prediction outcome, which indicates the probability of the sample x_i belonging to class y,

$$p_{iy} = \frac{1}{2} p(y|x_i;\theta) + \frac{1}{2M} \sum_{m=1}^{M} p(y|x_i^m;\theta). \tag{2}$$

Here half weight is assigned to the predicted probability on the unaugmented version of the sample. The reason is that the original data samples still carry most of the useful information and higher weights assigned to the original samples can make the prediction more stable and reliable.

Next, we propose to improve the pseudo-labels through the following augmented self-labeling:

$$\begin{aligned} &\min_{\{q_{iy}\}} -\sum_{i=1}^{N}\sum_{y=1}^{K} q_{iy} \log p_{iy} + \lambda \sum_{i=1}^{N}\sum_{y=1}^{K} q_{iy} \log q_{iy} \\ &s.t. \quad \forall i, y: \quad q_{iy} \in [0,1], \quad \sum_{y=1}^{K} q_{iy} = 1, \quad \sum_{i=1}^{N} q_{iy} = \frac{N}{K}. \end{aligned} \tag{3}$$

where p_{iy} is the augmented prediction term in Eq.(2), K is the number of classes, and $\{q_{iy}\}$ denote the soft-labels we aim to produce after taking the overall prediction situation into consideration

through constraints. The negative entropy term in the objective is added to get smoothed soft-labels while the hyperparameter λ controls the degree of smoothness. Following [25], the equipartition constraints, $\{\sum_i q_{iy} = N/K, \forall y\}$, are added to avoid degenerate solutions where the same arbitrary label is assigned to all the samples. The constraints enforce that each category contains similar number of samples, which is reasonable in class-balanced datasets. In general, it can be explained as enforcing maximization over the mutual information between the sample indices and labels [25].

This augmented self-labeling problem for pseudo-label refinement is actually an instance of the optimal transport problem [5]. To make it more clear, we convert the notations in Eq.(3) to matrix form by introducing Q with $[Q]_{iy} = q_{iy}$ as the label matrix with dimension of $N \times K$ and P with $[P]_{iy} = p_{iy}$ as the predicted probability matrix with dimension of $N \times K$. The objective in Eq. (3) can be rewritten as:

$$\min_{Q \in U(r,c)} \langle Q, -\log P \rangle - \lambda H(Q) \tag{4}$$

where $\langle \cdot \rangle$ denotes the Frobenius inner product, i.e. the sum of element-wise product between two matrices. The matrix Q is thus constrained to be an element of the transport polytope [5],

$$U(r,c) := \{Q \in \mathbb{R}_+^{N \times K} | Q\mathbf{1}_K = r, Q^\top \mathbf{1}_N = c\}. \tag{5}$$

where $r = \mathbf{1}_N$, and $c = \frac{N}{K}\mathbf{1}_K$; $\mathbf{1}_d$ denotes a column vector with all 1 values and length d. This is equivalent to the constraints of labels in Eq. (3). This problem can be solved via the fast version of Sinkhorn-Knopp algorithm [5], which has the following form of solution:

$$Q = \text{diag}(u) \cdot P^{1/\lambda} \cdot \text{diag}(v), \tag{6}$$

where u, v are vectors guaranteeing the constraints in Eq. (5) and can be computed with the Sinkhorn's fixed point iteration until convergence:

$$(u, v) \leftarrow (r./(P^{1/\lambda}v), c./([P^{1/\lambda}]^\top u)). \tag{7}$$

Specifically, we initialize v as normalized unit vector and then iteratively compute u and v until v converges. In practice, this iteration can converge in a few steps.

2.2 Model Re-training and Regularization

After pseudo-label refinement through the augmented self-labeling procedure above, we convert the soft-labels $\{q_{iy} \in [0, 1]\}$ into hard-labels $\{\hat{q}_{iy} \in \{0, 1\}\}$ by setting the pseudo-label of each instance as the class label with the maximum probability. Then we perform model re-training to update the target model $f(\cdot; \theta)$ by minimizing the following cross-entropy loss:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{K} \hat{q}_{iy} \log p(y|x_i; \theta). \tag{8}$$

In order to learn a target prediction model that generalizes well, we further incorporate several model regularization terms into the model re-training process. First, we adopt the following conditional entropy minimization [8] loss as a regularization term:

$$\mathcal{L}_{ent} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{K} p(y|x_i; \theta) \log p(y|x_i; \theta) \tag{9}$$

This conditional entropy regularization term is suitable for exploiting unlabeled data in a semi-supervised manner, and it pushes the decision boundary far away from the data dense regions to support the cluster assumption [22]. However, the conditional entropy above is empirically estimated using the available target data. According to [8, 22], this approximation holds only if the model is locally-Lipschitz. Therefore, we further add the following virtual adversarial loss [18] as a regularization term to guarantee the locally-Lipschitz constraint:

$$\mathcal{L}_{vat} = \mathbb{E}_x \left[\max_{\|r\| \leq \epsilon} D_{\text{KL}}(f(x) \| f(x + r)) \right], \tag{10}$$

where $D_{\text{KL}}(\cdot \| \cdot)$ is the Kullback-Leibler Divergence and $f(x) = [p(1|x; \theta), \cdots, p(K|x; \theta)]$.

Table 1: Classification accuracy (%) on Office-31 (ResNet-50)

Methods	A → D	A → W	D → A	D → W	W → A	W → D	Avg.
ResNet-50 [9]	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DANN [7]	79.7	82.0	68.2	96.9	67.4	99.1	82.2
ADDA [23]	77.8	86.2	69.5	96.2	68.9	98.4	82.9
CDAN+E [17]	92.9	94.1	71.0	98.6	69.3	100.	87.7
CDAN+BSP [4]	93.0	93.3	73.6	98.2	72.6	100.	88.5
CDAN+TransNorm [24]	94.0	95.7	73.4	98.7	74.2	100.	89.3
CAN [10]	95.0	94.5	78.0	99.1	77.0	99.8	90.6
SDDA [12]	85.3	82.5	66.4	**99.0**	67.7	99.8	83.5
SHOT [16]	93.1	90.9	74.5	98.8	74.8	**99.9**	88.7
3C-GAN [15]	92.7	93.7	75.3	98.5	**77.8**	99.8	**89.6**
ASL (Ours)	**93.4**	**94.1**	**76.0**	98.4	75.0	99.8	89.5

Moreover, as the optimal classifier would generalize well on both domains according to the theoretical analysis in [1], we add the following weight regularization term to prevent excessive deviation from the source model parameters θ_s:

$$\mathcal{L}_{wr} = \|\theta - \theta_s\|_2^2. \tag{11}$$

This weight regularization not only prevents the target hypothesis from getting far away from the initial source model and helps to preserve the source knowledge in the target model, but also stables the re-training of the target model during alternative updates. Finally, by integrating these regularization terms into the cross-entropy loss in Eq. (8), we have the following overall loss function for model re-training:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1(\mathcal{L}_{ent} + \mathcal{L}_{vat}) + \lambda_2\mathcal{L}_{wr}, \tag{12}$$

where λ_1 and λ_2 are trade-off parameters. The entropy loss and virtual adversarial loss empirically share the same trade-off parameter [22, 15].

3 Experiments

We conduct experiments on the standard benchmarks under the source-free unsupervised domain adaptation setting. We compare our ASL method[1] with the state-of-the-art source-free domain adaptation methods and report the comparison results.

Datasets We evaluate our method on the following benchmark datasets. **Office-31** [20] is a standard small-sized visual domain adaptation benchmark which contains images of 31 categories from three domains: Amazon (**A**), DSLR (**D**) and Webcam (**W**), each containing 2817, 498 and 795 images respectively. **VisDA-2017** [19] is a large-scale synthetic-to-real dataset with images in 12 categories from two domains, **Synthetic** and **Real**, each consists of 152,397 and 55,388 images respectively.

Implementation Details We use the same network architectures as the previous methods to achieve fair comparisons. On **Office-31**, we use ResNet-50 [9] as the backbone network. ResNet-101 [9] is utilized as the backbone module on **VisDA-2017**. The target model is optimized using the mini-batch SGD algorithm where batch size is set to be 32. The learning rate is fixed to be 10^{-4} for the backbone, and 10^{-3} for the bottleneck and FC layers. For trade-off parameters, we set $\lambda = 2, \lambda_1 = 1, \lambda_2 = 0.1, M = 4$ for the **Office-31** dataset and $\lambda = 100, \lambda_1 = 1, \lambda_2 = 0.01, M = 1$ for the **VisDA-2017** dataset.

Results on Office-31 Table 1 reports the results of different methods on the six domain adaptation tasks of this small sized dataset, where the top section includes results of the source-only and unsupervised domain adaptation (UDA) methods that use the source domain data and the bottom section presents the comparison results of the source-free unsupervised domain adaptation methods. We can see that our proposed ASL method outperforms most previous UDA methods that exploit the source data. Among the source-free UDA methods, our method achieves better performance than SHOT [16] and SDDA [12], and achieves similar performance as 3C-GAN [15], while 3C-GAN uses generative models to generate lots of labeled target data and is time-consuming and resource-extensive. In particular, our ASL method achieves the state-of-the-art performance on the first three

[1] Code is available at `https://github.com/cnyanhao/ASL`.

Table 2: Class-wise accuracy (%) on VisDA-2017 (ResNet-101)

Methods	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-class
ResNet-101 [9]	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
DANN [7]	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
MCD [21]	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
CDAN [17]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
CDAN+BSP [4]	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
SWD [13]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
CAN [10]	97.0	87.2	82.5	74.3	97.8	96.2	90.8	80.7	96.6	96.3	87.5	59.9	87.2
PPDA [11]	81.5	79.4	80.3	61.8	92.3	91.9	84.5	82.7	86.5	58.4	74.2	43.5	76.4
SHOT [16]	92.6	81.1	80.1	58.5	89.7	86.1	81.5	77.8	89.5	**84.9**	84.3	49.3	79.6
3C-GAN [15]	94.8	73.4	68.8	**74.8**	93.1	**95.4**	88.6	**84.7**	89.1	84.7	83.5	48.1	81.6
ASL (Ours)	**97.3**	**85.3**	**86.9**	70.7	**96.4**	72.8	**93.0**	80.1	**95.5**	78.1	**87.7**	**50.3**	**82.8**

Table 3: Ablation study: test accuracies (%) of different variants.

Methods	Office-31
Source Only	76.1
Naive Pseudo-Labeling (PL) [14]	76.7
Naive PL + $\mathcal{L}_{ent} + \mathcal{L}_{vat} + \mathcal{L}_{wr}$	83.3
Self-Labeling (SL)	83.8
SL + $\mathcal{L}_{ent} + \mathcal{L}_{vat} + \mathcal{L}_{wr}$	86.7
Augmented Self-Labeling (Asl)	88.0
Asl + $\mathcal{L}_{ent} + \mathcal{L}_{vat}$	88.4
ASL = Asl + $\mathcal{L}_{ent} + \mathcal{L}_{vat} + \mathcal{L}_{wr}$	89.5

tasks, i.e. A→D, A→W and D→A, under the source-free unsupervised domain adaptation setting. These results demonstrate the effectiveness of the proposed augmented self-labeling method.

Results on VisDA-2017 Table 2 reports the test accuracy results for different methods on this large scale benchmark. Again the top section reports the results of source-only and UDA methods and the bottom section reports the results of source-free UDA methods. We can see our ASL method achieves the state-of-the-art performance under the source-free unsupervised domain adaptation setting: it outperforms PPDA [11], SHOT [16] and 3C-GAN [15] in most classes and produces the best per-class average result. It also outperforms most previous unsupervised domain adaptation methods that exploit source data. This again validates the efficacy of the proposed ASL method.

Ablation Study As shown in Table 3, we investigated the contribution of different components of the proposed method and compared the variants with the naive pseudo-labeling (PL) method [14], which directly fine-tunes the source model with the pseudo-labeled target data. We can see that both self-labeling (SL) and augmented self-labeling (Asl) can easily outperform the naive PL method with or without the model regularization terms. The model regularization terms can improve the performance of all the methods: naive PL, SL, and Asl. Nevertheless, our proposed augmented self-labeling can promote the results by a large margin comparing with the self-labeling. This demonstrates the effectiveness of all the components of our overall ASL method for source-free unsupervised domain adaptation.

4 Conclusion

In this paper, we proposed a new Augmented Self-Labeling method for the source-free unsupervised domain adaptation, where only the source model and the unlabeled target data are available. We treated this problem as a joint optimization over the pseudo-labels and the prediction model, and solved it with two alternating steps, where augmented self-labeling improves the pseudo-labels and re-training improves the model with the self-labeled target data. We exploited data augmentation to improve the self-labeling quality by ensembling multiple prediction probability matrices corresponding to the augmented versions of samples. We also deployed model regularization terms to help model re-training. Experiments on the benchmarks validated the effectiveness of our proposed method for source-free unsupervised domain adaptation.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[2] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv:1902.01046*, 2019.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[4] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 2019.

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013.

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[8] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2005.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, 2019.

[11] Youngeun Kim, Donghyeon Cho, and Sungeun Hong. Towards privacy-preserving domain adaptation. *IEEE Signal Processing Letters*, 27:1675–1679, 2020.

[12] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *WACV*, 2021.

[13] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, 2019.

[14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

[15] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020.

[16] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

[17] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018.

[18] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[19] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017.

[20] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[21] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

[22] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T approach to unsupervised domain adaptation. In *ICLR*, 2018.

[23] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[24] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *NeurIPS*, 2019.

[25] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.