Active Adversarial Domain Adaptation

Jong-Chyi Su[1] Yi-Hsuan Tsai[2] Kihyuk Sohn[2] Buyu Liu[2]
Subhransu Maji[1] Manmohan Chandraker[2]

[1]University of Massachusetts, Amherst [2]NEC Laboratories America

1. Introduction

The *covariate shift* problem is common in many practical computer vision applications, where the training and test data are drawn from different distribution, *e.g.*, the seasonal distribution of natural species may change in a camera trap dataset. Many domain adaptation (DA) methods have been proposed to address this issue [3, 10, 19, 17, 11, 5, 18] by matching the marginal distributions of source and target domain. While domain adaptation provides a good starting point, the performance of unsupervised DA methods often fall far behind their supervised counterparts [16, 1]. In such cases, some labeled data from the target domain can bring in performance benefits. However, obtaining ground-truth annotations can be laborious and naïvely collecting annotated data could be inefficient. In this work, we aim to answer the following questions: 1) how to select data to label from the target domain effectively, and 2) how to perform adaptation given these labeled data from the target domain.

To this end, we propose an *Active Adversarial Domain Adaptation* (AADA) that exploits the relation between domain adaptation and active learning to answer those questions. Our approach explores a duality between two related problems: adversarial domain alignment and importance sampling for adapting models across domains. The former uses a domain discriminative model to align domains, while the latter utilizes it to weigh samples to account for distribution shifts. Specifically, our importance weight promotes samples with large uncertainty in classification and diversity from labeled examples, thus serves as a sample selection scheme for active learning. We show that these two views can be unified in one framework for domain adaptation and transfer learning when the source domain has many labeled examples while the target domain does not. AADA provides significant improvements over fine-tuning based approaches and other sampling methods when the two domains are closely related. The overall framework of our AADA is illustrated in Figure 1. We perform experiments on different domain adaptation tasks, including classification and object detection, and demonstrate that the advantage over baseline approaches is retained even after hundreds of examples being actively annotated.

2. Proposed Algorithm

Domain Adaptation. We adopt the domain adversarial neural network (DANN) [5], which is composed of three components: *feature extractor* G_f for the input x, *class predictor* G_y that predicts the class label $G_y(G_f(x)) \rightarrow \{1, ..., L\}$, and *discriminator* G_d that classifies the domain label $G_d(G_f(x)) \rightarrow \{0, 1\}$. We use 1 for the source domain and 0 for the target domain. The objective function of the discriminator G_d is defined as:

$$\begin{aligned}\mathcal{L}_d = \mathbb{E}_{x\sim p_{\mathcal{S}}(x)}\big[\log G_d(G_f(x))\big] \\ + \mathbb{E}_{x\sim p_{\mathcal{T}}(x)}\big[\log(1-G_d(G_f(x)))\big],\end{aligned} \tag{1}$$

where G_f, G_y, G_d are parameterized by $\theta_f, \theta_y, \theta_d$, respectively. To perform domain alignment, features generated from G_f should be able to fool the discriminator G_d, and hence we adopt an adversarial loss to form a min-max game:

$$\min_{\theta_f,\theta_y}\max_{\theta_d} \mathcal{L}_c(G_y(G_f(x)), y) + \lambda\mathcal{L}_d, \tag{2}$$

where $\mathcal{L}_c$ is the cross-entropy loss for classification, y is the class label, and λ is the weight between two losses.

Sample Selection. Given an unsupervised domain adaptation setting where labeled data is only available from the source domain, the goal of our sample selection is to find the most informative data from unlabeled target domain. We motivate the sample selection criteria from the idea of importance weighted empirical risk minimization (IWERM) [14], whose learning objective is defined as follows:

$$\min_{\theta_f,\theta_y} \mathbb{E}_{(x,y)\sim p_{\mathcal{S}}(x,y)}\Big[\frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)}\mathcal{L}_c\big(G_y(G_f(x)), y\big)\Big], \tag{3}$$

where $w(x) = \frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)}$ is an importance of each labeled data in the source domain. The formulation indicates which data is more important during optimization: 1) the data with higher empirical risk $\mathcal{L}_c\big(G_y(G_f(x)), y\big)$, and 2) the one with higher importance, *i.e.*, larger density in the target distribution $p_{\mathcal{T}}(x)$ but lower in the source $p_{\mathcal{S}}(x)$.

Unfortunately, applying this intuition to come up with a sample selection strategy is non-trivial. This is because the

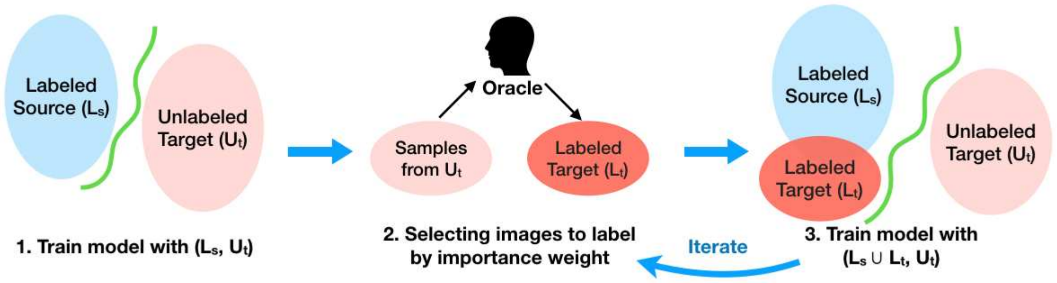


Figure 1: We start from unsupervised domain adaptation setting with labeled source L_s and unlabeled target U_t data and train the model with domain adversarial loss. In each following round, we first select samples using importance weight from unlabeled target domain to obtain annotations. We then re-train the model with labeled data $L_s \cup L_t$ and unlabeled data U_t.

Algorithm 1 AADA

```
Input: labeled source $L_s$; unlabeled target $U_t$;
       labeled target $L_t = \emptyset$; budget per round $b$
Model: $\mathcal{M}$=$\{G_f, G_y, G_d\}$; feature extractor $G_f$;
       class predictor $G_y$; discriminator $G_d$
Train $\mathcal{M}$ with $(L_s, U_t)$
for round $\leftarrow$ 1 to MaxRound do
    Compute $s(x)\ \forall x \in U_t$ via (5)
    Select a set of $b$ images $z$ from $U_t$ according to $s(z)$
    Get labels $y_z$ from oracle
    $L_t \leftarrow L_t \cup (z, y_z)$
    $U_t \leftarrow U_t \setminus (z, y_z)$
    Train $\mathcal{M}$ with $(L_s \cup L_t, U_t)$
```

target data is mostly unlabeled and the empirical risk cannot be computed before annotation. Another problem is that the importance estimation of high-dimensional data is difficult [15]. We take advantage of domain discriminator to resolve the second issue. Note that, with adversarial training, the optimal discriminator [7] is obtained at:

$$G_d^*(\hat{x}) = \frac{p_{\mathcal{S}}(x)}{p_{\mathcal{S}}(x) + p_{\mathcal{T}}(x)} \Rightarrow w(x) = \frac{1 - G_d^*(\hat{x})}{G_d^*(\hat{x})}, \qquad (4)$$

where $\hat{x} = G_f(x)$. Next, assuming cross-entropy as an empirical risk, we resolve the first issue by measuring the entropy of unlabeled data, which is a lower bound of the cross-entropy. Finally, our sample selection criterion $s(x)$ for unlabeled target data is written as follows:

$$s(x) = \frac{1 - G_d^*(G_f(x))}{G_d^*(G_f(x))} \mathcal{H}(G_y(G_f(x))). \qquad (5)$$

Two components in the measure are interpreted as follows: 1) *diversity* cue $(1 - G_d^*(G_f(x)))/G_d^*(G_f(x))$, and 2) *uncertainty* cue $\mathcal{H}(G_y(G_f(x)))$. The diversity cue allows us to select unlabeled target data which is less similar to the labeled ones in the source domain, while the uncertainty cue suggests data which the model cannot predict confidently. The overall algorithm is shown in Algorithm 1.

3. Experiments on Digit Classification

Our proposed method aims to address two questions: 1) how to select images to label from U_t to yield the most performance gain? and 2) how to train a classifier given $\{L_s, L_t, U_t\}$? We perform experiments comparing these two components in a mix-and-match way, on digit classification task from SVHN [12] to MNIST [9]. We explore the following training schemes: **1) Adversarial Training:** we train the classifier via (2) using $(L_s \cup L_t, U_t)$; **2) Joint Training:** we train the classifier in a supervised way using $L_s \cup L_t$; **3) Fine-tuning:** we train a classifier using L_s and then fine-tune it on L_t, both in a supervised way.

The sampling strategies we explored are: **1) Importance Weight:** we select samples based on the proposed importance weight $s(x)$ in (5); **2) K-means Clustering:** we perform k-means clustering on image features $G_f(x), \forall x \in U_t$, where the number of clusters is set to b in each round. The sample which is the closest to its center is selected in each cluster; **3) K-center (Core-set) [13]:** we use greedy k-center clustering to select b images z from U_t such that the largest distance between unlabeled data $U_t \setminus z$ and labeled data $L_t \cup z$ is minimized; **4) Diversity [4]:** for each unlabeled sample in U_t, we compute its distance to all samples in L_t and obtain the average distance. Then we rank unlabeled samples w.r.t. its average distance in descending order and select the top b samples. L2 distance is applied on features $G_f(x)$; **5) Best-versus-Second Best (BvSB) [8]:** we use the difference between the highest and the second highest class prediction as the uncertainty measure., *i.e.*, $\max_i G_{y_i}(\hat{x}) - G_{y_j}(\hat{x})$, where class j has the second highest prediction; **6) Random Selection:** we select samples uniformly at random from all the unlabeled target data U_t.

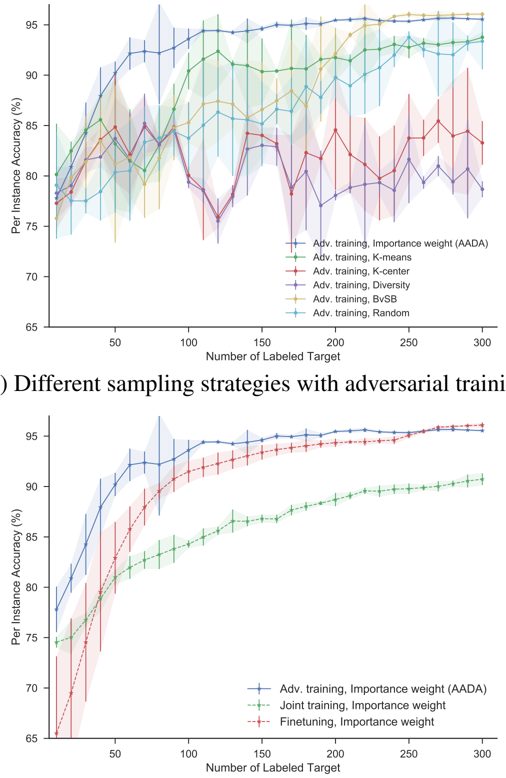
Our AADA uses importance weight for sample selection, and adversarial training as the training scheme. Note that, different sampling methods do not compete with AADA as it can be combined with our method. For example, BvSB can be used as an alternative uncertainty measurement as opposed to entropy in (5).

3.1. Comparison of Sampling Methods

As shown in Figure 2a, our importance weight performs favorably against other active sampling methods and can achieve 95% accuracy with 160 samples after 16 rounds, while the random selection baseline requires two times more annotations to achieve similar performance. Moreover, our proposed method consistently improves the performance when more samples are selected and annotated, whereas other baselines generate unstable performances.

3.2. Comparison of Training Schemes

We compare different training schemes and show the effectiveness of combining adversarial training with importance weight. As shown in Figure 2b, our AADA method demonstrates its effectiveness, especially when very few labeled targets L_t are available; on the other hand, when more and more labeled targets are available, fine-tuning becomes a competitive option as the benefit of leveraging information from source domain has decreased.



(a) Different sampling strategies with adversarial training.

(b) Different training schemes with importance weight.

Figure 2: Digit classification results (SVHN → MNIST). 10 images are selected to label in each round.

4. Experiments on Object Detection

Now we focus on object detection task adapting from KITTI [6] to Cityscapes [2]. We use the same setting as [1], which only considers the car object. We select $\{10, 10, 10, 20, 50, 100\}$ images in each round and assume that the cost of labelling one image is the same. We report our quantitative results in Table 1. Our baselines contain adversarial training with other sampling methods and different training schemes with random sampling. Note that BvSB is not included here since in the single object category detection scenario, it provides the same measurement as entropy. Overall, using adversarial training and importance weight (AADA) yields the best performance. Specifically, 60.4% accuracy can be achieved with 100 labeled target selected by AADA, while other baselines require about twice as much annotations to achieve similar performance. We further illustrate images selected with AADA within two rounds in Figure 3. Images selected in the third round have more cars and the semantic layouts are different w.r.t. that of the fourth round, showing that our method is able to select diverse samples.

Training	Sampling	Number of Labeled Target					
		10	20	30	50	100	200
Adversarial	Imp. weight	**49.4**	**53.3**	**54.6**	**57.4**	**60.4**	**62.3**
Adversarial	K-means	49.1	51.7	53.8	56.8	59.2	60.9
Adversarial	Entropy	48.9	50.9	52.3	54.3	58.1	61.0
Adversarial	Random	47.4	49.8	51.6	55.2	58.6	61.7
Joint	Imp. weight	48.5	52.1	53.5	56.2	58.6	60.5
Joint	Random	45.5	48.8	51.8	54.9	59.0	61.6
Fine-tuning	Random	41.0	46.0	48.7	51.4	56.0	59.8

Table 1: Object detection results (KITTI → Cityscapes). Our AADA method (first row) performs the best.

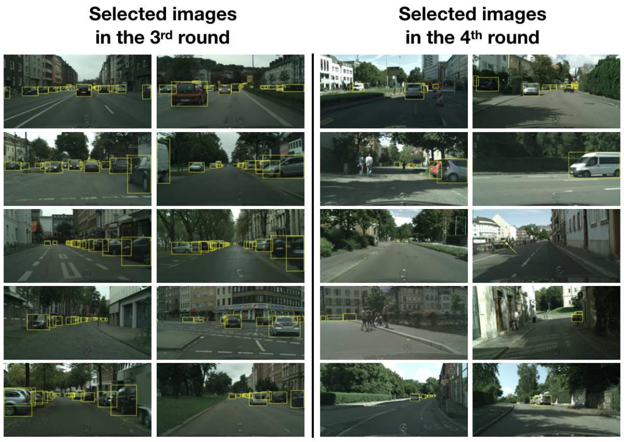

Selected images in the 3rd round
Selected images in the 4th round

Figure 3: Top 10 images selected in the third and the fourth rounds from the target domain (Cityscapes) using AADA.

5. Conclusion

We propose AADA, a unified framework for domain adaptation and active learning via adversarial training. When few labeled target are available, the domain adversarial model helps improve the classification; meanwhile, the discriminator can be utilized to obtain the importance weight for active sample selection in the target domain.

References

[1] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018. 1, 3

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[3] H. Daumé III, A. Kumar, and A. Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59. Association for Computational Linguistics, 2010. 1

[4] S. Dutt Jain and K. Grauman. Active image segmentation propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2864–2873, 2016. 2

[5] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1

[6] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[8] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. 2009. 2

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[10] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013. 1

[11] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016. 1

[12] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011. 2

[13] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. 2018. 2

[14] M. Sugiyama, M. Krauledat, and K.-R. MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007. 1

[15] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008. 2

[16] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[17] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 1

[18] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017. 1

[19] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1