A DIRT-T Approach to Unsupervised Domain Adaptation

Rui Shu
Stanford University

Hung Bui
Adobe Research

Stefano Ermon
Stanford University

Abstract

Domain adaptation refers to the problem of how to leverage labels in one source domain to boost up learning performance in a new target domain where labels are scarcely available or completely unavailable. In this paper, we address these issues through the lens of the cluster assumption, i.e., decision boundaries should not cross high-density data regions. We propose two novel and related models: (1) the Virtual Adversarial Domain Adaptation (VADA) model, which combines domain adversarial training with a penalty term that punishes the violation of the cluster assumption; (2) the Decision-boundary Iterative Refinement Training with a Teacher (DIRT-T) model, which takes the VADA model as initialization and employs natural gradient steps to further minimize the cluster assumption violation. Extensive empirical results demonstrate that the combination of these two models significantly improve the state-of-the-art performance on several visual domain adaptation benchmarks.

1 Introduction

In many tasks, direct access to vast quantities of labeled data to the task of interest (the target domain) is either costly or otherwise absent, but labels are readily available for related training sets (the source domain). However, the source data distribution is often dissimilar to the target data distribution, and the resulting significant covariate shift is often detrimental to the performance of the source-trained model when applied to the target domain [1]. Solving the covariate shift problem of this nature is commonly referred to as domain adaptation. In this paper, we consider a challenging setting of domain adaptation where 1) we are provided with fully-labeled source samples and completely-unlabeled target samples, and 2) the existence of a classifier in the hypothesis class with low error on both source and target distributions is not guaranteed. Borrowing approximately the terminology from [2], we refer to this setting as unsupervised, *non-conservative* domain adaptation.

To tackle unsupervised domain adaptation, [3] proposed to constrain the classifier to only rely on domain-invariant features. This is achieved by training the classifier to perform well on the source domain while minimizing the divergence between features extracted from the source versus target domains. To achieve divergence minimization, [3] employ domain adversarial training. However, a fundamental weakness of domain adversarial training is that it does not account for the case where good generalization on the source domain hurts target performance in the non-conservative setting.

[4] addressed these issues by replacing domain adversarial training with asymmetric tri-training (ATT), which relies on the assumption that target samples that are labeled by a source-trained classifier with high confidence *are* correctly labeled by the source classifier. In this paper, we consider an orthogonal assumption: the cluster assumption [5], that covariate distribution contains separated data clusters and that data samples in the same cluster share the same class label. This assumption introduces an additional bias where we seek decision boundaries that do not go through high-density regions. Based on this intuition, we propose two novel models: (1) the Virtual Adversarial Domain Adaptation (VADA) model which incorporates an additional virtual adversarial training [6] and

Workshop on Learning with Limited Labeled Data (NIPS 2017), Long Beach, CA, USA.

conditional entropy loss to push the decision boundaries away from the empirical data, and (2) the Decision-boundary Iterative Refinement Training with a Teacher (DIRT-T) model which uses natural gradient to further refine the output of the VADA model while focusing purely on the target domain. We demonstrate that

1. In conservative domain adaptation, where the classifier is trained to perform well on the source domain, VADA can be used to further constrain the hypothesis space by penalizing violations of the clustering assumption, thereby improving domain adversarial training.
2. In non-conservative domain adaptation, where we account for the mismatch between the source and target optimal classifiers, DIRT-T allows us to transition from a good joint (source and target) classifier (VADA) to a better target domain classifier. Interestingly, we demonstrate the advantage of natural gradients in DIRT-T refinement steps.

We report results for domain adaptation in digits classification (MNIST-M, MNIST, SYN DIGITS, SVHN), traffic sign classification (SYN SIGNS, GTSRB), and general object classification (STL-10, CIFAR-10). We show that, in nearly all experiments, VADA improves upon previous methods and that DIRT-T improves upon VADA, setting new state-of-the-art performance across a wide range of visual domain adaptation benchmarks. In adapting MNIST $\rightarrow$ SVHN, a very challenging task, we out-perform ATT by over 20%.

2 Related Work

Given the extensive literature on domain adaptation, we highlight the works most relevant to our paper. [3] proposed to project both source and target distributions into some feature space and encourage distribution matching in the feature space. To better perform non-conservative domain adaptation, [4] proposed to modify tri-training [7] for domain adaptation, leveraging the assumption that highly-confident predictions are correct predictions [8]. Both methods are based on [9]'s theoretical analysis of domain adaptation, which states the following,

Theorem 1 *[9] Let $\mathcal{H}$ be the hypothesis space and let (X_s, ϵ_s) and (X_t, ϵ_t) be the two domains and their corresponding generalization error functions. Then for any $h \in \mathcal{H}$,*

$$\epsilon_t(h) \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \epsilon_s(h) + \min_{h' \in \mathcal{H}} \epsilon_t(h') + \epsilon_s(h'), \tag{1}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ denotes the $\mathcal{H}\Delta\mathcal{H}$-distance between the domains X_s and X_t,

$$d_{\mathcal{H}\Delta\mathcal{H}} = 2 \sup_{h,h' \in \mathcal{H}} \left| \mathbb{E}_{x \sim \mathcal{D}_s} \left[h(x) \neq h'(x) \right] - \mathbb{E}_{x \sim \mathcal{D}_t} \left[h(x) \neq h'(x) \right] \right|. \tag{2}$$

Intuitively, $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the extent to which small changes to the hypothesis in the source domain can lead to large changes in the target domain. It is evident that $d_{\mathcal{H}\Delta\mathcal{H}}$ relates intimately to the complexity of the hypothesis space and the divergence between the source and target domains. For disjoint domains and infinite-capacity models, $d_{\mathcal{H}\Delta\mathcal{H}}$ is maximal.

3 Constraining via Conditional Entropy Minimization

In this paper, we apply the cluster assumption to domain adaptation. The cluster assumption assumes that the input distribution X contains density clusters and that points in the same cluster come from the same class. This assumption has been extensively studied and applied successfully to a wide range of classification tasks [5, 6, 10, 11, 12, 13, 14]. If the cluster assumption holds, the optimal decision boundaries should occur far away from data-dense regions in the space of $\mathcal{X}$ [5]. Following [10], we achieve this behavior via minimization of the conditional entropy with respect to the target distribution,

$$\mathcal{L}_c(\theta; \mathcal{D}_t) = -\mathbb{E}_{x \sim \mathcal{D}_t} \left[h_\theta(x)^\top \ln h_\theta(x) \right], \tag{3}$$

where $\mathcal{D}_t$ is the target domain data and h maps to the K-simplex. Intuitively, minimizing the conditional entropy forces the classifier to be confident on the unlabeled target data, which occurs if the classifier places its decision boundaries far from data-dense regions. The conditional entropy

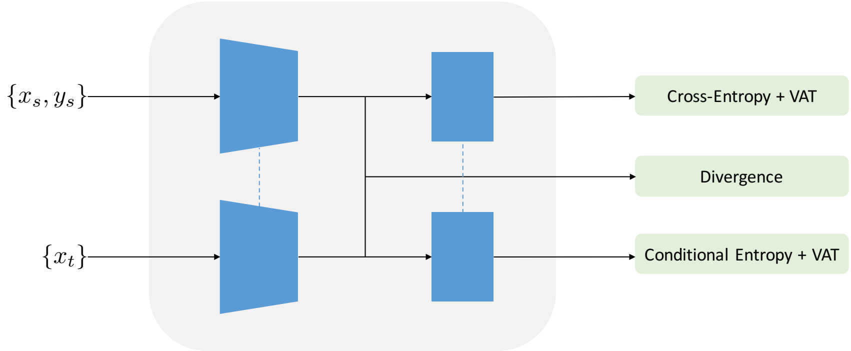

Figure 1: VADA improves upon domain adversarial training by additionally penalizing violations of the cluster assumption.

must be empirically estimated using the available data. However, [10] notes that this approximation breaks down if the classifier h is not locally-Lipschitz. To prevent this, we propose to incorporate the locally-Lipschitz constraint via virtual adversarial training [6] and add to the objective function the additional term

$$\mathcal{L}_v(\theta; \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}} \left[\max_{\|r\| \le \epsilon} \mathrm{D}_{\mathrm{KL}}(h_{\hat{\theta}}(x) \| h_\theta(x + r)) \right], \tag{4}$$

which enforces classifier consistency around the norm-ball neighborhood of each sample x, where $\hat{\theta}$ is a copy of θ. Note that virtual adversarial training can be applied with respect to either the target or source distributions. We can combine the conditional entropy minimization objective and domain adversarial training to yield

$$\min_\theta . \mathcal{L}_y(\theta; \mathcal{D}_s) + \lambda_d \mathcal{L}_d(\theta; \mathcal{D}_s, \mathcal{D}_t) + \lambda_s \mathcal{L}_v(\theta; \mathcal{D}_s) + \lambda_t \left[\mathcal{L}_v(\theta; \mathcal{D}_t) + \mathcal{L}_c(\theta; D_t)\right], \tag{5}$$

a basic combination of cross-entropy $\mathcal{L}_y$, domain adversarial $\mathcal{L}_d$ [3], and semi-supervised $(\mathcal{L}_v, \mathcal{L}_c)$ objectives. We refer to this as the Virtual Adversarial Domain Adaptation (VADA) model.

$\mathcal{H}\Delta\mathcal{H}$-Distance Minimization. VADA aligns well with the theory of domain adaptation provided in Theorem 1. Let the loss,

$$\mathcal{L}_t(\theta) = \mathcal{L}_v(\theta; \mathcal{D}_t) + \mathcal{L}_c(\theta; D_t), \tag{6}$$

be a proxy measure for the degree of violation of the target-side cluster assumption. For a reasonably small choice of λ_t, VADA can penalize models with high $\mathcal{L}_t$ while still enabling decently small source generalization error. This penalization effectively rejects hypotheses which egregiously violate the target-side cluster assumption. By rejecting such hypotheses from the hypothesis space $\mathcal{H}$, VADA reduces $d_{\mathcal{H}\Delta\mathcal{H}}$ and yields a tighter bound on the target generalization error. We verify empirically that VADA achieves significant improvements over existing models on multiple domain adaptation benchmarks (Table 1).

4 Decision-boundary Iterative Refinement Training

In non-conservative domain adaptation, we account for the following inequality,

$$\min_{h \in \mathcal{H}} \epsilon_t(h) < \epsilon_t(h^a) \text{ where } h^a = \arg\min_{h \in \mathcal{H}} \epsilon_s(h) + \epsilon_t(h), \tag{7}$$

where (ϵ_s, ϵ_t) are generalization error functions for the source and target domains. This means that, for a given hypothesis class $\mathcal{H}$, the optimal classifier in the source domain does not coincide with the optimal classifier in the target domain.

We assume that the optimality gap in Eq. (7) results from violation of the cluster assumption. In other words, we suppose that any source-optimal classifier drawn from our hypothesis space *necessarily* violates the cluster assumption in the target domain. Since VADA is still constrained to do well on the source domain (ensured by choosing a small λ_t), it will still violate the target-side cluster assumption to some extent.

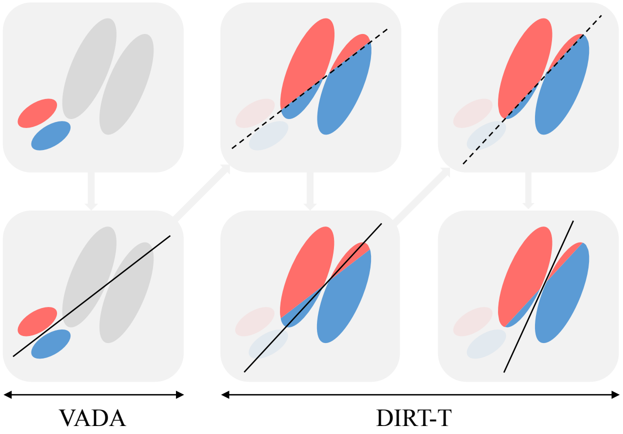

Figure 2: DIRT-T uses VADA as initialization. After removing the source training signal, DIRT-T minimizes cluster assumption violation in the target domain through a series of natural gradient steps.

Under this assumption, the natural solution is to initialize with the VADA model and then further minimize the cluster assumption violation in the target domain. In particular, we first use VADA to learn an initial classifier h_{θ_0}. Next, we incrementally push the classifier's decision boundary away from data-dense regions by minimizing the proxy target-side cluster assumption violation loss $\mathcal{L}_t$ in Eq. (6). We denote this procedure Decision-boundary Iterative Refinement Training (DIRT).

4.1 Decision-boundary Iterative Refinement Training with a Teacher

Stochastic gradient descent minimizes the proxy loss $\mathcal{L}_t$ by selecting gradient steps $\Delta\theta$ according to the following objective,

$$\min_{\Delta\theta}. \ \mathcal{L}_t(\theta + \Delta\theta) \tag{8}$$

$$\text{s.t. } \|\Delta\theta\| \leq \epsilon, \tag{9}$$

which defines the neighborhood in the parameter space. This notion of neighborhood is sensitive to the parameterization of the model; depending on the parameterization, a seemingly small step $\Delta\theta$ may result in a vastly different classifier. This contradicts our intention of incrementally and locally pushing the decision boundary to a local conditional entropy minimum, which requires that the decision boundary of $h_{\theta+\Delta\theta}$ stay close to that of h_θ. It is therefore important to define a neighborhood that is parameterization-invariant. Following [15], we instead select $\Delta\theta$ using the following objective,

$$\begin{aligned} &\min_{\Delta\theta}. \ \mathcal{L}_t(\theta + \Delta\theta) \\ &\text{s.t. } \mathbb{E}_{x\sim D_t}\left[\mathrm{D}_{\mathrm{KL}}(h_\theta(x) \| h_{\theta+\Delta\theta}(x))\right] \leq \epsilon. \end{aligned} \tag{10}$$

Each optimization step now solves for a gradient step $\Delta\theta$ that minimizes the conditional entropy, subject to the constraint that the Kullback-Leibler divergence between $h_\theta(x)$ and $h_{\theta+\Delta\theta}(x)$ is small for $x \sim \mathcal{X}_t$. The corresponding Lagrangian suggests that one can instead minimize a sequence of optimization problems

$$\min_{\theta_n}. \ \lambda_t \mathcal{L}_t(\theta_n) + \beta_t \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(h_{\theta_{n-1}}(x) \| h_{\theta_n}(x))\right], \tag{11}$$

that approximates the application of a series of natural gradient steps.

In practice, each of optimization problems in Eq. (11) can be solved approximately via a finite number of stochastic gradient descent steps. We denote the number of steps taken to be the refinement interval B. Similar to [14], we use the Adam Optimizer with Polyak averaging [16]. We interpret $h_{\theta_{n-1}}$ as a (sub-optimal) teacher for the student model h_{θ_n}, which is trained to stay close to the teacher model while seeking to reduce the cluster assumption violation. As a result, we denote this model as Decision-boundary Iterative Refinement Training with a Teacher (DIRT-T).

Weakly-Supervised Learning. This sequence of optimization problems has a natural interpretation that exposes a connection to weakly-supervised learning. In each optimization problem, the teacher model $h_{\theta_{n-1}}$ pseudo-labels the target samples with noisy labels. Rather than naively training the student model h_{θ_n} on the noisy labels, the additional training signal $\mathcal{L}_t$ allows the student model to place its decision boundaries further from the data. If the clustering assumption holds and the initial noisy labels are sufficiently similar to the true labels, conditional entropy minimization can improve the placement of the decision boundaries [17].

Domain Adaptation. An alternative interpretation is that DIRT-T is the *recursive* extension of VADA, where the act of pseudo-labeling of the target distribution constructs a new "source" domain (i.e. target distribution X_t with pseudo-labels). The sequence of optimization problems can then be seen as a sequence of non-conservative domain adaptation problems in which $X_s = X_t$ but $p_s(y \mid x) \neq p_t(y \mid x)$, where $p_s(y \mid x) = h_{\theta_{n-1}}(x)$ and $p_t(y \mid x)$ is the true conditional label distribution in the target domain. Since $d_{\mathcal{H}\Delta\mathcal{H}}$ is strictly zero in this sequence of optimization problems, domain adversarial training is no longer necessary. Furthermore, if $\mathcal{L}_t$ minimization does improve the student classifier, then the gap in Eq. (7) should get smaller each time the source domain is updated.

4.2 Model Evaluation and Conclusion

Source Target	MNIST MNIST-M	SVHN MNIST	MNIST SVHN	DIGITS SVHN	SIGNS GTSRB	CIFAR STL	STL CIFAR
MMD [18]	76.9	71.1	-	88.0	91.1	-	-
DANN [3]	81.5	71.1	35.7	90.3	88.7	-	-
DRCN [19]	-	82.0	40.1	-	-	66.4	58.7
DSN [20]	83.2	82.7	-	91.2	93.1	-	-
kNN-Ad [21]	86.7	78.8	40.3	-	-	-	-
ATT [4]	94.2	86.2	52.8	92.9	96.2	-	-
Π-model (aug) [22]	-	92.0	71.4	94.2	98.4	76.3	64.2
Without Instance-Normalized Input:							
Source-Only	58.5	77.0	27.9	86.9	79.6	76.3	63.6
VADA	97.7	97.9	47.5	94.8	98.8	**80.0**	73.5
DIRT-T	**98.9**	**99.4**	**54.5**	**96.1**	**99.5**	-	**75.3**
With Instance-Normalized Input:							
Source-Only	59.9	82.4	40.9	88.6	86.2	77.0	62.6
VADA	95.7	94.5	73.3	94.9	99.2	**78.3**	71.4
DIRT-T	**98.7**	**99.4**	**76.5**	**96.2**	**99.6**	-	**73.3**

Table 1: Accuracy. In all settings, both VADA and DIRT-T achieve state-of-the-art performance in all settings. DIRT-T omitted for CIFAR → STL since STL only has 5000 images in the training set.

Source Target	MNIST MNIST-M	SVHN MNIST	MNIST SVHN	DIGITS SVHN	SIGNS GTSRB	CIFAR STL	STL CIFAR
ATT	37.1	16.1	17.9	9.0	**20.5**	-	-
Π-model (aug)	-	3.7	18.1	**10.6**	1.0	**4.5**	7.4
DIRT-T	**40.4**	**22.4**	26.6	9.2	19.9	-	**11.7**
DIRT-T (W.I.N.I.)	38.8	17.0	**35.6**	7.6	13.4	-	10.7

Table 2: Additional comparison of the margin of improvement computed by taking the reported performance of each model and subtracting the reported source-only performance in the respective papers. W.I.N.I. indicates "with instance-normalized input."

We achieve state-of-the-art results across all tasks. Our experiments demonstrate that VADA achieves strong performance across several visual domain adaptation benchmarks, and DIRT-T further improves VADA performance. In four of the tasks (MNIST → MNIST-M, SVHN → MNIST, MNIST → SVHN, STL → CIFAR), we achieve substantial margin of improvement compared to previous models. In the remaining three tasks, our improvement margin over the source-only model is competitive against previous models. Our closest competitor is the Π-model. However, unlike the Π-model, we do not perform data augmentation. Our proposed models open up several possibilities for future work. One possibility is to apply DIRT-T to weakly supervised learning; another is to improve the natural gradient approximation via K-FAC [23] and PPO [24]. Given the strong performance of our models, we also recommend them for other downstream domain adaptation applications.

References

[1] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[2] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[3] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[4] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017.

[5] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *AISTATS*, pages 57–64, 2005.

[6] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017.

[7] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[8] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.

[9] 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.

[10] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[11] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

[12] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan Salakhutdinov. Good semi-supervised learning that requires a bad gan. *arXiv preprint arXiv:1705.09783*, 2017.

[13] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[14] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. 2017.

[15] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013.

[16] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[17] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[19] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[20] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016.

[21] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016.

[22] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

[23] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pages 2408–2417, 2015.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.