Adversarial target-invariant representation learning

Anonymous*

1 Introduction and Motivation

One of the most common assumptions in machine learning is that all examples used for training and testing are independently and identically drawn from the same distribution (1). However, in practical scenarios, this assumption can not be taken for granted as there might exist, for example, variations between the conditions in which training and test data were collected (2). Previous work attempted to enable machine learning models to compensate for the mismatch between training and test data distributions under different assumptions. Particularly, in the domain adaptation setting, one assumes that at training time an unlabeled sample from the test data distribution is available. Theoretical results for this setting (3) have shown that under the covariate shift assumption, the error on the target domain is bounded by the error on the source domain and the $\mathcal{H}$-divergence (4) between source and target distributions measured on a shared feature space. Methods based on this result attempt to learn domain-invariant representations while preserving task-relevant cues (5). Recent work (6; 7) leveraged deep neural networks effectiveness on learning useful representations so as to learn a representation space where both task error and the divergence between source and target domains are simultaneously minimized.

Despite the success presented by domain adaptation strategies in several tasks (8; 9), such techniques rely on the assumption that an unlabeled sample from a target distribution of interest is known at training time. In this work, we are interested in a more general setting where no samples from any target distribution are available for training a model. This problem is known as domain generalization and has recently drawn attention from the machine learning community. Previous work on domain generalization proposed to use data augmentation (10) at training time, meta-learning to simulate domain shift (11), adding a self-supervised task to encourage a feature extractor to learn better representations (12), or learning domain-invariant representations (2). In this work, we tackle domain generalization problems through building upon the domain adaptation results and previously proposed methods, often based on $\mathcal{H}$-divergence minimization. We show that, by minimizing pair-wise divergences across a set of training source domains, the feature extractor is encouraged to learn representations which are invariant across unseen target domains, under the assumption that samples from any target domain can be drawn from a mixture of all source domains. We show that minimizing an upper-bound on the pair-wise $\mathcal{H}$-divergence between source domains encourages the features to be invariant to any target domain. In summary, our main contributions are: i) we revisit and extend theoretical results from the domain adaptation literature to the domain generalization problem; ii) we devise an algorithm based on these results to learn domain-invariant representations and evaluate it on two domain generalization benchmarks.

2 Contribution and Experiments

Let $\mathfrak{D}$ be a meta-distribution composed by distributions $\mathcal{D}$ denominated domains, i.e. while sampling from $\mathfrak{D}$, one is actually sampling from one of the possible $\mathcal{D}_i \in \mathfrak{D}$. Further consider a training set $(x_m, y_m) \sim \mathfrak{D}$ and a test set $(x'_m, y'_m) \sim \mathfrak{D}$ are constructed, and the set of domains that represented in the training and test sets are referred to as source and target domains, respectively.

*Presenting author identifies as Latinx

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Now let $d_{\mathcal{H}}(\mathcal{D}_i^S, \mathcal{D}_k^S)$ denote the $\mathcal{H}$-divergence between the i-th and k-th source domains $\mathcal{D}_i^S$ and $\mathcal{D}_k^S$, and $d_{\mathcal{H}}(\mathcal{D}_i^S, \mathcal{D}_k^S) \leq \epsilon,\ \forall\ i, k \in \{1, \ldots, N_S\}$, where N_S is the number of source domains, be the maximum pair-wise $\mathcal{H}$-divergence between all source domains. We thus assume that all examples that appear at test time can be explained by a mixture of the source domains, i.e. any target domain $\mathcal{D}_j^T$ can be written as $\mathcal{D}_j^T(\cdot) = \sum_{i=1}^{N_S} \pi_{i,j} \mathcal{D}_i^S(\cdot)$, $\sum_{i=1}^{N_S} \pi_{i,j} = 1$. Consequently $d_{\mathcal{H}}(\mathcal{D}_j^T, \mathcal{D}_i^S) \leq \epsilon,\ \forall\ i \in \{1, \ldots, N_S\}$. Using the triangle inequality for the $\mathcal{H}$-distance, we have that for any target domains $\mathcal{D}_k^T$ and $\mathcal{D}_j^T$, $d_{\mathcal{H}}(\mathcal{D}_k^T, \mathcal{D}_j^T) \leq d_{\mathcal{H}}(\mathcal{D}_k^T, \mathcal{D}_i^S) + d_{\mathcal{H}}(\mathcal{D}_i^S, \mathcal{D}_j^T) = \epsilon + \epsilon = 2\epsilon$. Therefore, given the aforementioned assumptions, by minimizing the maximum pair-wise $\mathcal{H}$-divergence between source domains, we are also minimizing the $\mathcal{H}$-divergence between target domains. Taking such result into account, we propose a method to learn representations by minimizing an empirical estimation of ϵ, while attaining a good performance at the goal task. Our algorithm contains three main modules: a feature extractor F with parameters ϕ, a task classifier T with parameters θ_T, and the $\mathcal{H}$-divergence estimators D_j with parameters θ_j, $j \in \{1, \ldots, N_S\}$.

As shown in (3), the $\mathcal{H}$-divergence between two domains can be estimated by a discriminator responsible for distinguishing samples from one domain and the other. In practice, this could be implemented in different ways. For example, one could have an empirical $\mathcal{H}$-divergence estimator for each pair of source domains and train the feature extractor to minimize the maximum values between the estimates. However, if adopting this procedure, the number of $\mathcal{H}$-divergence estimators is $\mathcal{O}(n^2)$, where n is equal to N_S. Another possibility is to have instead a model responsible to discriminate examples from one source domain from all the others. This way, each discriminator is estimating an average of the empirical $\mathcal{H}$-divergence between one source domain and the remaining ones. As a result, we have one domain discriminator per source domain, which corresponds to a $\mathcal{O}(n)$ number of $\mathcal{H}$-divergence estimators. To avoid increasing the computational cost of our approach, we decided to estimate ϵ using the second approach. Hence, the procedure for estimating ϕ, θ_T, and all θ_j's can be formulated as the following multiplayer minimax game:

$$\min_{\phi, \theta_T} \max_{\theta_1, \ldots, \theta_{N_S}} \mathcal{L}_T(T(F(x; \phi); \theta_T), y_T) - \sum_{j=1}^{N_S} \mathcal{L}_j(D_j(F(x; \phi); \theta_j), y_j), \tag{1}$$

where $\mathcal{L}_T(\cdot)$ is the task-related loss, and each $\mathcal{L}_j(\cdot)$ represents the binary cross-entropy loss for one-versus-all domain classification, y_T corresponds to the task label, and y_j is equal to 1 in case x belongs to the j-th source domain, or 0 otherwise. Intuitively, the feature extractor attempts to minimize the task-specific module loss and minimize the estimated $\mathcal{H}$-divergence values by maximizing the losses provided by the domain discriminators, while each domain discriminator aims at improving its estimation of the empirical $\mathcal{H}$-divergence. We perform alternate updates on ϕ, θ_T and all θ_i's.

To assess the quality of the learned representations using our proposed approach we perform experiments on the PACS and VLCS domain generalization benchmarks and compare its performance with two state-of-the-art methods, namely Epi-FCR (13) and JiGen (12). We also consider for comparison two adversarial approaches CIDDG (11) and MMD-AAE (14). In addition, we report the performance of a convolutional neural network trained using all source domains without any mechanism to enforce domain generalization. This model is referred in the results as All Sources (AS). Each model was run with three different initializations on a leave-one-domain-out validation. The average accuracy on the test partition of the target domain is reported.

The PACS benchmark consists of images from 7 classes from four different source domains (Photo, Art painting, Cartoon, and Sketch). In Table 1 we show the results and observe that our proposed method achieves better average performance across all source domains than the compared methods. The VLCS benchmark is composed by 5 overlapping classes from the VOC2007 (15), LabelMe (16), Caltech-101 (17), and SUN (18) datasets. By observing the results presented in Table 2, we observe that our approach outperformed the compared methods in almost all cases, and, overall, it obtained better average accuracy across all domains.

Table 1: Results on the PACS benchmark.

Target	CIDDG	Epi-FCR	JiGen	AS	Proposed
P	78.65	86.10	89.00	90.02	88.12
A	62.70	64.70	67.63	64.86	66.60
C	69.73	72.30	71.71	70.18	73.36
S	64.45	65.00	65.18	61.40	68.03
Average	68.88	72.00	73.38	71.61	74.02

Table 2: Results on the VLCS benchmark.

Target	MMD-AAE	Epi-FCR	JiGen	AS	Proposed
V	67.70	67.10	70.62	73.44	71.14
L	62.60	64.30	60.90	60.44	67.63
C	94.40	94.10	96.93	97.88	95.52
S	64.40	65.90	64.30	67.92	69.37
Average	72.28	72.90	73.19	74.92	75.92

References

[1] T. Darrell, M. Kloft, M. Pontil, G. Rätsch, and E. Rodner, "Machine learning with interdependent and non-identically distributed data (dagstuhl seminar 15152)," in *Dagstuhl Reports*, vol. 5, no. 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

[2] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 624–639.

[3] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems*, 2007, pp. 137–144.

[4] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 180–191.

[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*, vol. 17, no. 1, pp. 2096–2030, 2016.

[6] R. Shu, H. H. Bui, H. Narui, and S. Ermon, "A dirt-t approach to unsupervised domain adaptation," in *Proc. 6th International Conference on Learning Representations*, 2018.

[7] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.

[8] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, p. 18, 2012.

[9] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant representations for noisy speech recognition," *arXiv preprint arXiv:1612.01928*, 2016.

[10] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Advances in Neural Information Processing Systems*, 2018, pp. 5334–5344.

[11] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[13] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," *arXiv preprint arXiv:1902.00113*, 2019.

[14] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[16] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[17] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[18] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 129–136.