Multi-domain Invariance by Embedding Alignment via Pseudo-Labels and Class Distances

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep neural networks often struggle when the domain changes between training and testing data. We consider the problem of learning a model on two labelled training domains and a third target domain for which we have no labels. To improve domain invariance, we propose to align the model's feature embeddings across the three domains for each class, where we use pseudo-labels on the unlabelled domain. For alignment, we minimize the element-wise distances between the feature embeddings of respective classes. We validate our method in a setup where we do not have access to the target domain (domain generalization). We contrast this to having access to the target domain's images without their labels (unsupervised domain adaptation). We find that the alignment increases domain invariance, especially when combined with pseudo-labels to align all three domains, thereby increasing the performance on the third target domain significantly.

1 Introduction

Deep neural networks often struggle when the domain changes between training and testing data. Most image classification networks operate on the assumption that training data and target data follow the same distribution [4]. In practice, this is a detrimental simplifying assumption. For instance, when detecting cows on sandy beaches instead of in the training data sets on grassy mountains, performance dropped heavily [3].

We consider the problem of learning a model on two labelled training domains and a third target domain for which we have no labels. This problem setting is often encountered in practice. After applying a model trained on a first domain, it is being applied in a second domain, where it does not work as well. As a consequence, several images of the second domain are labelled, after which it performs well on both domains. Next, the model needs to be applied to a third domain. We want to avoid labelling in this new third domain. To that end, we explore a multi-domain setup to improve domain invariance on a third unlabelled domain.

To improve domain invariance, we want to learn a model to generate similar feature embeddings for a particular class across the three domains. For this purpose, we focus on aligning the feature representations [9, 13, 23]. Following [7], we propose to minimize the element-wise distances between the feature embeddings of respective classes. To align all three domains, we need a class label per image for each domain. For the third domain, we have no labels. To solve this, we generate pseudo-labels. The novelty of our approach is the multi-domain setup combined with the pseudo-labels for the unlabelled target domain.

We validate the merit of pseudo-labels for three-domain alignment. We test our method in a setup where we do not have access to the third target domain (domain generalization). We assess the merit of aligning features versus no alignment, i.e., standard training on two domains and testing on the

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

third. For the alignment, we consider various distance measures. We show that alignment is helpful: performance on the third target domain improves. We contrast this to having access to the target domain's images without their labels (unsupervised domain adaptation). We find that the alignment increases especially when combined with pseudo-labels to align all three domains. We show that the feature embeddings across different domains have become more similar for the respective classes. More importantly, the performance on the third target domain has increased significantly. Together this indicates that our method improves multi-domain invariance.

2 Related Work

We take inspiration from two fields: causality and information theory. From the first, we learn the importance of features that are likely to be domain invariant. A classification head is guaranteed to make a domain invariant prediction if and only if all the features used in this prediction are the full set of domain invariant features [28]. From the second, we learn how we can generate these features that are likely to be domain invariant. Adding a domain invariant regularizer to the loss function is an effective method to promote domain invariance. We use such a regularizer to train the network to generate domain invariant features.

Causality. Using the invariant features for causal discovery and inference was first proposed by [11, 17]. They introduce *invariant causal prediction*, in which they state the importance of using invariant features for invariant prediction. Much research followed up on the invariant causal prediction work [14, 18, 21, 22]. One approach is to add a regularizer to the loss function [20]. They use a domain label as one of their target labels. In our case, the features are not given a-priori and are high-dimensional, so we must learn to approximate invariant features. The introduction of *invariant risk minimisation* [2], building on this causal inference framework, states that equivalent to learning features whose correlation with the target variables remain stable is the use of a domain invariant data representation and classifier which is optimal for all domains. There is discussion about the merit of invariant risk minimisation relative to *empirical risk minimisation* [1, 19, 24]. This trade-off between empirical and invariant risk minimisation represents the trade-off between distinctive ability and domain invariance. We use emperical risk minimisation as the baseline and monitor the influence of invariant risk minimisation with a scalar hyperparameter.

Information theory. One of the top performers on a benchmark data set called DomainNet [16] uses the mutual information between the trainable domains and an oracle domain to increase domain invariance [6]. This oracle is entirely domain generalised and optimised for any possible domain. They substitute the oracle with an ImageNet pre-trained ResNet50, which they assume is generalised enough and can approximate such an oracle. Achieving complete domain invariance in a dataset or network is often a trade-off with the domain-specific information. By regularizing the loss function with the mutual information between the model and their oracle, they ensure that the model weights favour domain invariant information while also training for classification accuracy. We also aim to influence the learning process with a domain invariant information source. We instead measure the distance between domains, which can be viewed as a surrogate for the oracle. It is domain invariant information and has no predictive use on its own. Letting the distance or dependency between the distributions regularise the loss function replaces the mutual information.

Embedding alignment. Feature alignment for domain invariance has been studied extensively [7, 9, 13, 23]. Various distances were used for the aligning features from a source to a target domain, such as KL-divergence [29], Maximum mean discrepancy [27] and the Wasserstein distance [25]. These approaches are promising but suffer from high variance in performance, shifting from domain to domain and accumulating errors using the predicted pseudo-labels in their training. That is why it is important to have adequate performance on a target domain before using the distance measures and to suppress the possible error accumulation. Using multiple source domains before predicting pseudo labels is hypothesised to increase this performance significantly, thus allowing for a better examination of the feature alignment technique. This examination will be one of our contributions.

3 Method

Learning to align. Our method operates on three domains, of which two are labelled (source) and one is unlabelled (target). First we learn from the two source domains, by training on two

batches, one from each domain. Alignment of embeddings from both domains is achieved by adding a distance minimization objective to the classification loss. The distance between embeddings of respective classes in both domains, is added to the loss function with a hyperparameter γ to balance the penalisation. These distances are calculated each batch with mirrored classes from the two domains to compare similar images. The full loss function then takes the following form:

$$Loss = L_{ce} + \gamma * L_{distance}. \tag{1}$$

This loss is back-propagated through the network, where the distance loss only influences the backbone as it is a function of the embeddings. This is the variant for domain generalization.

For unsupervised domain adaptation, we leverage the target domain's images. Now, we align with the third domain by having again two batches, where the first batch mixes the two source domains, and the second batch is selected from the target domain. Because for the target domain we only have images, we plug-in pseudo-labels as proxy for real labels. We study on the effect of using only a top κ section of pseudo-labels, sorted on prediction confidence. This reduces the amount of labelling errors. Our method is outlined in Figure 1.

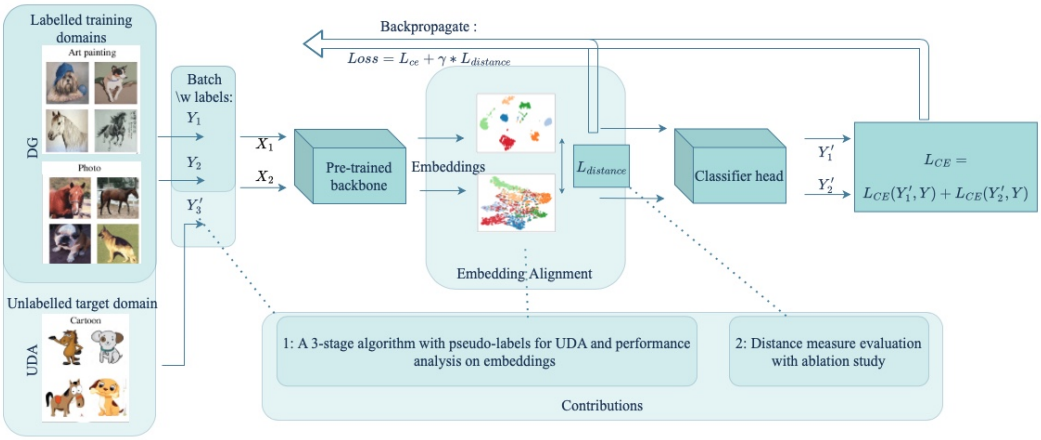

Figure 1

Distance measures. We consider three distance measures: the l^2-norm; the Hilbert-Schmidt Independence Criterion (HSIC) [10], since it is able to measure higher-level correlations; and the Wasserstein distance [15] since it is a proper measure between sample distributions. We use a normalised HSIC or nHSIC [5] implementation [12]. It measures a dependency between two distributions, which believe is present across domains if a feature-label relationship is invariant. We also use [8] as the wasserstein implementation, wich is an approximation of the optimal transport problem.

4 Experiments

We validate our method in a setup where we always have access to two domains and their labels. The first experiment is to measure the effect of the alignment versus standard training, both without access to the target domain, i.e., domain generalization (DG). The second experiment is to measure the added value of having access to the target domain's images without their labels, i.e., unsupervised domain adaptation (UDA). We test both the domain generalisation and unsupervised domain adaptation on three domains from the PACS dataset [26]: Photo, Art and Cartoon. We consider Art as the unlabelled target domain.

Implementation. We have chosen a ResNet18 pre-trained on ImageNet as a domain invariant backbone to improve. The classifier head consists of two hidden linear layers with two ReLU activation functions to promote non-linear prediction, for the non-linear dependency seen in the embeddings. After some tuning, the learning rate was set to 0.001 and we used Adam as optimiser.

Results. The performance of various settings is summarized in Table 1 for optimal values of hyperparameters, in order to compare the accuracy when aligning two labelled domains (domain generalization) and the merit of pseudo-labels from the third unlabelled domain (unsupervised domain

Table 1: An overview of the best results for resp. no alignment, best DG, best UDA per distance measure. *Optimal values. **Absolute gain with respect to the model performance before retraining with the pseudo-labels and alignment.

	Domain generalisation		Unsupervised domain adaptation		
	Accuracy %	γ*	Accuracy gain %**	γ*	κ*
Baseline	69.9 ± 1.9	-	4.4 ± 2.7	-	-
Norm	70.1 ± 2.2	0.03	4.4 ± 0.9	0.7	0.5
HSIC	**72.5** ± 1.4	1	**6.6** ± 1.8	0.3	0.5
OT	72.0 ± 3.7	0.7	5.9 ± 1.3	0.3	0.5

adaptation). For domain generalization, some improvement is visible in Table 1 using the domain alignment procedure in domain generalisation. The HSIC improves the performance the most from an accuracy of 69.9% in the baseline to 72.5%. For unsupervised domain adaptation, there is a positive effect of the pseudo-labels, regardless of using the distance measure. Leveraging the available information from the target domain already adds 4.4%, as much as with the norm distance, while the pseudo-labels are imperfect at $\pm 70\%$. Adding the distance minimisation to the second training pass results in another gain with HSIC as embedding distance reaching +6.6%.

We can increase domain invariance by increasing γ, however this comes at the cost of distinctive performance. To analyse this behaviour and find the optimal values we present table 2. Note the peak in accuracy at $\gamma = 1$ and $\gamma = 0.7$ for HSIC and optimal transport respectively.

Table 2: The results of the γ ablation study both HSIC and optimal transport distances for domain generalisation. There is a convex optimal value for both distances.

γ	0	0.3	0.5	0.7	1	1.5
HSIC (Acc. %)	69.9 ± 1.7	71.8 ± 1.2	70.8 ± 1.2	72.3 ± 0.9	**72.5** ± 1.4	72.4 ± 1.1
OT (Acc. %)	69.9 ± 1.7	70.0 ± 1.9	71.4 ± 3.1	**72.0** ± 3.3	69.3 ± 3.8	66.7 ± 1.9

Aligned embeddings. We analyze how much the domain invariance is improved by inspecting t-SNE plots of the embeddings of the target domain, see Figure 2. Optimal transport produced the best examples visually. The embeddings with alignment (middle) combined with pseudo-labels (right) show better alignment and separability between the classes.

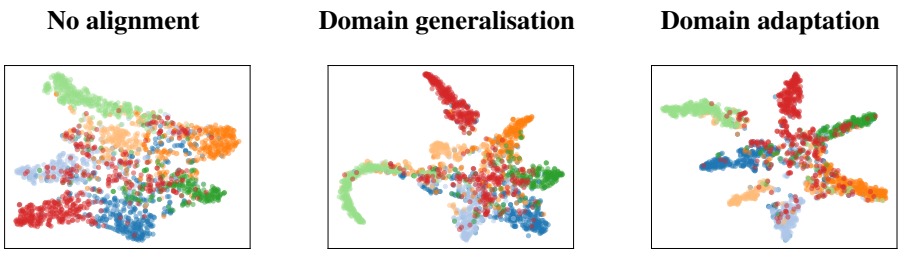

Figure 2: t-SNE plots of the feature embeddings of seven different classes: dog (blue), elephant (light blue), giraffe (orange), guitar (light orange), horse (green), house (light green) and person (red). Alignment and separability of the classes is improved significantly.

5 Conclusions

We have shown the potential of using distance minimisation for embedding alignment in domain generalisation and unsupervised domain adaptation in a three-domain setting. Optimizing our hyper parameters has shown that both optimal transport and HSIC are suitable candidates for both processes, where HSIC is the preferred choice. We have found that re-training the network with predicted pseudo-labels can be a good step towards adaptation to an unlabelled domain, especially when combined with embedding alignment.

References

[1] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or Invariant Risk Minimization? A Sample Complexity Perspective. 10 2020.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. 7 2019.

[3] Sara Beery, Grant Van Horn, and Pietro Perona Caltech. Recognition in Terra Incognita. Technical report.

[4] 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(1-2), 2010.

[5] M B Blaschko and a Gretton. A Hilbert-Schmidt Dependence Maximization Approach to Unsupervised Structure Discovery. *Proceedings of the 6th International Workshop on Mining and Learning with Graphs (MLG 2008)*, 2008.

[6] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain Generalization by Mutual-Information Regularization with Pre-trained Models. 3 2022.

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, 2019.

[8] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22, 2021.

[9] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test, 2012.

[10] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Scholkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3734 LNAI, 2005.

[11] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 9 2018.

[12] Wan Duo Kurt Ma, J. P. Lewis, and W. Bastiaan Kleijn. The HSIC bottleneck: Deep learning without back-propagation. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020.

[13] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, 2016.

[14] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. 7 2017.

[15] Gaspard Monge. Mémoire sur la théorie des déblais et de remblais. In *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*. 1781.

[16] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, 2019.

[17] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Technical Report 5, 2016.

[18] Mateo Rojas-Carulla, Bernhard Scholkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19, 2018.

[19] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The Risks of Invariant Risk Minimization. 10 2020.

[20] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 83(2):215–246, 4 2021.

[21] Bernhard Scholkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij JMOOIJ. On Causal and Anticausal Learning. Technical report.

[22] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 5 2021.

[23] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9915 LNCS, 2016.

[24] V Vapnik. Principles of Risk Minimization for Learning Theory. Technical report.

[25] Hanrui Wu, Yuguang Yan, Michael K. Ng, and Qingyao Wu. Domain-attention Conditional Wasserstein Distance for Multi-source Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(4), 7 2020.

[26] Jiaolong Xu, Liang Xiao, and Antonio M. Lopez. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7, 2019.

[27] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, 2017.

[28] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey J Gordon. On Learning Invariant Representations for Domain Adaptation. Technical report.

[29] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2015-January, 2015.