Deep Domain Adaptation without Access to a Specific Target Domain

Emiel Witting[1], Yasin Tepeli[1,*], and Joana P. Gonçalves[1,*]

[1]Department of Intelligent Systems, Delft University of Technology, Delft, Netherlands
{e.a.witting@student., y.i.tepeli@}tudelft.nl

Introduction. Unsupervised domain adaptation methods have garnered significant attention for their ability to mitigate feature distribution shifts between the source (training) and target (testing) domains in machine learning applications [1], which is usually caused by sample selection bias [2]. Deep domain adaptation [3] (DDA), specifically makes use of deep learning to extract high-level features that are shared across source and target domains to align their distributions. However, domain adaptation to specific target domains has its limitations. The alignment achieved might not effectively generalize to other target sets emerging from different domains, and the number of samples in the target set may not be sufficient for alignment. In cases where unlabeled samples are abundantly accessible however, these could be used to better capture the global underlying distribution of the data and can be used in an unsupervised or semi-supervised manner. The goal is to assess whether DDA methods can effectively align the distribution with the broader global sample set, and generalizing to unforeseen target sets under different scenarios of data shift.

Approach. For a comprehensive evaluation, we assessed the performance of DDA methods across various bias scenarios using covariate and concept shifts with both weak and strong variations. We utilized two prominent DDA methods: the Domain Adversarial Neural Network [4] (DANN) and a modified autoencoder-based [5,6] (AE) approach. The latter method extracts shared features between two sets and subsequently reduces residual domain shift in a second stage by minimizing a statistical domain distance measure [7]. Both methods simultaneously adapt two domains and learn a classification model.

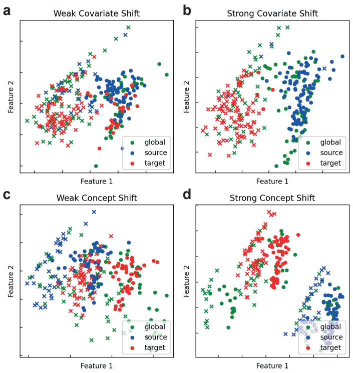

Fig. 1. Source, target, and global sets with shifts ×: Class 1, □: Class 2.

We compared the adaptation to global set (S→G) with DDA methods against other configurations such as supervised model trained on source set (S_{only}), or target set (T_{only}) and DDA methods but with adaptation to the target set (T→G). The accuracy on the target domain served as our primary performance metric. To test the effectiveness of DDA methods in a controlled environment, we generated synthetic binary classification datasets characterized by four clusters per class in three dimensions. These clusters were generated around the vertices of a 3-dimensional hypercube, and two additional dimensions that are added as a linear combinations of the original dimensions. The

*: Supervisor

datasets were divided into source, target, and global sets over 10 different runs. Source and target sets exhibited bias, while the global set remained unbiased.

Data Shift Scenarios. The first scenario was covariate shift that modify the feature space through biasing (shifting) while preserving aligned labels. We first partitioned the dataset randomly into source, target, and global sets. Then, we introduced distinct covariate shifts into the source and target sets by selecting samples based on their feature values, without altering labels which ensured that the original global decision boundary was maintained. The sampling was done proportional to distinct multivariate normal distributions over the features. Weak and strong variations of the shift were generated by varying the distance between source and target distribution (Fig. 1a-b).

Subsequently, we evaluated the performance of DDA methods in concept shift scenarios where both features and labels underwent joint translation. Our simulation of concept shift was performed by introducing distinctive pseudo-domains (or clusters) within the datasets and subjecting each of them to random translations of varying strengths. We then sampled the global set from all pseudo-domains, while the source and target sets were exclusively sampled from a specific random domain (Fig. 1c-d).

Results. The DDA methods could not improve the S_{only} model on weak covariate shift which could be attributed to the ineffectiveness of the induced bias as the performance difference between optimal T_{only} and the biased S_{only} was minimal (Fig. 2a). As the strength of the shift intensified, the performance disparity between T_{only} and S_{only} became more pronounced, although the adaptation to the target (S→T) demonstrated limited impact with both DDA methods. Conversely, both DDA methods decreased performance with global set (S→G) compared to S_{only}.

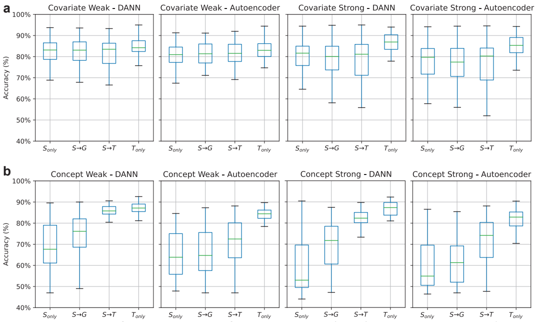

Fig. 2. Performance of DDA methods on **a)** covariate and **b)** concept shifts.

As for concept shift, a more substantial performance (median accuracy) gap was observed between the T_{only} and S_{only} supervised models (~20% in weak and 28% to 34% in strong shift: Fig. 2b). With DANN, S→G enhanced the performance of S_{only} (8.5% increase in weak, 19% in strong), albeit to a lesser degree than S→T. In the case of the AE, with S→T, there was a significant increase for both weak (19%) and strong (8.6%) shifts compared to S_{only}. However, S→G did not yield any improvement in weak shifts and demonstrated insignificant enhancement in strong shifts. Overall, DANN's domain-adversarial mechanism effectively harmonizes the distribution of training data with the broader global sample set, successfully bridging the adaptation gap and showcasing adaptability to intricate shifts while AE presents less consistent improvements.

To summarize, our study unveils that with a more extensive global sample set, DDA methods demonstrated enhanced generalizability and ability of bias mitigation, even though they may not be as proficient as using target sets for adaptation. We underscore the necessity of considering a wider spectrum of data distributions during adaptation, especially within real-world scenarios marked by diverse and unforeseen target sets.

References.

1. Kouw, W. M. and Loog, M.: A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3):766–785, 3 (2021)
2. Palak, H.M., and Nidhi, G.: Effect of selection bias on Automatic Colonoscopy Polyp Detection. Biomedical Signal Processing and Control 85 (2023)
3. Csurka, G.: A Comprehensive Survey on Domain Adaptation for Visual Applications. Domain Adaptation in Computer Vision Applications. Cham: Springer International Publishing, 1–35. (2017)
4. Sicilia, A., Zhao, X., and Hwang, S.J.: Domain Adversarial Neural Networks for Domain Generalization: When It Works and How to Improve. Machine Learning (2021)
5. Rifai, S. et al.: Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. Proceedings of the 28th International Conference on Machine Learning, ICML 2011 (2011)
6. Lu, C. et al.: Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. Signal Processing 130,377–388 (2017)
7. Li, X. et al.: Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation. Neurocomputing 383, 235–247 (2020)