ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Domain Adversarial Tangent Learning Towards Interpretable Domain Adaptation

Christoph Raab[1], Sascha Saralajew[2] and Frank-Michael Schleif[1]

1- University of Applied Science Würzburg-Schweinfurt - Department of Computer Science, Würzburg - Germany
2- Bosch Center for Artificial Intelligence, Renningen - Germany.

Abstract. Deep learning struggles to generalize well to an unseen target domain of interest. Current domain adaptation methods simultaneously learn a classifier and an adversarial game for invariant representations but inadequately align local structures, while the underlying process is hard to interpret. We propose a new interpretable adversarial domain architecture, matching local manifold approximations across domains. Evaluated against related networks, the approach is competitive, while the adaptation process can be visually verified.

1 Introduction

Deep Domain Adaptation (DDA) is a technique to learn a network capable of adapting from a training or source domain to an evaluation or target domain, assuming the domain data distributions are related but inevitably different. Joint Adversarial Domain Adaptation (JADA) [1] is the current state of the art in DDA and learns the following multi-task schema: a classifier, a domain discriminator, and a local adaptation divergence on top of a feature extractor network. The classifier is learned on labeled source data and should generalize well to the target domain. The discriminator learns to separate both domains while the feature extractor tries to fool it, playing a min-max game and learning global domain invariance [2]. Finally, local adaptation [3] is used to jointly align structures such as classes or clusters where pseudo labels [4] are frequently used.

However, network predictions [4] are unreliable, in a DDA setting, due to possible distribution shifts during adaptation [5]. Using trivial models such as moving average to approximate and align local data [3] neglects multi-modal domain structures [6] and are inappropriate. Additionally, these approaches use neural networks without an interpretable domain invariant class representation.

The Generalized Tangent Learning Vector Quantization (GTLVQ) [7] models the classification boundary implicitly by approximating the local data manifold structure via affine subspaces (tangent space approximations). Employing the GTLVQ as domain tangent discriminator is favorable to address the above issues because (i) provides a locally invariant and reliable model in the adaptation process by subspace and online learning, (ii) can capture multi-modal manifold structures, and (iii) provides an interpretable model by visualizing points from the affine subspace to verify the adaptation process.

Contributions: we introduce the GTLVQ as a domain discriminator and derive the Domain Adversarial Tangent Network (DATN) to match local domain

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

manifolds during domain adaptation (Sec. 3). We validate our DATN against related networks and show state-of-the-art performance (Sec. 4). Further, we interpret and visualize the adaptation processes by Siamese subspace samplings, which to our best knowledge, is the first approach to interpret DDA networks (Sec. 4).

2 Background and Related Work

In unsupervised deep domain adaptation [2, 6], we consider a labeled source dataset $\mathbf{D}_s = \{\mathbf{X}_s, \mathbf{Y}_s\} = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^n \overset{i.i.d.}{\sim} p(\mathcal{S})$ in the source domain $\mathcal{S}$ and an unlabeled target dataset $\mathbf{D}_t = \{\mathbf{X}_t\} = \{\mathbf{x}_j^t\}_{j=1}^m \overset{i.i.d.}{\sim} p(\mathcal{T})$ in the target domain $\mathcal{T}$ with same label space $\forall i, j : y_i, y_j \in \mathcal{Y}$ but different distributions $p(\mathcal{S}) \neq p(\mathcal{T})$. The overall goal is (still) to learn a classifier model, but additionally, it should generalize to a related target domain. The input feature space $\mathcal{X}$ is the initial representation of the source and target, i. e., $\mathbf{X}_s, \mathbf{X}_t \in \mathcal{X}$.

Adversarial Domain Adaptation (ADA): Initially, we consider a vanilla ADA network [2]: $f : \mathcal{X} \to \mathcal{F}$ with parameters θ_f as feature extractor, $g : \mathcal{F} \to \mathcal{Y}$ with parameters θ_g as classifier and $d : \mathcal{F} \to \mathcal{D} = \{-1, 1\}$ as a domain classifier with parameters θ_d, predicting the domain of a sample. The network learns by

$$\arg\min_{\theta_f, \theta_g, \theta_d} \mathbb{E}\left[\mathcal{L}_y\left(g\left(f\left(\mathbf{X}_s; \theta_f\right); \theta_g\right), \mathbf{Y}_s\right)\right] + \mathbb{E}[\mathcal{L}_d\left(d\left(R\left(f\left(\mathbf{X}; \theta_f\right), \lambda\right); \theta_d\right), \mathbf{Y}_d)], \quad (1)$$

where $\mathbf{Y}_d = [\mathbf{1}n, -\mathbf{1}m] \in \mathcal{D}$ are the domain labels and $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t]$ is the combined source and target data. $\mathcal{L}_y$ is the cross entropy classification loss given source data and $\mathcal{L}_d$ is the binary cross entropy considering all available data. The Gradient Reversal Layer (GRL) [2] is $R(x; \lambda) = x$ with $\frac{\partial R}{\partial x} = -\lambda \mathbf{I}$ and by multiplying the gradients of $d(\cdot)$ with $-\lambda \mathbf{I}$ leads to the min-max game, and the invariant representation. Trade-off is controlled by hyperparameter $\lambda \in [0, 1]$.

Joint Adversarial Domain Adaptation (JADA): The JADA networks extend Eq. (1) to capture the class structure of both domains within adversarial learning. The Conditional Domain Adaptation Network (CDAN) [6] utilizes a multi-linear map $g(\mathbf{X}) \otimes d(\mathbf{X})$ to model uncertainty and class affiliation into $\mathcal{L}_d$. The Joint Adversarial Adaptation Network [1] introduces two non-equal classifiers $g_1(\mathbf{X}_t)$, $g_2(\mathbf{X}_t)$ predicting target labels. The network is learned to maximize the difference between both predictions using L_1-norm, and by GRL, the feature extractor learns a classifier-level independent representation. The concept of JADA was recently generalized by introducing k classifiers, explicitly minimizing the classifier predictions via GRL [8]. Adversarial learning takes place separately. Finally, the Adversarial Semantic Consistency Network (ASC) [3] tries to approximate local structures by a class-wise mini-batch moving average. Target class means are found using pseudo labels. Source and target means are aligned together with the above adversarial learning.

Delimitation: The above methods share the following issues addressed by DATN: (i) they utilize classifier predictions in the adaptation process leading to degenerated features if pseudo labels are false [9]. DATN learns local domain

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

structures over multiple batches via domain loss. (ii) moving averages used by ASC are heavily biased towards the mini-batches and cannot capture multi-modality by a single mean. We do not restrict our model to a single subspace per domain and can capture multi-modal domain structures [6] given by the classes. (iii) all mentioned discriminators are uninterpretable with no guarantee or tool to verify working adaptation mechanics. Note that plotting a classification boundary is unreliable due to the possibility of mode collapse of the discriminator. We apply Siamese networks [10] to visualizes the discriminator model to see the invariant and relevant aspects of class data from both domains.

3 Model

Current JADA approaches use neural networks to align both domains via adversarial learning, but the prior discussion motivates various opportunities for GTLVQ. We will briefly summarize GTLVQ, reformulate it as a discriminator, and assemble the Domain Adversarial Tangent Network (DATN) as JADA network.

GTLVQ: Originated from prototype-learning GTLVQ [7] consists of q affine subspaces $W = \{\mathbf{v}_k + \mathbf{B}_k\boldsymbol{\theta} | \boldsymbol{\theta} \in \mathbb{R}^l, \mathbf{B}_k \in \mathbb{R}^{f\times l}\}_{k=1}^q$, where $\mathbf{B}_k$ is an orthonormal basis of the l-dimensional linear subspace of bottleneck space $\mathcal{F} \in \mathbb{R}^f$, $\mathbf{v}_k$ is a translation vector and $\boldsymbol{\theta}$ is a parameter vector. The relation of $\mathbf{x} \in \mathbb{R}^f$ to $w_k \in W$ is computed by the tangent distance $d(\mathbf{x}, w_k) = \min_{\boldsymbol{\theta}} d_e(\mathbf{x}, \mathbf{v}_k + \mathbf{B}_k\boldsymbol{\theta})$, where d_e is the euclidean distance. Assuming $\mathbf{B}_k\mathbf{B}_k^T = \mathbf{I}_f$, the optimal parameter vector for minimizing $d(\mathbf{x}, w_k)$ yields $\boldsymbol{\theta}^* = \mathbf{B}_k^T(\mathbf{x} - \mathbf{v}_k)$ so that the Tangent Distance can be simplified to

$$d(\mathbf{x}, w_k) = \sqrt{(\mathbf{x} - \mathbf{v}_k)^T\mathbf{P}_k(\mathbf{x} - \mathbf{v}_k)}, \tag{2}$$

where $\mathbf{P}_k = \mathbf{I}_f - \mathbf{B}_k\mathbf{B}_k^T \in \mathbb{R}^{f\times f}$. Note that w_k with label y_k is called set-prototype. Let $d^+(\mathbf{x}_i, W) = min\ d(\mathbf{x}_i, w_k)\ \forall w_k : y_k = y_i$ be the closest set-prototype with same label as $\mathbf{x}_i$ and $d^-(\mathbf{x}_i, W) = min\ d(\mathbf{x}_i, w_k)\ \forall w_k : y_k \neq y_i$ the closest set-prototype with different label. The GTLVQ is learned to have close same label set-prototypes, while different label ones are further away w.r.t $\mathbf{x}_i$. The method learns via projected stochastic gradient descent, consisting of vanilla SGD of w_k parameters and orthonormalization of $\mathbf{B}_k$ after every update.

DATN: We employ GTLVQ as Tangent Discriminator (TD) via $d(\cdot, W)$ in Eq. (1) to separate both domains with their respective manifold structure as two-class task. To have consistent loss functions over all models and to learn all set-prototypes simultaneously, we learn TD with the following cross entropy function

$$\mathcal{L}_d(\mathbf{x}, y_d; W) = -\sum_{k=1}^{q}(y_k = y_d)\log\left(\frac{\exp(-d(\mathbf{x}, w_k))}{\sum_{a=1}^{q}\exp(-d_j(\mathbf{x}, w_a))}\right) \tag{3}$$

where $y_d \in \mathcal{D}$ is the domain label of $\mathbf{x}$, $d(\mathbf{x}, w_k)$ is the tangent distance, and y_k is the label of k-th set-prototype. Minimizing Eq. (3) will lower the distance

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

to closest correct and increase it to incorrect set-prototypes. Hence, the model captures the multi-modality by learning to separate both domains via multiple set-prototypes. It is not biased by the current batch because every prototype is adapted by minimizing Eq. (3), without needing source or target pseudo labels. During learning, the TD gradients are reverse propagated by GRL to the feature extractor, resulting in merging both domain manifolds to fool the discriminator. By plugging Eq. (3) into Eq. (1) all models are trained simultaneously. The gradients of $\mathcal{L}_d$ w.r.t W are omitted due to space issues.

Local Adaptation Interpretability: We use the Siamese [10] concept. A translation $\mathbf{v}_k \in w_k$ is obtained by forwarding domain data from original space $\mathbf{v}_k^x \in \mathcal{X}$ to the bottleneck space via $\mathbf{v}_k = f(\mathbf{v}_k^x) \in \mathcal{F}$. We call $\mathbf{v}_k^x$ *Siamese Translation* (ST) and by setting $\theta = 0$, such an $\mathbf{v}_k \in w_k$ is a representative subspace sample. Recap that to classify both domains, multiple prototypes represent either source or target domain. We identify pairs of closest source and target translations in $\mathcal{F}$ for interpretation. Their Siamese counterparts are, for example, images or text in the input space $\mathcal{X}$. Hence, we can identify domain invariant features using activation heatmaps of adapted source and target representatives and interpret local adaptation success for the first time in DDA.

4 Experiments

The DATN approach is evaluated against related competitive approaches[1]. The results are shown in Tab. 1. In the following, we describe the experimental details:

Setup. The study follows the standard protocol for evaluating unsupervised deep domain adaptation [6] and utilizes all available labeled source data for learning and all unlabeled target data for knowledge transfer and evaluation. The performance is summarized as mean with standard deviation over three random runs. Related competitive methods are listed in the results table.

Dataset. The Office-31 dataset contains images from three separated domains, namely Amazon (**A**), Webcam (**W**), and DSLR camera (**D**) [6, 11]. Each domain has 31 classes with objects frequently located in the office. Since all three domains are acquired with different settings and photo cameras, the adaptation problem is to train on one domain and test on another. For example, a dataset combination is **A**→**D** (Amazon to DSLR).

Implementation Details. The DATN hyper-parameter, namely the subspace dimension, is optimized via grid-search and set to $l = 128$ for all dataset combinations. Feature extractor and classifier parameters are trained via SGD with a momentum of 0.9, while the Tangent Discriminator (TD) is learned via ADAM. The target STs are found by clustering with 31 centroids in $\mathcal{F}$ and randomly select an image from each cluster. The subspaces are the right singular vectors of an epoch of source and target data, respectively. The STs as original images with their subspaces are the initial TD model. We use progress based

[1] Code and data are published at `github.com/ChristophRaab/DATL`

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Dataset	**A→W**	**D→W**	**W→D**	**A→D**	**D→A**	**W→A**	Avg.
Resnet [12]	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DANN (2015) [2]	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
CDAN (2018) [6]	93.1±0.2	98.2±0.2	**100±0**	89.8±0.3	70.1±0.4	68.0±0.4	86.6
CAT (2019) [13]	94.4±0.1	98.0±0.2	**100±0**	**90.8±1.8**	72.2±0.6	70.2±0.1	87.6
CLDA (2020) [4]	78.5±0.3	**99.3±0.2**	99.8 ± 0.1	79.1±0.1	64.7±0.3	65.8±0.2	81.2
SCA (2020) [11]	93.6±0.1	98±0.2	**100±0**	89.5±0.1	**72.6±0.2**	72.4±0.3	**87.7**
DATN (ours)	**94.6±0.2**	98.4±0.2	**100±0**	90.3±0.7	71.4±0.2	71.6±2.3	**87.7**

Table 1: Mean prediction **accuracy** with standard deviation on the **Office-31**.

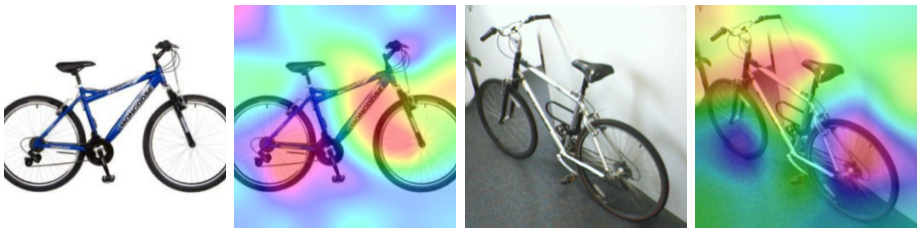

Fig. 1: Source bike translation (Left) with closest target translation (right) with score cam heatmap.

parameter scheduling for λ (GRL, see Sec. 2) as suggested for adversarial domain architectures [2, 6]. The batch size is 36, and the bottleneck dimension is 256.
Results. We compare our DATN against competitive domain adaptation networks (Tab 1). The baselines are Resnet [12] and DANN [2]. The results show the competitive performance of DATN against SCA [11] and CAT [13] with insignificant performance differences. DATN outperforms the remaining approaches. The results support the claim that GTLVQ as Tangent Discriminator is a reasonable substitute for neural network-based discriminators with similar performance compared to recent networks. However, in contrast to compared results, we can verify local adaptation by interpreting the GTLVQ model.
Interpretability. The Siamese source translation for a bike with their nearest target counterpart with the score class activation mapping are plotted in Fig. 1. The match is found by the pairwise euclidean distance of the source and target translations in $\mathcal{F}$. Because the target translations are found randomly given target clusters, as described above, the network can find the correct corresponding translation and match both local manifold approximations together. It is observable that the focus of DATN given the STs is on the switching gear and some parts of the wheel, invariant to rotations and the respective domain. Hence, we can identify the invariant regions and interpret the adaptation process.

5 Conclusion

We studied the effect of GTLVQ as a domain tangent discriminator in a JADA network. The proposed Deep Adversarial Tangent Network (DATN) learns all set-prototypes simultaneously to approximate local domain manifolds. The as-

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

sembled DATN is competitive to the state of the art unsupervised domain adaptation networks. In contrast to related work, the discriminator is interpretable to verify local adaptation success, making the approach favorable for application. Future work should target extensive derivation of the gradients of DATN and the local interpretability aspect, as well as more benchmark experiments.

References

[1] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. Joint Adversarial Domain Adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 729–737, New York, NY, USA, oct 2019. ACM.

[2] Yaroslav Ganin and Victor S Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1180–1189, 2015.

[3] Manliang Cao, Xiangdong Zhou, Yiming Xu, Yue Pang, and Bo Yao. Adversarial Domain Adaptation with Semantic Consistency for Cross-Domain Image Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 259–268, New York, NY, USA, nov 2019. ACM.

[4] Zhihai He, Bo Yang, Chaoxian Chen, Qilin Mu, and Zesong Li. CLDA: an adversarial unsupervised domain adaptation method with classifier-level adaptation. *Multimedia Tools and Applications*, apr 2020.

[5] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle Self-Training for Domain Adaptation. *arXiv*, pages 1–22, mar 2021.

[6] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 1647–1657, 2018.

[7] Sascha Saralajew and Thomas Villmann. Adaptive tangent distances in generalized learning vector quantization for transformation and distortion invariant classification learning. In *2016 International Joint Conference on Neural Networks (IJCNN)*, volume 2016-Octob, pages 2672–2679. IEEE, jul 2016.

[8] Sayan Rakshit, Ushasi Chaudhuri, Biplab Banerjee, and Subhasis Chaudhuri. Class Consistency Driven Unsupervised Deep Adversarial Domain Adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 657–666. IEEE, jun 2019.

[9] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3235–3246. Curran Associates, Inc., 2018.

[10] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, Roopak Shah, H. Jansen, M. P. Gallee, and F. H. Schroder. Signature Verification using a "Siamese" Time Delay Neural Network. In J D Cowan, G Tesauro, and J Alspector, editors, *Advances in Neural Information Processing Systems 6*, volume 18, pages 737–744. Morgan-Kaufmann, 1994.

[11] Weijian Deng, Liang Zheng, Yifan Sun, and Jianbin Jiao. Rethinking Triplet Loss for Domain Adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(8):1–1, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016.

[13] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster Alignment With a Teacher for Unsupervised Domain Adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 2019-Octob, pages 9943–9952. IEEE, oct 2019.