Unsupervised Domain Adaptation: State-Of-The-Art review and analysis

Roberto Mazzaro 229301

Filippo Daniotti 232087

Michele Yin 229359

Giovanni Ambrosi 232252

Andrea Bonora 232222

***Abstract*—Deep neural networks have greatly improved the performances of shallow models in many fields. Currently, we are heading to increasingly larger and deeper networks, which require similarly large datasets. However, while deep models have good performances on the dataset they are trained on, they also perform significantly worse on a different dataset. In *domain adaptation* we aim to mitigate this issue. Here we present a survey of some techniques that have been researched in the specific field of *unsupervised domain adaptation*.**

I. Unsupervised domain adaptation

In domain adaptation (DA) the goal is to minimize the domain loss that occurs when moving a model from a source, seen domain, X_s to a target X_t, unseen domain by minimizing the *domain shift*. In the particular case of unsupervised domain adaptation (UDA) only source labels are available in during training $X_s = \{x_s, y_s\}$, whereas target labels are not $X_t = \{x_t\}$. Here we present a non-comprehensive overview of the many methods available for UDA.

II. Discrepancy-based methods

In discrepancy-based methods we assume that the distributions of source and target domains are shifted by a reasonably small amount. In theory, it should be possible to align the distributions to reduce domain bias, therefore minimizing the distance in source and target features.

To align source and target domains, a correct distance metric has to be chosen. Some distribution metrics distance may be used (e.g. Kullback-Leibler); minimizing this metric while retaining the network capability to extract good features is key.

Some of the earliest ideas are MMD [1, Tzeng et al.] and CORAL [2, Sun et al.] loss, that aim to align the distributions of source and target using first or second order moments respectively. However, because they do not consider higher order moments, theoretically the approach might be insufficient. The distributions may have same first or second order moments, but not third or higher order statistics. This may be addressed by either explicitly considering higher order moments [3, Zellinger et al.], [4, Chen et al.], or by proposing a kernel [5, Wang et al] to map the features to a higher order Hilbert space before comparing them using a distance metric; this way it is possible to measure more than one moment using one metric.

Typically, those formulations are very simple to implement and can be generalized by appending the distance based loss to the objective, with a regulation factor λ: $L = L_{clf}(Y_s, \hat{X}_s) + \lambda L_{distance}(X_s, X_t)$.

Other works propose to add domain alignment layers, like DAN [6, Long et al]. They are proven to be more effective because the network has more layers, thus it can learn a transformation that aligns both source and target features.

One issue with discrepancy-based approaches is that the distribution of source and target classes may be different. Therefore, applying a distance based metric may not lead to an effective domain adaptation. Some metrics are specifically designed to address this imbalance, as in [7, Balaji et al.].

Also, it may be the case that target features are unevenly distributed on a decision boundary for source features. This hinders performances, leading to extremely uneffective domain adaptation results. To overcome this issue, it is possible to add a clustering term to the loss. The idea is to promote source features compactness and distance to other classes, which will lead to a similar representation in the target space. Some works use similar ideas based on confidence, entropy or others but the key concept is the same [8, Kang et al.], [9, Chen et al]. Also, pseudolabelling can be very effective if target and source domains have a small domain shift, like shown in DSAN [10, Zhu et al.]. Here, the authors perform pseudolabelling of target features and then align pseudolabels of target and ground truth on source for each class label. However, effective pseudolabelling may not improve results if source and target distributions have a high domain shift.

Lastly, while deep networks can be used effectively in many fields, as they extract features that are better than human generated features, like SIFT [11, Lowe et al.] or SURF [12, Bay et al.] etc. However, it has been proven that the network has to eventually transition from general features to domain specific features [13, Long et al.]. There is some work on learning two different backbones for source and target domains [14, Rozantsev et al.] based on finding the alignment from source to target.

III. Adversarial-based methods

The fundamental idea of this class of methods can be found in the paper of [15, Ganin at al.], where a domain discriminator is added. The classifier is trained to minimize the classification loss and the domain predictor to minimize the domain loss, while the feature extractor is trained to minimize the classification loss but maximize the domain loss. This adversarial game between the discriminator and the feature extractor is possible thank to the Gradient Reversal Layer and results in extracted features that are shared across both domains.

In the following years many different architectures that improve this idea have been presented. In [16, Tzeng at al.] a discriminative mapping of the target images to the source feature space is learned by fooling the domain discriminator; moreover, they introduce a GAN loss. In [17, Long at al.] they exploit both the domain specific feature representation and the discriminative information conveyed in the classifier prediction to condition the discriminator through a multilinear mapping.

Some works instead of introducing a discriminator introduce one or more GANs. In [18, Bousmalis at al.] they use a GAN that works at pixel level and tries to transform source images into target ones to train directly the task specific classifier, while in [19, Sankaranarayanan at al.] the generated images enter only in the discriminator and its feedback is used to train the feature extractor. Last trend in GAN based methods is to introduce cycles as is done by [20, Hoffman at al.], where source images are transformed into target ones and then back to source; moreover, many different losses are exploited to better adapt images. In [21, Russo at al.] they apply cycle mechanism for both for source and target images and source-like target images are annotated with pseudo-labels and used to train the classifier.

Recently, the research is concentrated in finding techniques that can be applied to already existing UDA techniques to improve them. In [22, Wang at al.] they propose to use a fine-grained discriminator that also includes the class information to obtain a better class alignment. In [23, Chen at al.] they remove the discriminator and the Nuclear-norm Wasserstein Discrepancy module is introduced; this module is attached to the classifier and can be used as a discriminator. The idea is that the self-correlation matrix produced by the classifier for the source images has high values only in the diagonal due to the supervised training, while for the target images the self-correlation matrix has also high values in the off-diagonal elements. This approach should encourage intra and inter-class correlation. In [22, Wang at al.] they use conditional entropy to reweight samples in order to give an higher weight to poorly aligned ones; triplet loss is also used. In [24, Chen at al.] they combine self-training to learn discriminative features, and adversarial training to align features distributions. In [25, Wei at al.] they propose a meta-learning scheme where one task is the meta-train task and while the other is used for validation. In the following epoch they are used in the other way around; this way, both tasks are optimized in a coordinated way. In [26, Gao at al.] they constrained gradients of two domains to have similar distributions and to do that pseudo labels are computed to calculate the target loss; moreover jacobian regularization is applied. In [27, Jin at al.] they relabel well-aligned target samples to source domain to make distributions more separable. In [28, Rangwani at al.] they discover that reaching a smoother minimum of the task loss leads to a better generalization, while this is not true for the adversarial loss; they also propose SDAT (Smooth Domain Adversarial Training), which is able to reach a smoother minimum with few additional gradient computation steps. The idea is that the Hessian matrix should have a low value for the maximum eigenvalue if the reached minimum is smooth.

IV. Optimal Transport

Optimal transport is the general problem of moving one distribution of mass to another as efficiently as possible. Let X and Y be two sets of data having the same number of data samples N. In a matching problem each data point x in X must be matched with exactly one data point y in Y. The cost for this match to occur is $c(x, y)$. These potential matches are N. The average of c across all data points define the cost of a specific match (the *transport*). The discrete OT problem is to find the transport with the lower the average cost. There are many ways to define a distance between P and Q, such as Total Variation, Hellinger, L_2, $\mathcal{X}^2$ and the Wasserstein distance (Eq.1), which is the most used in recent works.

$$W_p(P, Q) = \left(\inf_{J \in \Im(P,Q)} \int ||x - y||^p dJ(x, y) \right)^{1/p} \tag{1}$$

One on the main work based on OT is [29, Nguyen et al.], which uses a teacher-student architecture. The authors assume to have a well-qualified classifier (teacher) that gives accurate prediction on the data in the source domain. The goal is to learn a classifier (student) for the target domain by minimizing the proposed objective function composed of the loss of the teacher and a *OT-based* term computed using the Wasserstein distance.

V. Other methods

a) Incremental methods: The concept of incremental methods starts from self-labeling approaches, which are specific approaches in which a supervised model is trained from the labeled data and then used to automatically assign a pseudo-label to each unlabeled sample. The concept of self-labeling has been investigated by [30, Gallego et al.], which took the current DANN approach [15] and provided novel ways to enhance its capacity for incremental adaptation to the target domain. The main assumption of this work is that we can add the subset of target domain samples on which classifier is more confident about to the source-labeled domain while assuming the prediction as ground-truth label. Then, using the new training set, we may retrain the DANN network to fine-tune its weights. This process is repeated iteratively, moving the labeled samples with greater confidence from the target domain to the source domain after each iteration. We stop when there are no more samples left in the target domain to move. The main idea is that by including target domain information in the source domain, the DANN learns new domain-invariant features that better suit the eventual classification task.

b) Information-based methods: Mutual Information (MI) maximization has been shown as a promising approach in unsupervised learning. In particular, we can define the Mutual information (MI) of two random variables as a measure of the mutual dependence between the two variables. Recently, it has also been used in the more restricted context of hypothesis transfer learning (HTL). Specifically, [31, Lao et al.] propose a model that use MI maximization on the unlabeled target data to transfer knowledge from a set of source hypothesis, learned from the source domain, to a corresponding target set of target hypotheses. This is in contrast to the common approaches for HTL and UDA that tend to use a single hypothesis, failing to uncover different modes of the model distribution. The crucial drawback of multiple independent MI maximization is that different target hypotheses can be optimized in an unrestricted manner due to a lack of supervision. To overcome this limitation, an hypothesis disparity (HD) regularization is included to align the target hypothesis. This is done in a way that takes uncertainty expressed through different source hypothesis into consideration while marginalizing out the undesired disagreements.

c) Semi-Supervised learning methods: Some have explored the idea of using SSL methods in UDA: this is achieved by reframing SSL as a special case of UDA. In [32, Zhang et al.] they pointed that both UDA and SSL aims to use unlabelled (target) data as data-dependent regularizer to improve model performance over a labelled (source) baseline.

When it comes to reframing: in a SSL scenario labelled data are typically insufficient to represent the overall distribution P_{ssl}, i.e. labelled data only represent a sub-domain of that distribution. Now, consider the sub-domain with the smallest support set for the distribution P_{small} where all classes are represented; hence, we can consider P_{small} and P_{ssl} as source and target distribution in a UDA setting, thereby allowing to run SSL pipelines for UDA tasks. They tried several SOTA SSL methods and achieved notable results.

d) Self-Training based methods: In self-training we aim to discover pseudo-labels for the target distribution. We first train on source data only and the train model is used to classify unlabelled samples, picking the highest-confidence label as pseudo-label.

One limitation of self-training is that performances drop as domain divergence increases. In [33, Zhang et al.] they proposed to mitigate this problem by constructing intermediate datasets and models, starting from source only and progressively increasing the percentage of unlabelled data. The self-training procedure is run iteratively until a dataset with target only examples and possibly accurate pseudo-labels is constructed.

e) Self-Supervised methods: In self-supervised learning we artificially create labelled data out of unlabelled data, which we then use to train a model on auxiliary tasks (*pretexts*), under the assumption that the latent representation learned from pretexts will effectively translate to downstream real applications.

This idea has been explored in UDA in [34, Sun et al.]. They proposed a multi-head architecture, with a shared feature extractor, one head for the specific task and multiple heads for the pretexts. The task specific head should be trained on source data only, while the remaining heads should be trained on both source and target. They picked three pretexts: (i) rotation classification, (ii) horizontal flip detection, (iii) patch location detection. The idea is that these pretexts should force the model to learn more domain-invariant structural features, hence providing the model with improved accuracy on the target domain.

This idea is architecture-independent, thus easily adaptable to other UDA methods. For instance, [35, Xie et al.] revisited the CycleGAN [20] paradigm with self-supervision. In particular, they enriched the architecture with a siamese network S trained with a self-supervised patch detection pretext. The S network takes two patches and has (i) a content registration branch, accounting for conteng consistency in translation, (ii) and a domain classification branch, which determines whether the two patches are both from source, both translated or mixed. The idea is to locally disentangle domain information and image content.

f) Other methods: M-ADDA [36, Issam H. et al.] is an adversarial deep metric learning where we use a metric to perform target clustering. The clusters are encoded by a decoder that maps the target features in an embedding space. The training of the models is differentiated between the source model and target model. We use a triplet loss to optimize source network parameters by computing the square distance between a random chosen example and two other examples, one with the same label and one with different label. The target model is trained through two losses, one for the target encoder, which consists in an adversarial loss and one for the decoder, which is a magnet loss. The magnet loss computes simultaneously the distribution of each target class in the embedding space and reduces the overlap between clusters.
In Heuristic Domain Adaptation [37, Shuhao Cui et al.] they use a fundament network that learns the domain-invariance representation and an heuristic network that focuses on domain-specific features. We ideally want a model that has learned the distributions of the two domains without any domain specific bias. To ensure the convergence of the two networks they compute the cosine similarity between the generator function (difference of fundamental and heurstic functions) and the heuristic function itself.

Finally, in [38, Yunzhong H. et al.] they investigate what neural networks learn in domain adaptation. They proposed a source-free image translation (SFIT), a novel method that generates source-style images from original target images through a generator network, so that it mitigates and represents the knowledge difference between models. The entire procedure is built around two main concepts, which are *relationship preserving* and *knowledge distillation*. Through knowledge distillation they distill the knowledge of the generator by portraying the adapted knowledge in the target model with source model and generator combined. Relationship preserving indicates a successful depiction of the target model knowledge on the generated images and more in general it states the distributions' alignment.

References

[1] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014. [Online]. Available: https://arxiv.org/abs/1412.3474

[2] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," 2016. [Online]. Available: https://arxiv.org/abs/1607.01719

[3] W. Zellinger, B. A. Moser, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Robust unsupervised domain adaptation for neural networks via moment alignment," *Information Sciences*, vol. 483, pp. 174–191, may 2019. [Online]. Available: https://doi.org/10.1016%2Fj.ins.2019.01.025

[4] C. Chen, Z. Fu, Z. Chen, S. Jin, Z. Cheng, X. Jin, and X.-S. Hua, "Homm: Higher-order moment matching for unsupervised domain adaptation," 2019. [Online]. Available: https://arxiv.org/abs/1912.11976

[5] W. Wang, B. Li, S. Yang, J. Sun, Z. Ding, J. Chen, X. Dong, Z. Wang, and H. Li, "A unified joint maximum mean discrepancy for domain adaptation," 2021. [Online]. Available: https://arxiv.org/abs/2101.09979

[6] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015. [Online]. Available: https://arxiv.org/abs/1502.02791

[7] Y. Balaji, R. Chellappa, and S. Feizi, "Normalized wasserstein distance for mixture distributions with applications in adversarial learning and domain adaptation," 2019. [Online]. Available: https://arxiv.org/abs/1902.00415

[8] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," 2019. [Online]. Available: https://arxiv.org/abs/1901.00976

[9] C. Chen, Z. Chen, B. Jiang, and X. Jin, "Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation," 2018. [Online]. Available: https://arxiv.org/abs/1808.09347

[10] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1713–1722, apr 2021. [Online]. Available: https://doi.org/10.1109%2Ftnnls.2020.2988928

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008, similarity Matching in Computer Vision and Multimedia. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314207001555

[13] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015. [Online]. Available: https://arxiv.org/abs/1502.02791

[14] A. Rozantsev, M. Salzmann, and P. Fua, "Residual parameter transfer for deep domain adaptation," 2017. [Online]. Available: https://arxiv.org/abs/1711.07714

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," 2015. [Online]. Available: https://arxiv.org/abs/1505.07818

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," 2017. [Online]. Available: https://arxiv.org/abs/1702.05464

[17] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," 2017. [Online]. Available: https://arxiv.org/abs/1705.10667

[18] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1612.05424

[19] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," 2017. [Online]. Available: https://arxiv.org/abs/1704.01705

[20] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," 2017. [Online]. Available: https://arxiv.org/abs/1711.03213

[21] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: symmetric bi-directional adaptive gan," 2017. [Online]. Available: https://arxiv.org/abs/1705.08824

[22] H. Wang, T. Shen, W. Zhang, L. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," 2020. [Online]. Available: https://arxiv.org/abs/2007.09222

[23] L. Chen, H. Chen, Z. Wei, X. Jin, X. Tan, Y. Jin, and E. Chen, "Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation," 2022.

[24] M. Chen, S. Zhao, H. Liu, and D. Cai, "Adversarial-learned loss for domain adaptation," 2020. [Online]. Available: https://arxiv.org/abs/2001.01046

[25] G. Wei, C. Lan, W. Zeng, and Z. Chen, "Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation," 2021. [Online]. Available: https://arxiv.org/abs/2103.13575

[26] Z. Gao, S. Zhang, K. Huang, Q. Wang, and C. Zhong, "Gradient distribution alignment certificates better adversarial domain adaptation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8917–8926, 2021.

[27] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Re-energizing domain discriminator with sample relabeling for adversarial domain adaptation," 2021. [Online]. Available: https://arxiv.org/abs/2103.11661

[28] H. Rangwani, S. K. Aithal, M. Mishra, A. Jain, and R. V. Babu, "A closer look at smoothness in domain adversarial training," 2022. [Online]. Available: https://arxiv.org/abs/2206.08213

[29] T. Nguyen, T. Le, N. Dam, Q. H. Tran, T. Nguyen, and D. Phung, "Tidot: A teacher imitation learning approach for domain adaptation with optimal transport," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 2862–2868. [Online]. Available: https://doi.org/10.24963/ijcai.2021/394

[30] A. Gallego, J. Calvo-Zaragoza, and R. B. Fisher, "Incremental unsupervised domain-adversarial training of neural networks," *CoRR*, vol. abs/2001.04129, 2020. [Online]. Available: https://arxiv.org/abs/2001.04129

[31] Q. Lao, X. Jiang, and M. Havaei, "Hypothesis disparity regularized mutual information maximization," 2020. [Online]. Available: https://arxiv.org/abs/2012.08072

[32] Y. Zhang, H. Zhang, B. Deng, S. Li, K. Jia, and L. Zhang, "Semi-supervised models are strong unsupervised domain adaptation learners," 2021. [Online]. Available: https://arxiv.org/abs/2106.00417

[33] Y. Zhang, B. Deng, K. Jia, and L. Zhang, "Gradual domain adaptation via self-training of auxiliary models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09890

[34] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," 2019. [Online]. Available: https://arxiv.org/abs/1909.11825

[35] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, and Y. Zheng, "Self-supervised cyclegan for object-preserving image-to-image domain adaptation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 498–513.

[36] I. Laradji and R. Babanezhad, "M-adda: Unsupervised domain adaptation with deep metric learning," 2018. [Online]. Available: https://arxiv.org/abs/1807.02552

[37] S. Cui, X. Jin, S. Wang, Y. He, and Q. Huang, "Heuristic domain adaptation," *CoRR*, vol. abs/2011.14540, 2020. [Online]. Available: https://arxiv.org/abs/2011.14540

[38] Y. Hou and L. Zheng, "Visualizing adapted knowledge in domain transfer," *CoRR*, vol. abs/2104.10602, 2021. [Online]. Available: https://arxiv.org/abs/2104.10602