ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Shallow and Deep Models for Domain Adaptation problems

Siamak Mehrkanoon*,[1], Matthew B. Blaschko[2], Johan A.K. Suykens[1]

1- Department of Electrical Engineering ESAT-STADIUS,
KU Leuven, B-3001 Leuven, Belgium
2- ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
*Corresponding author, e-mail: siamak.mehrkanoon@esat.kuleuven.be

Abstract.
Manual labeling of sufficient training data for diverse application domains is a costly, laborious task and often prohibitive. Therefore, designing models that can leverage rich labeled data in one domain and be applicable to a different but related domain is highly desirable. In particular, domain adaptation or transfer learning algorithms seek to generalize a model trained in a source domain to a new target domain. Recent years has witnessed increasing interest in these types of models due to their practical importance in real-life applications. In this paper we provide a brief overview of recent techniques with both shallow and deep architectures for domain adaptation models.

1 Introduction

Often in many application domains huge volumes of unlabeled data are generated and made available, but the cost of obtaining data labels remains high. To overcome the burden of annotation, several attempts have been made in the literature in order to exploit the unlabeled data or data available in different but related domains. In particular, in this context semi-supervised learning as well as transfer learning have gained more attention recently due to their practical importance in many real life problems [1, 2].

The most common underlying assumption of many machine learning algorithms is that both training and test data exhibit the same distribution or same feature domains. However in many cases the data change from one domain to another or its statistical properties evolve in time [3]. For instance, in visual applications domain shifts can simply be caused by changing conditions, location, background, pose change among others.

The non-stationary nature of the data brings a new challenge for many existing learning algorithms, which are based on the stationary assumption. When there is a distributional, feature space and/or dimension mismatch between the two domains, the models learned with data in one domain would fail to predict the test data in the other. Domain Adaptation (DA) is a particular case of transfer learning (TL) that leverages labeled data in one or more related source domains, to learn a classifier for unseen or unlabeled data in a target domain. In general in DA it is assumed that the two domains share the same task i.e. class labels are shared between domains. In addition the source domains are assumed to be related to the target domain, but not identical.

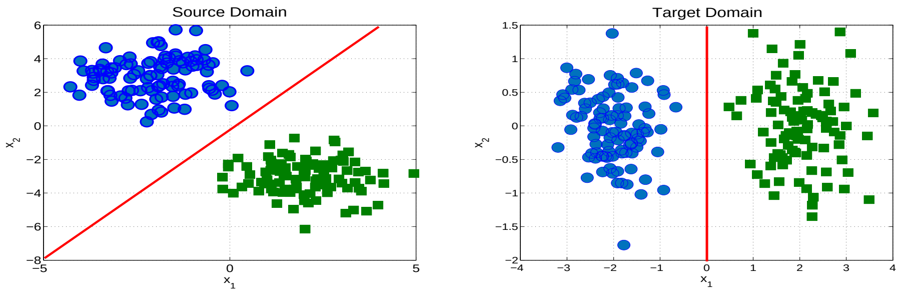
Domain adaptation can also be seen as a particular case of semi-supervised learning where one aims at leveraging unlabeled data to improve the model generalization performance when limited number of labeled training data is available. Therefore if one ignores the distributional mismatch, then the traditional semi-supervised models can be employed where data of the source and target domain provides the labeled and unlabeled instances. Depending on the availability of the labeled instances in both domains, three scenarios can be considered, i.e. unsupervised, supervised and semi-supervised domain adaptation [4]. Unsupervised domain adaptation approaches, do not take label information into consideration when learning the feature representation [5]. On the other hand, supervised domain adaptation approaches, only use labeled data from the source and target domains. In the semi-supervised setting, one learns from labeled source instances as well as a small fraction of the target labeled instances [4, 6, 7]. This paper is organized as follows. In Section 2, a brief overview of existing shallow domain adaptation methodologies is provided. Section 3, discusses the recent domain adaptation methods with deep architectures.

2 Shallow architectures for domain adaptation

In general, two types of domain adaptation problems have been addressed in the literature, i.e homogeneous and heterogeneous domain adaption. Let us assume that $\mathcal{X}_s$ and $\mathcal{X}_t$ denote the source and target domain data, and the marginal probability distribution of the source and target domains are $P(\mathcal{X}_s)$ and $P(\mathcal{X}_t)$ respectively. In the homogeneous case, the feature representation for the source and target domains is the same $\mathcal{X}_s = \mathcal{X}_t$ but $P(\mathcal{X}_s) \neq P(\mathcal{X}_t)$ (see Fig. 1 for an illustration). However, in domain adaptation across heterogeneous feature spaces, the distributions, feature domains or feature dimensions in source and target domains are different. The existing methodologies in the literature for the domain adaptation problem can be categorized into methods with shallow and deep architectures. In what follows we give a brief overview of some of the successful shallow domain adaptation methods.

- **Instance re-weighting methods:** In the sample reweighting approach, one assigns sample-dependent weights to the training data with the aim of minimizing the distribution discrepancy between the source and target data points in the reweighted space [8]. The mechanisms that are mostly used in the literature for the estimation of sample dependent weights are formulated as the density ratio between the probability densities of the two domains [9]. A selective Transfer Machines algorithm that jointly optimizes the weights as well as the classifier's parameters is also introduced in [10]. The authors in [11] proposed a method to infer re-sampling weights through maximum entropy density estimation.

- **Feature Transformation:** Another key research challenge in domain adaptation is how to learn a domain-invariant feature representation for



(a) (b)

Fig. 1: (a) Source dataset (b) Target dataset. Given source and target dataset, one usually aims at learning a classifier from the source domain and adapt it to the target domain which exhibits a different distribution than the source domain data (i.e. $P(\mathcal{X}_s) \neq P(\mathcal{X}_t)$).

both source and target domains. The adaptation then can be accomplished by learning a model on the new space. Depending on the availability of the target labeled data one can consider either unsupervised or semi-supervised feature transformation method. One of the first such DA method is the Transfer Component Analysis (TCA) [12] that proposes to discover common latent features having the same marginal distribution across the source and target domains, while maintaining the intrinsic structure (local geometry of the data manifold) of the original domain by a smoothness term.

The authors in [13] introduced a method to learn the feature transformation in order to produce a set of common transfer components across domains. The Structural Correspondence Learning method proposed in [13] learns a common feature space by identifying correspondence among features from different domains. A domain adaptation approach that uses the correlation subspace as a joint representation of the source and target data is introduced in [14]. In this approach the new representation is learnt using unlabeled data pairs in both source and target domains. A deep learning approach to learn new cross-domain feature representation from the source and target data is proposed in [15]. The Heterogeneous Feature Argumentation (HFA) [16] embeds the source and target data into a common latent space where the transformation metrics are computed by the minimization of the structural risk functional SVM expressed as a function of these projection matrices.

In many domain adaptation problems, side information in the form of correspondence instances (paired instances) are available for either unlabeled or labeled instances across domains. For instance consider the snapshots of the actions of the same persons in two different time instances shown in Fig. 2. In this case one can have access to paired labeled and unlabeled samples across domains.

In general the instance similarity constraints between domains, if available, can be used to enhance the performance of the classifier [6]. The authors in [17] proposed a method that adapts representations using a small number of paired synthetic and real views of the same object/scene. In their experiments, each real example is paired with a corresponding synthetic image in the same pose, and an additional unpaired synthetic images are also provided as training data.

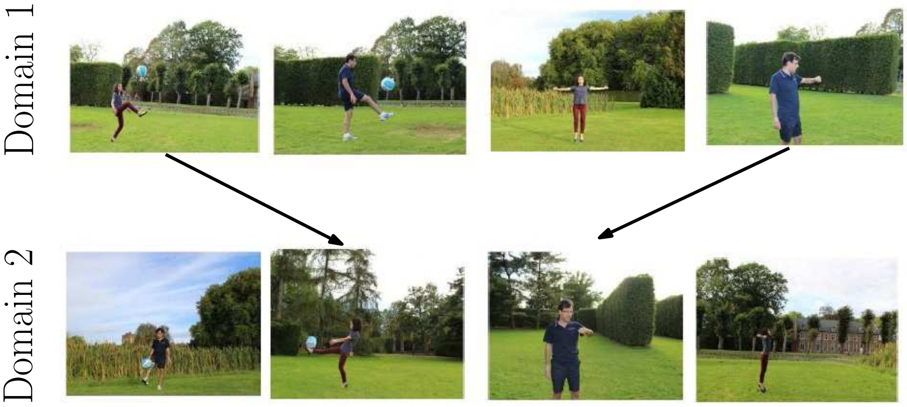


Fig. 2: Example of labeled and unlabeled paired instances (Image: adapted from Mehrkanoon et al. [4]). Objects that are present in both domains are paired instances. Among paired instances, some of them can be labeled as well.

Mehrkanoon et al. in [4] introduced a Regularized Semi-Paired Kernel Canonical Correlation Analysis (RSP-KCCA) formulation for learning a new representation of the data for the sake of the domain adaptation problem. The optimization problem is formulated in the primal-dual LS-SVM setting where side information are incorporated through regularization terms. The proposed model learns a joint representation of the data set across different domains by solving a linear system of equations in the dual. The approach is naturally equipped with out-of-sample extension property which plays an important role for model selection. Different types of instances ranging from unlabeled, labeled, paired and unpaired are seamlessly integrated to the model. Therefore the model can be employed in unsupervised, semi-supervised as well as supervised scenarios.

3 Deep architectures for domain adaptation

Deep Learning techniques have attracted many researchers due to their success in revolutionizing many application domains ranging from auditory to vision signal processing [18, 19]. Deep learning based models deal with complex tasks by learning from subtasks. In particular, several nonlinear modules are stacked in hierarchical architectures to learn multiple levels of representation (hierarchical features) from the raw input data. Each module transforms the representation at one level into a slightly more abstract representation at a higher level, i.e. the higher-level features are defined in terms of lower-level ones.

Recent advances in artificial neural networks and in particular deep learning have also shown promising results on domain adaptation problems [12, 20, 21]. In this context, one of the earliest deep model is the Stacked Denoising Autoencoders (SDAs) which aimed at adapting the sentiment classification between reviewers of different products [21]. The marginalized stacked denoising autoencoders (mSDA) is proposed in [20] to learn new representations for domain adaptation. As opposed to SDAs, mSDA does not require stochastic gradient descent or other optimization algorithms to learn the parameters and that there is a closed-form solution for obtaining the model parameters. The authors showed that the representations learned by mSDA are as effective as the traditional SDAs in benchmarked tasks. As suggested in [20], both domains not necessarily should use identical features and one can pad all input vectors with zeros to make both domains be of equal dimensionality.

Other research studies have shown that layers of deep convolutional networks can be fine-tuned to novel tasks [22]. It has been shown that the features learned by deep convolutional networks are more abstract and have discrimination power in the target domain even without any adaptation [22, 23, 24]. Further attempts have been made in the literature to exploit deep models for domain adaptation problems and in general these models can be categorized into three groups. The first group utilizes the CNN models to extract features which later is used by the shallow DA methods. These methods consider the deep network as feature extractor [25]. New representation for the input data are obtained by means of the activations of the layers in a deep architecture. In particular, the authors in [22] examined the features learned from a deep convolutional network trained using a large labeled fixed set of object recognition tasks on novel generic tasks which may differ significantly from the originally trained tasks. They demonstrate that by leveraging an auxiliary large labeled object database to train a deep convolutional architecture, one can learn features that have a good generalization ability to perform semantic visual discrimination tasks using simple linear classifiers. On the other hand, one then can use these Deep Convolutional Activation Features within shallow DA methods. For instance the authors in [22] used Deep features in several DA methods such as Geodesic Flow Kernel [5], Max-Margin Domain Transforms [26] and Feature Augmentation [27].

The methods in the second group try to adjust the the pre-trained network to the new task by fine-tuning on the source domain, and use the model to predict class labels for target instances [28, 29, 30]. However it should be noted that fine-tuning a network might also require a relatively large amount of labeled data which is not always available for the target domain. Therefore, in this case the model is usually fine-tuned using the labeled source data augmented with, if available, the few labeled target instances. When using these approaches, one should make sure that the model does not overfit to the source data in case only few labeled target were available.

Among methods in third category one can mention for instance the Stacked Denoising Autoencoders [21], the Stacked Marginalized Denoising Autoencoders [20], Domain Adaptive Neural Network [31]. The domain adaptation method introduced in [32] uses a few target samples to reconstruct the output of the filters that were found affected by the domain shift. Other deep models starts with two streams, representing the source and target domains which are then trained with a combination of a classification and a discrepancy loss [33, 34, 35]. In these types of models the discrepancy loss aims at diminishing the shift between the two domains while the classification loss relies on the labeled source data. The authors in [36] introduced Transfer Neural Trees to relate heterogeneous cross-domain data and jointly solved cross-domain feature mapping, adaptation, and classification. A deep learning model for domain adaptation by means of interpolating between domains is proposed in [37] where a predictively useful representation of the data is learned by taking into consideration the information from the distribution shift between the two domains.

Lifelong learning is a framework for continual adaptation to new domains. In this scenario, new tasks are trained in sequence, and models must be adapted such that they have good performance on the new task, while retaining high performance on previously seen tasks without retraining. [38] have extended theoretical works for transfer learning to show generalization bounds in the lifelong learning setting depending on the Kullback-Leibler divergence between task distributions. Practical algorithms address the setting by assuming the marginal distributions are equal $P(\mathcal{X}_s) = P(\mathcal{X}_t)$ [39], or by keeping a low-memory approximation to the previous marginal distributions by an auto-encoder [40] or a small number of carefully chosen samples [41].

Last but not least, it should be noted that most of the research works in the domain adaptation literature emphasis on image categorization tasks. Relatively few papers discuss the domain adaptation problem in other challenging tasks such as object detection, semantic segmentation, pose estimation, video event or action detection. It is expected that in near future more attentions will be given to address these challenging problems and the literature will witness novel methodologies and architectures in the context of domain adaptation.

Acknowledgments

This work was supported by • The Postdoctoral Fellowship of the Research Foundation-Flanders (FWO: 12Z1318N). • Research Council KUL: CoE PFV/10/002, PhD/Postdoc grants Flemish Government; FWO: projects: G0A4917N, G.088114N. Siamak Mehrkanoon is Postdoctoral Fellow of the Research Foundation - Flanders (FWO). Matthew Blaschko is a professor at the KU Leuven, Belgium. Johan Suykens is a full professor at the KU Leuven, Belgium.

References

[1] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning.* MIT Press, 2010.

[3] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning.* MIT Press, 2009.

[4] Siamak Mehrkanoon and Johan AK Suykens. Regularized semipaired kernel CCA for domain adaptation. *IEEE transactions on neural networks and learning systems*, 2017.

[5] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012.

[6] Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 668–675. IEEE, 2013.

[7] Min Xiao and Yuhong Guo. Feature space independent semi-supervised domain adaptation via kernel matching. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):54–66, 2015.

[8] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

[9] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[10] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3515–3522. IEEE, 2013.

[11] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[12] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[13] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics, 2006.

[14] Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Transactions on Image Processing*, 23(5):2009–2018, 2014.

[15] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[16] Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1134–1148, 2014.

[17] Eric Tzeng, Coline Devin, Judy Hoffman, Chelsea Finn, Xingchao Peng, Sergey Levine, Kate Saenko, and Trevor Darrell. Towards adapting deep visuomotor representations from simulated to real environments. *arXiv preprint arXiv:1511.07111*, 2015.

[18] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1627–1634. ICML, 2012.

[21] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[22] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[24] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[25] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[26] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013.

[27] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[28] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[29] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1717–1724. IEEE, 2014.

[30] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In *European Conference on Computer Vision*, pages 435–442. Springer, 2016.

[31] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[32] Rahaf Aljundi and Tinne Tuytelaars. Lightweight unsupervised domain adaptation by convolutional filter reconstruction. In *European Conference on Computer Vision*, pages 508–515. Springer, 2016.

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[34] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[35] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[36] Wei-Yu Chen, Tzu-Ming Harry Hsu, Yao-Hung Hubert Tsai, Yu-Chiang Frank Wang, and Ming-Syan Chen. Transfer neural trees for heterogeneous domain adaptation. In *European Conference on Computer Vision*, pages 399–414. Springer, 2016.

[37] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML Workshop on Challenges in Representation Learning*.

[38] Anastasia Pentina and Christoph H Lampert. Lifelong learning with non-i.i.d. tasks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1540–1548. 2015.

[39] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer, 2016.

[40] Amal Rannen Triki, Rahaf Aljundi, Matthew B. Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *International Conference on Computer Vision*, 2017.

[41] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *CVPR*, pages 5533–5542. IEEE Computer Society, 2017.