Transfer Learning: Current Status, Trends and Challenges

Ricardo Sousa[1]
rsousa@rsousa.org
Luis M. Silva[2]
lmas@up.pt
Luis A. Alexandre[3]
lfbaa@di.ubi.pt
Jorge Santos[4]
jms@isep.ipp.pt
Joaquim Marques de Sá[1]

[1] INEB, Porto, Portugal
[2] INEB, Porto, Dep. de Matemática at Universidade de Aveiro, Aveiro, Portugal
[3] IT, Universidade da Beira Interior, Covilhã, Portugal
[4] INEB, Porto, Instituto Superior Engenharia do Porto, Porto, Portugal

Abstract

Transfer Learning (TL) has gained significant interest in the Machine Learning (ML) community. Aiming to overcome standard ML learning models assumption that data distributions are the same independently of its origins, a wide variety of works have emerged to tackle this issue. Nowadays, with data available on large amounts and from different applications, and where labeling is a cumbersome task, it is counter productive to have intelligent systems for each specific real-world problem: DVD and home appliances rating, Amazon and medical reports analysis or histology imaging. In this paper we will review current work on TL, open issues and challenges that have not been addressed yet. We conclude this work with a set of remarks and guidelines for future TL methodologies.

1 Introduction

One common assumption in statistical learning theory when devising learning algorithms is that data from different problems are drawn from the same underlying distribution. However, this assumption fails in many real world problems. For the particular case of product reviews, models that perform recommendations of goods on the Amazon website cannot be straightforward applied in the IMDB website. In the same way, data (e.g., text statistics) present in the Wikipedia cannot correspond to the same information as data from the Reuters website. Moreover, despite the fact that textual data is presented in English – and the established grammatical rules are applied – a review that provides a positive feedback for a home appliance device cannot be derived in the same way as a positive review for a horror film. The interest on Transfer Learning (TL) is multifold. First and foremost, it alleviates the need of data labeling which is an expensive and cumbersome job; second, it often produces algorithms with good generalization capability for different learning problems. Finally, it has been claimed that TL provides learning models with good generalization performances in different problems with far less computational effort [4, 16, 17]. In a nutshell, it is the ability of *reusing* what was learned on one problem (also coined as source problem) onto another (target problem). Reusability will be the cornerstone of this manuscript. So, what is TL?

Definition 1 (TL): *TL is a Machine Learning (ML) research field whose goal is the development of algorithms capable of transferring the learning model obtained in a source problem to a target problem without the necessity of building a new model from scratch.*

For more than three decades there has been a significant amount of work on TL. Strangely enough, however, recent and classical works on TL have not been considered or thoroughly analyzed. Moreover, most of the times the concept of TL has been mixed with active, online and even sequential learning [32]; or, concepts from statistical classical learning theory have been used to define all possible TL scenarios. TL in fact shares ideas from areas such as *dataset shift* where the distribution of the data can change over time (and to which sequential algorithms may be applied). However, dataset shift is part of TL mostly because of the assumptions on the data distribution [25, Chapter 1]. Overall, some of these discrepancies have been hindering the evolvement of TL. In this work we will analyze the current developments of TL and open issues identified until now.

2 Transfer Learning: Current Status

To realize how much work has been done in TL, we have to go back some decades. In fact, this subject has been around since the 80's with considerable advancements since then (see for instance [4, 10, 14, 21, 23, 30, 31] and references therein). Probably, the foremost work that envisioned the concept of TL was Tom Mitchel [21] where the idea of bias learning was first presented. A first attempt to extend these ideas was soon performed in [24] where Neural Networks (NNs) where first used for TL. Their pragmatic approach consisted on training a NN on a source problem to be then retrained on a target problem. In a simple way, layer weights obtained using the source problem data were then reused and retrained to solve the target problem. At the time based on Decision Trees, Pratt adopted entropy measures to assess the quality of the hyperplanes. Soon after, [15] derived a framework to use (abstract) internal representations generated by NNs for TL problems.

After these pivotal works, a significant number of implementations and derivations of TL started to appear. In [29] a new learning paradigm was proposed for TL where one would inclemently learn concept after concept. Sebastien Thrun envisioned this approach to how humans learn: by stacking knowledge upon another (as building blocks) resulting in an extreme nested system of learning functions. At that time, a particular case of [29], coined as Multi-Task Learning (MTL), was presented [8, 9] along with their theoretical formulations [2]. In a nutshell, MTL solves all target problems all at once. However, this approach does not hold for our definition of TL (see **Definition 1**) since it learns a common representation for all of our data. These approaches assume that there is a significant amount of information overlapping all concepts that need to be learned, which sometimes is not the case.

In 2000 a specific formulation was introduced in [27] by Shimodaira. Although initially not contextualized in the domain of TL, its theoretical conclusions on how to learn a model on a target problem based on a source problem had a significant impact and its implications were only later realized. [27] described a weighted least squares based on the prior knowledge of the densities of source and target problems. At the time, Shimodaira only addressed the issue of data probabilities being different leading to what he had termed *covariate shift*. After that, [3, 10, 13] presented different algorithms to address the limitations of [27] such as the estimation of data probabilities leading to the rise of the *domain adaptation*[1]. In [28] an extension of Shimodaira's work was presented so that it could cope with the leave-one-out risk. The implications of this specific trend of TL was on the resolution of many Natural Processing Language (NPL) problems [3, 5] and genome sequence analysis [26]. Recently, an overview on TL was presented in [22] with a vast, but horizontal, analysis of the most recent works that tackle classification, regression and unsupervised learning for TL. [18] provided fundamental mathematical reasonings for TL by devising: 1) generalization bounds for max-margin algorithm such as SVMs and 2) its theoretical limits for error variability based on the leave-one-out risk [6]. [18], to the best of our knowledge, was one of the first to identify a gap in the literature of the theoretical limitations of algorithms on TL.

NNs are the majority of the chosen algorithms to perform TL. With the recently re-interest on NNs and the availability of more computational power along with new and faster algorithms, NN with deep architectures started to emerge to tackle TL. In [14] a framework for covariate shift with deep networks was presented. [1, 16, 17] widened the research line of [24] by addressing the following questions: How can we tailor deep neural networks for TL? How TL performs with reusing layers and with

[1] Although both terms are widely used in the literature, the underlying principles are strictly the same. For this reason we will opt to use instead covariate shift to avoid confusion on the terminology.

different types of data? As described so far, in spite of the immense different TL interpretations and definitions, concerns started to appear on how to unify this area of research on ML. Sharing these views, in [23] an unifying framework for many of the existing TL methods is presented. Here, concepts for covariate shift and, in more general, TL, are jointly defined.

3 Transfer Learning: Trends and Challenges

We will now focus on other pressing issues that ought be addressed in the near time soon. One important aspect is how to measure knowledge gains when doing TL. We will also briefly address the impact that differences among datasets (source and target problem) may have in the learning rates. Other open issues are concerned with the increasing availability of data. Knowing that is humanly impossible to analyze this amount of data we found that there are few learning models that address this; finally, we also identified a gap in public competitions to benchmark available methods that are presented in the literature.

Unification of TL: One main issue that has been hindering the advance of TL is the vast amount of formulations on TL. For instance, we have shown the work of [24, 29] that promote the idea of never-ending learning, the covariate shift of Shimodaira [27] and domain adaptation of [3, 4]. They share concepts from Shimodaira's covariate shift, but variants on the terminology are employed which inevitably leads to confusion. In fact, a first tentative for an unification on TL was proposed in [23].

Measuring Knowledge Gains: Bengio *et al*. in [14] analyzed until a certain extent how to quantify TL gains. Although overcoming some interpretation issues regarding performance results that can occur when dealing with multiple source domains, it is unknown how these measures behave in other TL methods besides covariate shift, particularly in situations where class sets are different between problems. Simpler approaches like mean squared error (MSE) or statistical inspired coefficients can provide further information such as class agreement. Moreover, measures as the ones employed by Bengio in [14] can lead to non definite results if one obtains a perfect baseline model (see [14]).

Dissimilar Datasets: How difficult is to do TL? It is important to state that according to [12] the quality of the results is related with Kullback-Leibler divergence measured on the datasets with different source/ target problem pair. In a straightforward reading it may seem that for different problems it may be infeasible to perform TL; Or, that TL models need to be more robust for heterogeneous problems; Or, that data features are not representative. Based on our review, it was not possible to identify works that try to make this analysis or at least to perform an attempt on that. Although these intuitive ideas have empirically present a relation between domain divergence and TL algorithms performance, a theoretical reason for these behaviors is still unknown [7, 17].

A new trend: Big Data With the emergence of evermore data it is infeasible for a human to analyze it in its life time. There has been a special interest in big data, particularly in the biology research fields [19]. However, the development of learning models for each specific problem may be a greedy and slow, but unnecessary, endeavor. Besides the fast-paced research for the development of new learning models to deal with big data, TL still seems to be very far behind. Due to the large size of these datasets, it is also a cumbersome and counter-productive task to pre-process it. Therefore, highly heterogeneous, with under-represented classes and contaminated with noise data may be difficult to process by an automatic learning algorithm. To overcome this, authors choose to use reduced, cleaner, versions of these datasets. See for instance [14]. We see that big data trend only strengthens the motivations and objectives for TL. One of the first works that tackle this was presented in 2014 in [11].

Public Competitions for Learning Algorithms Benchmark Contrary to many other research fields such as fingerprints[2] or object categorization[3], there has been a void in open challenges in TL where one can benchmark our method with others from the literature. Even though there have been workshops and special sessions in international top tier conferences such as 2013 and 2011 edition of NIPS[4], 2011 edition of ICML[5], or, 2009 edition of ECML PKDD[6] where several fundamental questions on TL were addressed and pushed forward the research in this field, to the best of our knowledge, the only competitions on TL were at ICML [20] and NIPS[7].

4 Conclusions

At this point it should be clear the interest on Transfer Learning (TL). We have described the limitations of standard machine learning (ML) approaches and how TL can aid in this matter. We have identified key, breakthrough, works on this subject but there are still aspects that need to be addressed. Most of the fundamental issues on TL such as the assumption on data distribution, generalization bounds or loss functions have started to be explored. In fact, some of these issues were raised in [20], but few authors have researched this. Despite of the amount of works that have been presented we have only seen a glimpse of what can be done on TL. Public competitions had always been a way to present breakthrough results and to assess in a homogeneous manner the state-of-the-art on various domains of research. But, only two competitions were conducted so far. In the future, it should be clearer in what TL consists by unifying its principles and concepts, by providing means to understand the impact of different pairs of source/target problems on the performance results, and finally, ways to measure it.

References

[1] Telmo Amaral, Luis M. Silva, Luis M. Alexandre, Chetak Kandaswamy, Joaquim Marques de Sá, and Jorge Santos. Transfer learning using rotated image data to improve deep neural network performance. In *ICIAR*, 2014.

[2] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.(JAIR)*, 12:149–198, 2000.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of Representations for Domain Adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[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):151–175, October 2009. ISSN 0885-6125. doi: 10.1007/s10994-009-5152-4.

[5] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[6] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[7] Lorenzo Bruzzone and Mattia Marconcini. Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):770–87, May 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.57.

[8] Caruana. Multitask Learning. *Machine Learning*, 75:41–75, 1997.

[9] Caruana. *Multitask Learning*. PhD thesis, 1997.

[10] Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi-supervised learning*, volume 2. MIT press Cambridge, 2006.

[11] Zhiyuan Chen and Bing Liu. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 703–711, 2014.

[12] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for Transfer Learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 193–200, New York, NY, USA, 2007. ACM. doi: 10.1145/1273496.1273521.

[13] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.(JAIR)*, 26: 101–126, 2006.

[14] 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.

[15] Nathan Intrator. Making a Low-dimensional Representation Suitable for Diverse Tasks. *Connection Science*, 8 (2):205–224, 1996. doi: 10.1080/095400996116884.

[16] Chetak Kandaswamy, Luis M. Silva, L. M. Alexandre, Jorge Santos, and JP Marques de Sá. Improving deep neural network performance by reusing features trained with transductive transference. In *ICANN*, 2014.

[17] Chetak Kandaswamy, Luis M. Silva, L. M. Alexandre, Ricardo Sousa, Jorge Santos, and JP Marques de Sá. Improving transfer learning accuracy by reusing stacked denoising autoencoders. In *Proceedings of the IEEE SMC Conference*, 2014.

[18] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *Proceedings of The 30th International Conference on Machine Learning*, pages 942–950, 2013.

[19] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

[20] Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian J Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *ICML Unsupervised and Transfer Learning*, pages 97–110, 2012.

[21] Tom M Mitchell. The need for biases in learning generalizations. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ., 1980.

[22] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.191.

[23] Novi Patricia and Barbara Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the Computer Vision and Pattern Recognition*, 2014.

[24] Lorien Y Pratt, LY Pratt, SJ Hanson, CL Giles, and JD Cowan. Discriminability-Based Transfer between Neural Networks. In S J Hanson, J D Cowan, and C L Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 204–211. 1992.

[25] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[26] Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, and Bernhard Schölkopf. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2009.

[27] 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.

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

[29] S. Thrun. Is Learning the *n*-th Thing Any Easier Than Learning the First? In D. Touretzky and M Mozer, editors, *Advances in Neural Information Processing Systems (NIPS) 8*, pages 640–646, Cambridge, MA, 1996. MIT Press.

[30] T Tommasi, F Orabona, and B Caputo. Learning Categories from Few Examples with Multi Model Knowledge Transfer. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.197.

[31] V N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 10(5):988–99, January 1999. ISSN 1045-9227. doi: 10.1109/72.788640.

[32] Peilin Zhao and Steven C Hoi. OTL: A framework of online transfer learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-2010)*, pages 1231–1238, 2010.

[2] FVC ongoing contest: https://biolab.csr.unibo.it/FVCOnGoing/

[3] VOC Challenge: http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[4] NIPS TL: http://nips.cc/Conferences/2013/Program/event.php?ID=3721, and https://sites.google.com/site/nips2011domainadap/

[5] ICML: http://clopinet.com/isabelle/Projects/ICML2011/

[6] ECML: http://www.ecmlpkdd2009.net/program/tutorials/transfer-learning-for-reinforcement-learning-domains/

[7] NIPS TL competition: https://sites.google.com/site/nips2011workshop/transfer-learning-challenge