Technology Review

University of Illinois at Urbana-Champaign

CS 410: Text Information Systems

Fall 2017

Andrew Nedilko

1. Overview and Motivation

Artificial intelligence in general and machine learning in particular are nowadays being researched and developed at an increasingly faster rate. They are transforming every large industry in the world. Most of the success in the industry comes from supervised learning, some – from more and more popular reinforcement learning, but both of these approaches have certain limitations associated with them.

A statistical algorithm cannot be applied with the same level of success to all the machine learning domains. This is because any domain knowledge captured by that algorithm is specific to the particular needs of analyzing data from this domain. Trying to apply the same algorithm to a different domain may very well result in a much less meaningful outcome. However, a new domain may have certain similarity with prior applications, and it would be reasonable to try and reuse the knowledge gained from prior models. This is where a "fast-evolving frontier of data science" [2] known as transfer learning comes into play. Historically, it has been also called knowledge consolidation, knowledge transfer, inductive transfer, learning to learn, context-sensitive learning, meta learning, domain adaptation, and so on.

In contrast to multi-task learning, when we try to identify common (latent) features which may be equally beneficial for each individual task, transfer learning is concerned with finding and applying knowledge from the so called source domains/tasks to target domain/tasks. This means that the tasks are no longer learned simultaneously, and the focus is more on the target tasks with a respective asymmetry of the two types of domain.

Unsupervised learning is the next step on the path of technology development which offers indisputably much broader opportunities as it combines the success of machine learning in general and supervised learning in particular with an inherent capability of the human brain – the ability to learn on its own. And transfer learning may very well become one of the major aids in unsupervised learning.

As an example, transfer learning can be used to resolve the following problems:

- <u>Data availability</u>: new domains may lack a sufficient amount of labeled training data of acceptable quality, and transfer learning can help to develop machine learning models using relevant training data from previous projects. Also, as mentioned in [1], transfer learning can help with mitigating training-data obsolescence which often happens in dynamic domains in which prior training data can become easily outdated, e.g. trying to estimate social sentiment;
- <u>Boosting productivity</u> and accelerating time on new modeling projects (e.g., an NLP algorithm classifying English-language technical documents in one scientific discipline should, in theory, be adaptable to classifying documents in another language in a related field);
- <u>Prediction refinement</u>: as state in [3], transfer learning can help data scientists mitigate the risks of machine-learning-based predictions in a domain which may be related to highly improbable events. If the modeled underlying conditions have radically changed, prior training data sets or feature models become inapplicable and transfer learning can help utilize useful training data and feature model subsets from related domains. Some good examples here include the unexpected results of Brexit or 2016 U.S. presidential election.

According to [1], transfer learning can be beneficial in many examples of data engineering. For instance, text categorization including web document classification into several predefined categories. Let us imagine that we have a source domain represented by labeled university web pages, and there is a new website with different data features or data distributions for which there is a lack of labeled training data. In this and similar cases, it may be helpful to transfer the classification knowledge gained on the basis of the university website to the new website.

Another example would be a binary sentiment analysis of product reviews. The distribution of review data among different types of products can be very different. To reduce the costs and effort for labeling reviews for various products, we may adjust a classification model trained on some products to help build classification models for other products.

The need for transfer learning may also arise when the the labeled data obtained in one time period may not follow the same distribution in a later time period.

2. Definition

Based on the information and document classification examples presented in [1], transfer learning can be defined as follows.

First of all, the main concepts in transfer learning are a domain and a task. A domain $\mathcal{D}$ contains a feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathrm{X})$ over the feature space, where $\mathcal{X}=x_1,\cdots,x_n\in\mathcal{X}$. For document classification based on the bag-of-words principle, $\mathcal{X}$ is the space of all documents, x_i is the i-th term vector corresponding to a document, and X is a training sample.

Given a domain, $\mathcal{D}=\{\mathcal{X},P(X)\}$, a task $\mathcal{T}$ consists of a label space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$ generally learned from the training data consisting of pairs $x_i\in\mathcal{X}$ and $y_i\in\mathcal{Y}$. In document classification, $\mathcal{Y}$ is the set of all True/False labels, and y_i is either *True* or *False*.

Given a source domain $\mathcal{D}_S$ with a respective source learning task $\mathcal{T}_S$, as well as a target domain $\mathcal{D}_T$ with a target learning task $\mathcal{T}_T$, the objective of transfer learning is to learn the target conditional probability distribution $P(Y_T|X_T)$ in $\mathcal{D}_T$ with the knowledge from $\mathcal{D}_S$ and $\mathcal{T}_S$ where $\mathcal{D}_S\neq\mathcal{D}_T$ or $\mathcal{T}_S\neq\mathcal{T}_T$. There may be a limited number of target examples with labels, but it should be extremely small in comparison with the source domain. Both the domain $\mathcal{D}$ and the task $\mathcal{T}$ are defined as tuples.

Given the above domains $\mathcal{D}_S$ and $\mathcal{D}_T$ such that $\mathcal{D}=\{\mathcal{X},P(X)\}$ and the tasks $\mathcal{T}_S$ and $\mathcal{T}_T$ such that $\mathcal{T}=\{\mathcal{Y},P(Y|X)\}$, four different cases of transfer learning are possible:

- $\mathcal{X}_S\neq\mathcal{X}_T$ - the feature spaces of the two domain are different, e.g. the sets of documents are provided in two different languages which is generally referred to as cross-lingual adaptation.
- $P(X_S)\neq P(X_T)$ - the marginal probability distributions of the two domain are different, e.g. the domains describe different topics which is known as domain adaptation.
- $\mathcal{Y}_S\neq\mathcal{Y}_T$ - the label are different for each task. This may, for example, correspond to a situation when one domain has binary document classes, and the other classifies its documents to more than two classes (e.g. 5 or 10). This can also occur along with case 4 below when the conditional probability distributions are different.
- $P(Y_S|X_S)\neq P(Y_T|X_T)$ - the conditional probability distributions of the two tasks are different, e.g. source and target documents are unbalanced with respect to their user-defined classes.

If there is an explicit or implicit relationship between the feature spaces of the two domains, they are considered related.

3. Practical Applications

When it comes to text information systems, transfer learning has been applied efficiently in many real-world situations. Below is a brief summary of some of them, along with a short description of new approaches in the field and some examples of using various toolkits enabling us to apply transfer learning.

3.1 Clustering

In [4] a concept of self-taught clustering is introduced as an instance of unsupervised transfer learning whose goal is to cluster a small collection of target unlabeled data using a large amount of auxiliary unlabeled data. The target and auxiliary data can vary in topic distribution.

[5] describes the idea of the so called “co-clustering based classification for out-of-domain documents” which aims at classifying documents across different domains. Labeled data is available in the source domain, and the target domain contains an unlabeled data set which is to be classified. The two domains are drawn from different distributions. This can occur when we, for example, consider two related web directories; one directory consists of documents about cars, and another one about trucks.

3.2 Text classification

The authors of [6] focus on linear text classification algorithms and consider the task of automatically learning a parameter function g for text classification. Given a set of example text classification problems, they try to “meta-learn” a new learning algorithm (as specified by the parameter function g), which may then be applied to new classification problems. As the meta-learning technique leverages data from a variety of related classification tasks to obtain a good classifier for new tasks, it is an instance of transfer learning. The authors believe that picking a good text classifier from the class of linear text classifiers is equivalent to finding the right parameter function for the available statistics.

Focusing on logistic regression, [7] presents a method of automatically constructing a multivariate Gaussian prior with a full covariance matrix for a given supervised learning task. The prior relaxes the independence assumption and allows parameters to be dependent. The method uses other “similar” learning problems to estimate the covariance of individual parameter pairs. Then, a semidefinite program is used to combine these estimates and learn a good prior for the learning task at hand.

3.3 Spam filtering

Transfer learning has been successfully used for spam filtering. One of the examples is described in [8] when a personalized spam filter is generalized across related learning tasks. This is done in a setting with several collections of emails that have no labeled training examples; instead, one common set of labeled data is given. The labeled data and the email collections are drawn from different distributions.

3.4 Sentiment classification

[9] states that sentiments are expressed differently in various domains, while annotating corpora for every possible domain is impractical. The work investigates domain adaptation for sentiment classifiers, focusing on online reviews for different types of products. First, the structural correspondence learning (SCL) domain adaptation algorithm is extended. A key step here is the selection of pivot features used to link the source and target domains. Secondly, the A-distance between domains is estimated as a measure of the loss due to adaptation from one domain to the other. This is done to select a subset of domains to label as sources.

3.5 Name-entity recognition

In [10] some current inductive and transductive approaches are adapted to the problem of transfer learning for name extraction. A novel maximum entropy based technique and Iterative Feature Transformation (IFT) are introduced. The work shows how simple relaxations, such as providing additional information like the proportion of positive examples in the test data, can significantly improve the performance of some of the transductive transfer learners.

3.6 Collaborative filtering

In [11] a shared rating-pattern mixture model known as a Rating-Matrix Generative Model (RMGM) is learned in terms of the latent user- and item-cluster variables. RMGM connects multiple rating matrices from different domains by mapping the users and items in each rating matrix onto the shared latent user and item spaces in order to transfer useful knowledge. In [12], co-clustering algorithms are applied on users and items in an auxiliary rating matrix. Then, a cluster-level rating matrix known as a codebook is constructed. By supposing that the target rating matrix (on one product) is related to the auxiliary one (on another product), the target domain can be rebuilt by expanding the codebook, completing the knowledge transfer process.

3.7 Cross-lingual

In [13] the authors study transfer learning for a cross-language classification problem for translating webpages from English to Chinese. The setting of the problem is such that there are plenty of labeled English text data whereas there is only a small number of labeled Chinese text documents. Transfer learning from one feature space to another is achieved by constructing a suitable mapping function as a bridge.

[14] also touches on the subject of transferring knowledge across different languages. It is done in the context of cross-lingual embedding models. Reliable cross-lingual adaptation methods can allow us to leverage the vast amounts of labeled data that we have in English and use them for any other language, particularly low-resource languages. Given the current state-of-the-art, there is still a long path to go, but some recent advances, e.g. zero-shot translation [15], are very promising. It should be noted that zero-shot learning is an example of taking transfer learning to the extreme aiming at learning from only a few, one or even zero instances of a class, hence the few-shot, one-shot, and zero-shot learning, respectively. This may be one of the hardest problems in machine learning. At the same time, it is something that comes naturally to humans which means that AI may eventually learn how to do this too.

3.9 Data Sets for Research

So far, several data sets have been published for transfer learning research [1]. In regard to text retrieval and text analytics, they include text mining, email spam-filtering, and sentiment analysis data sets.

3.10 New Approaches to Transfer Learning

The following methods may be cited as examples of relatively recently developed algorithms for transfer learning: Markov logic networks [16] (a complete MLN transfer system is offered that conducts mapping between the source MLN and the target domain and then revises the mapping results to further improve the accuracy) and Bayesian networks [17] (an algorithm for learning Bayes Net structures taking advantage of the similarity between tasks by biasing learning toward similar structures for each task).

3.11 Toolkits for transfer learning

This should probably be considered an active area of tool development. I believe the following two examples are worth mentioning here in the context of Python: PyTorch [18] and Keras [19]. The former is an example of training a CNN using transfer learning, and the latter is a transfer learning implementation using the Keras module.

4. Conclusion

The known problems with transfer learning include avoiding negative transfer between source domains / tasks and target domains / tasks. The main idea is to have suitable transferability measures to select relevant source domains. In order to do this we need to define the criteria to measure the similarity (distance) between domains or tasks based on which we can cluster domains. This may allow us to measure transferability.

Another important issue is the so called heterogeneous transfer learning when the domains have different feature spaces or when transfer occurs from multiple such source domains. Also, transfer learning methods have been mainly applied in small-scale applications with a limited variety. Hopefully, in the future it will be more broadly used to solve challenging tasks as well, e.g. social network analysis or logical inference.

Machine learning tends to be too specific to the data and requirements of the task at hand. Transfer learning is the act of abstracting away from a specific data set and reusing the knowledge obtained. In a way, this advanced approach can be compared to pioneer novel methods of building relational data models and other similar models in 1960s and 1970s when data was finally abstracted away from its storage. Here, we observe how data processing and analyzing techniques are abstracted away from the data itself.

Transfer learning refers to reuse of some or all of the training data, feature representations, neural-node layering, weights, training method, loss function, learning rate, and other properties of a prior model. However, transfer learning is more effective when it is used as a supplement to other types of learning, and not a replacement for them because together they all form the core of most data science practices. Generally, a data scientist uses transfer learning to identify statistical knowledge obtained on prior projects implemented through supervised, semi-supervised, unsupervised, or reinforcement learning.

Currently, transfer learning is associated with a breakthrough quest of the data science community to discover a “master learning algorithms” that would automatically obtain and reuse fresh contextual knowledge through deep neural networks and other forms of AI. As Andrew Ng, chief scientist at Baidu and professor at Stanford, said during his NIPS 2016 tutorial [2], transfer learning, after supervised learning, will be the next driver of the commercial success of machine learning. In his opinion, transfer learning will be a key driver of machine learning success in industry.

It is clear, that we have yet to walk a long path towards creating such a “superintelligence”, although some people already fear a robot-induced apocalypse. But it wouldn’t be too unreasonable to predict that, as the data science develops more and more statistical models based on practical knowledge, they will evolve into machine intelligence of superior performance. Transfer learning may become a tool that uses statistical knowledge to describe everything in our world.

There are different kinds of voices nowadays calling for further development of AI like Mark Zukenberg [22] and, on the contrary, making warnings about AI like the renowned scientist Hawking, tech entrepreneur Musk, and iconic geek Wozniak [20] who are afraid of weapons that could, within only a few years, select and destroy targets autonomously, without human intervention. "The endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow," they stated in a letter published at the International Joint Conference on Artificial Intelligence in Buenos Aires.

It would be unwise to deny the possibility of misuse of AI, machine learning, transfer learning, and the like. It is yet quite difficult to build the basics of ethics into machines (and, frankly speaking, today’s AI is really not at that advanced level of development when it is feasible to do so). Therefore, others believe that we are more likely to be hit by asteroids than to fall under the control of AI.

Despite the above difference in opinions, one thing remains true: there are many current and even more potential future areas where AI, along with transfer learning, may be used to effectively optimize the ongoing tasks. Transfer learning is a field of data science that attracts a lot of attention and may potentially bring very interesting and far-reaching results. We are yet to see how useful they are going to be.

References

1) Pan, S. J., & Yang, Q. A survey on transfer learning. 2010, IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359

2) Sebastian Ruder. Transfer Learning - Machine Learning's Next Frontier. 2017, http://ruder.io/transfer-learning/index.html

3) Kira Radinsky, Yoni Acriche. How to Make Better Predictions When You Don’t Have Enough Data. 2016, https://hbr.org/2016/12/how-to-make-better-predictions-when-you-dont-have-enough-data?platform=hootsuite

4) W. Dai, Q. Yang, G. Xue, and Y. Yu, “Self-taught clustering,” in Proceedings of the 25th International Conference of Machine Learning. ACM, July 2008, pp. 200–207

5) W. Dai, G. Xue, Q. Yang, and Y. Yu, “Co-clustering based classification for out-of-domain documents,” in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 2007

6) Chuong B. Do, Andrew Y. Ng. Transfer learning for text classification.

7) Rajat Raina, Andrew Y. Ng, Daphne Koller. Constructing Informative Priors using Transfer Learning. Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

8) Steffen Bickel. ECML-PKDD Discovery Challenge 2006

9) J. Blitzer, M. Dredze, and F. Pereira, “Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,” in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007, pp. 432–439

10) A. Arnold, R. Nallapati, and W. W. Cohen, “A comparative study of methods for transductive transfer learning,” in Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Washington, DC, USA: IEEE Computer Society, 2007, pp. 77–82

11) B. Li, Q. Yang, and X. Xue, “Transfer learning for collaborative filtering via a rating-matrix generative model,” in Proceedings of the 26th International Conference on Machine Learning, Montreal, Quebec, Canada, June 2009.

12) B. Li, Q. Yang, and X. Xue, “Can movies and books collaborate? - cross-domain collaborative filtering for sparsity reduction,” in Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 2009

13) X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu, “Can Chinese web pages be classified with English data source?” in Proceedings of the 17th International Conference on World Wide Web. Beijing, China: ACM, April 2008, pp. 969–978

14) Sebastian Ruder, Ivan Vulić, Anders Søgaard. A Survey of Cross-lingual Word Embedding Models. 2017

15) Johnson, M., Schuster, M., Le, Q. V, Krikun, M., Wu, Y., Chen, Z., Dean, J. Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2016

16) Mihalkova, Lilyana; Huynh, Tuyen; Mooney, Raymond J. (July 2007), "Mapping and Revising Markov Logic Networks for Transfer" (PDF), Learning Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-2007), Vancouver, BC, pp. 608–614, retrieved 2007-08-05

17) Niculescu-Mizil, Alexandru; Caruana, Rich (March 21–24, 2007), "Inductive Transfer for Bayesian Network Structure Learning" (PDF), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007), retrieved 2007-08-05

18) Transfer learning example base on Pytorch.
http://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html

19) Transfer learning example base on Keras.
https://medium.com/towards-data-science/transfer-learning-using-keras-d804b2e04ef8

20) James Kobielus. The Bogus Bogeyman of the Brainiac Robot Overlord.
http://www.dataversity.net/the-bogus-bogeyman-of-the-brainiac-robot-overlord/

21) James Kobielus. Transfer learning jump-starts new AI projects.
https://www.infoworld.com/article/3155262/analytics/transfer-learning-jump-starts-new-ai-projects.html

22) Killer robots? Musk and Zuckenberg escalate row over dangers of AI
https://www.theguardian.com/technology/2017/jul/25/elon-musk-mark-zuckerberg-artificial-intelligence-facebook-tesla