Unsupervised and Transfer Learning

Challenges in Machine Learning, Volume 7

Isabelle Guyon, Gideon Dror, Vincent Lemaire,
Graham Taylor, and Daniel Silver, editors

Nicola Talbot, production editor

Unsupervised and Transfer Learning

Challenges in Machine Learning, Volume 7

Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors

Nicola Talbot, production editor

Microtome Publishing
Brookline, Massachusetts
www.mtome.com

Unsupervised and Transfer Learning
Challenges in Machine Learning, Volume 7

Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors

Nicola Talbot, production editor

ISBN-13: 978-0-9719777-7-8

Unsupervised and Transfer Learning
Challenges in Machine Learning, Volume 7

Foreword

Over the years, machine learning has grown to be a major computer science discipline with wide applications in science and engineering. Despite its phenomenal success, a fundamental challenge to machine learning is the lack of sufficient training data to build accurate and reliable models in many practical situations. We may make the best choice of learning algorithms, but when quality data are in short supply, the resulting models can perform very poorly on a new domain.

The lack of training data problem can be taken as an opportunity to developing new theories of machine learning for developing models of complex phenonmena. As humans, we observe that we often have the ability to adapt what we have learned in one area to a new area, provided that these areas are somewhat related. We often learn in a continual fashion by standing on the shoulders of earlier models, rather than learning new models from scratch each time. We also come to recognize features that are commonly used across a set of related models. And finally, in concept learning, we often learn via not millions of examples, but through only a selected few. These phenomena call for new insights into unsupervised and transfer learning.

In the past, researchers in different subfields of machine learning have been making advances in their separate ways in the areas listed above. Recently, many of these advances have started to overlap and suggest synergistic opportunities for impact on the field. Holding a joint workshop in 2011 to bring together researchers in unsupervised and transfer learning researchers was timely. Exploring the intersection of these two fields holds strong promises for us to gain new insights, especially in their respective abilities to discover 'deep' features for domain representation. It is also an innovative idea to hold an international machine-learning contest together with the workshop, which inspires many new approaches to the difficult problem.

This volume represents a significant effort of the editors and workshop organizers, who not only put in much time in organizing the proceedings and papers, but also the challenge itself. The authors have made excellent effort in bringing us a high-quality collection with a wide coverage of topics. This collection starts with a survey paper by the editors on the state of the art of the field of unsupervised and transfer learning. It then presents papers related to theoretical advances in deep learning, model selection and clustering. The next part consists of articles by the Challenge winners on their approaches in solving the unsupervised and transfer learning contest problems. Finally, the last part consists of articles that cover various applications and specific approaches to unsupervised and transfer learning. All these articles give a complete picture of researchers' efforts for this important and challenging problem.

Looking forward, we can see several strands of emerging themes in unsupervised and transfer learning. As we march into the era of Big Data, questions on how to separate quality patterns from noise will become more pressing. The recent Google experiments on deep learning have given have show that it is possible to train a very large unsupervised neural network (16,000 computer processors with one billion connections) to automatically develop features for recognizing cat faces. The data sparsity problem associated with extremely large-scale recommendation systems provides us with strong motivation for finding new ways to transfer knowledge from auxiliary data sources. New questions about the scalability, reliability and adaptability of the unsupervised and transfer learning models will take central stage in much of ML research in the coming decade. Indeed, for unsupervised and transfer learning research, we live in very interesting times!

Qiang Yang, Huawei Noah's Ark Research Lab and Hong Kong University of Science and Technology

Preface

Machine learning is a subarea of artificial intelligence that is concerned with systems that improve with experience. It is the science of building hardware or software that can achieve tasks by learning from examples. While much of the efforts in this field over the past twenty years have been devoted to "supervised learning", that is learning under the supervision of a "teacher" providing guidance in the form of labeled data (input output pairs), recent advances in unsupervised and transfer learning have seen a complete paradigm shift in machine learning. Unsupervised learning considers the problem of discovering regularities, features or structure in unlabeled data. Transfer learning considers the use of prior knowledge, such as learned features, from one or more source tasks when developing a hypothesis for a new target task. While human beings are adept at transfer learning using mixtures of labeled and unlabeled examples, even across widely disparate domains, we have only begun to develop machine learning systems that exhibit the combined use of unsupervised learning and knowledge transfer.

This book is a result of an international challenge on Unsupervised and Transfer Learning (UTL) that culminated in a workshop of the same name at the ICML-2011 conference in Bellevue, Washington, on July 2, 2011; it captures the best of the challenge findings and the most recent research presented at the workshop.

The book is targeted for machine learning researchers and data mining practitioners interested in "lifelong machine learning systems" that retain the knowledge from prior learning to create more accurate models for new learning problems. Such systems will be of fundamental importance to intelligent software agents and robotics in the 21st century. The articles include new theories and new theoretically grounded algorithms applied to practical problems. It addressed an audience of experienced researchers in the field as well as Masters and Doctoral students undertaking research in machine learning.

The book is organized in three major sections that can be read independently of each other. The introductory chapter is a survey on the state of the art of the field of unsupervised and transfer learning providing an overview of the book articles. The first section includes papers related to theoretical advances in deep learning, model selection and clustering. The second section presents articles by the challenge winners. The final section consists of the best articles from the ICML-2011 workshop; covering various approaches to and applications of unsupervised and transfer learning.

This project was sponsored in part by the DARPA Deep Learning program and is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. The UTL Challenge was organized by ChaLearn (`http://www.chalearn.org/`); a society that works to stimulate research in the

field of machine learning through international challenges. Any opinions, findings, and conclusions or recommendations expressed in this book are those of the authors and do not necessarily reflect the views of the sponsors.

This volume reprints papers from JMLR W&CP volume 27.

November 2012

The Editorial Team:

Daniel Silver
Acadia University
danny.silver@acadiau.ca

Isabelle Guyon
Clopinet
isabelle@clopinet.com

Gideon Dror
Academic College of Tel-Aviv-Yaffo
gideon@mta.ac.il

Vincent Lemaire
Orange Labs
vincent.lemaire@orange-ftgroup.com

Graham Taylor
University of Guelph
gwtaylor@uoguelph.ca

Table of Contents

Foreword i

Introduction

ICML2011 Unsupervised and Transfer Learning Workshop 1
D.L. Silver, I. Guyon, G. Taylor, G. Dror & V. Lemaire; JMLR W&CP 27:1–16, 2012.

Fundamentals and theory

Deep Learning of Representations for Unsupervised and Transfer Learning 19
Y. Bengio; JMLR W&CP 27:17–36, 2012.

Autoencoders, Unsupervised Learning, and Deep Architectures 43
P. Baldi; JMLR W&CP 27:37–50, 2012.

Information Theoretic Model Selection for Pattern Analysis 57
J.M. Buhmann, M.H. Chehreghani, M. Frank & A.P. Streich; JMLR W&CP 27:51–64, 2012.

Clustering: Science or Art? 73
U. von Luxburg, R.C. Williamson & I. Guyon; JMLR W&CP 27:65–80, 2012.

Challenge contributions

Transfer Learning by Kernel Meta-Learning 91
F. Aiolli; JMLR W&CP 27:81–95, 2012.

Unsupervised and Transfer Learning Challenge: a Deep Learning Approach 109
G. Mesnil *et al*; JMLR W&CP 27:97–110, 2012.

Stochastic Unsupervised Learning on Unlabeled Data 125
C. Liu, J. Xie, Y. Ge & H. Xiong; JMLR W&CP 27:111–122, 2012.

Advances in transfer learning

Transfer Learning with Cluster Ensembles 139
A. Acharya, E.R. Hruschka, J. Ghosh & S. Acharyya; JMLR W&CP 27:123–132, 2012.

Divide and Transfer: an Exploration of Segmented Transfer to Detect Wikipedia Vandalism 151
S.-C. Chin & W.N. Street; JMLR W&CP 27:133–144, 2012.

Self-measuring Similarity for Multi-task Gaussian Process 165
K. Hayashi, T. Takenouchi, R. Tomioka & H. Kashima; JMLR W&CP 27:145–154, 2012.

Transfer Learning for Auto-gating of Flow Cytometry Data 177
G. Lee, L. Stoolman & C. Scott; JMLR W&CP 27:155–166, 2012.

Inductive Transfer for Bayesian Network Structure Learning 191
A. Niculescu-Mizil & R. Caruana; JMLR W&CP 27:167–180, 2012.

Unsupervised dimensionality reduction via gradient-based matrix factorization with two adaptive learning rates 207
V. Nikulin & T.-H. Huang; JMLR W&CP 27:181–194, 2012.

One-Shot Learning with a Hierarchical Nonparametric Bayesian Model 225
R. Salakhutdinov, J. Tenenbaum & A. Torralba; JMLR W&CP 27:195–206, 2012.

Multitask Learning in Computational Biology 239
C. Widmer & G. Rätsch; JMLR W&CP 27:207–216, 2012.

Transfer Learning in Sequential Decision Problems: A Hierarchical Bayesian Approach 251
A. Wilson, A. Fern & P. Tadepalli; JMLR W&CP 27:217–227, 2012.

Appendices

Datasets of the Unsupervised and Transfer Learning Challenge 265
Isabelle Guyon

ICML2011 Unsupervised and Transfer Learning Workshop

Daniel L. Silver DANNY.SILVER@ACADIAU.CA
Acadia University, Canada

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet, California, USA

Graham Taylor GWTAYLOR@CS.NYU.EDU
New York University, USA

Gideon Dror GIDEON@MTA.AC.IL
Academic College of Tel-Aviv-Yaffo, Israel

Vincent Lemaire VINCENT.LEMAIRE@ORANGE-FTGROUP.COM
Orange Labs, France

Editor: Neil Lawrence

Abstract

We organized a data mining challenge in "unsupervised and transfer learning" (the UTL challenge) followed by a workshop of the same name at the ICML 2011 conference in Bellevue, Washington[1]. This introduction presents the highlights of the outstanding contributions that were made, which are regrouped in this issue of JMLR W&CP. Novel methodologies emerged to capitalize on large volumes of unlabeled data from tasks related (but different) from a target task, including a method to learn data kernels (similarity measures) and new deep architectures for feature learning.

Keywords: transfer learning, unsupervised learning, metric learning, kernel learning, unlabeled data, challenges

1. Introduction

Unsupervised learning considers the problem of discovering regularities or structure in unlabeled data (*e.g.*, finding sub-manifolds or clustering examples) based on a representation of the domain. Transfer learning considers the use of prior knowledge (such as labeled training examples, or shared features) from one or more source tasks when developing a hypothesis for a new target task. While human beings are adept at transfer learning using mixtures of labeled and unlabeled examples, even across widely disparate domains, we have only begun to develop machine learning systems that exhibit the combined use of unsupervised learning and knowledge transfer.

To foster greater research in this area we organized a international challenge on Unsupervised and Transfer Learning that culminated in a workshop of the same name at the ICML-2011 conference in Bellevue, Washington, on July 2, 2011. This workshop

[1] http://clopinet.com/isabelle/Projects/ICML2011/.

addressed a question of fundamental and practical interest in machine learning: the development and assessment of methods that can generate data representations (features) that can be reused across domains of tasks.

This edition of JMLR W&CP presents the challenge results and a collection of outstanding contributed articles on the subject of transfer learning and unsupervised learning. This paper and the edition focuses on unsupervised and transfer learning for classification problems based on real-valued feature representations that are related more closely to data mining tasks. Methods of transfer learning have also been investigated for reinforcement learning (Ramon et al., 2007; Taylor and Stone, 2007), however these are outside the scope of this edition.

2. Overview of Transfer Learning and Unsupervised Learning

2.1. Transfer Learning

Transfer learning refers to use of knowledge for one or more source tasks to develop efficiently a more accurate hypothesis for a new target task. Transfer learning has most frequently been applied to sets of labeled data that have a supervised target value for each example. For instance, there would be significant benefit in using an accurate diagnostic model of one disease to develop a diagnostic model for a second related disease for which you have few training examples. While all learning involves generalization across problem instances, transfer learning emphasizes the transfer of knowledge across domains, tasks, and distributions that are similar but not the same. Inductive transfer has gone by a variety of names: bias learning, learning to learn, machine life-long learning, knowledge transfer, transfer learning, meta-learning, and incremental, cumulative, and continual learning.

Research in inductive transfer began in the early 1980s with discussions on inductive bias, generalization and the necessity of heuristics for developing accurate hypotheses from small numbers of training examples (Mitchell, 1980; Utgoff, 1986). This early research suggested that the accumulation of prior knowledge for the purposes of selecting inductive bias is a useful characteristic for any learning system. Following the first major workshop on inductive transfer (NIPS1995 Workshop, 1995) a series of articles were published in special issues of Connection Science (Lorien Pratt (Editor), 1996) and Machine Learning (Pratt and Sebastian Thrun (Editors), 1997), and a book entitled "Learning to Learn" (Thrun and Lorien Y. Pratt (Editors), 1997) .

Since that time, research on inductive transfer has occurred using traditional machine learning methods (Caruana, 1997; Baxter, 1997; Silver and Mercer, 1996; Heskes, 2000; Thrun and Lorien Y. Pratt (Editors), 1997; Bakker and Heskes, 2003; Ben-David and Schuller, 2003), statistical regression methods (Greene, 2002; Zellner, 1962; Breiman and Friedman, 1998), Bayesian methods involving constraints such as hyper priors (Allenby and Rossi, 1999; Arora et al., 1998; Bakker and Heskes, 2003), and more recently kernel methods such as support vector machines (SVMs) (Jebara, 2004; Allenby and Rossi, 2005). All of these approaches rely upon the development of a hypothesis for a target task under a constraint or regularization that characterizes a similarity or *relatedness* to one or more source tasks. In 2005, a second major workshop on inductive transfer occurred at NIPS. Papers from this workshop can be found in (Silver and Bennett, 2008) as well as at (NIPS2005 Workshop, 2005).

More recently, there has been work on inductive transfer in the areas of self-taught learning (Raina et al., 2007b), transductive learning (Arnold et al., 2007), *context-sensitive* multiple task learning (Silver et al., 2008), the learning of model structure (Niculescu-Mizil and Caruana, 2007), unsupervised transfer learning [Yu,Wang], and a variety of methods that mix unsupervised and supervised learning to be discussed in greater detail below.

2.2. Unsupervised Learning

Unsupervised learning refers to the process of finding structure in unlabeled data resulting in new data representations (including feature representations) and/or clustering data into categories of similar examples, based on such representations (Hinton and Sejnowski, 1999). The unlabeled data distinguishes unsupervised learning from supervised learning and reinforcement learning. Important recent progress has been made in purely unsupervised learning (Smola et al., 2001; Bengio et al., 2003; Globerson and Tishby, 2003; Ghahramani, 2004; Luxburg, 2007). However, these advances tend to be ignored by practitioners who continue using a handful of popular algorithms like PCA and ICA (for feature extraction and dimensionality reduction), and K-means, and various hierarchical clustering methods for clustering (Jain et al., 1999).

2.3. Combining Unsupervised and Transfer Learning

It is often easier to obtain large quantities of unlabeled data from databases and sources on the web, for example images of unlabeled objects. For this reason the idea of using unsupervised learning in combination with supervised learning has attracted interest for some time. Semi-supervised learning is a machine learning approach that is halfway between supervised and unsupervised learning. In addition to the labeled data for a given task of interest, the algorithm is provided with unlabeled data for the *same* task - typically a small amount of labeled data and a large amount of unlabeled data (Blum and Mitchell, 1998). Note that these approaches usually assume that the categories of the unlabeled data, even though unknown to the learning machine, are the same as the categories of the labeled data, *i.e.,* that the "tasks" are the same.

In contrast, in the transfer learning setting, the unlabeled data does not need to come from the same task. There has been considerable progress in the past decade in developing cross-task transfer using both discriminative and generative approaches in a wide variety of settings (Pan and Yang, 2010). These approaches include multi-layer structured learning machines from the "Deep Learning" family such as convolutional neural networks, Deep Belief Networks, and Deep Boltzmann Machines (Bengio, 2009; Gutstein, 2010; Erhan et al., 2010), sparse coding (Lee et al., 2007; Raina et al., 2007a), and metric or kernel learning methods (Bromley et al., 1994; WU et al., 2009; Kulis, 2010). The "Learning to learn" and "Lifelong Learning" veins of research have continued to provide interesting results in both machine learning and cognitive science in terms of short-term learning with transfer and long-term retention of learned knowledge (Silver et al., 2008). These references include recent evidence of the value of combining unsupervised generative learning with transfer learning to generate a rich set of representation (features) upon which to build related supervised discriminative tasks. The goal of the challenge we organized was to perform an evaluation of unsupervised and transfer learning algorithms free of inventor bias to help to identify and popularize algorithms that have advanced the state of the art.

3. Overview of the UTL Challenge

Part of the ICML workshop was devoted to the presentation of the results of the Unsupervised and Transfer Learning challenge (UTL challenge Guyon et al., 2011a,b). The challenge, which started in December 2010 and ended in April 2011, was organized in 2 phases. The aim of **Phase 1** was to benchmark **unsupervised learning** algorithms used as preprocessors for supervised learning, in the context of transfer learning problems. The aim of **phase 2** was to encourage researchers to exploit the possibilities offered by new cutting-edge cross-task transfer learning algorithms, which **transfer supervised learning knowledge from task to task**.

To that end, the competitors were presented with five datasets illustrating classification problems from different domains: handwriting recognition, video processing, text processing, object recognition, and ecology. Each dataset was split into 3 subsets: development, validation, and final evaluation sets. In phase 1, all subsets were provided without labels to the participants. The labels remained known only to the organizers throughout the challenge. The goal of the participants was to produce the best possible data representation for the final evaluation data. This representation was then evaluated by the organizers on supervised learning classification tasks by training and testing a linear classifier on subsets of the final evaluation data, such than a learning curve would be produced. The evaluation metric was the area under the learning curve, which is a means of aggregating performance results over a range of number of training examples considered.

To avoid the possibility of participants selecting their model based on final evaluation set performance, the final results remained secret until the end of the challenge. Rather, feed-back was provided on-line during the challenge on the performance obtained on validation data, and the final evaluation set data was used only for the final ranking. For both phases, the participants could either submit a data representation (for validation data and final evaluation data) or a matrix of similarity between examples (a kernel). Hence, the competition was equivalently a data representation learning challenge and a kernel learning challenge.

In contrast with a classical evaluation of unsupervised learning as a preprocessing, the three subsets (development, validation, and final evaluation sets) were **not drawn from the same distribution**. In fact, they all had different sets of class labels. Picture for instance a problem of optical character recognition (OCR), the development set could contain only lowercase alphabetical letters, the validation set could contain uppercase letters, and the final evaluation set, digits and symbols. This setting is typical of real world problems in which there is an abundance of data available for training from a source domain, which is distinct from the target domain of interest. For instance, in face recognition, there is an abundance of pictures from unknown strangers that are available on the Internet, compared to the few images of your close family members that you care to classify. The development set represents a source domain whereas the validation and final evaluation sets represent alternative target domains on which different sets of tasks can be defined[2].

In the second phase of the challenge, a few labels of the development set were provided, offering to the participants the possibility of using supervised learning in some way to produce better data representations for the validation and final evaluation sets. The setting remained otherwise unchanged.

One of the main findings of this challenge is the power of unsupervised learning as a preprocessing tool. For all the datasets of the challenge, unsupervised learning produced results significantly better than the baseline methods (raw data or simple normalizations). The participants exploited effectively the feed-back received on the validation set to select the best data representations. The skepticism around the effectiveness of unsupervised learning is justified when no performance on a supervised task is available. However, unsupervised learning can be the object of model selection using a supervised task, similarly to preprocessing, feature selection, and hyperparameter selection. An interesting new outcome of this challenge is that the supervised tasks used for model selection can be distinct from the tasks used for the final evaluation. So, even though the learning algorithms are unsupervised, transfer learning is happening at the model selection level. This setting is related to the "self-taught learning" setting proposed in (Raina et al., 2007a). Another interesting finding is that, perhaps the development set is not useful at all. The winners of phase 1 did not use it. They devised a method to select a cascade of preprocessing steps to be used to produce a new kernel. The same cascade was then applied to produce the kernel of the final evaluation set(Aiolli, 2012). The importance of the degree of resemblance of the validation task and final task remains to be determined.

In phase 1, there was a danger of overfitting by trying too many methods and relying too heavily on the performance on the validation set. One team for instance overfitted in phase 1, ranking 1st on the validation set, but only 4th on the final evaluation set. Possibly, criteria involving both the reconstruction error and the classification accuracy on the validation tasks may be more effective for model selection. This should be the object of further research. In phase 2, the participants had available "transfer labels" for a subset of the development data (for classification tasks distinct from the classification tasks of the validation set and the final evaluation set). Therefore, they had the

[2] In this paper, we call "domain" the input space (*e.g.*, a feature vector space) and we call "task" the output space (represented by labels for classification problems). We use the adjective "source" for an auxiliary problem, for which we have an abundance of data (*e.g.*, pictures of strangers in the Internet), and "target" for the problem of interest (*e.g.*, pictures of family members).

opportunity to use such labels to devise transfer learning strategies. The most effective strategy seems to have been to use the transfer labels for model selection again. None of the participants used those labels for learning.

Overall, an array of algorithms were used (Aiolli, 2012; Le Borgne, 2011; Liu et al., 2012; Mesnil et al, 2012; Saeed, 2011; Xu et al, 2011), including linear methods like Principal Component Analysis (PCA), and non-linear methods like clustering (K-means and hierarchical clustering being the most popular), Kernel-PCA (KPCA), non-linear auto-encoders and restricted Bolzmann machines (RBMs). A general methodology seems to have emerged. Most top ranking participants used simple normalizations (like variable standardization and/or data sphering using PCA) as a first step, followed by one or several layers of non-linear processing (stacks of auto-encoders, RBMs, KPCA, and/or clustering). Finally, "transduction" played a key role in winning first place: either the whole preprocessing chain was applied directly to the final evaluation data (this is the strategy of Fabio Aiolli who won first place in phase 1, Aiolli, 2012); or alternatively, the final evaluation data, preprocessed with a preprocessor trained on development+validation data, was post-processed with PCA (so-called "transductive PCA" used by the LISA team, who won the second phase, Mesnil et al, 2012).

4. Overview of Proceedings

The following provides an overview of the workshop proceedings including the tutorials, invited presentations, challenge winner articles and other refereed articles submitted to the workshop.

4.1. Tutorials

The workshop provided two foundational tutorials included in this proceeding. The morning tutorial covered *Deep Learning of Representations for Unsupervised and Transfer Learning* with Yoshua Bengio from the Université de Montréal (Bengio, 2012). Deep learning algorithms seek to exploit the unknown structure in the input distribution in order to discover good representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features. The paper focusses on why unsupervised pre-training of representations using autoencoders and Restricted Boltzmann Machines can be useful, and how it can be exploited in the transfer learning scenario, where we care about predictions on examples that are not from the same distribution as the training distribution.

The afternoon tutorial entitled *Towards Heterogeneous Transfer Learning* was presented by Qiang Yang, Hong Kong University of Science, co-author of an authoritative review of transfer learning (Pan and Yang, 2010). Transfer learning has focused on knowledge transfer between domains with the same or similar input spaces. The heterogeneous transfer approach considers the ability to use knowledge from very different task domains and input spaces. The authors demonstrated heterogeneous transfer learning between text classification and image classification domains even when there are no explicit feature mappings provided. They explained that the key is to identify and