Computer Vision and Image Understanding 000 (2016) 1–12

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Scalable greedy algorithms for transfer learning

Ilja Kuzborskij[a,b,d,*], Francesco Orabona[c], Barbara Caputo[a,d]

[a] Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland
[b] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[c] Stony Brook University, Dept. of Computer Science, Stony Brook, NY 11794, USA
[d] University of Rome La Sapienza, Dept. of Computer, Control and Management Engineering, Rome, Italy

ARTICLE INFO

Article history:
Received 8 October 2015
Revised 14 June 2016
Accepted 7 September 2016
Available online xxx

Keywords:
Transfer learning
Domain adaptation
Visual object detection
Greedy algorithms
Feature selection

ABSTRACT

In this paper we consider the binary transfer learning problem, focusing on how to select and combine sources from a large pool to yield a good performance on a target task. Constraining our scenario to real world, we do not assume the direct access to the source data, but rather we employ the source hypotheses trained from them. We propose an efficient algorithm that selects relevant source hypotheses and feature dimensions simultaneously, building on the literature on the best subset selection problem. Our algorithm achieves state-of-the-art results on three computer vision datasets, substantially outperforming both transfer learning and popular feature selection baselines in a small-sample setting. We also present a randomized variant that achieves the same results with the computational cost independent from the number of source hypotheses and feature dimensions. Also, we theoretically prove that, under reasonable assumptions on the source hypotheses, our algorithm can learn effectively from few examples.

1. Introduction

Over the last few years, the visual recognition research landscape has been heavily dominated by Convolutional Neural Networks, thanks to their ability to leverage effectively over massime amount of training data (Donahue et al., 2014). This trend dramatically confirms the widely accepted truth that any learning algorithm performs better when trained on a lot of data. This is even more true when facing noisy or “hard” problems such as large-scale recognition (Deng et al., 2009). However, when tackling large scale recognition problems, gathering substantial training data for all classes considered might be challenging, if not almost impossible. The occurrence of real-world objects follows a long tail distribution, with few objects occurring very often, and many with few instances. Hence, for the vast majority of visual categories known to human beings, it is extremely challenging to collect training data of the order of $10^4 - 10^5$ instances. The “long tail” distribution problem was noted and studied by Salakhutdinov et al. (2011), who proposed to address it by leveraging on the prior knowledge available to the learner. Indeed, learning systems are often not trained from scratch: usually they can be build on previous knowledge acquired over time on related tasks (Pan and Yang, 2010). The scenario of learning from few examples by *transferring* from what is already known to the learner is collectively known as Transfer Learning. The target domain usually indicates the task at hand and the source domain the prior knowledge of the learner.

Most of the transfer learning algorithms proposed in the recent years focus on the object detection task (binary transfer learning), assuming access to the training data coming from both source and target domains (Pan and Yang, 2010). While featuring good practical performance (Gong et al., 2012), they often demonstrate poor scalability w.r.t. the number of sources. An alternative direction, known as a Hypothesis Transfer Learning (HTL) (Ben-David and Urner, 2013; Kuzborskij and Orabona, 2013), consists in transferring from the *source hypotheses*, that is classifiers trained from them. This framework is practically very attractive (Aytar and Zisserman, 2011; Kuzborskij et al., 2013; Tommasi et al., 2014), as it treats source hypotheses as black boxes without any regard of their inner workings.

The goal of this paper is to develop an HTL algorithm able to deal effectively and efficiently with a large number of sources, where our working definition of large is at least 10^3. Note that this order of magnitude is also the current frontier in visual classification (Deng et al., 2009). To this end, we cast Hypothesis Transfer Learning as a problem of *efficient selection* and *combination* of source hypotheses from a large pool. We pose it as a subset selection problem building on results from the literature (Das and Kempe, 2008; Zhang, 2009a). We present[1] a greedy algorithm,

* Corresponding author.
E-mail addresses: ilja.kuzborskij@idiap.ch (I. Kuzborskij), francesco@orabona.com (F. Orabona), caputo@dis.uniroma1.it (B. Caputo).

[1] We build upon preliminary results presented in Kuzborskij et al. (2015).

http://dx.doi.org/10.1016/j.cviu.2016.09.003
1077-3142/© 2016 Elsevier Inc. All rights reserved.

Please cite this article as: I. Kuzborskij et al., Scalable greedy algorithms for transfer learning, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.09.003

GreedyTL, which attains state of the art performance even with a very limited amount of data from the target domain. Morever, we also present a randomized approximate variant of GreedyTL, called GreedyTL-59, that has a complexity *independent* from the number of sources, with no loss in performance. Our key contribution is a $L2$-regularized variant of the Forward Regression algorithm (Hastie et al., 2009). Since our algorithm can be viewed as a feature selection algorithm as well as an hypothesis transfer learning approach, we extensively evaluate it against popular feature selection and transfer learning baselines. We empirically demonstrate that GreedyTL dominates all the baselines in most small-sample transfer learning scenarios, thus proving the critical role of regularization in our formulation. Experiments over three datasets show the power of our approach: we obtain state of the art results in tasks with up to 1000 classes, totalling 1.2 million examples, with only 11 to 20 training examples from the target domain. We back our experimental results by proving generalization bounds showing that, under reasonable assumptions on the source hypotheses, our algorithm is able to learn effectively with very limited data.

The rest of the paper is organised as follows: after a review of the relevant literature in the field (Section 2), we cast the transfer learning problem in the subset selection framework (Section 3). We then define our GreedyTL, in Section 4, deriving its formulation, analysing its computational complexity and its theoretical properties. Section 5 describes our experimental evaluation and discuss the related findings. We conclude with an overall discussion and presenting possible future research avenues.

2. Related work

The problem of how to exploit prior knowledge when attempting to solve a new task with limited, if any, annotated samples is vastly researched. Previous work span from transfer learning (Pan and Yang, 2010) to domain adaptation (Ben-David et al., 2010; Saenko et al., 2010), and dataset bias (Torralba and Efros, 2011). Here we focus on the first. In the literature there are several transfer learning settings (Ben-David et al., 2010; Gong et al., 2012; Saenko et al., 2010). The oldest and most popular is the one assuming access to the data originating from both the source and the target domains (Ben-David et al., 2010; Duan et al., 2012; Gong et al., 2012; Kuzborskij et al., 2016; Saenko et al., 2010; Seah et al., 2011; Tommasi and Caputo, 2013). There, one typically assumes that plenty of source data are available, but access to the target data is limited: for instance, we can have many unlabeled examples and only few labeled (Patel et al., 2015). Here we focus on the Hypothesis Transfer Learning framework (HTL, (Ben-David and Urner, 2013; Kuzborskij and Orabona, 2013)). It requires to have access only to *source hypotheses*, that is classifiers or regressors trained on the source domains. No assumptions are made on how these source hypotheses are trained, or about their inner workings: they are treated as “black boxes”, in spirit similar to classifier-generated visual descriptors such as Classemes (Bergamo and Torresani, 2014) or Object-Bank (Li et al., 2010). Several works proposed HTL for visual learning (Aytar and Zisserman, 2011; Oquab et al., 2014; Tommasi et al., 2014), some exploiting more explicitly the connection with classemes-like approaches (Jie et al., 2011; Patricia and Caputo, 2014), demonstrating an intriguing potential. Although offering scalability, HTL-based approaches proposed so far have been tested on problems with less than a few hundred of sources (Tommasi et al., 2014), already showing some difficulties in selecting informative sources.

Recently, the growing need to deal with large data collections (Choi et al., 2010; Deng et al., 2009) has started to change the focus and challenges of research in transfer learning. Scalability with respect to the amount of data and the ability to identify and separate informative sources from those carrying noise for the task at hand have become critical issues. Some attempts have been made in this direction. For example, Lim et al. (2011); Vezhnevets and Ferrari (2014) used taxonomies to leverage learning from few examples on the SUN09 dataset. In Lim et al. (2011), authors attacked the transfer learning problem on the SUN09 dataset by using additional data from another dataset. Zero-shot approaches were investigated by Rohrbach et al. (2011) on a subset of the Imagenet dataset. Large-scale visual detection has been explored by Vezhnevets and Ferrari (2014). However, all these approaches assume access to all source training data. A slightly different approach to transfer learning that aimed to cirumvent this limitation, is reuse of a large convolutional neural network pre-trained on a large visual recognition dataset. The simplest approach is to use outputs of intermediate layers of such a network, such as DeCAF (Donahue et al., 2014) or Caffe (Jia et al., 2014). A more sophisticated way of reuse is fine-tuning, a kind of warm-start, that has been successfully exploited in visual detection (Girshick et al., 2015) and domain adaptation (Ganin and Lempitsky, 2015; Long et al., 2015).

In many of these works the use of richer sources of information has been supported by an increase in the information available in the target domain as well. From an intuitive point of view, this corresponds to having more data points than dimensions. Of course, this makes the learning and selection process easier, but in many applications it is not a reasonable hypothesis. Also, none of the proposed algorithms has a theoretical backing.

While not explicitly mentioned before, the problem outlined above can also be viewed as a learning scenario where the number of features is by far larger than the number of training examples. Indeed, learning with classeme-like features (Bergamo and Torresani, 2014; Li et al., 2010) when only few training examples are available can be seen as a Hypothesis Transfer Learning problem. Clearly, a pure empirical risk minimization would fail due to severe overfitting. In machine learning and statistics this is known as a feature selection problem, and is usually addressed by constraining or penalizing the solution with sparsity-inducing norms. One important sparsity constraint is a non-convex $L0$ pseudo-norm constraint $\|\boldsymbol{w}\|_0 \leq k$, that simply corresponds to choosing up to k non-zero components of a vector $\boldsymbol{w}$. One usually resorts to the *subset selection* methods, and greedy algorithms for obtaining solutions under this constraint (Das and Kempe, 2008; 2011; Zhang, 2009a; 2009b). However, in some problems introducing $L0$ constraint might be computationally difficult. There, a computationally easier alternative is a convex relaxation of $L0$, the $L1$ regularization. Empirical error minimization with $L1$ penalty with various loss functions (for square loss is known as Lasso) has many favorable properties and is well studied theoretically (Bühlmann and Van De Geer, 2011). Yet, $L1$ penalty is known to suffer from several limitations, one of which is poor empirical performance when there are many correlated features. Perhaps the most famous way to resolve this issue is an *elastic net* regularization which is a weighted mixture of $L1$ and squared $L2$ penalties (Hastie et al., 2009). Since our work partially falls into the category of feature selection, we have extensively evaluated the aforementioned baselines in our task. As it will be shown below, none of them achieves competitive performances compared to our approach.

3. Transfer learning through subset selection

Definitions. We will denote with small and capital bold letters respectively column vectors and matrices, e.g. $\boldsymbol{a} = [a_1, a_2, \ldots, a_d]^T \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$. The subvector of $\boldsymbol{a}$ with rows indexed by set S is $\boldsymbol{a}_S$, while the square submatrix of $\boldsymbol{A}$ with rows and columns indexed by set S is $\boldsymbol{A}_S$. For $\boldsymbol{x} \in \mathbb{R}^d$, the *support* of $\boldsymbol{x}$ is $\text{supp}(\boldsymbol{x}) = \{i \in \{1, \ldots, d\} : x_i \neq 0\}$. Denoting by $\mathcal{X}$ and $\mathcal{Y}$ re-

Please cite this article as: I. Kuzborskij et al., Scalable greedy algorithms for transfer learning, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.09.003