Transfer Learning through Greedy Subset Selection

Ilja Kuzborskij[1], Francesco Orabona[2] and Barbara Caputo[1,3]

[1] Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
ilja.kuzborskij@idiap.ch
[2] Yahoo! Labs, 229 West 43rd Street, 10036 New York, NY, USA
francesco@orabona.com
[3] University of Rome La Sapienza, Dept. of Computer, Control and Management Engineering, Rome, Italy
caputo@dis.uniroma1.it

Abstract. We study the binary transfer learning problem, focusing on how to select sources from a large pool and how to combine them to yield a good performance on a target task. In particular, we consider the transfer learning setting where one does not have direct access to the source data, but rather employs the source hypotheses trained from them. Building on the literature on the best subset selection problem, we propose an efficient algorithm that selects relevant source hypotheses and feature dimensions simultaneously. On three computer vision datasets we achieve state-of-the-art results, substantially outperforming transfer learning and popular feature selection baselines in a small-sample setting. Also, we theoretically prove that, under reasonable assumptions on the source hypotheses, our algorithm can learn effectively from few examples.

1 Introduction

It is a truth universally acknowledged that learning algorithms perform better when trained on a lot of data. This is even more true when facing noisy or "hard" problems such as large-scale visual recognition [7]. However, considering object detection tasks, access to training data might be restricted. As noted in [23], the distribution of real-world objects is highly skewed, with few objects occurring very often, and many with few instances. Moreover, learning systems are often not trained from scratch: usually they can be build on previous knowledge acquired over time on related tasks [21]. 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 assume access to the training data coming from both source and target domains [21]. While featuring good practical performance [11], and well understood theoretical guarantees [2], they often demonstrate poor scalability w.r.t. number of sources. An alternative direction, known as a Hypothesis Transfer Learning (HTL) [15, 3], consists in transfering from the *source hypotheses*, that is classifiers trained from them. This framework is practically very attractive [1, 25, 16], 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. 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 and build on results from the literature [6, 28]. We develop a greedy algorithm, `GreedyTL`, which attains the state of art performance given a very limited amount of data from the target domain, while able to scale well over the large number of sources. Our key contribution is a $L2$-regularized variant of the Forward Regression algorithm [13]. Since our algorithm can be viewed both as feature selection and hypothesis transfer learning algorithm, we extensively evaluate it against popular feature selection and transfer learning baselines. We empirically demonstrate that all baselines but `GreedyTL`, fail 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 a very limited data.

2 Related Work

In the literature there are several transfer learning settings [21, 2, 22, 11]. We focus on the Hypothesis Transfer Learning framework (HTL, [15, 3]). There, it is required 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, on the independence of their underlying distribution from that of the target, or about their inner workings: they are treated as "black boxes", in spirit similar to classifier-generated visual descriptors such as Classemes [4] or Object-Bank [17]. Several works proposed HTL for visual learning [1, 24, 20], some exploiting more explicitly the connection with classemes-like approaches [14], 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 [25], already showing some difficulties in selecting informative sources.

Recently, the growing need to deal with large data collections [7, 5] 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 [18, 26]. However, all these approaches assume access to all source training data. Moreover, 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. On the other hand, HTL-based approaches proposed so far have been tested only on problems with less than a few hundred of sources [25], already showing some difficulties in selecting informative sources.

3 Transfer Learning through Subset Selection

Definitions. We will denote with small and capital bold letters respectively column vectors and matrices, e.g. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_d]^T \in \mathbb{R}^d$ 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 $\text{supp}(\boldsymbol{x}) = \{i : x_i \neq 0, i \in \{1, \ldots, d\}\}$. Denoting by $\mathcal{X}$ and $\mathcal{Y}$ respectively the input and output space of the learning problem, the training set is $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, drawn i.i.d. from the probability distribution p defined over $\mathcal{X} \times \mathcal{Y}$. We will focus on the binary classification problem so $\mathcal{Y} = \{-1, 1\}$, and, without loss of generality, $\mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq 1, \boldsymbol{x} \in \mathbb{R}^d\}$.

To measure the accuracy of a learning algorithm, we have a non-negative *loss* function $\ell(h(\boldsymbol{x}), y)$, which measures the cost incurred predicting $h(\boldsymbol{x})$ instead of y. In particular, we will focus on the square loss, $\ell(h(\boldsymbol{x}), y) = (h(\boldsymbol{x}) - y)^2$, for its appealing computational properties. The *risk* of a hypothesis h, with respect to the probability distribution p, is then defined as $R(h) := \mathbb{E}_{(\boldsymbol{x},y)\sim p}[\ell(h(\boldsymbol{x}), y)]$, while the *empirical risk* given a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ is $\hat{R}(h) := \frac{1}{m}\sum_{i=1}^m \ell(h(\boldsymbol{x}_i), y_i)$. Whenever the hypothesis is a linear predictor, that is, $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, we will also use risk notation as $R(\boldsymbol{w}) = R(h_{\boldsymbol{w}})$ and $\hat{R}(\boldsymbol{w}) = \hat{R}(h_{\boldsymbol{w}})$.

Source Selection. Assume, that we are given a finite source hypothesis set $\{h_i^{\text{src}}\}_{i=1}^n$ and the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$. As in previous works [19, 25, 14], we consider the target hypothesis to be of the form

$$h^{\text{trg}}_{\boldsymbol{w},\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + \sum_{i=1}^n \beta_i h_i^{\text{src}}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{w}$ and $\boldsymbol{\beta}$ are found by the learning procedure. The essential parameter here is $\boldsymbol{\beta}$, that is the one controlling the influence of each source hypothesis. Previous works in transfer learning have focused on finding $\boldsymbol{\beta}$ such that minimizes the error on the training set, subject to some condition on $\boldsymbol{\beta}$. In particular, [25] have proposed to minimize the leave-one-out error w.r.t. $\boldsymbol{\beta}$, subject to $\|\boldsymbol{\beta}\|_2 \leq \tau$, which is known to improve generalization for the right choice of τ [15]. A slightly different approach is to use $\|\boldsymbol{\beta}\|_1 \leq \tau$ regularization for this purpose [25], which is known to prefer $\boldsymbol{\beta}$ with the most coefficients equal to 0, thus assuming that the optimal $\boldsymbol{\beta}$ is sparse. Nonetheless, it is not clear whether transfer learning tasks are always truly sparse in practice.

In this work we embrace a weaker assumption, namely, there exist up to k sources that collectively improve the generalization on the target domain. Thus, we pose the problem of the Source Selection as a minimization of the regularized empirical risk on the target training set, while constraining the number of selected source hypotheses.

k-Source Selection. *Given the training set* $\left\{\left([\boldsymbol{x}_i^\top, h_1^{src}(\boldsymbol{x}_i), \ldots, h_n^{src}(\boldsymbol{x}_i)]^\top, y_i\right)\right\}_{i=1}^m$ *we have the optimal target hypothesis* $h^{trg}_{\boldsymbol{w}^\star,\boldsymbol{\beta}^\star}$ *by solving,*

$$\begin{aligned}(\boldsymbol{w}^\star, \boldsymbol{\beta}^\star) = \underset{\boldsymbol{w},\boldsymbol{\beta}}{\arg\min}\ & \left\{\hat{R}(h^{trg}_{\boldsymbol{w},\boldsymbol{\beta}}) + \lambda\|\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\right\}, \\ \text{s.t}\ & \|\boldsymbol{w}\|_0 + \|\boldsymbol{\beta}\|_0 \leq k.\end{aligned} \tag{2}$$

Notably, the problem (2) is a special case of the *Subset Selection* problem [6]: choose a subset of size k from the n observation variables, which collectively give the best prediction on the variable of interest. However, the Subset Selection problem is **NP**-hard [6]. In practice we can resort to algorithms generating approximate solutions, for many of which we have approximation guarantees. Hence, due to the extensive practical and theoretical results, we will treat the k-Source Selection as a Subset Selection problem, building atop of existing guarantees.

We note that our formulation, (2), differs from the classical subset selection for the fact that it is $L2$-regularized. This technical difference practically and theoretically makes an essential difference and it is the crucial part of our algorithm. First, $L2$ regularization is known to improve the generalization ability of empirical risk minimization. Second, we show that regularization also improves the quality of the approximate solution in situations when the sources, or features, are correlated. At the same time, the experimental evaluation corroborates our theoretical findings: Our formulation substantially outperforms standard subset selection, feature selection algorithms, and transfer learning baselines.

4 Greedy Algorithm for k-Source Selection

In this section we state the algorithm proposed in this work, `GreedyTL`[4].

GreedyTL. *Let $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{y} \in \mathbb{R}^m$ be the standardized training set, $\{h_i^{src}\}_{i=1}^n$, source hypothesis set, and k and λ, regularization parameters. Then, denote $\boldsymbol{C} = \boldsymbol{Z}^\top \boldsymbol{Z}$ and $\boldsymbol{b} = \boldsymbol{Z}^\top \boldsymbol{y}$, where $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} & \begin{matrix} h_1^{src}(\boldsymbol{x}_1) & \cdots & h_n^{src}(\boldsymbol{x}_1) \\ \cdots & \cdots & \cdots \\ h_1^{src}(\boldsymbol{x}_m) & \cdots & h_n^{src}(\boldsymbol{x}_m) \end{matrix} \end{bmatrix}$, and select set S of size k as follows: (I) Initialize $S \leftarrow \varnothing$ and $U \leftarrow \{1, \ldots, n+d\}$. (II) Keep populating S with $i \in U$, that maximize $\boldsymbol{b}_S^\top ((\boldsymbol{C} + \lambda \boldsymbol{I})_S^{-1})^\top \boldsymbol{b}_S$, as long as $|S| \leq k$ and U is non-empty.*

Derivation of the Algorithm. We derive `GreedyTL` by extending the well known Forward Regression (FR) algorithm [6], which gives an approximation to the subset selection problem, the problem of our interest. FR is known to find good approximation as far as features are uncorrelated [6]. In the following, we build upon FR by introducing a Tikhonov ($L2$) regularization into the formulation. The purpose of regularization is twofold: first, it improves the generalization ability of the empirical risk minimization, and second, it makes the algorithm more robust to the feature correlations, thus opting to find better approximate solution.

First, we briefly formalize the subset selection problem. In a subset selection problem one tries to achieve a good prediction accuracy on the *predictor* random variable Y, given a linear combination of a subset of the *observation* random variables $\{X_i\}_{i=1}^n$. The least squares subset selection then reads as

$$\min_{|S|=k, \boldsymbol{w} \in \mathbb{R}^k} \mathbb{E}\left[(Y - \sum_{i \in S} w_i X_i)^2 \right].$$

[4] Source code is available at http://idiap.ch/∼ikuzbor/

Now denote the covariance matrix of zero-mean unit-variance observation random variables by $\boldsymbol{C}$, and the covariances between Y and $\{X_i\}_{i=1}^m$ as $\boldsymbol{b}$. By virtue of the analytic solution to least-squares and using the introduced notation, we can also state the equivalent *Subset Selection problem*: $\max_{|S|=k} \boldsymbol{b}_S^\top (\boldsymbol{C}_S^{-1})^\top \boldsymbol{b}_S$. However, our goal is to obtain the solution to (2), or a *$L2$-regularized* subset selection. Similarly to the unregularized subset selection, it is easy to get that (2) is equivalent to $\max_{|S|=k} \boldsymbol{b}_S^\top ((\boldsymbol{C}_S + \lambda \boldsymbol{I})^{-1})^\top \boldsymbol{b}_S$. As said above, the Subset Selection problem is **NP**-hard, however, there are number ways to approximate it in practice [13]. We choose FR for this task for its simplicity, appealing computational properties and provably good approximation guarantees. Now, to apply FR to our problem, all we have to do is to provide it with normalized $(\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}$ instead of $\boldsymbol{C}^{-1}$.

In the basic formulation, FR requires to invert the covariance matrix at each iteration of a greedy search. Clearly, this naive approach gets prohibitive with the growth of both the number of variables and desired subset size, since its computational complexity would be in $\mathcal{O}(k(d+n)^4)$. However, we note that in transfer learning one typically assumes that training set is much smaller than sources and feature dimension. For this reason we apply rank-one updates w.r.t. the dual solution of regularized subset selection, so that the size of the inverted matrix does not change. The computational complexity then improves to $\mathcal{O}(km^2(d+n)^2)$.

Theoretical Guarantees. We now focus on the analysis of the generalization properties of `GreedyTL` for solving k-Source Selection problem (2). Throughout this paragraph we will consider a truncated target predictor $h^{\text{trg}}_{\boldsymbol{w},\boldsymbol{\beta}}(\boldsymbol{x}) := \mathtt{T}\left(\boldsymbol{w}^\top \boldsymbol{x} + \sum_{i=1}^n \beta_i h_i^{\text{src}}(\boldsymbol{x})\right)$, with $\mathtt{T}(a) := \min\{\max\{a, -1\}, 1\}$. First we state the bound on the risk of an approximate solution returned by `GreedyTL`. [5]

Theorem 1. *Let* `GreedyTL` *generate the solution* $(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\beta}})$*, given the training set* $(\boldsymbol{X}, \boldsymbol{y})$*, source hypotheses* $\{h_i^{src}\}_{i=1}^n$ *with* $\tau_\infty^{src} := \max_i\{\|h_i^{src}\|_\infty^2\}$*, hyperparameters* λ *and* k*. Then with high probability,*

$$R\left(h^{trg}_{\hat{\boldsymbol{w}},\hat{\boldsymbol{\beta}}}\right) - \hat{R}\left(h^{trg}_{\hat{\boldsymbol{w}},\hat{\boldsymbol{\beta}}}\right) \leq \tilde{\mathcal{O}}\left(\frac{1+k\tau_\infty^{src}}{\lambda m} + \sqrt{\hat{R}^{src}\frac{1+k\tau_\infty^{src}}{\lambda m}}\right),$$

where $\hat{R}^{src} := \frac{1}{m}\sum_{i=1}^m \ell\left(y_i, \mathtt{T}\left(\sum_{j\in \text{supp}(\hat{\boldsymbol{\beta}})} \hat{\beta}_i h_j^{src}(\boldsymbol{x}_i)\right)\right)$.

This results in a generalization bound which tells us how close the performance of the algorithm on the test set will be to the one on the training set. The key quantity here is $\hat{R}^{\text{src}}$, which captures the quality of the sources selected by the algorithm. To understand its impact, assume that $\lambda = \mathcal{O}(1)$. The bound has two terms, a fast one of the order of $\tilde{\mathcal{O}}(k/m)$ and a slow one of the order $\tilde{\mathcal{O}}(\sqrt{\hat{R}^{\text{src}}k/m})$. When m goes to infinity and $\hat{R}^{\text{src}} \neq 0$ the slow term will dominate the convergence rate, giving us a rate of the order of $\tilde{\mathcal{O}}(\sqrt{\hat{R}^{\text{src}}k/m})$. If $\hat{R}^{\text{src}} = 0$ the slow term completely disappears, giving us a so called fast rate of convergence of $\tilde{\mathcal{O}}(k/m)$. On the other hand, for any finite m

[5] Proofs for theorems can be found in the supplementary material.

if $\hat{R}^{\text{src}}$ is small enough, in particular of the order of $\tilde{\mathcal{O}}(k/m)$, we still have a rate of the order of $\tilde{\mathcal{O}}(k/m)$. Hence, the quantity $\hat{R}^{\text{src}}$ will govern the finite sample and asymptotic behavior of the algorithm, predicting a faster convergence in both regimes when it is small. In other words, when the source and target tasks are similar, TL facilitates a faster convergence of the empirical risk to the risk. A similar behavior was already observed in [15, 3].

However, one might ask what happens when the selected sources are providing bad predictions? Since $\hat{R}^{\text{src}} \leq 1$, due to truncation, the empirical risk converges to the risk at the standard rate $\tilde{\mathcal{O}}(\sqrt{k/m})$, the same one we would have without any transfering from the sources classifiers.

We now present another result that upper bounds the difference between the risk of solution of the algorithm and the empirical risk of the optimal solution to the k-Source Selection problem.

Theorem 2. *In addition to conditions of Theorem 1, let $(\boldsymbol{w}^\star, \boldsymbol{\beta}^\star)$ be the optimal solution to* (2)*. Given a sample covariance matrix $\hat{\boldsymbol{C}}$, assume that $\hat{C}_{i,j\neq i} \leq \gamma < \frac{1+\lambda}{6k}$, and $\epsilon := \frac{16(k+1)^2\gamma}{1+\lambda}$. Then with high probability,*

$$R\left(h^{trg}_{\hat{\boldsymbol{w}},\hat{\boldsymbol{\beta}}}\right) - \hat{R}\left(h^{trg}_{\boldsymbol{w}^\star,\boldsymbol{\beta}^\star}\right) \leq (1+\epsilon)\hat{R}^{src}_\lambda + \tilde{\mathcal{O}}\left(\frac{1+k\tau^{src}_\infty}{\lambda m} + \sqrt{\hat{R}^{src}_\lambda \frac{1+k\tau^{src}_\infty}{\lambda m}}\right),$$

where $\hat{R}^{src}_\lambda := \min_{|S|\leq k}\left\{\frac{\lambda}{|S|} + \frac{1}{|S|}\sum_{i\in S}\hat{R}(h^{src}_i)\right\}$.

To analyze the implications of Theorem 2, let us consider few interesting cases. Similarly as done before, the quantity $\hat{R}^{\text{src}}_\lambda$ captures how well the source hypotheses are aligned with the target task and governs the asymptotic and finite sample regime. In fact, assume for any finite m that there is at least one source hypothesis with small empirical risk, in particular, in $\tilde{\mathcal{O}}(\sqrt{k/m})$, and set $\lambda = \tilde{\mathcal{O}}(\sqrt{k/m})$. Then we have that $R(h^{\text{trg}}_{\hat{\boldsymbol{w}},\hat{\boldsymbol{\beta}}}) - \hat{R}(h^{\text{trg}}_{\boldsymbol{w}^\star,\boldsymbol{\beta}^\star}) = \tilde{\mathcal{O}}\left(\sqrt{k/m}\right)$, that is we get the generalization bound as if we are able to solve the original **NP**-hard problem in (2). In other words, if there are useful source hypotheses, we expect our algorithm to perform similarly to the one that identifies the optimal subset. This might seem surprising, but it is important to note that we do not actually care about identifying the correct subset of source hypotheses. We only care about how well the returned solution is able to generalize. On the other hand, if not even one source hypothesis has low risk, selecting the best subset of k sources becomes meaningless. In this scenario, we expect the selection of any subset to perform in the same way. Thus the approximation guarantee does not matter anymore.

We now state the approximation guarantees of `GreedyTL` used to prove Theorem 2. In the following Corollary we show how far the optimal solution to the regularized subset selection is from the approximate one found by `GreedyTL`.

Corollary 1. *Let $\lambda \in \mathbb{R}^+$ and $k \leq n$. Denote* $\text{OPT} := \min_{\|\boldsymbol{w}\|_0=k}\left\{\hat{R}(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2\right\}$*. Assume that $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{b}}$ are normalized, and $\hat{C}_{i,j\neq i} \leq \gamma < \frac{1+\lambda}{6k}$. Then, FR algorithm generates an approximate solution $\hat{\boldsymbol{w}}$ to the regularized subset selection problem that satisfies $\hat{R}(\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_2^2 \leq \left(1 + \frac{16(k+1)^2\gamma}{1+\lambda}\right)\text{OPT} - \frac{16(k+1)^2\gamma\lambda}{(1+\lambda)^2}$.*

Apart from being instrumental in the proof of Theorem 2, this statement also points to the secondary role of regularization parameter λ: unlike in FR, we can control the quality of approximate solution even if the features are correlated.

5 Experiments

In this section we present experiments comparing `GreedyTL` to several transfer learning and feature selection algorithms. As done previously, we considered the object detection task and, for all datasets, we left out one class considering it as the target class, while the remaining classes were treated as sources [25]. We repeated this procedure for every class and for every dataset at hand, and averaged the performance scores. In the following, we refer to this procedure as *leave-one-class-out*. We performed the evaluation for every class, reporting averaged class-balanced recognition scores.

We used subsets of Caltech-256 [12], Imagenet [7] and SUN09 [5]. The largest setting considered involves 1000 classes, totaling in 1.2M examples, where the number of training examples of the target domain varies from 11 to 20. Our experiments aimed at verifying two claims: (I) the importance of regularization when using greedy feature selection as a transfer learning scheme; (II) in a small-sample regime `GreedyTL` is more robust than alternative feature selection approaches, such as $L1$-regularization.

Datasets and features. We used the whole Caltech-256, a public subset of Imagenet containing 10^3 classes and all the classes of SUN09 that have more than 1 example, which amounts to 819 classes. For Caltech-256 and Imagenet, we used as features the publicly-available 1000-dimensional SIFT-BOW descriptors, while for SUN09 we extracted 3400-dimensional PHOG descriptors.

We composed a negative class by merging 100 held-out classes (*surrogate* negative class). We did so for each dataset, and we further split it into the *source* negative and the *target* negative class as $90\% + 10\%$ respectively, for training sources and the target. The training sets for the target task were composed by $\{2, 5, 10\}$ positive examples, and 10 negative ones. Following [25], the testing set contained 50 positive and 50 negative examples for Caltech-256 and Imagenet. For the skewed SUN09 dataset we took one positive and 10 negative training examples, with the rest left for testing. We drew each target training and testing set randomly 10 times, averaging the results over them. This procedure, commonly used in the literature, helps us avoiding cases of overfitting when building the source hypotheses.

Algorithms. We chose a linear SVM to train the source classifiers [10]. This allows us to compare fairly with relevant baselines (like Lasso) and is in line with recent trends in large scale visual recognition and transfer learning [8]. The source classifiers were trained for each class in the dataset, combining all the positive examples of that class and the source negatives. On average, each source classifier was trained using 10^4 examples for the Caltech-256, 10^5 for Imagenet and 10^3 for the SUN09 dataset. The models were selected by 5-fold cross-validation having regularization parameter $C \in \{10^{-4}, 10^{-3}, \cdots, 10^4\}$. In addition to trained source classifiers, for the Caltech-256, we also evaluated transfer from Classemes [4] and Object Bank [17], which are very similar in spirit to source classifiers. At the same time, for Imagenet, we evaluated transfer from DeCAF convolutional neural network [8].

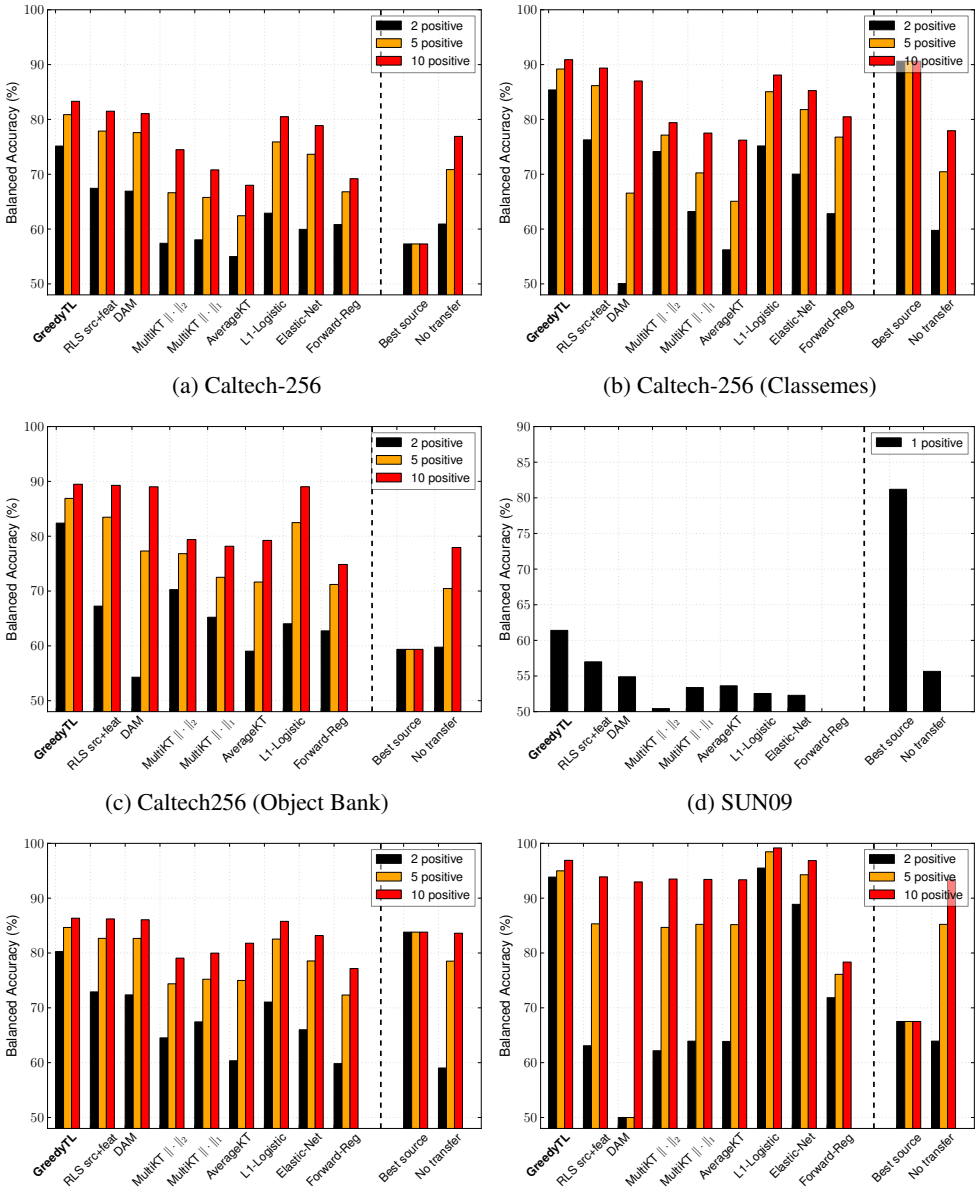
We divided the baselines into two groups - the linear transfer learning baselines that do not require access to the source data, and the feature selection baselines. We included the second group of baselines due to GreedyTL's resemblance to a feature selection algorithm. We focus on the linear baselines, since we are essentially interested in the feature selection in high-dimensional spaces from few examples. In that scope, most feature selection algorithms, such as Lasso, are linear. In particular, amongst TL baselines we chose: *No transfer*: Regularized Least Squares (RLS) algorithm trained solely on the target data; *Best source*: indicates the performance of the best source classifier selected by its score on the testing set. This is a pseudo-indicator of what an HTL can achieve; *AverageKT*: obtained by averaging the predictions of all the source classifiers; *RLS src+feat*: RLS trained on the concatenation of feature descriptors and source classifier predictions; *MultiKT* $\|\cdot\|_2$: HTL algorithm by [25] selecting $\boldsymbol{\beta}$ in (1) by minimizing the leave-one-out error subject to $\|\boldsymbol{\beta}\|_2 \leq \tau$; *MultiKT* $\|\cdot\|_1$: similar to previous, but applying the constraint $\|\boldsymbol{\beta}\|_1 \leq \tau$; *DAM*: An HTL algorithm by [9], that can handle selection from multiple source hypotheses. It was shown to perform better than a well known and similar ASVM [27] algorithm. For the feature selection baselines we selected well-established algorithms involving sparsity assumption: *L1-Logistic*: Logistic regression with $L1$ penalty [13]; *Elastic-Net*: Logistic regression with mixture of $L1$ and $L2$ penalties [13]; *Forward-Reg*: Forward regression – a classical greedy feature selection algorithm.

Results. Figure 1 shows the leave-one-class-out performance w.r.t. all considered datasets. In addition, Figures 2b, 2c, 2f show the performance when transferring from off-the-shelf classemes, object-bank feature descriptors, and DeCAF neural network activations. Whenever any baseline algorithm has hyperparameters to tune, we chose the ones that minimize the leave-one-out error on the training set. In particular, we selected the regularization parameter $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$. MultiKT and DAM have an additional hyperparameter that we call τ with $\tau \in \{10^{-3}, \ldots, 10^{3}\}$. Kernelized algorithms were supplied with a linear kernel. Model selection for GreedyTL involves two hyperparameters, that is k and λ. Instead of fixing k, we let GreedyTL select features as long as the regularized error between two consecutive steps is larger than δ. In particular, we set $\delta = 10^{-4}$, as in preliminary experiments we have not observed any gain in performance past that point. The λ is fixed to 1. Even better performance could be obtained tuning it.

We see that GreedyTL dominates TL and feature selection baselines throughout the benchmark, rarely appearing on-par, especially in the small-sample regime. In addition, on two datasets out of three, it manages to identify the source classifier subset that performs comparably or better than the Best source, that is the single best classifier selected by its performance on the testing set. The significantly stronger performance achieved by GreedyTL w.r.t. FR, on all databases and in all settings, confirms the importance of the regularization in our formulation.

Notably, GreedyTL outperforms RLS src+feat, which is equivalent to GreedyTL selecting all the sources and features. This observation points to the fact that GreedyTL successfully manages to discard irrelevant feature dimensions and sources. To investigate this important point further, we artificially add 10, 100 and 1000 dimensions of pure noise sampled from a standard distribution. Figure 2 compares feature selection

Fig. 1: Performance on the Caltech-256, subsets of Imagenet (1000 classes) and SUN09 (819 classes). Averaged class-balanced accuracies in the leave-one-class-out setting.



(a) Caltech-256

(b) Caltech-256 (Classemes)

(c) Caltech256 (Object Bank)

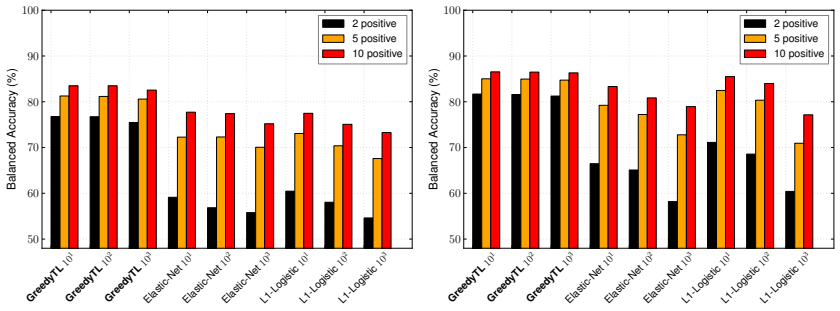
(d) SUN09

(e) Imagenet (1000 classes)

(f) Imagenet (DeCAF, 1000 classes)

methods to `GreedyTL` in robustness to noise. Clearly, in the small-sample setting,

Fig. 2: Baselines and number of additional noise dimensions sampled from a standard distribution. Averaged class-balanced recognition accuracies in the leave-one-class-out setting.



(a) Caltech-256 (b) Imagenet

`GreedyTL` is tolerant to large amount of noise, while $L1$ and $L1/L2$ regularization suffer a considerable loss in performance. We also draw attention to the failure of $L1$-based feature selection methods and MultiKT with $L1$ regularization to match the performance of `GreedyTL`.

6 Conclusions

In this work we studied the transfer learning problem involving hundreds of sources. The kind of transfer learning scenario we consider assumes no access to the source data directly, but through the use of the source hypotheses induced from them. In particular, we focused on the efficient source hypothesis selection and combination, improving the performance on the target task. We proposed a greedy algorithm, `GreedyTL`, capable of selecting relevant sources and feature dimensions at the same time. We verified these claims by obtaining the best results among the competing feature selection and TL algorithms, on the Imagenet, SUN09 and Caltech-256 datasets. At the same time, comparison against the non-regularized version of the algorithm clearly show the power of our intuition. We support our empirical findings by showing theoretically that under reasonable assumptions on the sources, the algorithm can learn effectively from few target examples.

Acknowledgments

I.K. is supported by the Swiss National Science Foundation Sinergia project Ninapro and Idiap Research Institute.

References

1. Aytar, Y., Zisserman, A.: Tabula rasa: Model transfer for object category detection. In: ICCV (2011)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. Machine Learning (2010)

3. Ben-David, S., Urner, R.: Domain adaptation as learning with auxiliary information. In: New Directions in Transfer and Multi-Task - Workshop @ NIPS (2013)
4. Bergamo, A., Torresani, L.: Classemes and other classifier-based features for efficient object categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2014)
5. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: CVPR (2010)
6. Das, A., Kempe, D.: Algorithms for subset selection in linear regression. In: STOC (2008)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML (2014)
9. Duan, L., Tsang, I.W., Xu, D., Chua, T.S.: Domain adaptation from multiple sources via auxiliary classifiers. In: ICML (2009)
10. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research (2008)
11. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR (2012)
12. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. rep., Caltech (2007)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements Of Statistical Learning. Springer (2009)
14. Jie, L., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: ICCV (2011)
15. Kuzborskij, I., Orabona, F.: Stability and hypothesis transfer learning. In: ICML (2013)
16. Kuzborskij, I., Orabona, F., Caputo, B.: From N to N+1: Multiclass Transfer Incremental Learning. In: CVPR (2013)
17. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS (2010)
18. Lim, J.J., Torralba, A., Salakhutdinov, R.: Transfer learning by borrowing examples for multiclass object detection. In: NIPS (2011)
19. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain Adaptation with Multiple Sources. In: NIPS (2009)
20. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR (2014)
21. Pan, S.J., Yang, Q.: A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions on (2010)
22. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV (2010)
23. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detection. In: CVPR (2011)
24. Tommasi, T., Caputo, B.: The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: BMVC (2009)
25. Tommasi, T., Orabona, F., Caputo, B.: Learning categories from few examples with multi model knowledge transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013)
26. Vezhnevets, A., Ferrari, V.: Associative embeddings for large-scale knowledge transfer with self-assessment. In: CVPR (2014)
27. Yang, J., Yan, R., Hauptmann, A.: Cross-Domain Video Concept Detection Using Adaptive SVMs. In: ACMM (2007)
28. Zhang, T.: Adaptive forward-backward greedy algorithm for sparse learning with linear models. In: NIPS (2008)