Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Learning by Transferring from Unsupervised Universal Sources

Yu-Xiong Wang and **Martial Hebert**
Robotics Institute, Carnegie Mellon University
{yuxiongw, hebert}@cs.cmu.edu

Abstract

Category classifiers trained from a large corpus of annotated data are widely accepted as the sources for (hypothesis) transfer learning. Sources generated in this way are tied to a particular set of categories, limiting their transferability across a wide spectrum of target categories. In this paper, we address this largely-overlooked yet fundamental *source* problem by both introducing a systematic scheme for generating *universal* source hypotheses and proposing a principled, scalable approach to automatically tuning the transfer process. Our approach is based on the insights that expressive source hypotheses could be generated *without any supervision* and that a sparse combination of such hypotheses facilitates recognition of novel categories from few samples. We demonstrate improvements over the state-of-the-art on object and scene classification in the small sample size regime.

Introduction

Learning from few samples has now attracted wide interest in large-scale object recognition, given the intrinsic long-tailed distribution of real-world objects (Zhu, Anguelov, and Ramanan 2014) and customized categories in personal image collections (Kienzle and Chellapilla 2006; Wang and Hebert 2015). Such scenarios are typically addressed in transfer learning (TL), which benefits from transfer of prior knowledge from related tasks to new ones, in the majority of cases either on a data or instance level, or on a feature or parameter level (Pan and Yang 2010). Despite featuring well established theoretical guarantees, these approaches often suffer from great practical constraints and limitations (Pan and Yang 2010; Kuzborskij, Caputo, and Orabona 2015): they require reusing data originating from the source domains and extensive supervised retraining on the target task, which is prohibitively expensive for large source data.

Hypothesis transfer learning (HTL) (Kienzle and Chellapilla 2006; Yang, Yan, and Hauptmann 2007a; 2007b; Yang and Hauptmann 2008; Duan et al. 2009; Chattopadhyay et al. 2011; Aytar and Zisserman 2011; 2012; Kuzborskij and Orabona 2013; Kuzborskij, Orabona, and Caputo 2013; Tommasi, Orabona, and Caputo 2014; Kuzborskij, Caputo, and Orabona 2015; Kuzborskij and Orabona 2015) has been recently considered an alternative, which transfers directly on a model level by reusing source hypotheses — classifiers or models trained from source data. This framework is practically appealing, since it requires neither the availability of the source data nor any knowledge on how the source models relate to each other. HTL is also efficient especially with small target samples, in which source hypotheses are generated in advance and treated as black boxes without any consideration of their inner workings at transfer stages.

Much attention in HTL (and also TL) has been focused on integrating the source information into the target task in different ways. Unfortunately, very little work has addressed the generation of useful source models. In most cases, sources are simply category classifiers well-trained from large amounts of labeled samples. This however might be infeasible for real-world applications: we focus on learning from few samples for target categories, whereas we have to train good classifiers of related categories as sources from enough labeled data in advance. For instance, if we are interested in recognizing Père David's deer* from few samples, following the conventional HTL practice, we might need to first obtain well-trained source classifiers of camels, cows, donkeys, and deer for transfer. Furthermore, sources generated in this way are tied to a specific set of categories due to its supervised nature, making it difficult to apply them across a wide spectrum of target categories.

It is thus unsurprising that the current HTL (and TL) algorithms are usually evaluated under well-controlled experimental setups: (1) use small-scale well-trained classifiers as sources, at most several hundred (Yang, Yan, and Hauptmann 2007b; Aytar and Zisserman 2011; 2012); (2) split a dataset with a portion of categories as sources and the rest as targets, which implicitly reduces the impact of dataset bias, e.g., leave-one-class-out (Tommasi, Orabona, and Caputo 2014; Kuzborskij, Caputo, and Orabona 2015); (3) transfer between visually similar categories with ideal sources (Hoffman et al. 2014).

To address this largely-overlooked yet fundamental problem, we introduce a systematic scheme for generating *uni-*

*A Père David's deer is a species of deer that has the neck of a camel, the hoofs of a cow, the tail of a donkey, and the antlers of a deer.

versal and expressive source hypotheses in an *unsupervised* fashion, which frees the recognition from ties to a particular set of categories and which generalizes well for broad novel target classes. Our key insight is that hypotheses that are informative across categories could be generated without any supervision because of the information implicit in the density structure of the feature space. More precisely, each hypothesis now lies in a region of low density and the combined hypotheses constitute a joint partition of the feature space. Partitions satisfying such property are explored in discovery of predictable discriminative binary codes (PBCs) (Rastegari, Farhadi, and Forsyth 2012), which focuses on learning binary codes as image representations for efficient image retrieval — a problem different from ours. Given that each bit in PBCs can be viewed as a split of the feature space induced by discrimination and learnability, our crucial observation is the equivalence between source hypotheses generation and binary codes discovery. We then modify the original supervised version of PBCs to be estimated in an unsupervised manner, leading to a library of unsupervised universal sources (UUS) with widespread visual/attribute coverage.

This unprecedented large-scale source pool, with two orders of magnitude more hypotheses than previous works, poses additional scalability challenges to the existing HTL approaches. These algorithms adopt a discriminative SVM framework (usually with a quadratic loss), in which a new target classifier is learned through adaptation by imposing closeness between the target classifier and a linear combination of the source hypotheses as regularizer. The weight associated to each source is either predefined for known transfer relationship (Kienzle and Chellapilla 2006), or determined by designing heuristic meta-level features (Yang, Yan, and Hauptmann 2007b), or estimated based on the conditional probability distribution of large amounts of unlabeled target data (Chattopadhyay et al. 2011). In other cases, the weight is obtained by minimizing empirical error with solely ℓ_2 norm (Yang and Hauptmann 2008; Aytar and Zisserman 2012; Kuzborskij, Orabona, and Caputo 2013) or sparsity-inducing (ℓ_0, ℓ_1) norm regularization (Tommasi, Orabona, and Caputo 2014; Kuzborskij, Caputo, and Orabona 2015). These frameworks have been tested on problems with less than a few hundred sources, but have already showed some difficulty in selecting informative sources due to severe over-fitting (Tommasi, Orabona, and Caputo 2014; Kuzborskij, Caputo, and Orabona 2015). To resolve this issue, we propose a scalable model transfer SVM (MT-SVM) approach by combining an elastic net regularization and biased SVM with a hinge loss. The relatedness among the tasks is autonomously evaluated through a principled optimization problem without extra validation, unlabeled samples, or a predefined ontology.

Our contribution is three-fold. First, we show how a universal library of source hypotheses, UUS, based on *unsupervised* PBC classifiers, is generated without bias to a particular set of categories. Second, we provide a *principled, scalable* HTL algorithm, MT-SVM, that selects a set of source hypotheses and uses them to infer the target model. Finally, we show how informative hypotheses are selected and transferred on novel classes with few samples.

An Unsupervised Universal Source Library

Source hypotheses are generated as prior knowledge or regularization for guiding learning on new tasks, while binary codes are inferred to encode high-dimensional image descriptors as compact binary strings. Although originating from different applications, each hypothesis and each bit can be both viewed as a partition of the feature space. To the best of our knowledge, this paper is the first work that constructs a bridge between these two. Partitions of our interest are those informative across categories, which satisfy certain discrimination and learnability properties and which could be intuitively interpreted as semantic or discriminative attributes. While largely overlooked in HTL, such partitions could be produced by predictable discriminative binary codes (PBCs) (Rastegari, Farhadi, and Forsyth 2012). In particular, we extend the original supervised PBCs to be estimated in an unsupervised manner, by first obtaining pseudo-labeled data via a series of sampling steps and then using these (pseudo-)labeled samples to learn PBCs. Learning a library of unsupervised universal sources (UUS) includes the following steps:

Pseudo-Classes Generation via Sampling. Given a large collection of N unlabeled images with feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, denoted as $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}^\dagger$, we first need to generate pseudo-labels that are stand-ins for plausible categories. To be specific, we want samples with the same pseudo-label to be similar in feature space (constraints within pseudo-classes), while those with different pseudo-labels should be very dissimilar (constraints between pseudo-classes). To achieve this, we first draw an M-subset $\mathcal{A}_\mathcal{S}$ from $\mathcal{D}$ by random subsampling. Within $\mathcal{A}_\mathcal{S}$, we create an initial skeleton by sampling C random seed points that are spread out (Max-step). We then augment each seed point to a pseudo-class by adding its $K-1$ nearest neighbors (Min-step). By this Max-Min sampling (Dai and Gool 2013), we have then generated a prototype set $\mathcal{B}_{\mathcal{PL}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{K\times C}$, where $y_i \in \{1, \ldots, C\}$ are the pseudo-labels.

Hypotheses Generation by PBCs Learning. On pseudo-labeled set $\mathcal{B}_{\mathcal{PL}}$, we generate a set of S-splits, represented by S weight vectors $\boldsymbol{w}^s$ by using the max-margin formulation introduced in (Rastegari, Farhadi, and Forsyth 2012):

$$\min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{L},\boldsymbol{B}} \frac{1}{2} \sum_{c\in\{1:C\}} \sum_{u,v\in c} d\left(\boldsymbol{B}_u, \boldsymbol{B}_v\right) + \eta \sum_{s\in\{1:S\}} \|\boldsymbol{w}^s\|^2 \qquad (1)$$

$$+ \lambda_1 \sum_{\substack{i\in\{1:K\times C\}\\ s\in\{1:S\}}} \xi_i^s - \frac{\lambda_2}{2} \sum_{\substack{c'\in\{1:C\}\\ p\in c'}} \sum_{\substack{c''\in\{1:C\}\\ q\in c'', c'\neq c''}} d\left(\boldsymbol{B}_p, \boldsymbol{B}_q\right)$$

$$s.t. \quad l_i^s \left(\boldsymbol{w}^{sT} \boldsymbol{x}_i\right) \geq 1 - \xi_i^s, b_i^s = \left(1 + sign\left(\boldsymbol{w}^{sT} \boldsymbol{x}_i\right)\right)\Big/2,$$

$$\xi_i^s > 0, \quad \forall i \in \{1 : K \times C\}, s \in \{1 : S\}.$$

Here for the s-th split, ξ_i^s is the slack variable of $\boldsymbol{x}_i$. For notational simplicity, $\boldsymbol{x}_i$ already includes a constant 1 as the last element and $\boldsymbol{w}$ includes the bias term. $l_i^s \in \{-1, 1\}$

[†]Notation: we denote column vectors and matrices with small and capital bold letters, i.e., $\boldsymbol{a}=[a_1, a_2, \ldots, a_N]^T \in \mathbb{R}^N$ and $\boldsymbol{A}=[A_{ij}] \in \mathbb{R}^{M\times N}$, respectively.

is the training label of $\boldsymbol{x}_i$ to be learned to indicate which side of the s-th split $\boldsymbol{x}_i$ should appear. $b_i^s \in \{0,1\}$ is the actual prediction indicating which side of the s-th split (trained with l_i^s) $\boldsymbol{x}_i$ actually lies. $\boldsymbol{B}_i = [b_i^1, \ldots, b_i^S]$ is the stacked binary code of $\boldsymbol{x}_i$ from all the splits, and d is the Hamming distance. Note that our introduction of pseudo-labels as supervisory information is crucial here, since the max-margin formulation (1) does not apply in the unsupervised settings. We then (pseudo-)label G more samples to each pseudo-class from the unlabeled data pool $\mathcal{C}_{\mathcal{UL}}$ as in (Choi et al. 2013). Based on this augmented dataset $\mathcal{D}_{\mathcal{AUG}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{(K+G)\times C}$, where $y_i \in \{1, \ldots, C\}$, we retrain a new set of S-split PBCs by using Eqn. (1). To ensure diversity, we repeat the subsampling procedure T times and generate $J = S \times T$ source hypotheses in total, which could be potentially large in practice.

Discussions. One crucial issue is why our candidate hypotheses generalize well for novel categories and what kind of information is transferred. In addition to the unsupervised aspect, we use PBC classifiers to group our pseudo-classes into a set of *abstract classes* and obtain *attribute-like hypotheses*. Such sources are more generic, untied to a specific set of categories. This is related to the observation in the supervised case that meta-classes outperform classemes (Bergamo and Torresani 2014). From the principle of Structural Risk Minimization, our UUS hypotheses provide an alternative mechanism to encode prior knowledge and control model capacity. This is related to the use of Universum (i.e., unlabeled examples that do not belong to the concerned classes, sometimes called "non-examples") in addition to labeled data for capacity control, which proved to be helpful in various learning tasks (Weston et al. 2006). When facing a large collection of non-examples, our UUS can be viewed as compressing the original source data while implicitly modeling a general distribution and preserving relevant information for classification.

Our UUS can be also considered as distinctive subdomains automatically discovered in a large source domain. Conventional discovery of latent domains for domain adaptation (Gong, Grauman, and Sha 2013; Hoffman et al. 2012) is supervised, in which object category labels are used to constrain feasible subdomain separations on source datasets. However, our hypotheses are generated in an entirely unsupervised manner without requiring any labeled data. Moreover, (Gong, Grauman, and Sha 2013; Hoffman et al. 2012) need to explicitly model the distribution on different subdomains, and measure the distance between distributions. However, modeling the distribution of high-dimensional image features on large datasets is typically more difficult than classifying them. Hence, our approach is more flexible, scalable, and broadly applicable in practice.

Model Transfer Support Vector Machine

Once we obtain the J source hypotheses $\{\boldsymbol{w}_j^{src}\}_{j=1}^J$, the original training samples used to build them are no longer used. We now consider a new target task with a small labeled training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^L$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ are the training samples and $y_i \in \{-1, 1\}$ are the corresponding labels. Hypothesis transfer learning (HTL) attempts to infer the target hypothesis $\boldsymbol{w}$ from both $\{\boldsymbol{w}_j^{src}\}_{j=1}^J$ and $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^L$ that generalizes better than the one produced only from $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^L$. HTL algorithms proposed so far are developed under a discriminative SVM framework modified by regularizing the distance between $\boldsymbol{w}$ and a linear combination of the sources $\boldsymbol{w}^{src}$. To identify useful sources, it is recast as a variable selection problem by constraining the combination weights with either ℓ_0, ℓ_1, or ℓ_2 norm (Aytar and Zisserman 2012; Tommasi, Orabona, and Caputo 2014; Kuzborskij, Caputo, and Orabona 2015). However, a single type of norm has its own pros and cons. Especially, in our scenario with only few target samples and large-scale generic weak sources, these existing approaches would be very noisy due to severe overfitting and would induce negative transfer.

As a well-known recipe, an elastic net regularization, combining a weighted mixture of ℓ_1 and squared ℓ_2 penalties, offers several desirable benefits: (1) ℓ_2 regularization is known to improve the generalization ability of empirical risk minimization (Kuzborskij, Caputo, and Orabona 2015); (2) ℓ_1 norm, as a convex relaxation of ℓ_0 norm, always converges to a good solution in practice, avoiding potential bad local minima when using a greedy scheme (Kuzborskij, Caputo, and Orabona 2015) to directly solve ℓ_0 problems; (3) joint ℓ_1 and ℓ_2 enjoys a similar sparsity of representation and encourages a grouping effect (Zou and Hastie 2005); (4) it is particularly useful in our case that the number of predictors is much bigger than the number of observations (Zou and Hastie 2005).

Formulation

By using the new regularization to rank the prior sources and introducing them as reference into SVM, we then obtain the objective function for our model transfer SVM (MT-SVM):

$$\min_{\boldsymbol{w},\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{w} - \sum_{j=1}^J \beta_j \boldsymbol{w}_j^{src}\right\|^2 + \frac{\alpha}{2}\sum_{j=1}^J \beta_j^2 + \gamma \sum_{j=1}^J |\beta_j| \quad (2)$$
$$+ \lambda \sum_{i=1}^L \left[1 - y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i\right)\right]_+ .$$

The last term represents the data fit on the L training samples, measured by the hinge loss; it is the new information from the target domain. The first term is similar to the max-margin principle in standard SVMs, with the only difference being the bias towards the linear combination of the generic source hypotheses $\sum_{j=1}^J \beta_j \boldsymbol{w}_j^{src}$ instead of 0, in which β_j's are transfer weights; it is the prior information from the source domains. In order to automatically select the best subset of known hypotheses from which to transfer, the second and third terms are introduced as an elastic net regularization that favors sparse $\boldsymbol{\beta}$. Here, α, γ, and λ are the regularization parameters to control the trade-off between the error term and regularization terms.

Following the duality derivation analogous to standard SVMs, the optimal solution to Eqn. (2) satisfies

$$\boldsymbol{w} = \sum_{j=1}^J \beta_j \boldsymbol{w}_j^{src} + \sum_{i=1}^L \mu_i y_i \boldsymbol{x}_i, \quad (3)$$

where μ_i's are Lagrange multipliers. The final target model is then conceptually straightforward: it linearly combines the

contribution from both the pre-trained generic models and target specific data, i.e., support vectors from both source and target domains. Combining Eqns. (2) and (3), as $\alpha \to \infty$ and $\gamma \to \infty$, β_j's will be forced to be zero and we will get back the standard SVM, i.e., no transfer. As $\lambda \to 0$, $\boldsymbol{w}$ will be forced to be purely constructed as a weighted combination of $\{\boldsymbol{w}_j^{src}\}$'s, i.e., maximum transfer. As $\alpha \to 0$, it becomes LASSO regression while ridge regression as $\gamma \to 0$. Hence by tweaking α, γ, and λ we obtain an intermediate solution with a decision boundary close to those of the auxiliary classifiers while separating the labeled examples well.

Optimization

The objective in Eqn. (2) can be optimized by alternating minimization of two subproblems:

- With fixed $\boldsymbol{w}$, the objective function of finding transfer weights $\boldsymbol{\beta}$ becomes an elastic net regularized least-squares minimization subproblem:

$$f(\boldsymbol{\beta}) = \frac{1}{2}\left\| \boldsymbol{w} - \sum_{j=1}^{J} \beta_j \boldsymbol{w}_j^{src} \right\|^2 + \frac{\alpha}{2}\sum_{j=1}^{J}\beta_j^2 + \gamma \sum_{j=1}^{J}|\beta_j|. \quad (4)$$

- With fixed $\boldsymbol{\beta}$, the objective function of learning target hypothesis $\boldsymbol{w}$ becomes a bias regularized SVM subproblem:

$$f(\boldsymbol{w}) = \frac{1}{2}\left\| \boldsymbol{w} - \sum_{j=1}^{J} \beta_j \boldsymbol{w}_j^{src} \right\|^2 + \lambda \sum_{i=1}^{L} \left[1 - y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i \right) \right]_+ . \quad (5)$$

Source Selection by Modified Feature-Sign Search. We solve Eqn. (4) by extending the feature-sign search (FS) algorithm (Lee et al. 2006), one of the state-of-the-art techniques for efficient sparse coding (i.e., ℓ_1 regularized least-squares) (Liu et al. 2014), to our case of elastic net regularization (i.e., joint ℓ_1 and ℓ_2 regularized least-squares). FS searches and maintains an optimal active set of potentially nonzero coefficients and sets other coefficients zero. Although it was developed in the context of dictionary learning and sparse coding, FS still fits our scenario, in which we could view the source hypotheses $\{\boldsymbol{w}_j^{src}\}_{j=1}^{J}$ as known dictionary bases and rearrange them into the matrix form of dictionary $\boldsymbol{W}^{src}$. The equivalent optimization problem of Eqn. (4) is then

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{W}^{src}\boldsymbol{\beta}\|^2 + \frac{\alpha}{2}\|\boldsymbol{\beta}\|^2 + \gamma\|\boldsymbol{\beta}\|_1. \quad (6)$$

Since the only difference lies in the extra ℓ_2 regularization term that is differentiable, we could easily modify the key update of $\widehat{\boldsymbol{\beta}}_{new}$ in a series of "feature-sign steps" with $\left(\widehat{\boldsymbol{W}}^{srcT}\widehat{\boldsymbol{W}}^{src} + \alpha \boldsymbol{I}\right)^{-1}$ instead of $\left(\widehat{\boldsymbol{W}}^{srcT}\widehat{\boldsymbol{W}}^{src}\right)^{-1}$ in (Lee et al. 2006). ($\widehat{\cdot}$ represents the active set.)

Model Transfer via Adaptive SVM. The optimal $\boldsymbol{w}$ in Eqn. (5) can be obtained by the Adaptive SVM algorithm (Yang, Yan, and Hauptmann 2007a; 2007b), which solves a quadratic program to maximize its Lagrange dual objective function. With small samples in our case, the problem can be efficiently solved by (modified) sequential minimal optimization (Kienzle and Chellapilla 2006; Yang, Yan, and Hauptmann 2007b). In addition, we initialize $\boldsymbol{w}$ using the standard SVM without bias on the given target training set. We then iteratively infer $\boldsymbol{\beta}$ and refine $\boldsymbol{w}$. Given the convexity of the problem, this block coordinate descent algorithm will converge to the global minimum.

Experimental Evaluation

In this section, we present experimental results evaluating our unsupervised sources (UUS) as well as our HTL approach (MT-SVM) on standard recognition benchmarks, comparing several state-of-the-art methods, and validating across tasks and categories the generality of our sources.

Implementation Details

For the feature space, consistent with recent work, we use the convolutional neural network (CNN) features pre-trained on ILSVRC 2012 (Krizhevsky, Sutskever, and Hinton 2012; Donahue et al. 2014; Russakovsky et al. 2015) without fine-tuning. For each resized image, we extract a $d = 4{,}096$-D feature vector $fc7$, taken from the last hidden layer of the network. There is no restriction on the corpus of unlabeled data to generate our pool of hypotheses. Here, for purpose of reproducibility, we simply use the ILSVRC 2012 training dataset without access to the label information, leading to $N = 1.2M$ unlabeled images $\mathcal{D}$.

To generate UUS, at each iteration, we subsample $M = 20K$ data to form $\mathcal{A}_{\mathcal{S}}$. We use the same setup and default parameters for the Max-Min sampling and PBCs learning procedures as in (Dai and Gool 2013; Rastegari, Farhadi, and Forsyth 2012; Choi et al. 2013). Using an augmented pseudo-labeled dataset $\mathcal{D}_{\mathcal{AUG}}$ of $C = 30$ pseudo-classes with $K{+}G{=}6 + 50$ samples per pseudo-class, we generate $S = 10$ split PBCs. Repeating $T = 2{,}000$ subsampling in parallel, we have generated $J = 20K$ source hypotheses in total.

In term of MT-SVM, for λ, we use the default value 1 as in Adaptive SVM (Yang, Yan, and Hauptmann 2007a; 2007b). For α and γ, in a preliminary experiment, we tested the ImageNet categories as targets and our UUSs for transfer. Empirically, we found that keeping the number of selected sources to be around $100 \sim 200$ yields good results. After searching α on a small grid $(0, 0.01, 0.1, 1, 10, 100)$ as suggested in (Zou and Hastie 2005), we found that $\alpha = 10$ roughly achieved the desired stable solution. For all our experiments, we then fixed $\alpha = 10$, and tuned γ to minimize the leave-one-out-error.

Comparison with Supervised Sources

Naturally, the most critical question to answer is whether our UUS indeed facilitates generalization to novel categories with few samples, compared to their supervised counterparts (i.e., category models, SS). To this end, we evaluate them on the Office dataset (Saenko et al. 2010), a standard domain adaptation benchmark for multiclass object recognition.

Datasets. The Office dataset contains 31 classes with a total of $4{,}652$ images from three distinct domains: Amazon, DSLR, and Webcam. In our experiment for a fair comparison between UUS and SS, we follow a similar experimental setup as in (Donahue et al. 2014; Hoffman et al. 2014): we use Webcam as the target domain since it was shown to be the most challenging shift domain (Hoffman et al. 2014). We view the ILSVRC 2012 training dataset as the source domain, on which our UUSs are generated. This scenario exemplifies the transfer from online web images to real-world images taken in typical office/home environments.

Transfer Scenario	Method	Acc (%)
Non-Transfer	SVM (source only) (Hoffman et al. 2014)	59.15 ± 1.1
	SVM (target only) (Hoffman et al. 2014)	64.97 ± 1.8
	SVM (source and target) (Hoffman et al. 2014)	66.93 ± 1.3
Transfer with Source Data	GFK (Gong et al. 2012)	67.97 ± 1.4
	SA (Fernando et al. 2013)	66.08 ± 1.4
	Daumé III (Daumé III 2007)	71.39 ± 1.5
HTL with Supervised Sources	PMT (Aytar and Zisserman 2011)	69.81 ± 1.8
	MMDT (Hoffman et al. 2013)	67.75 ± 1.4
	Late Fusion (Max) (Hoffman et al. 2014)	68.86 ± 1.2
	Late Fusion (Lin. Int. Avg) (Hoffman et al. 2014)	66.45 ± 1.1
HTL with Unsupervised Sources	Clustering+MT-SVM	67.13 ± 1.2
	UUS+MT-SVM (Ours)	$\mathbf{74.83 \pm 1.2}$
HTL–Upper Bound	*Late Fusion (Lin. Int. Oracle) (Hoffman et al. 2014)*	76.76 ± 1.3

Table 1: Performance comparison between HTL with supervised (SS) and unsupervised (USS) source hypotheses generated from ILSVRC for one-shot learning in the Subset A (16 common classes) on the Webcam domain of the Office dataset. We also include for completeness the results of transfer learning with source data. Using a large library of unsupervised sources, ours yields performance superior to other state-of-the-art HTL methods with well-trained source category models, and even close to the oracle with an ideal source and the optimal transfer weight on the test set (performance upper bound).

Source Hypotheses. We use our generated library of $20K$ UUSs as unsupervised sources. Moreover, for comparison, we also generate another $20K$ sources by a naïve unsupervised approach denoted as clustering, which creates hypotheses by clustering the data and produces classifiers between clusters. For supervised sources (SSs), we use the labeled samples from the 1,000 categories on ILSVRC as source data, with approximately 1,200 examples per category. With the same CNN features, we then train source SVM classifiers in one-vs.-all fashion, leading to 1,000 category models on these labeled samples.

Target Tasks. To better understand the transfer process, we group the 31 target classes into two subsets. **Subset A:** we focus on the 16 common classes between Webcam and ILSVRC as our target categories as in (Hoffman et al. 2014). 1 labeled training and 10 testing images per category are randomly selected on the Webcam domain, i.e., one-shot transfer and a balanced test set across categories. Therefore, each test split has 160 examples. **Subset B:** we also test the other 15 non-overlapping classes as our target categories in the similar one-shot transfer scenario. For each subset, we evaluate the two types of sources, independently calculate the multiclass accuracy, and report the average performance and standard errors over 20 random train/test splits, as shown in Table 1 and Table 2.

Baselines. We compare against three types of baselines. **Type I non-transfer**: SVM (source only), SVM (target only), and SVM (source and target). They are category SVMs trained on only labeled source, only target, and both source and target data, respectively. For completeness we also include **Type II transfer learning based on (labeled or unlabeled) source data**: GFK (Gong et al. 2012), SA (Fernando et al. 2013), and Daumé III (Daumé III 2007). For instance, Daumé III retrains SVMs on the augmented source and target data using tripled augmented feature, resulting in a relatively expensive procedure given the potentially large size of the source data and high feature dimensionality (Hoffman et al. 2014). Note that some of these baselines are only available for Subset A, since they require that the source comes from the same category as the target.

Transfer Scenario	Method	Acc (%)
Non-Transfer	SVM (target only)	63.34 ± 2.1
HTL with SS	Multi-KT (2014)	65.28 ± 1.3
	DAM (2009)	66.13 ± 1.4
	GreedyTL (2015)	68.72 ± 1.8
	SS+MT-SVM	70.30 ± 1.2
HTL with USS	Clustering+MT-SVM	64.92 ± 1.2
	UUS+MT-SVM (Ours)	$\mathbf{74.19 \pm 1.3}$

Table 2: Performance comparison between HTL with SS and USS for one-shot learning in the Subset B (15 non-overlapping classes) on the Webcam domain.

Type III Baselines of HTL with Supervised Sources. For Subset A, transfer becomes a domain adaptation problem: the transfer is largely dominated by the source category corresponding to the target as the single most relevant one, making other categories uninvolved in the transfer process. We then transfer the corresponding learned category models without a source selection step, including (1) PMT (Aytar and Zisserman 2011), which regularizes the angle between the target and source hyperplanes; (2) MMDT (Hoffman et al. 2013), which jointly optimizes over a feature transformation mapping target points and classifier weights to the source feature space; (3) Late Fusion, which independently trains a source and a target category classifier, and sets the final score for each example by choosing the maximum (Max) or linear interpolation (Lin. Int.) of source and target classifier scores. We report the performance of linear interpolation both averaged across linear combination hyperparameter settings (Avg) and with its best possible setting on the test set per experiment (Oracle). Importantly, the latter case is the best achievable performance for HTL (upper bound), which is equivalent to an ideal source (the same cat-

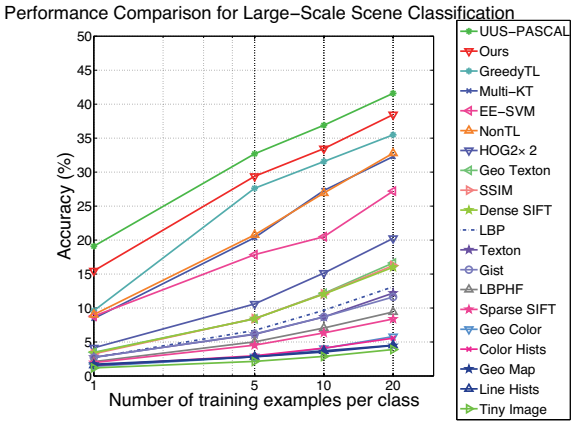

Figure 1: Performance comparison between our MT-SVM and state-of-the-art HTL approaches with our UUS generated on ILSVRC for multiclass scene classification from few samples on the SUN-397 dataset. We include for completeness the results of other modern features and approaches reported from (Xiao et al. 2014). We also conduct dataset sensitivity analysis by providing results with MT-SVM and UUS generated on PASCAL 2007.

egory as the target and with a large amount of training examples) transferred with the optimal transfer weight. These results are reported from (Hoffman et al. 2014). For Subset B, without explicit category correspondence, we use all 1,000 supervised sources and transfer the relevant ones by state-of-the-art HTL approaches, including Multi-KT (Tommasi, Orabona, and Caputo 2014), DAM (Duan et al. 2009), and GreedyTL (Kuzborskij, Caputo, and Orabona 2015).

Table 1 and Table 2 show that our transfer with unsupervised source hypotheses outperforms non-transfer and other state-of-the-art techniques of transfer with source data and transfer with supervised source hypotheses. Notably, in Table 1 ours achieves significant performance *close to the oracle*. Moreover, the naïve unsupervised clustering approach works poorly here. This verifies our assumption that information across categories is actually intrinsic in the data even without any supervision and could be effectively identified by our UUS. With such unsupervised nature, our approach reduces the effort of collecting large amounts of labeled data and training accurate relevant source category models, as is normally the case in previous transfer learning works.

Self-Taught Scene Classification

To show that our UUSs are informative across categories and tasks, we consider using them for large-scale scene classification on the SUN-397 dataset (Xiao et al. 2014). It has 108,754 images of 397 scene categories. This is a very challenging task given the strong domain shift between the object-centric source ILSVRC dataset and the scene-centric target SUN-397 dataset. At a high level, this transfer scenario is close to self-taught learning (Raina et al. 2007); however, we transfer on a model level while (Raina et al. 2007) transfers on a feature level. We evaluate performance as a function of the number of training examples per class

m	0	0.1	0.2	0.3	0.4
Acc (%)	10.51	13.83	14.47	14.72	15.01
m	0.5	0.6	0.7	0.8	0.9
Acc (%)	15.16	15.33	15.54	15.48	15.23

Table 3: Parameter sensitivity analysis with varied γ (reparameterized by m) for one-shot learning on SUN-397.

following the standard experimental setup (Xiao et al. 2014): a subset of the dataset with 50 training and 50 testing images per class is used for evaluation, averaging over 10 fixed and publicly available partitions. We focus on small-sample learning scenario by using the first 1, 5, 10, 20 images out of the 50 training images per class for training, and use all the same 50 testing images per class for testing. Fig. 1 summarizes the average performance over these 10 splits.

Baselines. We compare our MT-SVM against state-of-the-art HTL approaches with our UUS, including Non-Transfer, Multi-KT (Tommasi, Orabona, and Caputo 2014), Enhanced E-SVM (EE-SVM) (Aytar and Zisserman 2012), and GreedyTL (Kuzborskij, Caputo, and Orabona 2015).

Again, Fig. 1 shows that our approach performs significantly better than the non-transfer baseline for small-sample learning in large-scale scene classification. This indicates that our UUSs demonstrate *expressive and universal* capability for novel categories with considerable domain shift. More importantly, Fig. 1 shows that ours outperforms state-of-the-art HTL approaches with the same unsupervised sources. While they work under well-trained category source classifiers, EE-SVM and Multi-KT (with solely ℓ_2 or ℓ_1 regularization to learn transfer weights) perform poorly here due to generic weak sources and induced negative transfer. Ours is also superior to GreedyTL (with joint ℓ_2 and ℓ_0 regularization) by avoiding potential bad local minima compared to using greedy schemes to solve ℓ_0 problems. This observation reveals that our system manages to select informative candidate sources while discarding irrelevant ones, making such an approach preferable in ultra-large-scale scenarios.

Parameter Sensitivity Analysis. We also conduct a sensitivity experiment for the case of 1 training example per class. We fix $\lambda = 1, \alpha = 10$, and vary γ. For convenience, we parameterize γ by $m = \gamma/(\alpha + \gamma)$, so that m is always valued within $[0, 1]$. As shown in Table 3, there is a fairly smooth and flat region around Acc $= 15.2\%$.

Dataset Sensitivity Analysis. In the previous experiments, we used UUSs generated on ILSVRC for purpose of reproducibility without introducing extra data. To test the robustness of the pool of source hypotheses to the choice of dataset, we produce another library of $20K$ UUSs on PASCAL 2007, denoted as UUS-PASCAL. Given that PASCAL is 2 orders of magnitude smaller than ILSVRC, we first generate around $2K$ region proposals for each image on PASCAL using selective search (Uijlings et al. 2013), and extract their CNN features. In the feature space constructed by unlabeled proposals, we produce UUSs as before. As shown in Fig. 1, UUS-PASCAL outperforms UUS-ILSVRC. By using another large-scale dataset, we basically have more data beyond the original ILSVRC. Similar to the case of training

models on the training dataset and tuning their parameters on another validation dataset, it would potentially prevent over-fitting and provide more generalization ability.

Conclusions

We have drawn attention to a largely-overlooked yet fundamental problem with respect to *sources* in hypothesis transfer learning. We addressed a challenging transfer scenario and introduced a systematic scheme for generating a large library of source hypotheses in an unsupervised and discriminative way by bridging hypotheses with binary codes, two previously distinct areas. Without a bias to a particular set of categories, the produced hypotheses encode the intrinsic structure of the visual space. We also proposed a principled, scalable approach to automatically selecting informative sources and incorporating them to infer the target model. Our key technical contribution is the use of max-margin formulations with proper regularizations as a principled way for both source generation and transfer learning. The resulting models are accurate in recognition performance and efficient in transfer process and size of training data. This thus suggests promising future work towards integrating both pre-trained features and pre-trained models on unlabeled data.

Acknowledgments. We thank Liangyan Gui, David Fouhey, and Carl Doersch for valuable and insightful discussions. This work was supported in part by U.S. Army Research Laboratory (ARL) under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016, and by an AWS in Education Coursework Grant.

References

Aytar, Y., and Zisserman, A. 2011. Tabula rasa: Model transfer for object category detection. In *ICCV*.

Aytar, Y., and Zisserman, A. 2012. Enhancing exemplar SVMs using part level transfer regularization. In *BMVC*.

Bergamo, A., and Torresani, L. 2014. Classemes and other classifier-based features for efficient object categorization. *IEEE TPAMI* 36(10):1988–2001.

Chattopadhyay, R.; Ye, J.; Panchanathan, S.; Fan, W.; and Davidson, I. 2011. Multisource domain adaptation and its application to early detection of fatigue. In *SIGKDD*.

Choi, J.; Rastegari, M.; Farhadi, A.; and Davis, L. 2013. Adding unlabeled samples to categories by learned attributes. In *CVPR*.

Dai, D., and Gool, L. 2013. Ensemble projection for semi-supervised image classification. In *ICCV*.

Daumé III, H. 2007. Frustratingly easy domain adaptation. In *ACL*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*.

Duan, L.; Tsang, I. W.; Xu, D.; and Chua, T.-S. 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*.

Fernando, B.; Habrard, A.; Sebban, M.; and Tuytelaars, T. 2013. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*.

Gong, B.; Grauman, K.; and Sha, F. 2013. Reshaping visual datasets for domain adaptation. In *NIPS*.

Hoffman, J.; Kulis, B.; Darrell, T.; and Saenko, K. 2012. Discovering latent domains for multisource domain adaptation. In *ECCV*.

Hoffman, J.; Rodner, E.; Donahue, J.; Darrell, T.; and Saenko, K. 2013. Efficient learning of domain-invariant image representations. In *ICLR*.

Hoffman, J.; Tzeng, E.; Donahue, J.; Jia, Y.; Saenko, K.; and Darrell, T. 2014. One-shot adaptation of supervised deep convolutional models. In *ICLR*.

Kienzle, W., and Chellapilla, K. 2006. Personalized handwriting recognition via biased regularization. In *ICML*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Kuzborskij, I., and Orabona, F. 2013. Stability and hypothesis transfer learning. In *ICML*.

Kuzborskij, I., and Orabona, F. 2015. Fast rates by transferring from auxiliary hypotheses. arXiv:1412.1619.

Kuzborskij, I.; Caputo, B.; and Orabona, F. 2015. Transfer learning through greedy subset selection. In *ICIAP*.

Kuzborskij, I.; Orabona, F.; and Caputo, B. 2013. From N to N+1: Multiclass transfer incremental learning. In *CVPR*.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. 2006. Efficient sparse coding algorithms. In *NIPS*.

Liu, B.-D.; Wang, Y.-X.; Shen, B.; Zhang, Y.-J.; and Hebert, M. 2014. Self-explanatory sparse representation for image classification. In *ECCV*.

Pan, S., and Yang, Q. 2010. A survey on transfer learning. *IEEE TKDE* 22(10):1345–1359.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. 2007. Self-taught learning: Transfer learning from unlabeled data. In *ICML*.

Rastegari, M.; Farhadi, A.; and Forsyth, D. 2012. Attribute discovery via predictable discriminative binary codes. In *ECCV*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115(3):211–252.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.

Tommasi, T.; Orabona, F.; and Caputo, B. 2014. Learning categories from few examples with multi model knowledge transfer. *IEEE TPAMI* 36(5):928–941.

Uijlings, J.; van de Sande, K.; Gevers, T.; and Smeulders, A. 2013. Selective search for object recognition. *IJCV* 104(2):154–171.

Wang, Y.-X., and Hebert, M. 2015. Model recommendation: Generating object detectors from few samples. In *CVPR*.

Weston, J.; Collobert, R.; Sinz, F.; Bottou, L.; and Vapnik, V. 2006. Inference with the universum. In *ICML*.

Xiao, J.; Ehinger, K.; Hays, J.; Torralba, A.; and Oliva, A. 2014. Sun database: Exploring a large collection of scene categories. *IJCV* 1–20.

Yang, J., and Hauptmann, A. G. 2008. A framework for classifier adaptation and its applications in concept detection. In *MIR*.

Yang, J.; Yan, R.; and Hauptmann, A. 2007a. Adapting SVM classifiers to data with shifted distributions. In *ICDM Workshops*.

Yang, J.; Yan, R.; and Hauptmann, A. 2007b. Cross-domain video concept detection using adaptive SVMs. In *ACM MM*.

Zhu, X.; Anguelov, D.; and Ramanan, D. 2014. Capturing long-tail distributions of object subcategories. In *CVPR*.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *JRSSB* 67(2):301–320.