Learning to Learn: Model Regression Networks for Easy Small Sample Learning

Yu-Xiong Wang(✉) and Martial Hebert

Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
{yuxiongw,hebert}@cs.cmu.edu

Abstract. We develop a conceptually simple but powerful approach that can learn novel categories from few annotated examples. In this approach, the experience with already learned categories is used to facilitate the learning of novel classes. Our insight is two-fold: (1) there exists a *generic, category agnostic* transformation from models learned from few samples to models learned from large enough sample sets, and (2) such a transformation could be effectively learned by high-capacity regressors. In particular, we automatically learn the transformation with a deep model regression network on a large collection of model pairs. Experiments demonstrate that encoding this transformation as prior knowledge greatly facilitates the recognition in the small sample size regime on a broad range of tasks, including domain adaptation, fine-grained recognition, action recognition, and scene classification.

Keywords: Small sample learning · Transfer learning · Object recognition · Model transformation · Deep regression networks

1 Motivation

Over the past decade, large-scale object recognition has achieved high performance levels due to the integration of powerful machine learning techniques with big annotated training data sets [38,51,52,62,79,83,84]. In practical applications, however, training examples are often expensive to acquire or otherwise scarce [30]. Visual phenomena follow a long-tail distribution, in which a few subcategories are common while many are rare with limited training data even in the big-data setting [105,106]. More crucially, current recognition systems assume a set of categories known a priori, despite the obviously dynamic and open nature of the visual world [12,32,64,96].

Such scenarios of learning *novel categories from few examples* pose a multitude of open challenges for object recognition in the wild. For instance, when operating in natural environments, robots are supposed to recognize unfamiliar objects after seeing only few examples [50]. Humans are remarkably able to grasp a new category and make meaningful generalization to novel instances from just a short exposure to a single example [30,81]. By contrast, typical machine learning tools require tens, hundreds, or thousands of training examples and often break down for small sample learning [7,40].

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 616–634, 2016.
DOI: 10.1007/978-3-319-46466-4_37

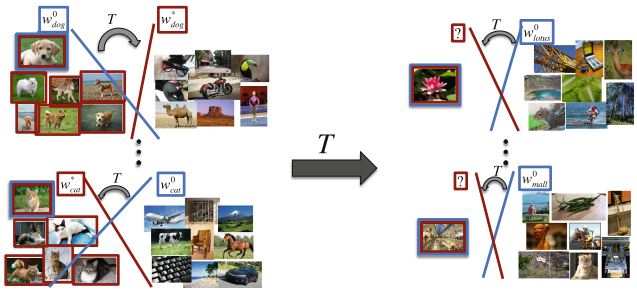

Fig. 1. Our main hypothesis is that there exists a generic, category agnostic transformation T from classifiers $\mathbf{w}^0$ learned from few annotated samples (represented as blue) to the underlying classifiers $\mathbf{w}^*$ learned from large sets of samples (represented as red). We estimate the transformation T by learning a deep regression network on a large collection of model pairs, i.e., a model regression network. For a novel category/task (such as scene classification and fine-grained object recognition), we introduce the learned T to construct the target model and thus facilitate its generalization in the small sample size regime (Color figure online)

In this paper, we explore a novel *learning to learn* approach that leverages the knowledge gained when learning models in large sample sets to facilitate recognizing novel categories from few samples. From a discriminative machine learning perspective, object recognition is basically a process that learns an object category classifier to separate annotated positive and negative examples in a feature space. We assume a *fixed, discriminative* feature space, which is reasonable especially considering the recent learned feature representations via deep convolutional neural networks. We now take the model such as SVM classifiers and make important modification. The central issue can be reduced to the following: How to estimate a classifier that would be learned from a large set of samples (on the order of hundreds or thousands of) based on its corresponding classifier learned from few annotated samples (as few as one and up to a hundred)?

Our main hypothesis is that there exists a *generic, category agnostic* transformation from small-sample models to the underlying large-sample models. This hypothesis is validated empirically in Sect. 4. Intuitively, a model can be viewed as a separating hyperplane in the feature space.[1] Small training examples already constrain the search space by pointing to an initial hyperplane not far from the desired hyperplane produced by a large training set. When gradually introducing additional examples, the initial hyperplane is progressively subject to a series of transformations until it converges as illustrated in Fig. 1.

We suspect that this transformation, or at least certain components of it, is fairly generic. In a machine learning context, a learner needs to be biased in some way for it to generalize well [9,30,40,81]. Consequently, there might exist some systematic bias from a small-sample model to its large-sample version. In essence,

[1] A kernel model can be viewed as a separating hyperplane in the lifted feature space.

this transformation potentially captures the natural intra-class variability in a discriminative manner and represents how sparse samples change to a category cluster. Hence, we view the model transformation as a form of shared structure and, when available, it can be re-purposed for novel categories.

A desirable goal, then, is to find ways of automatically learning such a transformation. We achieve this by learning a deep regression network on a large collection of model pairs, which we term as a *model regression network*. The network explicitly regresses between the small-sample classifiers (as input) and their corresponding large-sample classifiers (as ground-truth) on a variety of known categories. The deep learning framework enables us to learn the transformation without imposing strong priors. Now, for a novel category/task, we introduce the learned transformation to construct the target model and thus facilitate its generalization in the small sample size regime.

Our approach is inspired by the recent observation in deep learning based object recognition that features extracted from deep convolutional neural networks trained on a large set of particular object categories exhibit attractive transferability [4,20,76,104]. They could thus serve as universal feature extractors for novel categories/tasks. Our key insights then are that such generality would also hold on a model level and that it would be learnable in a similar fashion as on the feature level. This is also suggested by the duality perspective between the feature space and the classifier space [91]. Eventually, the transformation can be also viewed to be imposed on features but parametrized in a model fashion.

Our contribution is three-fold: First, we show how to construct a training "model set" by generating a large collection of model pairs that are learned from small and large sample sets respectively on various categories (Sect. 3.1). Second, we show how a model regression network, based on deep neural networks and this training model set, is learned and a generic transformation between these two types of models is identified by the regressor (Sects. 3.2 and 3.3). Finally, we show how our regression network is used to facilitate the recognition of novel categories from few samples, leading to significantly improved performance on a broad range of tasks, including domain adaptation, fine-grained recognition, action recognition, and scene classification (Sects. 3.4 and 4).

2 Related Work

It remains a fundamental challenge to understand how to recognize novel categories from few examples for both humans and machines. This line of research is generally addressed in the fields of one/few-shot learning [26], inductive transfer or transfer learning [70], multi-task learning [14], learning to learn [86], and meta-learning [82]. Because of high-dimensionality of feature spaces, successful generalization from small training samples typically requires strong and appropriately tuned "inductive biases" using additional available information [9,40].

A natural source of information comes from additional data via "data manufacturing" [7] in various ways. For instance, (1) obtain more examples of categories of interest from large amounts of unlabeled data as in semi-supervised

learning [15,107] and active learning [73], (2) augment the available examples by performing simple image transformations including jittering and noise injection as commonly used in deep learning [16,22,52], (3) borrow examples from other relevant categories [61], (4) introduce Universum examples (i.e., unlabeled examples that do not belong to the concerned classes) for max-margin regularization [98], and (5) synthesize new virtual examples, either rendered explicitly with computer graphics techniques or created implicitly through compositional representations [18,21,66,67,71,106]. These approaches can significantly improve recognition performance if a generative model that accounts for the underlying, natural intra-class variability is known. Unfortunately, such a model is usually unavailable [7] and the generation of additional real or artificial examples often requires substantial effort.

In a broad sense, learning novel categories is addressed by exploiting and transferring knowledge gained from familiar categories [14,70,72,77,86,87]. This is to imitate the human ability of adapting previously acquired experience when performing a new task [74]. In particular, inter-class transfer [40] and cross-generalization [7] are achieved by discovering shared feature representations: (1) captured by linear or nonlinear feature transformations [1,14,31,48,63,85,94], (2) obtained by feature selection [27,59,60] or regularization [37], (3) described by similarities between novel classes and familiar classes [8], (4) encoded as a distance metric by metric learning [10,11,29,75,92,100] or kernel learning [40], and (5) learned by boosting approaches [69,89,101]. Recently, there has been growing interest in learning deep convolutional neural networks in fully supervised, semi-supervised, or unsupervised fashions to extract generic features and then to transfer them to different tasks [19,22,33,35,46,49,52,65,83,95,99].

Another type of knowledge transfer focuses on modeling (hyper-)parameters that are shared across domains, typically in the context of generative statistical modeling [25,58,78]. A variational Bayesian framework is first developed by incorporating previously learned classes into the prior and combining with the likelihood to yield a new class posterior distribution [25,26]. Gaussian processes [57,78] and hierarchical Bayesian models [81] are also employed to allow transferring in a non-parametric Bayesian way. The recently proposed hierarchical Bayesian program learning utilizes the principles of compositionality and causality to build a probabilistic generative model of visual objects [54–56]. In addition, adaptive SVM and its variants present SVM-based model adaptation by combining classifiers learned on related categories [2,3,23,47,53,88,97,102]. Other approaches transfer the knowledge across different modalities [6,32,36]. Despite many notable successes, it is still unclear what kind of underlying structures are shared across a wide variety of categories and are useful for transfer.

Different from the previous work, we propose a plausible alternative for transferring inter-class structure from a model perspective. This paper is the first to show that there exists certain generic, category agnostic transformation between small-sample and large-sample models on a wide spectrum of categories. In addition, such a transformation could be effectively learned by high-capacity regressors, such as deep neural networks, in a model-level big-data setting. Our

approach could also be seen as an alternative parametric way of doing model distillation that relies on the connection between different models [5,13,41].

3 Model Regression Networks

We are given a fixed, discriminative feature space $\mathcal{X}$ of dimensionality d, such as the current deep convolutional neural network features.[2] For an object category c of interest, we generate a model or classifier $h(\mathbf{x})$ that discriminates between its positive and negative instances $\mathbf{x} \in \mathcal{X}$. We consider, for example, the linear SVM classifier commonly used for object recognition tasks, which is a separating hyperplane in the feature space. The classifier $h(\cdot)$ can then be represented as a weight vector $\mathbf{w}$ belonging to the model parameter space $\mathcal{W}$.

Let $\mathbf{w}^0$ indicate a classifier learned from few annotated samples *without any additional information*. Let $\mathbf{w}^*$ indicate the corresponding *underlying* classifier learned from a large set of annotated samples of the same category. Our goal is to generate $\mathbf{w}$ (or equivalently, $h(\cdot)$) that generalizes well from these few training examples, i.e., to make $\mathbf{w}$ as close as to the desired $\mathbf{w}^*$. The key assumption is that there exists a generic non-linear transformation $\widetilde{T} : \mathcal{W} \rightarrow \mathcal{W}$ for a broad range of categories, so that for $\mathbf{w}^0$ and $\mathbf{w}^*$ in any category c, we have $\mathbf{w}^* \approx \widetilde{T}\left(\mathbf{w}^0\right)$. That is, there is a set of large-sample models and $\widetilde{T}$ is the projection into that set (with $\mathbf{w}^*$ being a fixpoint of $\widetilde{T}$). Once the transformation $\widetilde{T}$ is available, we could easily improve the classifier generalization.

Inspired by recent progress in deep learning, it is possible to estimate this transformation $\widetilde{T}$ from a large set of known categories. A straightforward approach then is to learn a regression function T parameterized by Θ based on a large collection of "annotated" model pairs $\left\{\left(\mathbf{w}_j^0, \mathbf{w}_j^*\right)\right\}_{j=1}^J$ from these categories. That is, $\mathbf{w}_j^* \approx T\left(\mathbf{w}_j^0, \Theta\right)$ for any small-sample model $\mathbf{w}_j^0$ and its large-sample model $\mathbf{w}_j^*$ learned on the same category. We employ multi-layer neural networks as regressors, which are well-known to learn complex, non-linear functions with minimal human design. By doing so, we avoid an explicit description of the space of transformations. We then use the obtained transformation in learning models for novel categories.

3.1 Generation of Model Pairs

We start from large amounts of labeled data from a variety of categories, denoted as $\{(\mathbf{x}_i, y_i)\}_{i=1}^L$. Here $\mathbf{x}_i \in \mathbb{R}^d$ is the ith data sample in the feature space $\mathcal{X}$, $y_i \in \{1, \ldots, C\}$ is the corresponding label, and C is the number of categories. Different from conventional recognition systems that directly learn from the data and label pairs, we learn on a model level. To this end, we produce a collection of model pairs $\left\{\left(\mathbf{w}_j^0, \mathbf{w}_j^*\right)\right\}_{j=1}^J$ as our *training model set* using the original training

[2] Notation: We use boldface letters for vectors and matrices and italicized capital letters for transformation functions. For notational simplicity, $\mathbf{x}$ already includes a constant 1 as the last element and thus $\mathbf{w}$ includes the bias term.

data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$. Each model is generated as a binary classifier focused on separating a single category from all the remaining categories in a manner inspired by the one-vs.-all strategy in multi-class classification.

Specifically, for each category c, we first learn $\mathbf{w}^{c,*}$ from a large sample set. We treat $\mathbf{w}^{c,*}$ as the *ground-truth model*. Let the positive examples $\{\mathbf{x}_i^{c,pos}\}_{i=1}^{L_c}$ be all the data points of category c, where L_c is the total number of samples whose labels are c. We obtain negative examples $\{\mathbf{x}_i^{c,neg}\}_{i=1}^{M}$ by randomly sampling M data points from other categories not in category c. We train a binary SVM classifier $\mathbf{w}^{c,*}$ on the training set $\mathcal{P}^c = \{(\mathbf{x}_i^{c,pos}, +1)\}_{i=1}^{L_c} \cup \{(\mathbf{x}_i^{c,neg}, -1)\}_{i=1}^{M}$.

We now learn the small-sample model $\mathbf{w}^{c,0}$ for category c. Consistent with the few-shot scenario that consists of few positive examples, we randomly sample $N \ll L_c$ data points $\{\mathbf{x}_i^{c,pos}\}_{i=1}^{N}$ out of the L_c positive examples of category c. We train a binary SVM classifier $\mathbf{w}^{c,0}$ on the reduced training set $\mathcal{Q}^c = \{(\mathbf{x}_i^{c,pos}, +1)\}_{i=1}^{N} \cup \{(\mathbf{x}_i^{c,neg}, -1)\}_{i=1}^{M}$.

Note that we have many ways of choosing the small sample set for a given $\mathbf{w}^{c,*}$ to learn $\mathbf{w}^{c,0}$. This indicates that we could repeat the sampling procedure S times, leading to S small-sample models $\left\{\mathbf{w}_j^{c,0}\right\}_{j=1}^{S}$ learned from different small-sample sized training subset $\left\{\mathcal{Q}_j^c\right\}_{j=1}^{S}$ of $\mathcal{P}^c$. Since they correspond to the unique ground-truth model, we thus obtain a series of model pairs for category c as $\left\{\left(\mathbf{w}_j^{c,0}, \mathbf{w}^{c,*}\right)\right\}_{j=1}^{S}$. Including the learned model pairs from all the C categories, we generate the desired training model set $\left\{\left(\mathbf{w}_j^0, \mathbf{w}_j^*\right)\right\}_{j=1}^{J}$, where $J = S \times C$. Due to sub-sampling, the size of the training model set could be potentially large, with many orders of magnitude larger than the number of categories.

3.2 Regression Network

Given the training model set $\left\{\left(\mathbf{w}_j^0, \mathbf{w}_j^*\right)\right\}_{j=1}^{J}$ with one to one model correspondence, we aim to learn a mapping: $\mathbf{w}^0 \rightarrow \mathbf{w}^*$. We parametrize the transformation as a regression function $T\left(\mathbf{w}^0, \Theta\right)$, such that $\mathbf{w}^* \approx T\left(\mathbf{w}^0, \Theta\right)$. We simply use the square of the Euclidean distance to quantify the quality of the approximation. For each model $\mathbf{w}_j^0$, we have the corresponding small sample set $\mathcal{Q}_j = \left\{\left(\mathbf{x}_i^j, y_i^j\right)\right\}_{i=1}^{M+N}$ used to learn the model as well. To make the regression more robust, we include the performance on these samples as an additional loss, which is standard in the transfer learning approaches with model parameter sharing [97,102]. Our final loss function then is

$$L\left(\Theta\right) = \sum_{j=1}^{J} \left\{ \frac{1}{2} \left\| \mathbf{w}_j^* - T\left(\mathbf{w}_j^0, \Theta\right) \right\|_2^2 + \lambda \sum_{i=1}^{M+N} \left[1 - y_i^j \left(T\left(\mathbf{w}_j^0, \Theta\right)^T \mathbf{x}_i^j \right) \right]_+ \right\}. \quad (1)$$

The second term represents the data fitting on the training samples. Here, the performance loss is measured by a hinge loss, and it could be other types of losses such as a logistic loss as well.

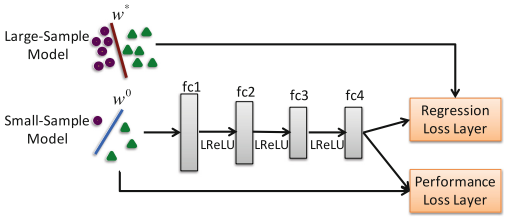

Fig. 2. The architecture of our model regression network. Given a model $\mathbf{w}^0$ learned from few samples as input, it is passed though four fully-connected layers with leaky ReLU. On the loss layer, a model regression loss and a classification performance (e.g., hinge) loss on the training data is minimized jointly

Consistent with recent work, we use a multi-layer feed-forward neural network as the regression function for its high capacity. As shown in Fig. 2, our regression network consists of $F = 4$ fully-connected layers where the fth layer applies a non-linear transformation G, which is an affine transformation followed by a non-linear activation function. We use leaky ReLU. For the purpose of regression capacity, the number of units in the first two layers is larger than the dimensionality of the input classifier weight vectors. The desired transformation T is then represented as a series of transformations G layer by layer.

3.3 Implementation Details

For the feature space, consistent with recent work, we use the Caffe Alexnet convolutional neural network (CNN) feature pre-trained on ILSVRC 2012 [20,45,52]. All the weights of the CNN are frozen to those learned on ILSVRC without fine-tuning on any other datasets. For each image, we extract the feature on the center 224×224 crop of the 256×256 resized image. It is a $d = 4{,}096$-dim feature vector $fc6$ taking from the penultimate hidden layer of the network, unless otherwise specified.

To generate the training model set, we use the ILSVRC 2012 training data set for purpose of reproducibility. There are 1,000 object categories with 600 to 1,300 images per category and 1.2 million images in total. We use Liblinear [24] to train linear SVM models $\mathbf{w}^0$ and $\mathbf{w}^*$. For each category, using all the positive images and randomly sampled negative images, we train $\mathbf{w}^*$ with the optimal SVM regularization parameter obtained by 10-fold cross-validation. We then randomly sample $N = 1, 2, \ldots, 9, 10, 15, 20, \ldots, 100$ positive images. For each N, we repeat random sub-sampling $S = 5$ times, and use different SVM regularization parameters from $10^{\{-2,-1,0,1,2\}}$ to train the SVM model $\mathbf{w}^0$ from few samples. These are essentially valid ways of doing "data augmentation" [52] for training the regression network, which mimic in practice how $\mathbf{w}^0$ changes. Hence, the number of the generated model pairs is 700 for each category, and the size of the training model set is 700,000. Finally, we randomly split the set

with 685 model pairs as training and the remaining 15 pairs as validation per category.

We then use Caffe [45] to train our model regression network on the generated training model set and the corresponding training data set. The number of units from $fc1$ to $fc4$ are 6144, 5120, 4097, and 4097, respectively. We use 0.01 as the negative slope for leaky ReLU. λ is set to 1. We implement the loss function as two loss layers in Caffe, with one loss layer focusing on the model regression accuracy and the other focusing on the performance loss on the training data. We train the network using standard SGD and batch normalization [44].

3.4 Learning Target Models for Novel Categories

We now consider recognizing a novel category from a small labeled training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{K}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a data sample and $y_i \in \{-1, 1\}$ is the corresponding label. By leveraging the obtained generic model transformation T as informative prior knowledge, we aim to infer the target model $\mathbf{w}$ that generalizes better than the one produced only from the few training examples. We use a coarse-to-fine procedure that learns the target model in three steps: initialization, transformation, and refinement.

Initialization. In this first step, we directly learn the target model $\mathbf{w}^0$ on the small training sample set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{K}$.

Transformation. Using $\mathbf{w}^0$ as input to our learned model regression network, after forward propagation, we obtain the output model $T\left(\mathbf{w}^0, \Theta\right)$. This thus encodes the prior knowledge about $\mathbf{w}$ being preferable.

Refinement. We then introduce $T\left(\mathbf{w}^0, \Theta\right)$ as biased regularization into the standard SVM max-margin formulation to retrain the model by minimizing

$$R\left(\mathbf{w}\right) = \frac{1}{2}\left\|\mathbf{w} - T\left(\mathbf{w}^0, \Theta\right)\right\|_2^2 + \eta \sum_{i=1}^{K} \left[1 - y_i\left(\mathbf{w}^T \mathbf{x}_i\right)\right]_+. \tag{2}$$

Equation (2) is similar to the standard SVM formulation, with the only difference being the bias towards $T\left(\mathbf{w}^0, \Theta\right)$ instead of 0. η is the regularization parameter used to control the trade-off between the regularization term and data fitting term. We thus obtain an intermediate solution with a decision boundary close to the regressed classifier while separating the labeled examples well.

4 Experimental Evaluation

In this section, we explore the use of our learned model regression network on a number of supervised learning tasks with limited data, including domain adaptation, fine-grained recognition, action recognition, and scene classification. We begin with a sanity check of the regression network for the 1,000 training categories on the ILSVRC validation data set. We then evaluate the network

Fig. 3. Performance sanity check of the model regression network by comparing small-sample models $\mathbf{w}^0$, large-sample models $\mathbf{w}^*$ (learned on thousands of examples), and regressed models $T(\mathbf{w}^0)$ on the held-out ILSVRC validation data set. X-axis: number of positive training examples. Y-axis: average binary classification accuracy. Our network effectively identifies a generic model transformation

for one-shot domain adaptation and compare with state-of-the-art adaptation approaches. We further evaluate our approach for novel fine-grained, action, and scene categories. Finally, we present experimental results evaluating the impact of different feature spaces and model types.

4.1 Sanity Check

Our model regression network is learned from the 1,000 categories on the ILSVRC training data set. As a sanity check, the first question to answer is whether the learned transformation indeed improves generalization of the small-sample models for these categories. To answer this question, we evaluate the models on the held-out ILSVRC validation data set, which contains the same 1,000 categories with 50 images per category and has no overlap with the ILSVRC training data.

Consistent with the way the models are generated, we evaluate them in a binary classification scenario. For each category, we construct a test set consisting of all these 50 positive images and 50 randomly sampled negative images from other categories. We compare the three types of models: small-sample models $\mathbf{w}^0$, large-sample models $\mathbf{w}^*$ (as ground-truth), and regressed models $T(\mathbf{w}^0)$ (without the refinement step). We evaluate how performance varies with the number of positive training examples N when used to learn $\mathbf{w}^0$. We average the classification accuracy over the models corresponding to the same N but with different sampled training data and SVM regularization parameters. Figure 3 summarizes the average performance over the 1,000 categories.

As expected, Fig. 3 shows that $T(\mathbf{w}^0)$ significantly improves the generalization of $\mathbf{w}^0$. In the one-shot learning case, there is a notable 20 % performance improvement of $T(\mathbf{w}^0)$ over $\mathbf{w}^0$, whose performance is only a little bit higher than chance (50 % for binary classification). With increased number of training

Table 1. Performance comparison between our model transformation with state-of-the-art approaches that adapt other types of prior knowledge gained on the ILSVRC source domain in manners of data, feature, model parameter, and joint fine-tuning for one-shot learning on the Webcam domain of the Office dataset

Source prior knowledge type	Method	Acc (%)
NA	SVM (target only) [43]	62.28
Data	SVM (source only) [43]	53.51
	SVM (source and target) [43]	56.68
Feature	GFK [34]	65.16
	SA [28]	59.30
	Daumé III [17]	59.21
	MMDT [42]	59.21
Model parameter	PMT [2]	66.30
	Late fusion (Max) [43]	59.59
	Late fusion (Lin. Int. Avg) [43]	60.64
Joint	Fine-tuning [43]	61.13
Model transformation	Model regression network (Ours)	**68.47**

examples, the performance of $T(\mathbf{w}^0)$ gradually converges to that of $\mathbf{w}^*$ trained on thousands of examples. This verifies the existence of a generic transformation from small-sample to large-sample models for these 1,000 categories, which is effectively identified by our model regression network. In the following experiments, we will show that the learned transformation applies to other novel categories as well.

4.2 One-Shot Adaptation

Our approach can be viewed as transferring certain prior knowledge gained from the source domain (ILSVRC) to new tasks. It is thus interesting to compare different types of prior knowledge, including those on data, feature, and model parameter levels. To this end, we provide a comprehensive evaluation in the scenario of domain adaptation, in which the target images come from the same set of source categories but are drawn from a different distribution. Due to the common categories between source and target domains, this experimental setup allows us to best identify the possible shared domain structure and compare with state-of-the-art adaptation approaches without learning additional category correspondence, which turns to be another difficult problem.

Datasets and Tasks. We evaluate on the Office dataset [80], a standard domain adaptation benchmark for multi-class object recognition. The Office dataset is a collection of 4,652 images from three distinct domains: Amazon, DSLR, and Webcam. We use Webcam as the target domain since it was shown to be the most challenging shifted domain [43]. Of the 31 categories in the dataset, 16

overlap with the categories presented in the 1,000-category ILSVRC. We focus on these common classes as our target (i.e., 16-way classification), as is customary in [43]. Following a similar experimental setup in [43], 1 labeled training and 10 test images per category are randomly selected on Webcam. We report average multi-class accuracy over 20 random train/test splits in Table 1.

Baselines. In addition to the SVM (target only) baseline that directly trains SVM classifiers on the target data, we compare against four other types of baselines that transfer prior knowledge on the ILSVRC source domain gained in manners of data, feature, model parameters, and joint fine-tuning. **Type I data level**: SVM classifiers trained on only source data and both source and target data, respectively. **Type II feature level**: geodesic flow kernel (GFK) [34], subspace alignment (SA) [28], Daumé III [17], and max-margin domain transforms (MMDT) [42], which seek common feature spaces using learned feature embedding, augmentation, or transformation. **Type III model parameter level**: projective model transfer (PMT) [2] and late fusion [43], which adapt the parameters of the pre-trained source classifier to construct the target classifier. **Type IV joint level**: fine-tune the weights of the pre-trained CNN on the 16-way target classification task. These results are reported from [43].

Table 1 shows that our model transformation provides an alternative, competitive way to encode the shared structure and prior knowledge. It is on par with or outperforms other types of prior knowledge and adaption approaches. Notably, ours achieves significantly better performance than fine-tuning, the standard transfer strategy for CNNs, in this one-shot learning scenario. Fine-tuning requires a considerable amount of labeled target data and actually reduces performance in the very sparse label regime.

4.3 Learning Novel Categories

We now evaluate whether our learned model regression network facilitates the recognition of novel categories from few samples. For multi-class classification on the target datasets, we test how performance varies with the number of training samples per category. Following the standard practice, we train linear SVMs in a one-vs.-all fashion with default settings in Liblinear [24]. After obtaining the regressed models, we then incorporate them to retrain each one-vs.-all classifier.

Datasets and Tasks. We evaluate on standard benchmark datasets for fine-grained recognition: Caltech-UCSD Birds (CUB) 200-2011 [93] and Oxford 102 Flowers [68], for action recognition (compositional semantic recognition): Stanford-40 actions [103], and for scene classification: MIT-67 [90]. We follow the standard experimental setups (e.g., the train/test splits) for these datasets: **CUB**200-2011 contains 11,788 images of 200 bird species; 5,994 images are used for training (29 or 30 images per class) and 5,794 for testing. 102 **Flowers** contains 102 flower classes and each class consists of 40–258 images; 10 images per class are used as training data and the rest are used as test data. **Stanford-**40 contains $9,532$ images of humans performing 40 actions with 180–300 images per action class; 100 images per class are used as training data and the rest are used

as test data. **MIT-**67 contains 15,620 images spanning 67 indoor scene classes; the provided split for this dataset consists of 80 training and 20 test images per class. In our experiments, due to the lack of published protocols for small-sample learning, we randomly generate the small-sample version of training images as shown in Fig. 4 and use all the same test images for testing.

Baselines. Due to the CNN training procedure, the original models directly learned from target samples can be viewed as transfer learning with feature sharing. We also include the transfer learning baseline with model parameter sharing on Stanford-40 and MIT-67, which transfers the 1,000 ILSVRC category models using [88]. Moreover, we report an additional CNN fine-tuning baseline on MIT-67, which is the best fine-tuning result we have achieved following [39].

Figure 4 summarizes the average performance over 10 random splits on these datasets. The performance of the model transfer is similar to the original models learned from few samples due to the dissimilarity between source and target tasks. In our case of limited target data, the standard fine-tuning approach leads to degraded performance due to over-fitting. The models refined by our regression network, however, significantly outperform them for a broad range of novel categories. Our approach has particularly large performance boosts in one-shot learning scenarios. For example, there is a nearly 15 % boost on MIT-67.

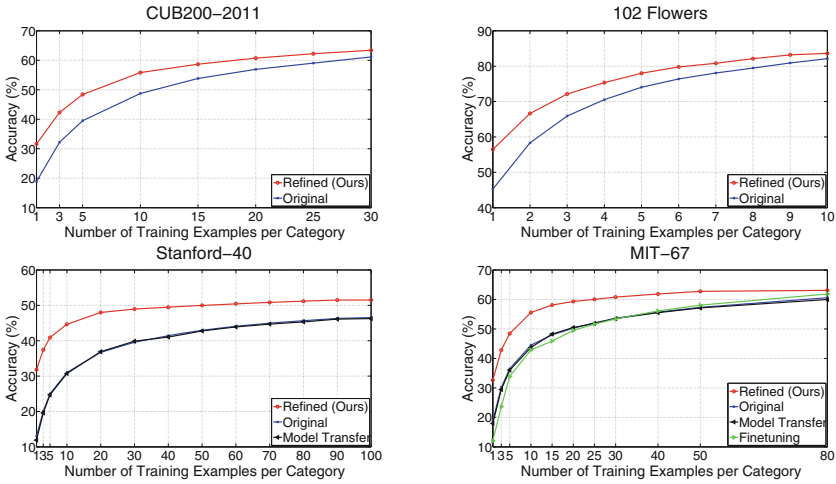

Fig. 4. Performance comparison between models learned from few samples and models refined by our model regression network for fine-grained recognition, action recognition, and scene classification on four benchmark datasets. For completeness, we also include additional baselines of transfer learning with model parameter sharing and CNN fine-tuning on certain datasets. The Alexnet CNN is used as the feature space. X-axis: number of training examples per class. Y-axis: average multi-class classification accuracy. Since they benefit from the learned generic model transformation, ours significantly outperform all the baselines for small sample learning

4.4 Evaluation of Different Feature Spaces

In the previous experiments, we used the Alexnet CNN as the feature. To test the robustness of our model regression network to the choice of the feature space, here we evaluate two additional features: the more powerful VGG19 CNN [83] $fc7$, pre-trained on ILSVRC 2012, and the unsupervised CNN [95] $fc6$, pre-trained on YouTube videos. We keep the other design choices the same (e.g., the way of generating the training model set and the regression network structure). In a similar way as before, we train our network and evaluate the recognition performance on the target tasks with few samples. Figure 5 validates the benefit of our approach in different feature space settings. Importantly, it shows that the data used to estimate the model transformation (ILSVRC) is not necessarily the same as the data used to learn the feature representation (YouTube).

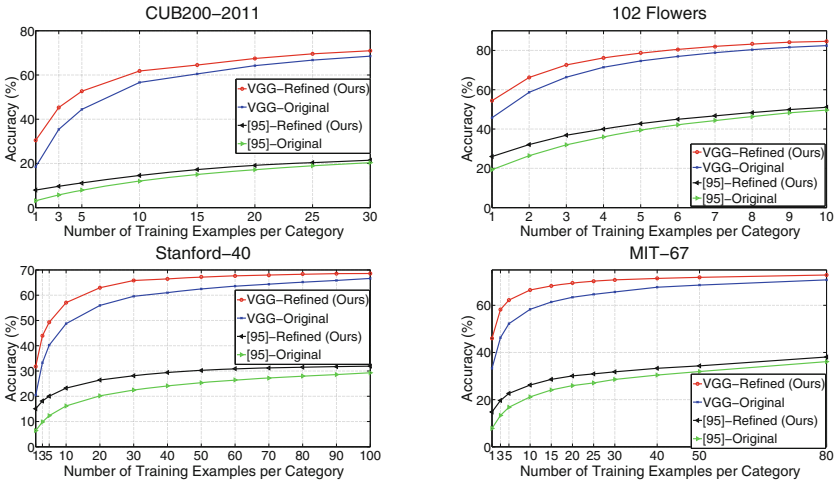

Fig. 5. Feature space evaluation between models learned from few samples and models refined by our model regression network on these four benchmark datasets. The stronger VGG CNN [83], pre-trained on ILSVRC, and the unsupervised CNN [95], pre-trained on YouTube, are used as the feature space, respectively. Ours show consistent performance improvements over the original models for small sample learning in different feature spaces

4.5 Evaluation of Different Types of Classification Models

In the previous experiments, we focused on SVM classifiers. In fact, the models do not need to come from max-margin classifiers and could be other set of weights learned in different fashions. To verify this, we test a widely used alternative classifier, logistic regression, and keep the other design choices the same (e.g., the way of generating the training model set and the regression network

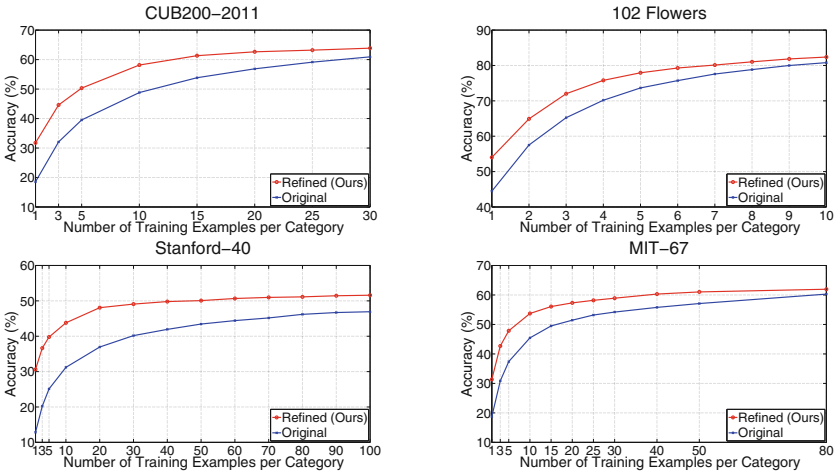

Fig. 6. Model type evaluation between models learned from few samples and models refined by our model regression network on these four benchmark datasets. We evaluate the logistic regression as the model of interest. The robust performance shows generic transformations for different types of models

structure). Naturally, we change the hinge loss to the logistic loss. In a similar way as before, we train our network and evaluate the recognition performance on the target tasks with few samples as shown in Fig. 6. Combining with Fig. 4, the logistic regression demonstrates comparable performance to SVM, and the refined logistic regression classifiers generalize better as well.

5 Conclusions

Even though it has long been believed that learning algorithms should be able to induce general functions not only from examples but also from experience as humans, it is still unclear what types of knowledge are shared across tasks and crucial for transfer. In this work we proposed a conceptually simple but powerful approach to address the problem of small sample learning in this context of learning to learn. Our approach is based on the insight that there exists a generic, category agnostic transformation T from small-sample models to the underlying large-sample models. In addition, such a transformation could be effectively learned by high-capacity regressors on a large collection of model pairs and could be later used as informative prior for learning novel categories. This work opens up several interesting questions and could be explored further. While we focused on the existence of the transformation here, it would be interesting to design the best network architecture and other types of regressors (e.g., kernelized ridge regression) to learn the transformation. Also, we have assumed that the transformation T is independent of the sample size whereas, in general, one would

envision that T would change when the number of samples increases dramatically all the way to T = identity for very large training sample sets. Finally, while we assumed a fixed representation, it would be interesting to extend this approach for use of a loss to inform modification of features as well.

Acknowledgments. We thank Liangyan Gui, David Fouhey, and Deva Ramanan for valuable and insightful discussions. This work was supported in part by ONR MURI N000141612007 and U.S. Army Research Laboratory (ARL) under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016. We also thank NVIDIA for donating GPUs and AWS Cloud Credits for Research program.

References

1. Amit, Y., Fink, M., Srebro, N., Ullman, S.: Uncovering shared structures in multiclass classification. In: ICML (2007)
2. Aytar, Y., Zisserman, A.: Tabula rasa: model transfer for object category detection. In: ICCV (2011)
3. Aytar, Y., Zisserman, A.: Enhancing exemplar SVMs using part level transfer regularization. In: BMVC (2012)
4. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: From generic to specific deep representations for visual recognition. In: CVPR Workshops (2015)
5. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: NIPS (2014)
6. Ba, J., Swersky, K., Fidler, S., Salakhutdinov, R.: Predicting deep zero-shot convolutional neural networks using textual descriptions. In: ICCV (2015)
7. Bart, E., Ullman, S.: Cross-generalization: learning novel classes from a single example by feature replacement. In: CVPR (2005)
8. Bart, E., Ullman, S.: Single-example learning of novel classes using representation by similarity. In: BMVC (2005)
9. Baxter, J.: A Bayesian/information theoretic model of learning to learn via multiple task sampling. Mach. Learn. **28**(1), 7–39 (1997)
10. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. arXiv preprint arXiv:1306.6709 (2013)
11. Ben-David, S., Schuller, R.: Exploiting task relatedness for multiple task learning. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 567–580. Springer, Heidelberg (2003)
12. Bendale, A., Boult, T.: Towards open world recognition. In: CVPR (2015)
13. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: KDD (2006)
14. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
15. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised Learning. Adaptive Computation and Machine Learning. The MIT Press, Cambridge (2006)
16. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
17. Daumé III, H.: Frustratingly easy domain adaptation. In: ACL (2007)
18. Denton, E.L., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a laplacian pyramid of adversarial networks. In: NIPS (2015)
19. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)

20. 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)
21. Dosovitskiy, A., Springenberg, J.T., Brox, T.: Learning to generate chairs with convolutional neural networks. In: CVPR (2015)
22. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
23. Duan, L., Tsang, I.W., Xu, D., Chua, T.S.: Domain adaptation from multiple sources via auxiliary classifiers. In: ICML (2009)
24. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. JMLR **9**, 1871–1874 (2008)
25. Fei-Fei, L., Fergus, R., Perona, P.: A Bayesian approach to unsupervised one-shot learning of object categories. In: ICCV (2003)
26. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. TPAMI **28**(4), 594–611 (2006)
27. Ferencz, A., Learned-Miller, E.G., Malik, J.: Building a classification cascade for visual identification from one example. In: ICCV (2005)
28. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV (2013)
29. Fink, M.: Object classification from a single example utilizing class relevance metrics. In: NIPS (2005)
30. Fink, M.: Acquiring a new class from a few examples: learning recurrent domain structures in humans and machines. Ph.D. thesis, The Hebrew University of Jerusalem (2011)
31. Fleuret, F., Blanchard, G.: Pattern recognition from one example by chopping. In: NIPS (2005)
32. Fu, Y., Sigal, L.: Semi-supervised vocabulary-informed learning. In: CVPR (2016)
33. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
34. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR (2012)
35. Goroshin, R., Bruna, J., Tompson, J., Eigen, D., LeCun, Y.: Unsupervised learning of spatiotemporally coherent metrics. In: ICCV (2015)
36. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2016)
37. Hariharan, B., Girshick, R.: Low-shot visual object recognition. arXiv preprint arXiv:1606.02819 (2016)
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
39. Held, D., Thrun, S., Savarese, S.: Robust single-view instance recognition. In: ICRA (2016)
40. Hertz, T., Hillel, A.B., Weinshall, D.: Learning a kernel function for classification with small training samples. In: ICML (2006)
41. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Workshops (2014)
42. Hoffman, J., Rodner, E., Donahue, J., Darrell, T., Saenko, K.: Efficient learning of domain-invariant image representations. In: ICLR (2013)
43. Hoffman, J., Tzeng, E., Donahue, J., Jia, Y., Saenko, K., Darrell, T.: One-shot adaptation of supervised deep convolutional models. In: ICLR Workshops (2014)
44. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML (2015)

45. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM MM (2014)
46. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV (2016)
47. Kienzle, W., Chellapilla, K.: Personalized handwriting recognition via biased regularization. In: ICML (2006)
48. Kim, J., Collomosse, J.: Incremental transfer learning for object recognition in streaming video. In: ICIP (2014)
49. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Workshops (2015)
50. Krause, E.A., Zillich, M., Williams, T.E., Scheutz, M.: Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In: AAAI (2014)
51. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalanditis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332 (2016)
52. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
53. Kuzborskij, I., Orabona, F., Caputo, B.: From N to N+1: multiclass transfer incremental learning. In: CVPR (2013)
54. Lake, B.M., Salakhutdinov, R., Gross, J., Tenenbaum, J.B.: One shot learning of simple visual concepts. In: CogSci (2011)
55. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: One-shot learning by inverting a compositional causal process. In: NIPS (2013)
56. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
57. Lawrence, N.D., Platt, J.C., Jordan, M.I.: Extensions of the informative vector machine. In: Winkler, J.R., Niranjan, M., Lawrence, N.D. (eds.) Deterministic and Statistical Methods in Machine Learning. LNCS (LNAI), vol. 3635, pp. 56–87. Springer, Heidelberg (2005)
58. Lee, S.I., Chatalbashev, V., Vickrey, D., Koller, D.: Learning a meta-level prior for feature relevance from multiple related tasks. In: ICML (2007)
59. Levi, K., Fink, M., Weiss, Y.: Learning from a small number of training examples by exploiting object categories. In: CVPR Workshops (2004)
60. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: CVPR (2004)
61. Lim, J.J., Salakhutdinov, R., Torralba, A.: Transfer learning by borrowing examples for multiclass object detection. In: NIPS (2011)
62. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014)
63. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: CVPR (2000)
64. Misra, I., Wang, Y.-X., Hebert, M.: Learning object models from few examples. In: SPIE Unmanned Systems Technology XVIII (2016)
65. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV (2016)

66. Movshovitz-Attias, Y.: Dataset curation through renders and ontology matching. Ph.D. thesis, Carnegie Mellon University (2015)
67. Movshovitz-Attias, Y., Yu, Q., Stumpe, M.C., Shet, V., Arnoud, S., Yatziv, L.: Ontological supervision for fine grained classification of street view storefronts. In: CVPR (2015)
68. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008)
69. Opelt, A., Pinz, A., Zisserman, A.: Incremental learning of object detectors using a visual shape alphabet. In: CVPR (2006)
70. Pan, S.J., Yang, Q.: A survey on transfer learning. TKDE **22**(10), 1345–1359 (2010)
71. Park, D., Ramanan, D.: Articulated pose estimation with tiny synthetic videos. In: CVPR (2015)
72. Patricia, N., Caputo, B.: Learning to learn, from transfer learning to domain adaptation: a unifying perspective. In: CVPR (2014)
73. Patterson, G., Van Horn, G., Belongie, S., Perona, P., Hays, J.: Tropel: crowdsourcing detectors with minimal training. In: HCOMP (2015)
74. Pinker, S.: How the mind works. Ann. N. Y. Acad. Sci. **882**(1), 119–127 (1999)
75. Quattoni, A., Collins, M., Darrell, T.: Transfer learning for image classification with sparse prototype representations. In: CVPR (2008)
76. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: CVPR Workshops (2014)
77. Rodner, E.: Visual transfer learning: informal introduction and literature overview. arXiv preprint arXiv:1211.1127 (2012)
78. Rodner, E., Denzler, J.: One-shot learning of object categories using dependent gaussian processes. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) Pattern Recognition. LNCS, vol. 6376, pp. 232–241. Springer, Heidelberg (2010)
79. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
80. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 213–226. Springer, Heidelberg (2010)
81. Salakhutdinov, R., Tenenbaum, J., Torralba, A.: One-shot learning with a hierarchical nonparametric Bayesian model. In: ICML Workshops (2012)
82. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: One-shot learning with memory-augmented neural networks. In: ICML (2016)
83. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
84. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
85. Thrun, S., Mitchell, T.M.: Learning one more thing. In: IJCAI (1995)
86. Thrun, S., Pratt, L.: Learning to Learn. Springer Science & Business Media, New York (2012)
87. Tommasi, T.: Learning to learn by exploiting prior knowledge. Ph.D. thesis, École Polytechnique Fédérale de Lausanne (2013)
88. Tommasi, T., Orabona, F., Caputo, B.: Learning categories from few examples with multi model knowledge transfer. TPAMI **36**(5), 928–941 (2014)
89. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. TPAMI **29**(5), 854–869 (2007)

90. Torralba, A., Quattoni, A.: Recognizing indoor scenes. In: CVPR (2009)
91. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
92. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. arXiv preprint arXiv:1606.04080 (2016)
93. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology (2011)
94. Wan, J., Ruan, Q., Li, W., Deng, S.: One-shot learning gesture recognition from RGB-D data using bag of features. JMLR **14**(1), 2549–2582 (2013)
95. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
96. Wang, Y.-X., Hebert, M.: Model recommendation: generating object detectors from few samples. In: CVPR (2015)
97. Wang, Y.-X., Hebert, M.: Learning by transferring from unsupervised universal sources. In: AAAI (2016)
98. Weston, J., Collobert, R., Sinz, F., Bottou, L., Vapnik, V.: Inference with the universum. In: ICML (2006)
99. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: ICML (2008)
100. Wolf, L., Hassner, T., Taigman, Y.: The one-shot similarity kernel. In: ICCV (2009)
101. Wolf, L., Martin, I.: Robust boosting for learning from few examples. In: CVPR (2005)
102. Yang, J., Yan, R., Hauptmann, A.: Adapting SVM classifiers to data with shifted distributions. In: ICDM Workshops (2007)
103. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: ICCV (2011)
104. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS (2014)
105. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. In: CVPR (2014)
106. Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? IJCV **119**(1), 76–92 (2016)
107. Zhu, X.: Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison (2005)