
One-Shot Adaptation of Supervised Deep Convolutional Models

Judy Hoffman, Eric Tzeng, Jeff Donahue
 UC Berkeley, EECS & ICSI
 {jhoffman, etzeng, jdonahue}@eecs.berkeley.edu

Yangqing Jia*
 Google Research
 jiayq@google.com

Kate Saenko
 UMass Lowell, CS & ICSI
 saenko@cs.uml.edu

Trevor Darrell
 UC Berkeley, EECS & ICSI
 trevor@eecs.berkeley.edu

Abstract

Dataset bias remains a significant barrier towards solving real world computer vision tasks. Though deep convolutional networks have proven to be a competitive approach for image classification, a question remains: have these models have solved the dataset bias problem? In general, training or fine-tuning a state-of-the-art deep model on a new domain requires a significant amount of data, which for many applications is simply not available. Transfer of models directly to new domains without adaptation has historically led to poor recognition performance. In this paper, we pose the following question: is a single image dataset, much larger than previously explored for adaptation, comprehensive enough to learn general deep models that may be effectively applied to new image domains? In other words, are deep CNNs trained on large amounts of labeled data as susceptible to dataset bias as previous methods have been shown to be? We show that a generic supervised deep CNN model trained on a large dataset reduces, but does not remove, dataset bias. Furthermore, we propose several methods for adaptation with deep models that are able to operate with little (one example per category) or no labeled domain specific data. Our experiments show that adaptation of deep models on benchmark visual domain adaptation datasets can provide a significant performance boost.

1 Introduction

Supervised deep convolutional neural networks (CNNs) trained on large-scale classification tasks have been shown to learn impressive mid-level structures and obtain high levels of performance on contemporary classification challenges [3, 23]. These models generally assume extensive training using labeled data, and testing is limited to data from the same domain. In practice, however, the images we would like to classify are often produced under different imaging conditions or drawn from a different distribution, leading to a domain shift. Scaling such models to new domains remains an open challenge.

Deep CNNs require large amounts of training data to learn good mid-level convolutional models and final fully-connected classifier stages. While the continuing expansion of web-based datasets like ImageNet [3] promises to produce labeled data for almost any desired category, such large-scale supervised datasets may not include images of the category across all domains of practical interest. Earlier deep learning efforts addressed this challenge by learning layers in an unsupervised fashion using unlabeled data to discover salient mid-level structures [6, 8]. While such approaches are

*This work was completed while Yangqing Jia was a graduate student at UC Berkeley

appealing, they have heretofore been unable to match the level of performance of supervised models, and unsupervised training of networks with the same level of depth as [17] remains a challenge.

Unfortunately, image datasets are inherently biased [21]. Theoretical [2, 4] and practical results from [20, 21] have shown that supervised methods’ test error increases in proportion to the difference between the test and training input distribution. Many visual domain adaptation methods have been put forth to compensate for dataset bias [7, 22, 1, 20, 18, 16, 13, 12, 14, 15], but are limited to shallow models. Evaluation for image category classification across visually distinct domains has focused on the Office dataset, which contains 31 image categories and 3 domains [20]. Recently, [9] showed that using the deep mid-level features learned on ImageNet, instead of the more conventional bag-of-words features, effectively removed the bias in some of the domain adaptation settings in the Office dataset [20]. However, [9] limited their experiments to small-scale source domains found only in Office, and evaluated on only a subset of relevant layers.

Yet until now, almost none of the previous domain adaptation studies used ImageNet as the *source* domain, nor utilized the full set of parameters of a deep CNN trained on source data. Recent work by Rodner et al. [19] attempted to adapt from ImageNet to the SUN dataset, but did not take advantage of deep convolutional features.

In this paper, we ask the question: will deep models still suffer from dataset bias when trained with all layers of the CNN and a truly large scale source dataset? Here, we provide the first evaluation of domain adaptation with deep learned representations in its most natural setting, in which all of ImageNet is used as source data for a target category. We use the 1.2 million labeled images available in the 2012 ImageNet 1000-way classification dataset [3] to train the model in [17] and evaluate its generalization to the Office dataset. This constitutes a three orders of magnitude increase in source data compared to the several thousand images available for the largest domain in Office.

We find that it is easier to adapt from ImageNet than from previous smaller source domains, but that dataset bias remains a major issue. Fine-tuning the parameters on the small amount of labeled target data (we consider one-shot adaptation) turns out to be unsurprisingly problematic. Instead, we propose a simple yet intuitive adaptation method: train a final domain-adapted classification “layer” using various layers of the pre-trained network as features, without any fine-tuning its parameters. We provide a comprehensive evaluation of existing methods for classifier adaptation as applied to each of the fully connected layers of the network, including the last, task-specific classification layer. When adapting from ImageNet to Office, it turns out to be possible to achieve target domain performance on par with source domain performance using only a single labeled example per target category.

We examine both the setting where there are a few labeled examples from the target domain (*supervised adaptation*) and the setting where there are no labeled target examples (*unsupervised adaptation*). We also describe practical solutions for choosing between the various adaptation methods based on experimental constraints such as limited computation time.

2 Background: Deep Domain Adaptation Approaches

For our task we consider adapting between a large source domain and a target domain with few or no labeled examples. A typical approach to domain adaptation or transfer learning with deep architectures is to take the representation learned via back-propagation on a large dataset, and then transfer the representation to a smaller dataset by fine-tuning, i.e. backpropagation at a lower learning rate [11, 23]. However, fine-tuning requires an ample amount of labeled target data and so should not be expected to work well when we consider the very sparse label condition, such as the *one-shot learning* scenario we evaluate below, where we have just one labeled example per category in the target domain.

In fact, in our experiments under this setting, fine-tuning actually reduces performance. Specifically, on the ImageNet→Webcam task reported in Section 4, using the final output layer as a predictor in the target domain received 66% accuracy, while using the final output layer after fine tuning produced a degraded accuracy of 61%.

A separate method that was recently proposed for deep adaptation is called Deep Learning for domain adaptation by Interpolating between Domains (DLID) [5]. This method learns multiple unsupervised deep models directly on the source, target, and combined datasets and uses a representation

which is the concatenation of the outputs of each model as its adaptation approach. While this was shown to be an interesting approach, it is limited by its use of unsupervised deep structures.

In general, unsupervised deep convolutional models have been unable to achieve the performance of supervised deep CNNs. However, training a supervised deep model requires sufficient labeled data. Our insight is that the extensive labeled data available in the source domain can be exploited using a supervised model without requiring a significant amount of labeled target data.

Therefore, we propose using a supervised deep source model with supervised or unsupervised adaptation algorithms that are applied to models learned on the target data directly. This hybrid approach will utilize the strong representation available from the supervised deep model trained on a large source dataset while requiring only enough target labeled data to train a shallow model with far fewer parameters. Specifically, we consider training a convolutional neural network (CNN) on the source domain and using that network to extract features on the target data that can then be used to train an auxiliary shallow learner. For extracting features from the deep source model, we follow the setup of Donahue et al. [9], which extracts a visual feature *DeCAF* from the ImageNet-trained architecture of [17].

3 Adapting Deep CNNs with Few Labeled Target Examples

We propose a general framework for selectively adapting the parameters of a convolutional neural network (CNN) whose representation and classifier weights are trained on a large-scale source domain, such as ImageNet. Our framework adds a final domain-adaptive classification “layer” that takes the activations of one of the existing network’s layers as input features. Note that the network cannot be effectively fine-tuned without access to more labeled target data. This adapted layer is a linear classifier that combines source and target training data using an adaptation method. To demonstrate the generality of our framework, we select a representative set of popular linear classifier adaptation approaches that we empirically evaluate in Section 4. We separate our discussion into the set of supervised and unsupervised adaptation settings.

Below we denote the features extracted over the source domain as $\mathbf{X}$ and the features extracted over the target domain as $\tilde{\mathbf{X}}$. Similarly, we denote the source domain image classifier as θ and the target domain image classifier as $\tilde{\theta}$.

3.1 Unsupervised Adaptation

Many unsupervised adaptation techniques seek to minimize the distance between subspaces that represent the source and target domains. We denote these subspaces as U and $\tilde{U}$, respectively.

GFK [12] The Geodesic Flow Kernel (GFK) method [12] is an unsupervised domain adaptation approach which seeks embeddings for the source and target points that minimize domain shift. Inputs to the method are U and $\tilde{U}$, lower-dimensional embeddings of the source and target domains (e.g. from principal component analysis). The method constructs the geodesic flow $\phi(t)$ along the manifold of subspaces such that $U = \phi(0)$ and $\tilde{U} = \phi(1)$. Finally, a transformation G is constructed by computing $G = \int_0^1 \phi(t)\phi(t)^\top dt$ using a closed-form solution, and classification is performed by training an SVM on the source data $\mathbf{X}$ and transformed target data $G\tilde{\mathbf{X}}$.

SA [10] The Subspace Alignment (SA) method [10] also begins with low-dimensional embeddings of the source and target domains U and $\tilde{U}$, respectively. It seeks to minimize in M , a transformation matrix, the objective $\|UM - \tilde{U}\|_F^2$. The analytical solution to this objective is $M^* = U^\top \tilde{U}$. Given M^* , an SVM is trained on source data $\mathbf{X}$ and transformed target data $UM^*\tilde{\mathbf{X}}$.

3.2 Supervised Adaptation

Late Fusion Perhaps the simplest supervised adaptation method is to independently train a source and target classifier and combine the scores of the two to create a final scoring function. We call this approach Late Fusion. It has been explored by many for a simple adaptation approach. Let us

denote the score from the source classifier as v_s and the score from the target classifier as v_t . For our experiments we explore two methods of combining these scores, which are described below:

- **Max:** Produce the scores of both the source and target classifier and simply choose the max of the two as the final score for each example. Therefore, $v_{\text{adapt}} = \max(v_s, v_t)$.
- **Linear Interpolation:** Set the score for a particular example to equal the convex combination of the source and target classifier scores, $v_{\text{adapt}} = (1 - \alpha)v_s + \alpha v_t$. This method requires setting a hyperparameter, α , which determines the weights of the source and target classifiers.

Late Fusion has two major advantages: it is easy to implement, and the source classifier it uses may be precomputed to make adaptation very fast. In the case of the linear interpolation combination rule, however, this method can potentially suffer from having a sensitive hyperparameter. We show a hyperparameter analysis in Section 4.

Daumé III [7] This simple feature replication method was proposed for domain adaptation by [7]. The method augments feature vectors with a source component, a target component, and a shared component. Each source data point $\mathbf{x}$ is augmented to $\mathbf{x}' = (\mathbf{x}; \mathbf{x}; \mathbf{0})$, and each target data point $\tilde{\mathbf{x}}$ is augmented to $\tilde{\mathbf{x}}' = (\tilde{\mathbf{x}}; \mathbf{0}; \tilde{\mathbf{x}})$. Finally, an SVM is trained on the augmented source and target data—a relatively expensive procedure given the potentially large size of the source domain and the tripled augmented feature dimensionality.

PMT [1] This classifier adaptation method, Projective Model Transfer (PMT), proposed by [1], is a variant of adaptive SVM. It takes as input a classifier θ pre-trained on the source domain. PMT-SVM learns a target domain classifier $\tilde{\theta}$ by adding an extra term to the usual SVM objective which regularizes the angle $\alpha(\tilde{\theta}, \theta) = \cos^{-1} \left(\frac{\theta^\top \tilde{\theta}}{\|\theta\| \|\tilde{\theta}\|} \right)$ between the target and source hyperplanes. This results in the following loss function:

$$\mathcal{L}_{\text{PMT}}(\tilde{\theta}) = \frac{1}{2} \|\tilde{\theta}\|_2^2 + \frac{\Gamma}{2} \|\tilde{\theta}\|_2^2 \sin^2 \alpha(\tilde{\theta}, \theta) + \ell_{\text{hinge}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}; \tilde{\theta}) , \quad (1)$$

where $\ell_{\text{hinge}}(\mathbf{X}, \mathbf{Y}; \theta)$ denotes the SVM hinge loss of a data matrix $\mathbf{X}$, label vector $\mathbf{Y}$, and classifier hyperplane θ , and Γ is a hyperparameter which, as it increases, enforces more transfer from the source classifier.

MMDT [15] The Max-margin Domain Transforms (MMDT) method from [15] jointly optimizes an SVM-like objective over a feature transformation matrix A mapping target points to the source feature space and classifier parameters θ in the source feature space. In particular, MMDT minimizes the following loss function (assuming a binary classification task to simplify notation, and with ℓ_{hinge} defined as in PMT):

$$\mathcal{L}_{\text{MMDT}}(\theta, A) = \frac{1}{2} \|\theta\|_2^2 + \frac{1}{2} \|A - I\|_F^2 + C_s \ell_{\text{hinge}}(\mathbf{X}, \mathbf{Y}; \theta) + C_t \ell_{\text{hinge}}(A\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}; \theta) , \quad (2)$$

where C_s and C_t are hyperparameters controlling the importance of correctly classifying the source and target points (respectively).

4 Evaluation

4.1 Datasets

The Office [20] dataset is a collection of images from three distinct domains: Amazon, DSLR, and Webcam. The 31 categories in the dataset consist of objects commonly encountered in office settings, such as keyboards, file cabinets, and laptops. Of these 31 categories, 16 overlap with the categories present in the 1000-category ImageNet classification task¹. Thus, for our experiments, we limit ourselves to these 16 classes. In our experiments using Amazon as a source domain, we follow the standard training protocol for this dataset of using 20 source examples per category [20, 12], for a total of 320 images.

¹ The 16 overlapping categories are *backpack, bike helmet, bottle, desk lamp, desktop computer, file cabinet, keyboard, laptop computer, mobile phone, mouse, printer, projector, ring binder, ruler, speaker, and trash can*.

ImageNet [3] is the largest available dataset of image category labels. We use 1000 categories’ worth of data (1.2M images) to train the network, and use the 16 categories that overlap with Office (approximately 1200 examples per category or $\approx 20K$ images total) as labeled source classifier data.

4.2 Experimental Setup & Baselines

For our experiments, we use the fully trained deep CNN model described in Section 2, extracting feature representations from three different layers of the CNN. We then train a source classifier using these features on one of two source domains, and adapt to the target domain.

The source domains we consider are either the Amazon domain, or the corresponding 16-category ImageNet subset where each category has many more examples. We focus on the Webcam domain as our target (test) domain, as Amazon-to-Webcam was shown to be the only challenging shift in [9] (the DSLR domain is much more similar to Webcam and did not require adaptation when using deep mid-level features). This combination exemplifies the shift from online web images to real-world images taken in typical office/home environments. Note that, regardless of the source domain chosen to learn the classifier, ImageNet data from all 1000 categories was used to train the network.

In addition, for the supervised adaptation setting we assume access to only a single example per category from the target domain (Webcam).

Each method is then evaluated across 20 random train/test splits, and we report averages and standard errors for each setting. For each random train/test split we choose one example for training and 10 other examples for testing (so there is a balanced test set across categories). Therefore, each test split has 160 examples. The unsupervised adaptation methods operate in a transductive setting, so the target subspaces are learned from the unlabeled test data.

Non-adaptive Baselines In addition to the adaptation methods outlined in Section 3, we also evaluate using the following non-adaptive baselines.

- **SVM (source only):** A support vector machine trained only on source data.
- **SVM (target only):** A support vector machine trained only on target data.
- **SVM (source and target):** A support vector machine trained on both source and target data. To account for the large discrepancy between the number of training data points in the source and target domains, we weighted the data points such that the constraints from the source and target domains effectively contribute equally to the optimization problem. Specifically, each source data point receives a weight of $\frac{n_t}{n_s + n_t}$, and each target data point receives a weight of $\frac{n_s}{n_s + n_t}$, where n_s, n_t denote the number of data points in the source and target, respectively.

Many of the adaptation methods we evaluate have hyperparameters that must be cross-validated for use in practice, so we set the parameters of the adaptation techniques as follows.

First, the C value used for C-SVM in the classifier for all methods is set to $C = 1$. Without any validation data we are not able to tune this parameter properly, so we choose to leave it as the default value. Since all methods we report require setting of this parameter, we feel that the relative comparisons between methods is sound even if the absolute numbers could be improved with a new setting for C. For Daumé III and MMDT, which look at the source and target data simultaneously, we use the same weighting scheme as we did for the source and target SVM. Late Fusion with the linear interpolation combination rule is reported across hyperparameter settings in Figure 1(a) to help understand how performance varies as we trade off emphasis between the learned classifiers from the source and target domains. Again, we do not have the validation data to tune this parameter so we report in the tables the performance averaged across parameter settings. The plot vs α indicates that there is usually a best parameter setting that could be learned with more available data. For PMT, we choose $\Gamma = 1000$, which corresponds to allowing a large amount of transfer from the source classifier to the target classifier. We do this because the source-only classifier is stronger than the target-only classifier (with ImageNet source). For the unsupervised methods GFK and SA, again we evaluated a variety of subspace dimensionalities and Figure 1(b) shows that the overall method performance does not vary significantly with the dimensionality choice.

Adaptation Method	Training Data	DeCAF ₆	DeCAF ₇
SVM (source only)	Amazon	50.28 $\pm$ 1.8	54.08 $\pm$ 1.7
SVM (target only)	Webcam	62.28 $\pm$ 1.8	64.97 $\pm$ 1.8
GFK [12]	Amazon	53.13 $\pm$ 1.1	53.39 $\pm$ 1.1
SA [10]	Amazon	51.74 $\pm$ 1.2	53.86 $\pm$ 1.0
SVM (source and target)	Amazon+Webcam	62.91 $\pm$ 1.8	65.82 $\pm$ 1.4
Late Fusion (Max)	Amazon+Webcam	65.35 $\pm$ 1.7	58.42 $\pm$ 1.1
Late Fusion (Lin. Int. Avg)	Amazon+Webcam	63.23 $\pm$ 1.4	64.29 $\pm$ 1.3
Daumé III [7]	Amazon+Webcam	68.89 $\pm$ 1.9	72.09 $\pm$ 1.4
PMT [1]	Amazon+Webcam	64.84 $\pm$ 1.5	65.63 $\pm$ 1.8
MMDT [15]	Amazon+Webcam	65.47 $\pm$ 1.8	68.10 $\pm$ 1.5
Late Fusion (Lin. Int. Oracle)	Amazon+Webcam	71.1 $\pm$ 1.7	72.82 $\pm$ 1.4

Table 1: Amazon $\rightarrow$ Webcam adaptation experiment. We show here multiclass accuracy on the target domain test set for both supervised and unsupervised adaptation experiments across the two fully connected layer features (similar to [9], but with one labeled target example). The best performing unsupervised adaptation algorithms are shown in blue and the best performing supervised adaptation algorithms are shown in red.

4.3 Effect of Source Domain Size

Previous studies considered source domains from the Office dataset. In this section, we ask what happens when an orders-of-magnitude larger source dataset is used.

For completeness we begin by evaluating Amazon as a source domain. Preliminary results on this setting are reported in [9], here we extend the comparison here by presenting the results with more adaptation algorithms and more complete evaluation of hyperparameter settings. Table 1 presents multiclass accuracies for each algorithm using either layer 6 or 7 from the deep network, which corresponds to the output from each of the fully connected layers.

An SVM trained using only Amazon data achieves 78.6% in-domain accuracy (tested on the same domain) when using the DeCAF₆ feature and 80.2% in-domain accuracy when using the DeCAF₇ feature. These numbers are significantly higher than the performance of the same classifier on Webcam test data, indicating that even with the DeCAF features, there is still a domain shift between the Amazon and Webcam datasets.

Next, we consider an unsupervised adaptation setting where no labeled examples are available from the target dataset. In this scenario, we apply two state-of-the-art unsupervised adaptation methods, GFK [12] and SA [10]. Both of these methods make use of a subspace dimensionality hyperparameter. We show the results using a 100-dimensional subspace and leave the discussion of setting this parameter until Section 4.6. For this shift the adaptation algorithms increase performance when using the layer 6 feature, but offer no additional improvement when using the layer 7 feature.

We finally assume that a single example per category is available in the target domain. As the bottom rows of Table 1 show, supervised adaptation algorithms are able to provide significant improvement regardless of the feature space chosen, even in the one-shot scenario. For this experiment we noticed that using the second fully connected layer (DeCAF₇) was a stronger overall feature in general.

4.4 Adapting with a Large Scale Source Domain

We next address one of the main questions of this paper: Is there still a domain shift when using a large source dataset such as ImageNet? To begin to answer this question we follow the same experimental paradigm as the previous experiment, but use ImageNet as our source dataset. The results are shown in Table 2.

Again, we first verify that the source only SVM achieves higher performance when tested on in-domain data than on Webcam data. Indeed, for the 16 overlapping labels, the source SVM produces 62.50% accuracy on ImageNet data using DeCAF₆ features and 74.50% accuracy when using

Adaptation Method	Training Data	DeCAF ₆	DeCAF ₇
SVM (source only)	ImageNet	53.51 $\pm$ 1.1	59.15 $\pm$ 1.1
SVM (target only)	Webcam	62.28 $\pm$ 1.8	64.97 $\pm$ 1.8
GFK [12]	ImageNet	65.16 $\pm$ 1.1	67.97 $\pm$ 1.4
SA [10]	ImageNet	59.30 $\pm$ 1.4	66.08 $\pm$ 1.4
SVM (source and target)	ImageNet+Webcam	56.68 $\pm$ 1.2	66.93 $\pm$ 1.3
Late Fusion (Max)	ImageNet+Webcam	59.59 $\pm$ 1.3	68.86 $\pm$ 1.2
Late Fusion (Lin. Int. Avg)	ImageNet+Webcam	60.64 $\pm$ 1.3	66.45 $\pm$ 1.1
Daumé III [7]	ImageNet+Webcam	59.21 $\pm$ 1.7	71.39 $\pm$ 1.5
PMT [1]	ImageNet+Webcam	66.30 $\pm$ 2.1	69.81 $\pm$ 1.8
MMDT [15]	ImageNet+Webcam	59.21 $\pm$ 1.3	67.75 $\pm$ 1.4
Late Fusion (Lin. Int. Oracle)	ImageNet+Webcam	71.65 $\pm$ 2.0	76.76 $\pm$ 1.3

Table 2: ImageNet $\rightarrow$ Webcam adaptation experiment. Comparison of unsupervised and supervised adaptation algorithms on the ImageNet to Webcam domain shift. Results are computed using the outputs of each of the fully connected layers as features. The best supervised adaptation performance is indicated in red and the best unsupervised adaptation performance is highlighted in blue.

Adaptation Method	Training Data	Source=ImageNet	Source=Amazon
SVM (source only)	Source	66.23 $\pm$ 0.8	53.23 $\pm$ 1.6
SVM (target only)	Webcam	63.13 $\pm$ 1.9	63.13 $\pm$ 1.9
GFK [12]	Source	68.73 $\pm$ 1.1	54.56 $\pm$ 1.2
SA [10]	Source	66.08 $\pm$ 1.1	55.98 $\pm$ 1.0
SVM (source and target)	Source+Webcam	75.13 $\pm$ 1.1	63.20 $\pm$ 1.7
Late Fusion (Max)	Source+Webcam	71.77 $\pm$ 1.4	62.25 $\pm$ 0.8
Late Fusion (LinInt Avg)	Source+Webcam	70.56 $\pm$ 1.2	64.56 $\pm$ 1.3
Daumé III [7]	Source+Webcam	77.15 $\pm$ 1.1	70.51 $\pm$ 1.7
PMT [1]	Source+Webcam	70.28 $\pm$ 1.8	66.77 $\pm$ 2.1
MMDT [15]	Source+Webcam	73.96 $\pm$ 1.2	66.23 $\pm$ 1.4
Late Fusion (Lin. Int. Oracle)	Source+Webcam	76.61 $\pm$ 1.5	71.49 $\pm$ 1.3

Table 3: ImageNet $\rightarrow$ Webcam and Amazon $\rightarrow$ Webcam adaptation experiments using DeCAF₈, the label activations of the CNN trained on the full ImageNet data. Again, we compare multiclass accuracy of various unsupervised and supervised adaptation methods. The best performing unsupervised adaptation algorithm is shown in blue and the best performing supervised adaptation algorithms are shown in red.

DeCAF₇ features. Compare this to the 54% and 59% for Webcam evaluation and a dataset bias is still clearly evident.

Note that when using ImageNet as a source domain, overall performance of all algorithms improves. In addition, unsupervised adaptation approaches are more effective than for the smaller source domain experiment.

4.5 Adapting a Pre-trained Classifier to a New Label Set

DeCAF₈ differs from the other DeCAF features in that it constitutes the 1000 activations corresponding to the 1000 labels in the ImageNet classification task. In the CNN proposed by [17], these activations are fed into a softmax unit to compute the label probabilities. We instead experiment with using the DeCAF₈ activations directly as a feature representation, which is akin to training another classifier using the output of the 1000-way CNN classifier.

Table 3 shows results for various adaptation techniques using both ImageNet and Amazon as source domains. We use the same setup as before, but instead use DeCAF₈ as the feature representation.

The ImageNet results are uniformly better with DeCAF₈ than with DeCAF₆ or DeCAF₇, likely due to the fact that DeCAF₈ was explicitly trained on ImageNet data to effectively discriminate between ImageNet categories. Because it can more effectively classify images from the source domain, it is able to better adapt from the source domain to the target domain.

However, we see a negligible difference in performance for Amazon, with performance actually decreasing with respect to DeCAF₇ for certain adaptation methods. We believe this is because the final activation vector is too specific to the 1000-way ImageNet task, and that DeCAF₇ provides a more general representation that is better suited to the Amazon domain. This, in turn, results in improved adaptation. In general, however, the difference between the various DeCAF representations with Amazon as a source are small enough to be insignificant.

4.6 Analysis and Practical Considerations

Our adaptation experiments show that, despite its large size, even ImageNet is not large enough to cover all domains, and that traditional domain adaptation methods go a long way in increasing performance and mitigating the effects of this shift. Depending on the characteristics of the problem at hand, our results suggest different methods may be most suitable.

If no labels exist in the target domain, then there are unsupervised adaptation algorithms that are easy to use and fast to compute at adaptation time, yet still achieve increased performance over source-only methods. For this scenario, we experimented with two subspace alignment based methods that both require setting a parameter that indicates the dimensionality of the input subspaces. Figure 1(b) shows the effect that changing the subspace dimensionality has on the overall method performance. In general, we noticed that these methods were not particularly sensitive to this parameter so long as the dimensionality remains larger than the number of categories in our label set. Below this threshold, the subspace is less likely to capture all important discriminative information needed for classification.

In the case where we have a large source dataset and a limited number of labeled target examples, it may be preferable to compute source classifier parameters in advance, then examine only the source parameters and the target data at adaptation time. Examples of these kinds of methods are Late Fusion and PMT. These methods are unaffected by the number of data points in the source domain at adaptation time, and can thus be applied quickly. In our experiments, we found that a properly tuned Late Fusion classifier with linear interpolation was the fastest and most effective approach. Figure 1(a) shows the performance of linear interpolation Late Fusion as we vary the hyperparameter α . Although the method is sensitive to α , we found that for both source domains, the basic strategy of setting α around 0.8 provides a close approximation to optimal performance. This setting can be interpreted as trusting the target classifier more than the source, but not so much as to completely discount the information available from the source classifier. In each table we report both the performance of linear interpolation both averaged across hyper parameter settings $\alpha \in [0, 1]$ as well as the performance of linear interpolation with the best possible setting of α per experiment – this is denoted as “Oracle” performance.

If there are no computational constraints and there are very few labels in the target domain, the best-performing method seems to be the “frustratingly easy” approach originally proposed by Daumé III [7] and applied again for deep models in [5].

Finally, we found that feature representation can have a significant impact on adaptation performance. Our results show that ImageNet as source performs best with the DeCAF₈ representation, whereas Amazon as source performs best with the DeCAF₇ representation. This, combined with our intuition, seems to indicate that for adaptation from source domains other than ImageNet, an intermediate representation other than DeCAF₈ is more powerful for adaptation, whereas ImageNet classification works best with the full representation that was trained on it.

5 Conclusion

In this paper, we presented the first evaluation of domain adaptation from a large-scale source dataset with deep features. We demonstrated that, although using ImageNet as a source domain generalizes

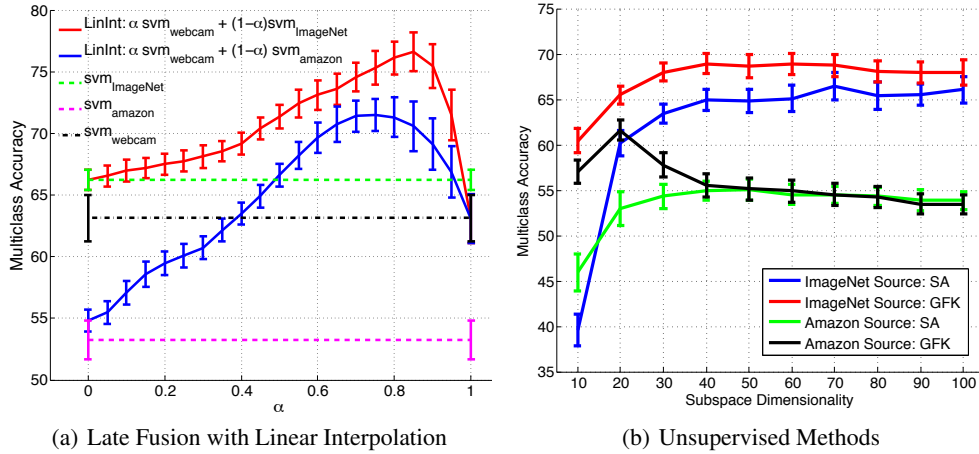


Figure 1: Evaluation of hyperparameters for domain adaptation methods. (a) Analysis of the combination hyperparameter α for Late Fusion with linear interpolation. (b) Analysis of the subspace dimensionality for the unsupervised adaptation algorithms

better than other smaller source domains, there is still a domain shift when adapting to other visual domains.

Our experimental results show that deep adaptation methods can go a long way in mitigating the effects of this domain shift. Based on our results, we also provided a set of practical recommendations for choosing a feature representation and adaptation method accounting for constraints on runtime and accuracy.

There are a number of interesting directions to take given our results. First we notice that though DeCAF₈ is the strongest feature to use for learning a classifier on ImageNet data, DeCAF₇ is actually a better feature to use with the Amazon source domain and the Webcam target domain. This could lead to a hybrid approach where one uses different feature representations for the various domains and produces a combined adapted model. Another interesting direction that should be explored is to integrate the adaptation algorithms into the deep models explicitly and even allow for feedback between the two stages. Current deep models although allow information flow between the final classifier and the representation learning architecture. We feel that the next step is to have a separate task specific adaptable layer that does not simply learn a new final layer, but instead learns a separate, but equivalent final layer, that is regularized by the final layer learned on the source dataset.

This future work is a natural extension of the result we have shown in this paper: that pre-trained deep representations with large source domains can be effectively adapted to new target domains using only shallow, linear adaptation methods, and that in cases where the target data is limited, this approach is the best way to mitigate dataset bias.

References

- [1] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *Proc. ICCV*, 2011.
- [2] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Proc. NIPS*, 2007.
- [3] A. Berg, J. Deng, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge 2012. 2012.
- [4] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Proc. NIPS*, 2007.
- [5] S. Chopra, S. Balakrishnan, and R. Gopalan. DLID: Deep learning for domain adaptation by interpolating between domains. In *ICML Workshop on Challenges in Representation Learning*, 2013.
- [6] A. Coates, A. Karpathy, and A. Ng. Emergence of object-selective features in unsupervised feature learning. In *Proc. NIPS*, 2012.

- [7] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.
- [8] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng. Large scale distributed deep networks. In *Proc. NIPS*, 2012.
- [9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *arXiv e-prints*, 2013.
- [10] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proc. ICCV*, 2013.
- [11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv e-prints*, 2013.
- [12] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, 2012.
- [13] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proc. ICCV*, 2011.
- [14] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering latent domains for multisource domain adaptation. In *Proc. ECCV*, 2012.
- [15] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *Proc. ICLR*, 2013.
- [16] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *Proc. ECCV*, 2012.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.
- [18] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.
- [19] Erik Rodner, Judy Hoffman, Jeff Donahue, Trevor Darrell, and Kate Saenko. Towards adapting imagenet to reality: Scalable domain adaptation with implicit low-rank transformations. *CoRR*, abs/1308.4200, 2013.
- [20] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.
- [21] A. Torralba and A. Efros. Unbiased look at dataset bias. In *Proc. CVPR*, 2011.
- [22] J. Yang, R. Yan, and A. Hauptmann. Adapting SVM classifiers to data with shifted distributions. In *ICDM Workshops*, 2007.
- [23] M. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. *ArXiv e-prints*, 2013.