Optimizing One-Shot Recognition with Micro-Set Learning

Kevin D. Tang
Cornell University

Marshall F. Tappen
University of Central Florida

Rahul Sukthankar
Intel Labs Pittsburgh & Carnegie Mellon

Christoph H. Lampert
Max Planck Institute for Biological Cybernetics

Abstract

For object category recognition to scale beyond a small number of classes, it is important that algorithms be able to learn from a small amount of labeled data per additional class. One-shot recognition aims to apply the knowledge gained from a set of categories with plentiful data to categories for which only a single exemplar is available for each. As with earlier efforts motivated by transfer learning, we seek an internal representation for the domain that generalizes across classes. However, in contrast to existing work, we formulate the problem in a fundamentally new manner by optimizing the internal representation for the one-shot task using the notion of micro-sets.

A micro-set is a sample of data that contains only a single instance of each category, sampled from the pool of available data, which serves as a mechanism to force the learned representation to explicitly address the variability and noise inherent in the one-shot recognition task. We optimize our learned domain features so that they minimize an expected loss over micro-sets drawn from the training set and show that these features generalize effectively to previously unseen categories. We detail a discriminative approach for optimizing one-shot recognition using micro-sets and present experiments on the Animals with Attributes and Caltech-101 datasets that demonstrate the benefits of our formulation.

1. Introduction

Numerous papers have shown that recognition performance increases as more training examples of each object type become available. This makes one-shot recognition, where the system is only given one example of each object type, one of the most difficult types of object recognition tasks.

With the training data of each class limited to one example in the one-shot problem, this paper will show how a large number of examples from similar types of objects can be used to improve one-shot recognition performance. Effectively, the system presented here "learns how to learn" from the small set of images presented to the one-shot recognition system.

Using animals as an example, similar to [14], consider the one-shot recognition task of distinguishing between images of ostriches and hyenas, with only one ostrich image and one hyena image. Inspired by work on transfer learning, we argue that learning an intermediate representation that makes it easier to perform a similar task, say distinguishing between a goose and a dog, is helpful for solving the original problem.

In this paper, we propose a novel, straightforward, learning-based approach for finding this intermediate representation. Assuming that relatively large amounts of data are available for the related tasks, such as the dog and goose problem from above, we show how training over *micro-sets* makes it possible to directly simulate and optimize the usefulness of the intermediate representation for one-shot recognition.

The remainder of this paper is organized as follows. Section 2 reviews work in object recognition with a particular focus on one-shot recognition. Section 3 formulates our problem. Section 4 details the micro-set framework. Section 5 describes our experimental methodology Sections 6 and 7 presents experimental results on the *Animals with Attributes* [14] and Caltech-101 [16] datasets, respectively. Section 8 discusses the implications of our experiments and Section 9 concludes the paper.

2. Related Work

Object category recognition has been an active research area in computer vision for several decades (see [3] for a recent comprehensive survey). However, despite significant evidence that humans can learn concepts from just a few exemplars, there has been comparatively little work in the area of one-shot recognition. Our approach is most related to transfer learning for recognition [6, 16, 7, 13, 22], learning predictive structures from multiple tasks [1], learning (pseudo-)metrics for recognition [7, 25], learning discriminative representations for efficient retrieval [10, 11, 24, 26] and recent efforts in exploiting semantic attributes for recognizing novel classes [5, 14, 20]. We briefly discuss these below.

The phrase "one-shot learning" was popularized by Fei-

Fei *et al*. [16], where parametric class models are learned in a Bayesian framework. A class-independent prior is learned from the familiar classes and then applied to the rare classes to bias parameters towards values observed in the familiar classes. However, since this prior is not category specific, it is biased towards popular features common across all objects of interest. Bart and Ullman [2] employ cross-generalization to transfer semantically-related patches between familiar and rare classes in a class-specific manner. Even though the method is evaluated on a large number of classes, the formulation only considers one novel category at a time in a leave-one-out manner; this is a binary task (to determine whether the novel category is present) whereas we directly address the more challenging task of assigning a test image to one of many novel classes, for each of which we have only a single training instance.

Miller *et al*. [17] and Fink [7] explore one-shot learning in the context of letter/digit recognition. The former introduces the notion of generating synthetic data to augment the single example using data-driven (rather than ad hoc) transforms while the latter learns a pseudo-metric that preserves class membership in the presence of intra-class variation. The results on one-vs-one discrimination are promising but generalizing the methods beyond the handwritten character domain is not straightforward.

At a high level, our approach is also related to the feature learning approaches investigated in image [10, 26], music [11] and pose [24] retrieval. Their goal is to learn a compact, discriminative representation using a set of training instances that generalizes to unseen media collections. Such representations, often learned using a pairwise variant of boosting, can be equivalently considered as binary attributes, bit strings or distance functions in Hamming space that classify membership in local neighborhoods. A key difference is that our proposed approach seeks intermediate representations that generalize over variations in object appearance due to intra-class variation while approaches such as [11] primarily aim to reliably retrieve the correct object under challenging imaging conditions. Thus, our work is more philosophically aligned to Nowak and Jurie [18], who exploit knowledge gleaned from pairs of "same" and "different" objects to learn a similarity function that enables them to compare images of previously unseen objects.

Our work is also inspired by several recent papers that learn semantic attributes and apply these to recognizing instances from novel classes [5, 14, 20]. Such approaches train a set of classifiers to recognize pre-defined attributes (e.g., whether the image shows an animal that lives in water) and combine these with side semantic information about the novel class (e.g., dolphins live in water) to recognize novel categories for which no training examples exist. Our proposed method differs from these in two important aspects: (1) we seek to learn our semantic attributes rather than employing pre-defined ones; (2) in exchange for a single exemplar of the object, we eschew any side information about the semantic relationships between object categories.

In relation to the prior work, the novel aspects of our proposed approach include:

1. **New framework for learning intermediate representations directly from data.**
   This framework is based on directly modeling the generalization performance in the one-shot recognition task, while previous approaches have focused on heuristics based on separation. The advantages of this new framework are discussed below in Section 8.
2. **Specific focus on the one-shot recognition task.**
   In contrast to the work discussed above, our work focuses on the one-shot recognition task in the *training* as well as the testing phase. Specifically, the one-shot recognition task requires the classifier to operate under extremely noisy conditions, particularly when a given exemplar is not sufficiently representative of its category. The micro-set framework forces the internal representation to cope with the noise and variability of the problem by imposing the one-shot conditions during training. Furthermore, the framework is inherently discriminative and multi-category, ensuring that the objective of the training phase matches that of the final multi-class problem.
3. **Evaluation on real-world images with many novel categories.**
   Unlike earlier approaches that are evaluated using a leave-one-out approach to novel categories, our experiments employ a minimum of 10 novel categories in each run. We present results on multiple, standardized image datasets using an experimental methodology suggested by Lampert *et al*. [14].

3. Formulating the Learning Problem

The first step in learning the intermediate representation is deciding on a classification methodology for framing the one-shot recognition problem. In previous work, there has been some variation in how this problem is posed. For example, work such as Fink [7], poses the problem in a standard multi-class classification formulation, where the goal is to correctly classify each image as depicting one of several object types. An alternative approach, espoused by Fei-Fei *et al*. [15, 16], is to separate images depicting the object of interest from those without the object using a binary classifier, given a single instance of the object.

Our work focuses on the multi-class classification approach. Each of the test images will be assigned a single label denoting the dominant object in the scene. The classification itself will be posed in a nearest-neighbor framework; with only one example per object type available, the nearest-neighbor paradigm is a natural choice for implementing this classification. For a problem with K classes, the training images will be represented by a set of feature vectors, $\mathbf{t}_1, \ldots, \mathbf{t}_K$. The object in a novel test image is then classified by comparing its feature vector $\mathbf{x}$ against the

features for each of the K object classes and assigning the novel image to the class of the most similar exemplar. Denoting $\mathcal{C}$ as the estimated class, this can be expressed formally as

$$\mathcal{C} = \arg \min_{i \in 1 \ldots K} ||\mathbf{x} - \mathbf{t}_i||_2, \quad (1)$$

where the distance is measured using the Euclidean (ℓ_2) norm in this case.[1]

3.1. Defining and Incorporating Attributes

Having defined this basic, nearest-neighbors formulation, the next step is to incorporate the intermediate representation into the classifier. For the purposes of this work, we consider intermediate representations consisting of the output of a vector function $\mathbf{r}(\mathbf{x})$, applied to the original feature vector $\mathbf{x}$:

$$\mathcal{C} = \arg \min_{i \in 1 \ldots K} ||\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{t}_i)||_2, \quad (2)$$

This intermediate representation, $\mathbf{r}(\mathbf{x})$, should be chosen to make the nearest-neighbor classification perform as well as possible. We have explored a variety of intermediate representation types, ranging from linear projections,

$$\mathbf{r}(\mathbf{x}) = \mathbf{A}\mathbf{x}, \quad (3)$$

to complicated non-linear functions such as multi-layer perceptrons. The latter often overfit to the training data and thus performed poorly on the one-shot task. In our experiments, we have observed the best results using a basic non-linear function consisting of a logistic function that is applied in an element-wise fashion to the results of a linear projection with a matrix $\mathbf{A}$:

$$\mathbf{r}(\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{A}\mathbf{x})}. \quad (4)$$

In the experiments described in Section 5, we refer to this type of internal representation as a logistic projection.

4. Learning the Intermediate Representation

To optimize the intermediate representation function, $\mathbf{r}(\mathbf{x})$, we adapt the criterion underlying Neighborhood Components Analysis (NCA), proposed by Goldberger *et al*. [9]. This criterion can be thought of as using the softmax function to convert distances between points to probabilities.

In the original NCA criterion, the probability that point i selects point j as a neighbor, denoted p_{ij}, is expressed as

$$p_{ij} = \frac{\exp(-||\mathbf{r}(\mathbf{x}_i) - \mathbf{r}(\mathbf{x}_j)||^2)}{\sum_{k \neq i} \exp(-||\mathbf{r}(\mathbf{x}_i) - \mathbf{r}(\mathbf{x}_k)||^2)}, \quad (5)$$

[1] This paper focuses on ℓ_2 since our experiments have shown no significant improvements from employing other norms.

where the summation in the denominator is over all data points, except point i.

In one-shot recognition, this probability can instead be expressed as just p_k, which denotes the probability that the test image, $\mathbf{x}$, contains the object with label k. The mathematical form of p_k is very similar to Equation (5):

$$p_k = \frac{\exp(-||\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{t}_k)||^2)}{\sum_{j=1}^{K} \exp(-||\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{t}_j)||^2)}, \quad (6)$$

where $\mathbf{t}_1, \ldots \mathbf{t}_K$ are the exemplars of each class, as described above.

4.1. Learning with micro-sets

Standard supervised learning cannot be applied to learning $\mathbf{r}(\mathbf{x})$ for the one-shot task, as only a single example per class is available. Instead, we argue that optimizing $\mathbf{r}(\mathbf{x})$ on a related task, where large amounts of labeled data is available, will enhance one-shot recognition performance; this is similar in spirit to the work on transfer learning in the machine-learning community.

The process for training $\mathbf{r}(\mathbf{x})$ is formulated to improve one-shot recognition as much as possible by directly simulating the one-shot recognition problem in the training criterion. The basic idea is to transform the training set into a huge number of different one-shot recognition problems and to optimize over the recognition performance averaged across these many problems. This enables the system to find an $\mathbf{r}(\mathbf{x})$ function that can best cope with the uncertainty and difficulty that comes from having only one example per object category.

For training, the labeled corpus of data generates a series of *micro-sets*, as depicted in Figure 1. In each micro-set, one image per class is identified as a training example for that class, while the remaining images serve as testing proxies. For a particular micro-set, given index i, let the function $\mu Tr(i,k)$ denote the image index of the exemplar for class k. Likewise, $\mu Te(i,k)$ will denote the set of test images for class k in micro-set i. Equation 7 can then be adapted to model the probability of correctly performing the one-shot recognition task in micro-set i as

$$p_{\mu_i} = \sum_{k=1}^{K} \left[\sum_{i \in \mu Te(i,k)} \frac{\exp(-||\mathbf{r}(\mathbf{t}_i) - \mathbf{r}(\mathbf{t}_{\mu Tr(i,k)}||^2)}{\sum_{j=1}^{K} \exp(-||\mathbf{r}(\mathbf{t}_i) - \mathbf{r}(\mathbf{t}_{\mu Tr(i,j)})||^2)} \right]. \quad (7)$$

In a one-shot recognition system, one of the major sources of variability is the choice of the exemplar image for each object category. One of the goals of the learning process is to find an intermediate representation that is resilient to this variation. Thus, the learning criterion sums over the classification accuracy of as many micro-sets as possible:

Figure 1. As shown above, a simple training set consisting of three example images for the two object classes, can be used to generate nine micro-sets. Each micro-set is one of the one-shot recognition problems that could be generated with the training set. The intermediate representation function, $\mathbf{r}(\mathbf{x})$, from Equation (5), is found by optimizing the probability of correct one-shot recognition across all of these sets.

$$L = \sum_{i=1}^{M} p_{\mu_i} \tag{8}$$

where M is the number of micro-sets, which should be as large as possible.

5. Experimental Methodology

To evaluate the impact of different learned intermediate representations on classification accuracy, we adopt a variant of the experimental methodology proposed for zero-shot recognition recently proposed in [14]. We present results on two large, publicly-available datasets: *Animals with Attributes* [14] and Caltech-101 [16]. Both of these standard datasets contain many images from a large set of classes, enabling us to evaluate one-shot recognition performance over a wide variety of configurations.

Our methodology is summarized as follows. First, we hold out ten categories to serve as test classes; these are used to evaluate how well an intermediate representation enhances classification performance on a novel recognition problem. Next, we train intermediate representations (using different algorithms) with images from the remaining categories. Finally, we evaluate the multi-class classification performance using a series of independent one-shot tasks that are averaged together to determine overall accuracy.

Specifically, for each one-shot task we randomly sample *one image* from each of the ten held-out categories; these serve as the single-exemplar "training set" for the task. The remaining images from the ten held-out categories are used as the "testing set". Since there are varying numbers of testing images for each held-out category, the errors obtained for each class are normalized by the number of testing images in that class. The classification error for the given one-shot task is computed using the nearest-neighbor rule (Eqn. 2). To obtain an accurate estimate of the intermediate representations' overall classification performance, we repeat this one-shot task 10,000 times and present averaged results.

The following subsections describe the feature representation and implementation details for the learning procedure employed in our experiments.

5.1. Image features

Each image in these data sets is represented by a vector of local features. To facilitate replication of our experiments and to enable future researchers to directly compare their one-shot recognition algorithms with ours, we restrict ourselves to publicly-available pre-computed features. A more judicious selection of low-level features could lead to better recognition accuracy but that is orthogonal to the primary contribution on this paper, which is to propose an effective general framework for learning intermediate representations in the context of one-shot recognition.

5.1.1 Animals with attributes dataset

The *Animals with Attributes* dataset introduced by Lampert *et al*. [14] consists of 30,475 images of 50 animals, accompanied by several pre-extracted features. The dataset[2] also contains semantic side information (e.g., whether a given animal lives in water) in the form of an 85-dimensional Osherson's attribute [19, 12] vector for each animal category that we specifically *do not use* in our experiments.

Our primary motivation in using this dataset is to directly compare our intermediate representation that was trained without side information against semantic attributes [14] in the context of one-shot recognition. In other words, we seek to understand the degree to which semantic side information helps in transfer learning to novel animal classes.

[2] Downloaded from http://attributes.kyb.tuebingen.mpg.de/.

The features pre-extracted by Lampert *et al*. [14] were used as the low-level image features. These include six different feature types: RGB color histograms, SIFT, rgSIFT, PHOG, SURF, and local self-similarity histograms. Concatenated together, the features form a 10,950-dimensional vector. We processed the low-level image features as follows. First, we performed a feature-wise normalization, followed by a dimensionality reduction to 500 dimensions using Principal Components Analysis (PCA). The resulting 500-dimensional vectors were then normalized to ensure that the range of each feature fell between -1 to 1.

5.1.2 Caltech-101 dataset

The Caltech-101 dataset [16] consists of 9,146 images from 101 distinct object categories, along with images in a background category, which we do not use. Our primary motivations for using this popular dataset, despite its noted deficiencies [21], are that Caltech-101 includes a broader range of object categories compared to the *Animals with Attributes* dataset and that publicly-available low-level features for Caltech-101 will make it easier for others to perform direct comparisons against our proposed methods.

In our experiments, we employ the features computed by Gehler and Nowozin [8].[3] In particular, we used the PHOG, LBP, and bag-of-words SIFT features, and also used additional USIFT features we extracted ourselves. Concatenated together, the features form a 2,881-dimensional vector. We also applied the normalization and dimensionality reduction described above to the Caltech-101 features to generate a feature vector of 500 dimensions for each image.

5.2. Learning $\mathbf{r}(\mathbf{x})$ using micro-sets

The procedure for learning the intermediate representation is the same for both datasets. Once the ten categories used for testing have been removed, the remaining object categories are used to learn $\mathbf{r}(\mathbf{x})$. In both sets of experiments, a minimum of 40 classes are used to learn $\mathbf{r}(\mathbf{x})$. Because our micro-set framework simulates one-shot recognition during training, we could draw from an enormous number of potential micro-sets; for instance, given a dataset with N images per category, there are N^{40} ways of selecting a training micro-set containing one exemplar from each of the 40 categories. To make good use of this diversity, we employ stochastic gradient descent during our training procedure, as follows.

Each iteration of the stochastic gradient procedure can be viewed as an instance of Efron's bootstrap [4], where the pool of samples is much larger than the size of the training set required for micro-set learning. Specifically, we sample a new micro-set (consisting of a single exemplar from each of the available classes and ten test examples that are used to compute recognition accuracy) during each iteration; we compute a gradient from the micro-set and use it to update $\mathbf{r}(\mathbf{x})$. Empirically, we observe that the optimization converges after approximately 500 iterations.

[3] Downloaded from http://www.vision.ee.ethz.ch/~pgehler/projects/iccv09/.

10-Class Accuracy				
Method	Mean	Variance	Min	Max
Random chance	10%			
Identity transformation (Raw)	14.1%	0%	7.2%	22.7%
Intermediate representation trained with micro-sets				
Linear projection	23.7%	0.1%	11.4%	35.2%
Logistic projection	27.2%	0.1%	11.1%	37.8%
Using Osherson's attributes from [14]				
One-shot	29.0%	0.14%	10.7%	42.4%
Manually defined (Zero-shot)	40.5%			

Table 1. Results for one-shot classification for the 10-class recognition task on the *Animals with Attributes* dataset. Notice that the logistic projection intermediate representation, trained using the micro-set method, performs nearly as well as the semantic attributes manually defined for the task of recognizing animals.

6. Experiments with *Animals with Attributes*

For the experiments with the *Animals with Attributes* dataset, our division of the 50 animal classes into 40 training and 10 testing categories is the same as that suggested in [14]. Tables 1 and 2 summarize one-shot recognition accuracy for two types of classification problems. In the 10-class problem, the testing images are classified into one of the 10 object categories that were held out while optimizing $\mathbf{r}(\mathbf{x})$. In the 50-class problem, each image can be labeled with any of the 50 categories in the data set; here, since all of the query images are still drawn from the 10 held-out classes, the 40 classes in the training set serve as distractors. We discuss both of these in greater detail in the following sub-sections.

6.1. One-Shot Accuracy on 10-Class Recognition

The key observations from Table 1 regarding the 10-class one-shot problem include:

- **The best intermediate representation trained using the micro-sets approach almost doubles the classification accuracy obtained using the raw features.** The benefits of micro-set training are clear, both for the linear and the logistic projection. The latter nearly doubles the recognition rate from 14% to 27%.
- **This learned representation performs comparably to the human-identified attributes, in a one-shot recognition framework.** For this comparison, we represent each image using the 85-dimensional vector of its Osherson attributes, as provided in the *Animals with Attributes* dataset. Those

were generated by Lampert *et al*. [14] using semantic attribute detectors trained using human-specified side information for each of the 50 animal categories, such as "does this animal eat fish?" The accuracy of this baseline is denoted as as "Osherson's (one-shot)".
Given that the Osherson attributes were designed to discriminate between animals, we are pleased to see that our intermediate representation, trained using micro-sets without any human-specified side information, achieves comparable accuracy in the one-shot recognition task.

- **Representing the category using manually-specified semantic attributes significantly outperforms one-shot recognition.**
For this baseline, denoted "Osherson's (zero-shot)" in Table 1, we employ the human-generated category-based representation made available by Lampert *et al*. [14] in the place of our one-shot exemplars. Specifically, rather than representing an animal category using the semantic attributes derived from a single exemplar, we represent each class with the Osherson semantic attributes for that animal (i.e., its ground truth attributes). This representation eliminates both the variability introduced by the idiosyncrasies of the few training samples in the one-shot framework and the uncertainty of estimating the semantic attributes from those images.
Unsurprisingly, the recognition accuracy achieved using these human-specified category descriptions is much higher than that obtained using our learned representation on the 10-class recognition problem. However, this baseline shows that the dataset is challenging even when ground-truth category-level information is available. We also note that this baseline performs very poorly (worse than chance!) on the 50-class version of the same task; this interesting observation is discussed in more detail in Section 6.2, below.
- **Max performance on one-shot can match zero-shot performance**
Clearly, one-shot recognition accuracy is sensitive to the choice of the single exemplar in the training set. Although the variance in our experimental results is low, we do observe that the *min* and *max* columns show a significant range. In the best case (max), a favorable choice of exemplars enables one-shot recognition accuracy to attain the performance of Lampert *et al*.'s zero-shot training. On the other hand, an unfortunate choice of exemplars, where the selected image is not representative of its category [23], leads to poor performance (min).

6.2. One-Shot Accuracy on 50-class Recognition

Moving to the 50-class recognition problem (see Table 2) decreases recognition performance across all experiments, which is not surprising given that the number of possible labels for each image has increased by a factor of five.

50-Class Accuracy				
Method	Mean	Variance	Min	Max
Random chance	2%			
Identity transformation (Raw)	5.38%	0.1%	0%	22%
Intermediate representation trained with micro-sets				
Linear projection	8.4%	0.07%	1.34%	18.3%
Logistic projection	7.5%	0.06%	2.07%	18.3%
Using Osherson's attributes from [14]				
One-shot	0.09%	0.008%	0%	5%
Manually defined (Zero-shot)	1%			

Table 2. Results for one-shot classification for the 50-class recognition task on the *Animals with Attributes* dataset. When considering all 50 classes, the learned representation significantly outperforms the semantic attributes.

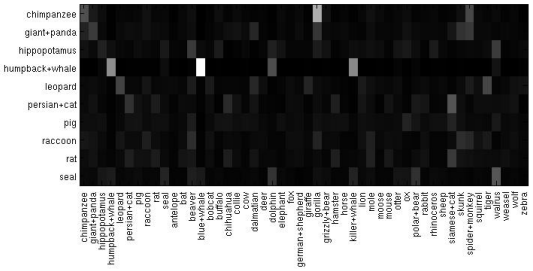
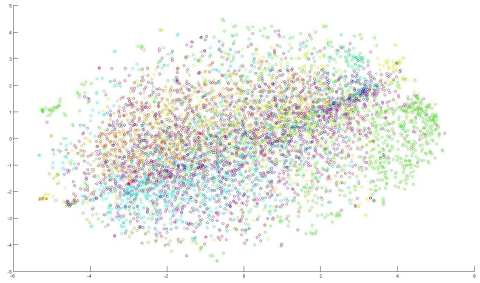


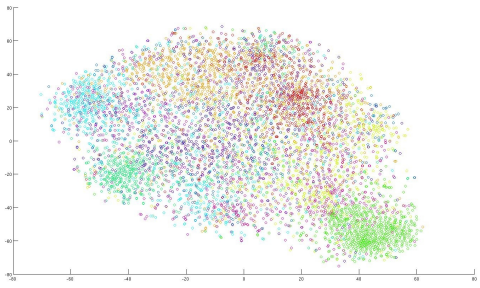
Figure 2. Confusion matrices for classification accuracies of the 50-class one-shot recognition evaluation on the *Animals with Attributes* data-set. Many of the mistakes are with similar image classes, for instance "persian+cat" is frequently classified with "siamese+cat". With only one example per category, it is difficult for the system to make fine distinctions of this sort.

The most notable result is that the recognition performance for the Osherson's (zero-shot) baseline drops below chance levels. In contrast, the degradation in performance when using the intermediate representation learned using the micro-sets approach is much less; in fact, accuracy relative to chance levels improves from $2.7\times$ to $3.74\times$ chance. We also note that linear projection outperforms logistic projection in this setting, indicating that the former is less prone to overfitting on the 40 training classes.

It should be noted that the low recognition rates in the 50-class recognition problem do not indicate that the representation completely fails. As can be seen in the confusion matrix in Figure 2, the misclassifications are primarily between animals from similar categories. For example, the humpback whale class is misclassified most as blue whale, dolphin, or killer whale.



(a) Raw features (with PCA)



(b) Learned representation (with micro-sets)

Figure 3. These two figures show the benefit of the learned intermediate representation, even after reducing the data to two dimensions. The points in each of these plots were computed using the t-SNE reduction technique [27]. Each color represents a different animal class from the ten held-out categories. As shown in (b), after image feature vectors are processed with the intermediate representation, clusters corresponding to different animal categories can still be seen in the 2D plot.

6.3. Visualizing the Improvement

Surprisingly, the benefit of this representation can be seen even after reducing the dimensionality of the image features to two dimensions. Figure 3 is a t-SNE visualization [27] of: (a) the raw feature vectors reduced using PCA to 500 dimensions; and (b) the intermediate representation learned using micro-sets. Each dot in the plot corresponds to an image from the ten held-out categories of the dataset, with color denoting the animal category. Note that very little category-relevant structure is visible in the first graph but several distinct clusters, each corresponding to an animal, are visible in the visualization of the micro-set based representation. Interestingly, these clusters survive despite having been reduced in dimension from 500 to 2.

7. Experiments with Caltech-101

The experiments detailed in the previous section were also conducted on the Caltech-101 dataset [16]. Our goal was to evaluate how representations optimized using micro-sets perform on datasets with a broader range of categories than animals alone, and with a greater number of object classes.

	Accuracy		
Method	Mean	Min	Max
Random Chance	10%		
Raw Features (PCA)	30.4%	10.69%	47.3%
Trained on 40 classes	43.0%	19.5%	59.8%
Trained on 65 classes	48.30%	27.7%	64.5%
Trained on 91 classes	52.41%	30.62%	68.6%

Table 3. Results for 10-class one-shot classification on Caltech-101, using the logistic projection as the intermediate representation. The same 10 testing categories were used for all methods, and the PCA vectors were taken from the 40 class case. Training on more categories improves the one-shot recognition performance.

Again, the images from 10 classes were held out as testing categories, while the remaining images were available to train $\mathbf{r}(\mathbf{x})$, using the logistic projection from Equation 4, as the intermediate representation. To better understand the impact of training classes on our internal representation, we repeated the training procedure using 40, 65, and 91 classes. These were designed so that the larger sets contained all of the object categories present in the smaller sets.

These three different training classes make it possible to measure the benefit of using a large number of classes to learn $\mathbf{r}(\mathbf{x})$. As shown in Table 3, increasing the number of classes leads to an increased recognition performance, with almost a 10% difference between the models trained on 45 and 91 categories. This confirms that exposing the micro-set training process to more categories generates an intermediate representation that is better able to separate novel image categories when presented with the testing classes.

It should be noted that even though our model is optimized for one-shot recognition, it is straightforward to extend the classification process to use more examples per class. As an example, by using three exemplars per class, recognition accuracy improves 5%.

8. Discussion

As mentioned in Section 2, the proposed approach for learning the intermediate representation is unique in that the learning criterion directly simulates the generalization performance of the one-shot classifier. In each micro-set, the NCA criterion measures the probability that a nearest-neighbor classifier assigns the ground-truth label to each image in the test-set for that micro-set. This effectively measures how well the intermediate representation enables a nearest-neighbor classifier to generalize from the training examples in that micro-set. Averaging these probabilities over large numbers of micro-sets leads to a better estimate of the true generalization performance.

A major advantage of the micro-set approach is that it

is both straightforward to implement and scales well to large data-sets. Other transfer-learning approaches, such as [22, 7] require computationally-expensive global steps. By contrast, our approach can be implemented with basic gradient-descent techniques, using a criterion that is straightforward to differentiate. Because the overall criterion is computed independently across micro-sets, the training parallelizes easily in a many-core or cluster architecture. The micro-set approach also scales well to large data-sets using a stochastic gradient descent implementation, as in the experiments discussed in Section 5.

In contrast to transfer learning approaches such as [22], our framework does not employ unlabeled data. This is because the amount of available labeled data for common categories already overwhelms our current (and projected) computing resources during the training phase. In fact, unlabeled data exacerbates rather than alleviates the problem. For these reasons, we argue that exploiting unlabeled data is less important than taking advantage of the large quantities of labels that are already available for related classes. Our focus is on showing how these can be effectively leveraged for the task of one-shot recognition.

9. Conclusion

In this paper, we have introduced a novel approach for learning an intermediate representation for one-shot recognition. Our approach directly simulates the generalization performance of a nearest-neighbor classifier by optimizing a NCA-like criterion over a large number of micro-sets. Each micro-set consists of a different one-shot recognition problem generated from the large set of training images. As we have shown, this approach makes it possible to learn an intermediate representation that significantly improves recognition performance for one-shot recognition of novel categories.

Acknowledgments

The authors thank Dean Pomerleau and Mark Palatucci for helpful discussions and Peter Gehler for making the sets of features available for download. Much of this work was completed while K.D.T. was participating in an NSF REU program at UCF (IIS-0851841). M.F.T. was supported by a grant from the NGA (HM1582-08-1-0021).

References

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6, Nov 2005. 1

[2] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005. 2

[3] S. Dickinson. The evolution of object categorization and the challenge of image abstraction. In *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press, 2009. 1

[4] B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92, 1997. 5

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2

[6] A. Farhadi, D. Forsyth, and R. White. Transfer learning in sign language. In *CVPR*, 2007. 1

[7] M. Fink. Object classification from a single example utilizing class relevance metrics. In *NIPS*, 2004. 1, 2, 8

[8] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009. 5

[9] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2004. 3

[10] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *CVPR*, 2004. 1, 2

[11] Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In *CVPR*, 2005. 1, 2

[12] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006. 4

[13] K. Lai and D. Fox. 3D laser scan classification using web data and domain adaptation. In *Proc. Robotics: Science and Systems*, 2009. 1

[14] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 4, 5, 6

[15] F.-F. Li, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *ICCV*, 2003. 2

[16] F.-F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. PAMI*, 28(4), 2006. 1, 2, 4, 5, 7

[17] E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In *CVPR*, 2000. 2

[18] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. In *CVPR*, 2007. 2

[19] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15(2), 1991. 4

[20] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Predicting novel classes using semantic knowledge. In *NIPS*, 2009. 1, 2

[21] J. Ponce et al. Dataset issues in object recognition. In J. Ponce et al., editors, *Towards Category-Level Object Recognition*. 2006. 5

[22] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008. 1, 8

[23] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and categorization*. 1978. 6

[24] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *ICCV*, 2003. 1, 2

[25] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML*, 2004. 1

[26] A. B. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *CVPR*, 2008. 1, 2

[27] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2008. 7