CLEAR: Cumulative LEARning for One-Shot One-Class Image Recognition

Jedrzej Kozerawski   Matthew Turk
University of California, Santa Barbara
{jkozerawski, mturk}@ucsb.edu

Abstract

This work addresses the novel problem of one-shot one-class classification. The goal is to estimate a classification decision boundary for a novel class based on a single image example. Our method exploits transfer learning to model the transformation from a representation of the input, extracted by a Convolutional Neural Network, to a classification decision boundary. We use a deep neural network to learn this transformation from a large labelled dataset of images and their associated class decision boundaries generated from ImageNet, and then apply the learned decision boundary to classify subsequent query images. We tested our approach on several benchmark datasets and significantly outperformed the baseline methods.

1. Introduction

Thanks to a combination of advanced machine learning algorithms and access to an abundance of annotated data sets, object recognition tasks have become easier to solve and can achieve very high accuracy [28, 21, 17, 18]. However, in many cases there may not be enough data to sufficiently train a classifier for a given object category. There are many reasons why this might be the case, such as data being unavailable, prohibitively expensive, or time consuming to obtain. Data availability forces a change in the methodology for classification tasks and has been a major factor in the need for novel approaches to recognition.

Table 1 presents a division of classification tasks based on the amount of available positive and negative examples in the training data. Reading from the top left to the bottom right, the first is presented the optimal recognition problem setting where we have access to an abundance of positive and negative examples to train on [18]. To the right is the one-shot/few-shot scenario, which occurs when we want to quickly learn a novel category with only one or a few examples [19, 15, 7, 37]. Next is the zero-shot classification problem, where we learn to recognize an object without seeing any examples, usually by utilizing the verbal description or some additional attribute information. This subset of clas-

sification problems is quickly gaining the attention of the research community [30, 27, 11, 13, 20, 16, 6, 22, 8]. Another significant direction is one-class classification, where we have access to a large number of positive examples but only a handful of (or no) negatives. This is the outlier detection problem, illustrated, for example, by the problem of detecting when a nuclear power plant is not working appropriately, while having only data describing its normal operation. It also has been widely studied [29, 32, 33, 23, 35].

Both one-shot and one-class classification have received much attention from the machine learning community [1, 19, 14], but the intersection of these scenarios – the one-shot one-class (OSOC) problem – remains rather unexplored, as does the overlap of the zero-shot and one-class (ZSOC) problems. This work focuses on tackling the particular problem of one-shot one-class (OSOC) classification, which occurs when it is necessary to build a classifier for a novel category having only a single example and no negatives.

The motivation behind this work is based on the way people learn new skills. While "practice makes perfect" – i.e., repetition or multiple attempts at the same task lead to mastering it – there is a additional aspect of learning: the more similar skills you already know, the easier it is to grasp a new one. For example, in learning foreign languages, a third language is usually easier to master than the second one, as we are more experienced in the action of learning itself – this is relevant to the concept of learning-to-learn [36]. In a similar fashion, we should be able to impart to an image recognition system the ability to improve its recognition skill – the more object categories it has already learned, the better and faster it will learn the next category.

Data availability		Positives		
		Many	*One*	*Zero*
Negatives	*Many*	Recognition	One-shot	Zero-shot
	None	One-class	**OSOC**	ZSOC

Table 1: Taxonomy of classification problems based on the amount of available training data
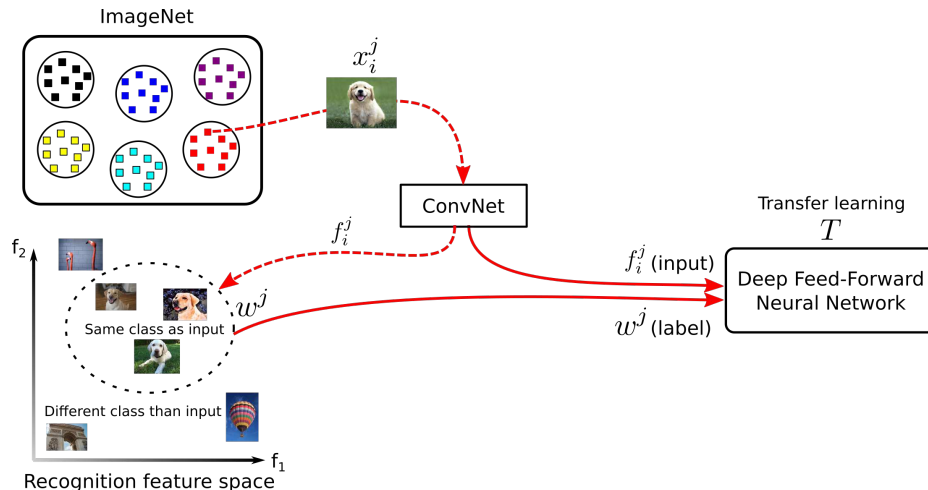
Figure 1: An overview of training the transfer learning model for one-shot one-class recognition. **Dashed lines:** Graphical representation of generating the training data. Using feature vectors f extracted from all the images in ImageNet, we create J decision boundaries w, one for each object category. **Solid lines:** The resulting training data pairs $\{f_i^j, w^j\}$ are then used to train the Deep Feed-Forward Neural Network.

In this paper we propose a "cumulative learning" approach to OSOC classification (CLEAR), inspired by this human-like way of learning, by accumulating the experiences of learning so far. Instead of learning-to-learn from only the most similar example [3], we exploit all of the knowledge acquired in the past training processes. Building on previous work by Wang and Hebert [39] on boundary transformations with neural networks working as regressors, we demonstrate that a deep network can perform the more advanced regression task of estimating a classification decision boundary from an image representation.

Figure 1 shows a high-level view of the training process for our method. To acquire the needed category-learning experience, we use ImageNet data from ILSVRC 2012 [28] and deploy an end-to-end model that takes an RGB image as input and provides as an output a one-vs-rest (OVR) decision boundary for the category the input image represents. The method workflow is depicted in Figure 2. We use a Support Vector Machine as the OVR classifier, as it is widely used in one-class problem settings (Section 2.2), but to demonstrate the general nature of our approach we evaluated it on different classification models (Section 4.4). Because of the limited amount of available training data, we chose to test classification models with a number of parameters similar to the dimensionality of the image representation. We do not use deep learning as classification method, since teaching the network to regress all the weights would require a tremendous amount of training data.

Our model with the vast category-creating experience collected on ImageNet can be directly tested on benchmark datasets without any fine-tuning or retraining using those benchmark datasets. We tested our approach on five datasets: Caltech-256 [9], two fine grain recognition sets (Oxford Flowers [24] and Caltech-UCSD Bird-200-2011 [38]), one scene recognition dataset (MIT Indoor scene recognition [26]), and one attribute recognition dataset (SUN attribute database [25]). We compared our method with two baselines: random chance classification (with 50% chance of image belonging to the same class) and a One-Class Support Vector Machine. Our proposed approach outperformed the baselines in all of the test sets.

The main contributions of this paper are as follows:

- We explore a novel problem, one-shot one-class classification, and present a working method for tackling this problem.

- We propose a learning-to-learn approach with the use of a Deep Neural Network and test it on multiple benchmark datasets and our approach outperformed the baseline methods.

- We demonstrate that generalized and robust SVM decision boundary transformations can be learned from only a single example and that our transfer learning model is applicable to any novel input image.

2. Related work

2.1. One-shot classification

One-shot classification is an important problem investigated by many machine learning researchers. A subgroup of methods approaching this problem tries to identify known instances (or their parts) in images of unseen objects. Lake
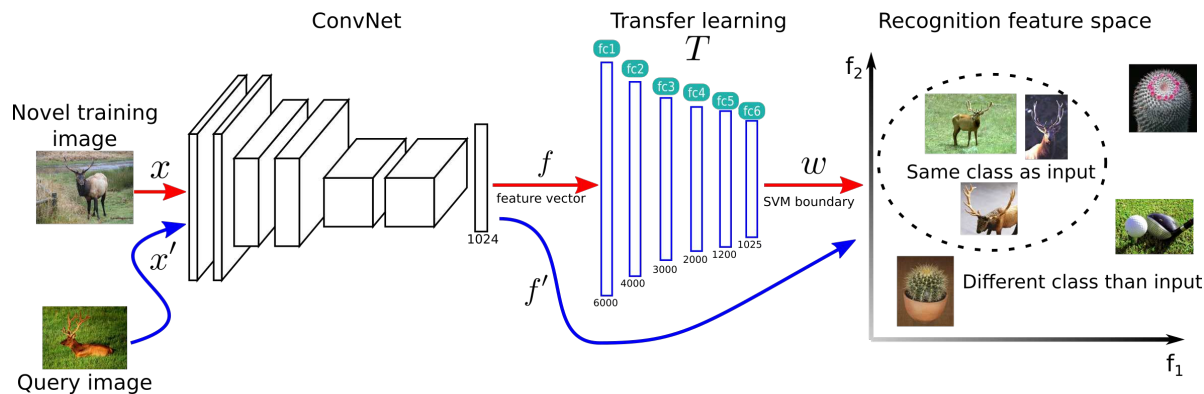
Figure 2: An overview of our OSOC recognition method. A novel training image is fed to the ConvNet and the output representation is input to our transfer learning model, which produces a OSOC SVM boundary. This decision function is used to classify any subsequent query image x' based on its feature representation f' as belonging to the same category as the novel training image or not. The ConvNet and the transfer learning network can be trained on independent datasets; in this case, we use ILSVRC 2014 and ILSVRC 2012, respectively. Five different image datasets were used for testing (see Section 4).

et al. [19] learned a generative stroke model to better understand how a newly seen character is created and was thus able to increase the accuracy of recognizing the same character based on the single example presented. That work mimics well how humans see objects (including characters) as compositions of different, already known objects. However the transition to learning visual categories in the same manner as characters is not as feasible and was addressed. Fei-Fei et al. [7] used a Bayesian Inference approach to quantify a probability that new images come from the same class as a single query image. They built a generative model to predict which subsets of features would make two images belong to the same class. This method involves a non-CNN feature extraction step. An approach involving Genetic Programming was used by Al-Sahaf et al. [1] to learn similarities between Local Binary Patterns (LBP) histograms of images. On a smaller scale, Aytar and Zisserman [3] used transfer learning to modify the SVM boundary for a novel class of objects by extracting information from a single most similar class. The drawback of the method is that it requires the presence of this similar class, utilizing information only from this single similar category; information accumulated in all the other categories is not exploited and treated as irrelevant.

Another approach by We et al. [40] used an elaborate technique to fuse temporal and spatial information from RGBD images and then transform data with a vector embedding. This allowed them to successfully classify gestures based on a small number of training examples for the CHALEARN gesture challenge [10].

One of the more interesting methods was presented by Koch et al. [15], who proposed a Siamese CNN architecture

that takes as an input two images and outputs the probability that both of them come from the same class. Such an approach shows the generalization capabilities of learning-to-learn approaches. The slight drawback is that one has to do such comparisons between every pair of images (query image, testing image) to determine the true class of every image in the test set.

Vinyals et al. [37] used Matching Networks to learn classifiers for novel classes in one-shot scenarios based on a mapping from a small support set of examples (input-label pairs) to a classifier for the given example. The method still requires examples in the testing phase coming from each of the novel classes at once (N-way one-shot classification). Very promising is the work of Burgess et al. [4], where they used Bayesian approach to update weights of hidden layers in the pretrained Deep CNN to create a classifier from just a single image input (instead of regular abundance of examples required to train such network).

2.2. One-class classification

A valuable taxonomy of one-class classification (OCC) problems was presented by Khan and Madden [14], who divide OCC problems into three categories based on the available data: training with just positive data, training with positive and scarce negative data, training with positive and unlabelled data. In our case, we focus on the first, hardest scenario of having just positive data and no negative or unlabelled data. In the same paper they break down OCC methods into SVM and non-SVM techniques, showing how prevalent SVM methods are for dealing with this scenario.

The first approach to One-Class Support Vector Machines (OC-SVM) was presented by Schölkopf et al. [29],

where, faced with examples coming only from one class, the algorithm attempts to maximize the margin between the origin and these samples. A similar approach was shown by Chen et al. [5], where they also used a One-Class SVM to better fit the target class of images in an image retrieval problem. Tax and Duin [32, 33] presented a novel method for dealing with OCC problems – Support Vector Data Description (SVDD) – and later [34] enriched their methodology by generating artificial outliers in lieu of optimizing the OCC problem, posing a balance between over- and under-fitting to the training data. In all of the methods of Schölkopf et al. and Tax and Duin [29, 32, 33] the origin played a very important role.

Using non-SVM methods, Manevitz and Yousef [23] presented an approach to learn a OCC with a feed-forward neural network as a classifier with just positive data. Tax [35] provided a novel method, called Nearest Neighbor Description (NN-d), for using a Nearest Neighbor classifier to deal with the OCC problem. Here a test object is accepted as coming from the class when its local density is greater than or equal to the local density of its nearest neighbor in the training set.

According to Khan and Madden [14] one-class classifiers have a very important feature working to their advantage – mainly that they can represent well the concept of "none of the above."

2.3. Transfer learning for boundary transformation

Another significant factor for this research is the concept of learning-to-learn [36] that has become an important direction in modern machine learning approaches, especially when dealing with few-shot and one-shot scenarios, and with methods involving decision boundary or model transformation. As mentioned in Section 2.1, Aytar and Zisserman [3] used transfer learning to modify an SVM boundary by using information from the most similar class to the novel one. This approach relied on having a classifier for a similar class and rejecting information from other categories. Moreover, it required negative examples to create a classifier (that will undergo this transformation) in the first place. The work of Burgess et al. [4] on Deep CNN hidden layers transformation (described in 2.1) is also very promising. A recent approach by Wang and Hebert [39] shows that there is an underlying learnable transformation between the decision boundaries of classifiers learned with few examples and the decision boundaries of classifiers learned with many examples. Similarly to the previous case, this method relies on having negative examples in the first place to create the weak classifier for the transformation.

3. Method

The CLEAR method for OSOC recognition (see Figure 2) consists of three main components: (1) a pretrained Convolutional Neural Network working as a feature extractor or image representation, (2) a binary classifier to achieve one-class classification, and (3) a transfer learning component – a Deep Feed-Forward Neural Network working as a regressor to transform the image representation into a classification decision boundary.

3.1. CNN representation and a classifier

Convolutional Neural Networks perform exceedingly well on classification tasks with a fixed number of possible outputs and plentiful training examples [18, 31]. When the amount of training data might be not enough for some classification task, two common techniques are to use a network such as [18, 31] pretrained on a large labelled dataset (frequently ImageNet [28]) and fine-tune the weights of the CNN to suit this novel recognition task, or to use the CNN as a representation (or feature extractor) and train a classifier such as a Support Vector Machine (SVM) on those extracted features, as in work by Athiwaratkun and Kang [2]. Both of these approaches work well when faced with a known a priori number of visual categories. Unfortunately, when the number of visual categories changes while training, the architecture of the CNN (including the number of output neurons) must change and the training process started all over again. The same corresponds to the SVM classifier requiring to be retrained with the new data and for the new number of possible categories.

Our method uses features extracted with the use of a CNN and an SVM classifier on top of that. Typical multiclass classification creates decision boundaries between all known J categories, so when a new category ($J + 1$) is introduced, it is necessary to incorporate that information and retrain the classifier (using data coming from the previous J classes as negative examples). To achieve a more robust approach that does not require such retraining, we employ a one-class type of classification, which creates separate classifier for every category with no negative examples – therefore when a new category is introduced, we simply create a classifier for that category based on only positive examples from that novel class, with no dependence on data from the previous J known categories.

3.2. Single image to decision boundary regression

We apply transfer learning to model the desired SVM decision boundary for a given single image input, based on generated {image, decision boundary} pairs for training. Wang and Hebert [39] demonstrated that it is possible to learn an underlying transformation from an SVM decision boundary learned from a small number of positive examples to an SVM decision boundary learned from a large number of positive examples. We extend this SVM-SVM regression to an image-SVM regression, learning the decision boundary from the novel image. The main differences between

their work and ours are as follows:

- Our method does not require creating a weak SVM classifier to provide an input; we work on image data directly.

- Our method is one-shot one-class, so it does not require any negative examples coming from other classes.

- Our transfer learning approach is trained on an entirely different dataset that we use for testing.

- Our method does not require retraining when the number of categories in the test set changes.

Our approach demonstrates that it is possible to learn an underlying, general transformation from an image to a classification decision boundary for the class it represents. We use a linear-SVM classifier to simplify the learning process for the regression network, as it is not required to learn a kernel. A kernel-based SVM can also be used here, although with more parameters for the network to learn, so the dimensions of the layers of our network would grow as well, and we would need much more training data to properly train all the weights.

Let us introduce some notation. Let f_i^j be a representation (or feature vector) describing the i-th image belonging to the class j, and let w^j be a linear-SVM decision boundary (a weight vector and a bias) for class j. We define a transformation T that maps f_i^j to w^j for all j. Using a large number of training pairs $\{f_i^j; w^j\}_N^{j=1}{}_{i=1}$ created from annotated data (where J is a number of classes and N is a number of images per class), we are able to learn the transitions T effectively.

3.3. Generation of training pairs

To generate pairs $\{f_i^j; w^j\}_N^{j=1}{}_{i=1}$ first we need to extract the representation of the input image (the feature vector) using the CNN. From every image x_i^j we extract a feature vector f_i^j, where j is a ground truth label of the image i. Next, for every category $\{c^j\}_{j=1}^J$ we create a classifier using all N available data points f_i^j belonging to class c^j as positive examples and M data points randomly sampled from other categories as negative examples. This step allows us to create a w^j decision boundary for every class j.

3.4. Implementation details

An overview of our method is depicted in Figure 2 and its training process in Figure 1. The transfer learning step, involving the neural network as a representation for the transformation T, consists of six fully connected layers with leaky ReLU activation functions.

We use the Caffe [12] GoogLeNet [31] CNN as a feature extractor. All the GoogLeNet weights are frozen to those learned on ILSVRC 2014. We resize each image to 224 × 224 pixels. The resized image is fed to GoogLeNet and we take the output of the final pooling layer "pool5/7x7_s1" as our 1024-dimensional feature vector describing the image. We used standard data augmentation in terms of image mirroring to increase the number of available training examples.

We use the ImageNet ILSVRC 2012 [28] dataset as a source to create a large number of training pairs $\{f_i^j; w^j\}_N^{j=1}{}_{i=1}$ (see Section 3.3). For each of 1000 categories in ILSVRC 2012 we create 1200 training pairs that we split 70-30 into training and validation sets. That gives us a total of 1,200,000 pairs to train the network.

The architecture of the neural network consists of six fully connected layers with dimensions from $fc1$ to $fc6$ of 6000, 4000, 3000, 2000, 1200, and 1025. We use the Caffe framework to train the network.

4. Experimental results

4.1. Accuracy metrics used

For each dataset we calculate the mean average precision (MAP) and F1-score measures, shown in Tables 2 and 3. We also present the ROC (Receiver Operating Characteristic) curves and their respecting AUC (Area Under Curve) scores for every test set in Figure 4. Additionally, we show the average precision and recall measures for every test set in Table 5.

It is important to note that in order to properly test the one-shot one-class method, only a single query image is presented (as a one-way one-shot classification instead of J-way one-shot classification). This is done to treat every classification problem as a one-class classification without any information about other $J - 1$ classes. For each class $c_{j=1}^J$ within the data set we subsampled K instances to create the testing set. Next, we created a classifier $w_{j=1}^J$ for each class using a single sample from the class (not in the test set) and then applied our method to it.

For every category we randomly select $K = 20$ images for testing and one image to create a classifier. We repeated the experiments 20 times, presenting average results to minimize the impact of any outliers.

4.2. Datasets used

We tested our method on five datasets: one for general object recognition, two for fine-grain image recognition, one for scene recognition and one for attribute recognition.

Caltech-256 [9] is a benchmark image recognition dataset consisting of 30,607 images coming from 256 different visual categories. The images have high intra-class variability and high object location variability. The images represent a diverse set of lighting conditions, poses, backgrounds and sizes.

Method	Caltech-256		CUB-200-2011		Flowers 102	
	MAP	F1	MAP	F1	MAP	F1
Threshold	0.0004	0.008	0.0035	0.0003	0.03	0.0029
Chance	0.005	0.008	0.007	0.01	0.013	0.019
One-Class SVM	0.061	0.011	0.051	0.01	0.068	0.016
CLEAR [ours]	**0.364**	**0.176**	**0.068**	**0.037**	**0.193**	**0.087**

Table 2: Object recognition datasets: Comparison of MAP and F1 scores between baselines and our method. Our method outperforms the baselines on all datasets, usually by at least one order of magnitude.

Method	MIT Indoor 67		SUN attributes	
	MAP	F1	MAP	F1
Threshold	0.001	0.0001	0.006	0.0006
Chance	0.019	0.029	0.015	0.021
One-Class SVM	0.025	0.003	0.009	0.001
CLEAR [ours]	**0.209**	**0.111**	**0.081**	**0.041**

Table 3: Attribute-oriented recognition datasets: Comparison of MAP and F1 scores between baselines and our method. Our method outperforms baselines on all datasets, usually by at least one order of magnitude.

Oxford flowers 102 [24] is a benchmark dataset for fine-grained image recognition containing 8,189 images of flowers belonging to 102 different categories. Each class consists of between 40 and 258 images.

Caltech-UCSD Birds-200-2011 (CUB-200-2011) [38] is also a benchmark for fine-grained image recognition. It consists of 11,788 images from 200 categories of birds.

The MIT Indoor 67 scene recognition dataset [26] contains 67 indoor categories and a total of 15,620 images. There are at least 100 images per category.

The SUN attribute database [25] is a large-scale scene attribute database with 102 discriminative attributes (e.g., natural, man-made, open, enclosed, etc.) of 14,340 images, for fine-grained scene understanding.

4.3. Discussion of results

As one-shot one-class recognition is an underexplored problem, we have compared our method with three baseline methods – threshold method ($\|f_i - f_j\|^2 \leq thresh$), random chance classification, where there is a 50% chance for every test image to be classified as the same class as a query and with a One-Class SVM created with the single query image (its feature extracted with GoogLeNet). We have also tested the Nearest Neighbor technique as a baseline, but for one-shot one-class setting it always categorizes all examples in the test set as positives (due to presence of only one training example), so we disregarded such approach as it is impractical. We divided our test sets into two cate-

gories: first called "object-based" recognition (Caltech-256, Oxford Flowers and Birds) and the second called "attribute-based" recognition (MIT Indoor and SUN attribute). We have done that to distinguish difficulty of the task of a simple object recognition from difficulty of a more complex scene understanding (where it is crucial to identify certain attributes and recognize important combination of different objects in the image). We present results of experiments in Table 2 for object recognition test sets and in Table 3 for the attribute-based test sets. Our proposed approach outperforms baselines for all test sets. For the Caltech-256 set, the gain in accuracy (compared to the random chance baseline) is a 72.8 times increase for MAP and 22 times for the F1 score. The average gain per dataset for the other object-based datasets is 12.3 for MAP and 4.1 for F1. For the attribute-based datasets the average gains per set are 8.2 for MAP and 2.9 for F1.

Class granularity problem: An important factor behind the differences in the performance of our method among the datasets is the granularity of the data in the test sets – i.e., the magnitude of the differences between the classes in a test set. For example, in Caltech-256 the classes represent different visual objects, frequently not similar to each other, such as "American flag" or "saturn," whereas Oxford Flowers consists only of categories representing different types of flowers, some of which are quite similar. As a measure of the dataset granularity we calculated the average Euclidean distance between centroids of categories within each dataset (the mean Inter Class Euclidean Distance) according to the following equation:

$$mICED = \frac{1}{2 \times N \times (N-1)} \times \sum_{i=1}^{N} \sum_{j=1}^{N} ||c_i - c_j|| \quad (1)$$

where $mICED$ is the mean Inter Class Euclidean Distance, N is the number of categories in a dataset, and c_i is the centroid of a category i.

This granularity metric relates directly to how far on average an SVM boundary should be from a given sample point. Calculated mICED metrics for all the datasets are presented in the Table 4, where the numbers in parenthesis are the differences between the granularity of ILSVRC 2012
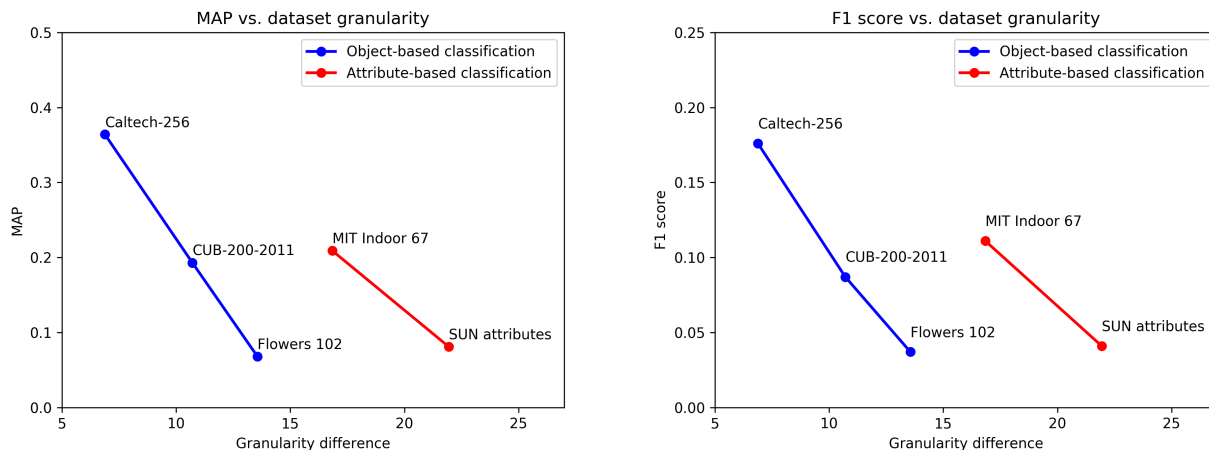
Figure 3: OSOC recognition accuracy as a function of dataset granularity.

ILSVRC 2012	Caltech-256	CUB-200-2011	Flowers 102	MIT Indoor 67	SUN attributes
35.15	28.27 (6.88)	21.59 (13.56)	24.44 (10.71)	18.31 (16.84)	13.21 (21.94)

Table 4: Mean Inter Class Euclidean Distance metrics (within a dataset) as a dataset granularity metric. Given in parentheses are differences between ILSVRC 2012 granularity and dataset granularity.
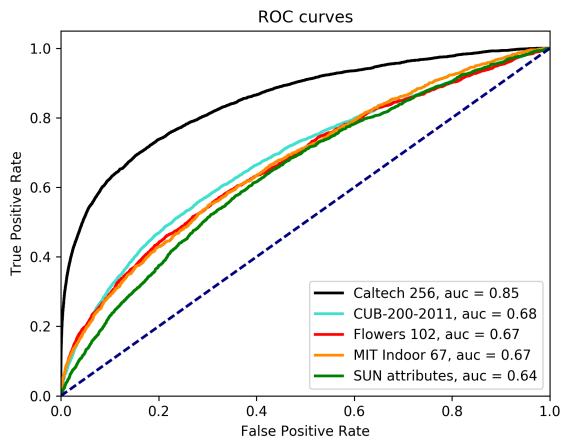

Figure 4: ROC curves and their AUC scores for every test dataset.

(on which the network was trained) and the subsequent test sets.

In one-shot one-class problems, the class granularity problem is unavoidable – based on only a single image, one cannot guess at what hierarchy level the class should be. For example, given an image of a dog, is the intent to recognize an animal of the same species, a mammal, a non-flying animal, or should the classification be based on some attribute like color or size? The class granularity problem explains why our method improves recognition accuracy far more for datasets with granularity similar to that of ImageNet ILSVRC 2012 data. As a comparison, we show in Figure 3 how the difference in granularity (compared to that of ILSVRC) influences the MAP and F1 scores for datasets (left and right, respectively). The figure compares separately the accuracies for object-based (blue) and attribute-based (red) test sets. Based on those figures we can see a linear dependency suggesting how the accuracy for a given test set decreases as the granularity difference increases.

Despite some dependence on dataset granularity, our approach shows robustness and generalization abilities. It performs well on the MIT Indoor 67 data, which has categories like "mall," "dental office," and "meeting room," and on the SUN attribute database with attribute-based categories. Both MAP and F1 scores are higher for MIT Indoor than for Oxford Flowers or CUB-200-2011 datasets; such high recognition accuracy for a recognition task much less related to the ILSVRC 2012 than, for example, Caltech-256 (in terms that it is capturing much more complex ideas than just a single object recognition) demonstrates that cumulative knowledge transfer learning might be used to alleviate more difficult problems such as attribute-based classification. We are confident that with additional context regarding the granularity of a classification problem (perhaps sup-

Metric	Caltech-256	CUB-200-2011	Flowers 102	MIT Indoor 67	SUN attributes
Precision	0.15	0.011	0.037	0.072	0.037
Recall	0.3	0.49	0.4	0.3	0.099

Table 5: Precision and recall values for the test data sets.

Metric	Caltech-256	CUB-200-2011	Flowers 102	MIT Indoor 67	SUN attributes
MAP	0.31	0.067	0.092	0.16	0.058
F1	0.17	0.034	0.045	0.08	0.025

Table 6: MAP and F1 accuracy measures for our CLEAR method with the logistic regression as classification.

plied by the user), such tasks will be more feasible with this transfer learning approach.

Figure 4 shows ROC curves for the five test sets compared with the ROC curve for random classification (dashed line). High AUC scores demonstrate valuable recognition abilities of our method. In Table 5 we present precision and recall measures for all five test sets. The results confirm that the obtained classifiers generalize well and are robust, since all classifiers within a data set have high recall values, but the precision decreases as the granularity difference increases (Table 4) due to progressively more false positives in data sets.
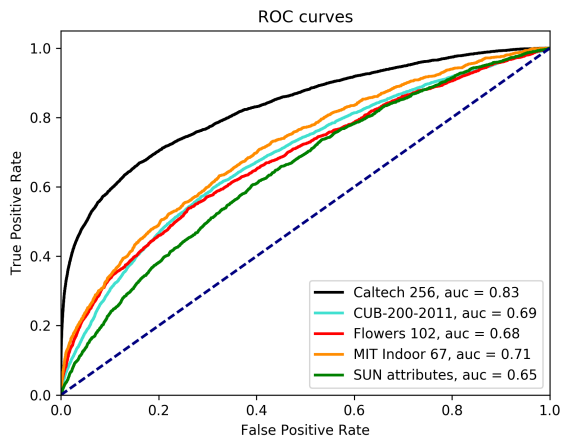

Figure 5: Evaluation of logistic regression as classification method – ROC curves and their AUC scores for every test dataset.

4.4. Evaluation of different classification models

We have focused on using SVMs as classifiers, but this approach can work with other classification models as well; all it requires is a set of weights to be learned and enough training data to learn those weights. To verify this, we

added a logistic regression classifier in place of an SVM. As logistic regression also provides an $D + 1$ dimensional decision function (from an D-dimensional feature space), the architecture of our network did not require any changes. In Table 6 we present MAP and F1 accuracy measures for our CLEAR method with logistic regression as a classifier. The results suggest that our proposed approach can be used with classification methods other than an SVM. Results for CLEAR with logistic regression shows improvement in comparison to the baselines, but both the MAP and F1 measures are lower than for CLEAR with SVM. In Figure 5 we present ROC curves for the five test datasets compared to the ROC curve of random classification (dashed line). High AUC scores demonstrate that our method is working well also with the logistic regression as a classification method.

5. Conclusion

This work addressed the novel problem of one-shot one-class classification. Our proposed method involves using transfer learning to understand how to obtain a classification decision boundary when given just a single image. This paper demonstrates the ability of a neural network to function as a model regressor. We have tested our approach on five benchmark datasets and our results outperformed the baseline comparisons. Our experiments reveal that the method has potential to generalize well and works robustly even on attribute-based datasets. Considering high AUC results for our method we expect that using one-shot one-class approaches might be helpful in the image retrieval applications as well. Further research in the OSOC area can help with more advanced attribute-based classification and can lead to a better understanding of different granularity levels of categories.

References

[1] H. Al-Sahaf, M. Zhang, M. Johnston, S. Cranefield, and A. Nayak. A one-shot learning approach to image classifica-

tion using genetic programming. In *Australasian Conference on Artificial Intelligence*, pages 110–122. Springer, 2013.

[2] B. Athiwaratkun and K. Kang. Feature representation in convolutional neural networks. *arXiv preprint arXiv:1507.02313*, 2015.

[3] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2252–2259. IEEE, 2011.

[4] J. Burgess, J. R. Lloyd, and Z. Ghahramani. One-shot learning in discriminative neural networks. *arXiv preprint arXiv:1707.05562*, 2017.

[5] Y. Chen, X. S. Zhou, and T. S. Huang. One-class svm for learning in image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 34–37. IEEE, 2001.

[6] Z. Ding, M. Shao, and Y. Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2050–2058, 2017.

[7] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[8] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong. Recent advances in zero-shot recognition. *arXiv preprint arXiv:1710.04837*, 2017.

[9] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[10] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante. The chalearn gesture dataset (cgd 2011). *Machine Vision and Applications*, 25(8):1929–1951, 2014.

[11] N. Hussein, E. Gavves, and A. W. Smeulders. Unified embedding and metric learning for zero-exemplar event detection. *arXiv preprint arXiv:1705.02148*, 2017.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] N. Karessli, Z. Akata, A. Bulling, and B. Schiele. Gaze embeddings for zero-shot image classification. *arXiv preprint arXiv:1611.09309*, 2016.

[14] S. S. Khan and M. G. Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014.

[15] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition.

[16] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Cognitive Science Society*, volume 33, 2011.

[20] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. *arXiv preprint arXiv:1703.05002*, 2017.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[22] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. *arXiv preprint arXiv:1705.01782*, 2017.

[23] L. M. Manevitz and M. Yousef. Document classification on neural networks using only positive examples (poster session). In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 304–306. ACM, 2000.

[24] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.

[25] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

[26] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[27] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588, 2000.

[30] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[32] D. M. Tax and R. P. Duin. Data domain description using support vectors. In *ESANN*, volume 99, pages 251–256, 1999.

[33] D. M. Tax and R. P. Duin. Support vector domain description. *Pattern recognition letters*, 20(11):1191–1199, 1999.

[34] D. M. Tax and R. P. Duin. Uniform object generation for optimizing one-class classifiers. *Journal of machine learning research*, 2(Dec):155–173, 2001.

[35] D. M. J. Tax. One-class classification. 2001.

[36] S. Thrun and L. Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[37] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[39] Y.-X. Wang and M. Hebert. Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision*, pages 616–634. Springer, 2016.

[40] D. Wu, F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 7–12. IEEE, 2012.