Low Shot Learning for Face Recognition

Ramya Sarma
University of Massachusetts
Amherst, MA
rsarma@umass.edu

Ananya Ganesh
University of Massachusetts
Amherst, MA
aganesh@umass.edu

1. Introduction

Face recognition is the task of uniquely identifying a person, given an image of their face. It is a relevant problem in security systems, access control, automated authentication, etc. This can be modeled as a classification problem where each class is a person whose identity is known. With enough example images for each person, a classifier can be trained using handcrafted features [1] or using learned features [4]. Deep learning approaches to face recognition are found to be very effective, but a significant drawback is that they require large labeled datasets in order to achieve good performance. This is a problem because access to new images of faces might be limited or expensive.

To combat this problem, we explore a low-shot learning approach to face recognition. Low-shot learning is the ability to learn to distinguish classes from a small number of examples. We focus on low-shot visual recognition where the AI systems are taught to recognize different objects from images using very few examples. In the low-shot learning set-up, there are a fixed number of base classes, for which a large number of training examples are available, and then there are novel classes, for which a limited number of training examples are available. The classifier is then evaluated based on its ability to correctly classify even the novel classes. For object recognition, a low-shot learning method based on shrinking and hallucinating features was proposed by Hariharan and Girshick in [3], which gave good results on ImageNet. We will be extending their approach to the task of face recognition to evaluate if it is applicable here. Solving the low-shot recognition problem will enable us to apply AI based methods to many real-world tasks.

We use a subset of images of celebrities available on the web as training data. The rich information provided by the base classes data helps to conduct disambiguation and improve the recognition accuracy in cases where we have lesser training data, that is the novel classes. We hope that this technique contributes to various real-world applications, such as image captioning and news video analysis. For this report, we describe our dataset and the preprocessing involved, the technical approach we will follow, baseline results, and our next steps.

2. Problem Statement

We propose to investigate the problem of low-shot visual recognition for face recognition. Low-shot visual learning refers to the ability to learn from very few examples.
Current supervised recognition systems require large labeled datasets to train classifiers for a fixed set of classes. This is a problem in face recognition because labeled data is not as widely available for human faces as for other subjects. This problem is interesting because the human visual system can recognize faces with just one example, and that's what we are moving towards.
Our aim: Given a novel class with very few training examples, the classifier should still be able to recognize the person in the novel class.

3. Datasets

We used a subset of the 1M Celebrities dataset from Microsoft Research [2]. As described in the introduction, it comprises of face images collected from the internet. The complete dataset comprises of complex and unprocessed images.
The data distribution within this celebrity 1M dataset is not uniform. Also, the images are not aligned and cropped. Hence, the dataset originally provided for the low-shot learning benchmark task warrants us to run face detection along with face recognition algorithms. It has images of 21,000 persons and an average of 200 images per class. The dataset is divided into two subsets:

- **Base set** - There are 20,000 persons in the base set. Each person has 50-100 images for training and about 5 images for testing.
- **Novel set** - There are 1,000 persons in the novel set.

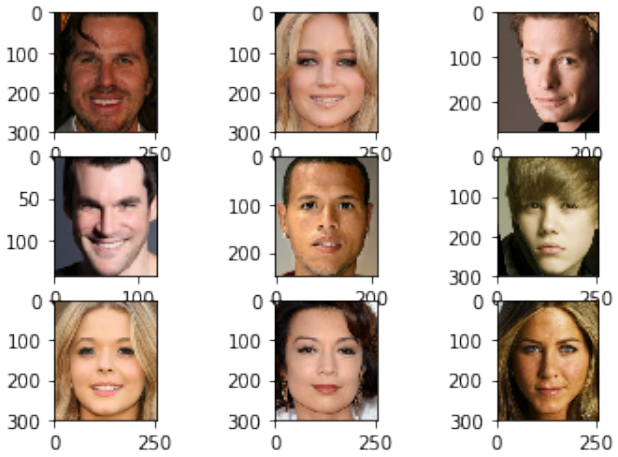

Figure 1. Sample image from each base class

Each person has 1-5 images for training, and 20 images for testing.

Our dataset is a cleaner subset of the 1 Million Celebrities dataset which has been cropped and aligned. We use this dataset instead of the 1 Million Celebrities dataset to eliminate the challenges presented by scalability and noisy samples and focus on low-shot learning.

We create an artificial low-shot learning scenario as described below.

3.1. Dataset Construction

We sample 5 classes randomly and consider these as our novel classes and another 9 classes are considered as base classes. Sample images from the base classes are shown in Figure 1.

The base and the novel classes data is split into two partitions, a training pool set and a test set. The training pool data will be used to sample the training data. The low-shot training data for the novel classes is generated by sampling images from new classes. The feature extractor is learnt on the base class data. The classifier is learnt on the base class training data and the sampled novel class training data. The final evaluation is done on a test set created from the novel and base classes. The evaluation is described in detail in the next section.

4. Technical Approach

4.1. Low shot learning

The idea is to augment the training set with generated examples of the novel classes. Traditionally, this is done by naive approaches such as jittering the pixels, horizontal flipping, changing the contrast of the images, etc., in order to get new examples from the existing images. However, a more effective approach as described in [3] is to hallucinate examples of the novel classes based on features extracted from the base classes. Thus, the process can be described in two phases:

- Representation learning - Using the base categories, a feature extractor is trained. We currently use Cross-Entropy Loss on the class labels, whereas [3] use Squared Gradient Magnitude loss, which we will experiment with.
- Low shot learning - Using features from the base classes, a generator is trained which will apply a transformation on a novel class image to get a new image. This transformation is derived from the training images in the base class.

Specifically, in the generation phase, a generator G is trained using transformation analogies from the base classes. For example, if the base class has an image of an non occluded face, and another example where a hand is resting on the face, we would identify this as a transformation that can be applied to the novel class. The analogies are collected by clustering the examples in the base class and represented using centroids from different clusters which have high similarity. Once analogies are collected, the generator is trained using a combination of classification loss and mean squared error loss.

Following this, a 'seed' example from the novel class is given to the trained generator, along with a randomly sampled analogy. which will output a hallucinated image of the novel class. This is then added to the training set with the label of the novel class, and will be learned by the classifier. We expect this to improve the original performance of the classifier.

4.2. Baseline

For our baseline model, we train a convolutional neural network that consists of three convolutional layers, one max-pool layer, followed by three fully connected layers. We use ReLU activation and use CrossEntropy loss for the classifier. We train the network using AdaGrad optimization. As mentioned in the dataset section, we have 9 base classes with 50 examples each and 5 novel classes with 5 examples each. First, we train the network using only the base classes and observe the results on the test set, which has 10 examples per class. Then, we train it on both the base classes and novel classes together and evaluate on test set. The results are reported in Table 1.The baseline model is meant to show that there is a drop in performance when novel categories with very few examples are included. We aim to beat these results using the low shot approach described above.

5. Preliminary Results

Since our dataset is fairly small, we are able to achieve 100% training accuracy on the network trained on base classes alone. We find that test accuracy is not so high and reaches 80%. This could possibly be due to overfitting and we will add dropout layers in future to improve performance. Also, when novel classes are added, the training accuracy still reaches 100% but test accuracy falls to 56.8%.

Class	Train Accuracy(%)	Test Accuracy (%)
Base	100	82
Base + Novel	100	56.83

Table 1. Train and Test accuracies for classes

Following this, we will first implement a naive generator that augments data by jittering, and then implement the hallucination based low shot learning approach.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004.

[2] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer, 2016.

[3] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy*, 2017.

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.