One-shot Face Recognition by Promoting Underrepresented Classes

Yandong Guo, Lei Zhang
Microsoft
{yandong.guo, leizhang}@microsoft.com

Abstract

We study in this paper the problem of one-shot face recognition, with the goal to build a large-scale face recognizer capable of recognizing a substantial number of persons. Given that for face recognition one can leverage a large-scale dataset to learn good face representation, our study shows that the poor generalization ability of the one-shot classes is mainly caused by the data imbalance problem, which cannot be effectively addressed by multinomial logistic regression that is widely used as the final classification layer in convolutional neural networks. To solve this problem, we propose a novel supervision signal called underrepresented-classes promotion (UP) loss term, which aligns the norms of the weight vectors of the one-shot classes (a.k.a. underrepresented-classes) to those of the normal classes. In addition to the original cross entropy loss, this new loss term effectively promotes the underrepresented classes in the learned model and leads to a remarkable improvement in face recognition performance. The experimental results on a benchmark dataset of $21,000$ *persons show that the new loss term significantly helps improve the recognition coverage rate from* 25.65% *to* 77.48% *at the precision of* 99% *for underrepresented classes, while still keeps an overall top-1 accuracy of* 99.8% *for normal classes.*

1. Introduction

In this paper, we study the problem of one-shot face recognition, with the goal to build a large-scale face recognizer. Being capable of recognizing a substantial number of individuals with high precision and high recall is of great value to many practical applications, such as surveillance, security, photo tagging, and celebrity recognition.

Building a large-scale face recognizer is a non-trivial effort. One of the major challenges is that, for some of the persons to be recognized, there might be very limited number of training samples, or even only one sample for each of them. This challenge naturally exists in many real scenarios, especially when the number of persons to be recognized is very large. Although recent years have witnessed great processes in deep learning and visual recognition, computer vision systems still lack the capability of learning visual concepts from just one or a very few examples [12].

Figure 1: The three images in the leftmost column are used for *training*. The rest images (in the right panel) are the corresponding images for *testing* (partially selected from the test set). **With only one image for each person, our model can recognize all these test images from hundreds of thousands of other testing images.** More detailed results are presented in the experimental results section. Note that we select three typical challenging cases: faces with occlusion, drawings, and low resolution image.

To study this problem, we design a benchmark task and provide the associated dataset consisting of $21,000$ persons

each with 50-100 images of high accuracy. Similar to [6], We divide this dataset into two sets, $20,000$ persons in *base set* and $1,000$ persons in *novel set*. The task is to study if tens of images are given for each person in the base set while only **one** image is given for each person in the novel set, how to develop an algorithm to recognize the persons in both the data sets. In particular, we mainly focus on the recognition accuracy for persons in the novel set as it shows the one-shot learning capability of a vision system, while we also check the recognition accuracy for those in the base set to ensure not to hurt their performance.

Our method for this benchmark task is to train a good face representation model and build a classifier on top of that. The base set is used to train a face representation model, which has good generalization performance on the novel set. Containing about one million images for $20,000$ persons with high accuracy makes our base set one of the largest public datasets [9, 8, 22, 3, 5, 17]. We use a standard residual network with 34 layers [7] to train a classification model for these $20,000$ persons. In order to evaluate the generalization performance of this model, we use the last pooling layer as the face feature and achieve a single model accuracy of 98.88% on the LFW verification task [9, 8], which is close to the state of the art. Note that the base set does not include any person in LFW by design. Recently, we have seen a considerable amount of research results reported on LFW, which are obtained by using different algorithms on different datasets [1, 17, 16, 18, 15, 19, 20]. To help advance the research in this field, we publish this dataset and hope it can help researchers compare different algorithms with the same training data to emphasize the contribution from the perspective of algorithms.

Despite the face representation model obtained with the base set, one has to solve the technical challenge caused by the highly imbalanced training data when building classifiers to recognize persons in both the base and novel sets. As the most widely used and efficient multiclass classifier in deep convolutional neural networks, multinomial logistic regression (MLR) has shown great performance on various visual recognition problems. However, it receives much less attention in the low-shot learning problem, where each novel class has only very few but each base class has much more training samples. In our experiments, we have observed very poor performance of MLR in recognizing persons in the novel set since the persons in the novel set have much less images per person compared with the persons in the base set.

A further analysis in Section 3 shows that a novel class with only one training sample can only claim a much smaller partition in the feature space. And we reveal that there is a close connection between the volume of a class partition in the feature space and the norm of the weight vector of this class in the multinomial linear regression model. Based on this finding, we propose to add a new loss term to the original cross-entropy loss for MLR, serving as a prior for the weight vectors in multinomial logistic regression. This new loss term is based on our empirical assumption and observation that on average, each person in the novel set should occupy a space of similar volume in the feature space, compared with the persons in the base set. We call this term the Underrepresented-classes Promotion (UP) loss. For comparison, we also explore other different options on the priors of the weight vectors.

To quantitatively evaluate the performance, we adopt the multi-class classification setup (close-domain face identification) with test images from both the base set ($100,000$ images, 5 images/person) and the novel set ($20,000$ images, 20 images/person). We mainly focus on the classification performance on the novel set to evaluate how well the computer can learn novel visual concepts with only one example, while also monitor the performance on the base set to ensure the performance gain on the novel set is not obtained by sacrificing the performance on the base set.

Our experimental results clearly demonstrate the effectiveness of the proposed method. With the UP term, we can recognize 77.48% of the test images in the novel set, with a high precision of 99%, while all the other methods can only recognize up to 25.65% of the test images at the same precision. Note that our UP term does not affect the accuracy on the base set. The classifier with the proposed UP term still has an overall top-1 accuracy of 99.8% on the base set.

Our contributions are highlighted as follows.

- We set up a benchmark task for one-shot face recognition, and provide the associated data set. This data set is divided into base set (many persons, many training images/person) and novel set (many persons, one training image/person). This benchmark task simulates a lot of real scenarios.
- The base set we provide is one of the largest published face data sets in the literature. With very minimal effort, one can learn a good face feature with face verification accuracy of 98.88% on LFW [9, 8].
- We reveal that the deficiency of multinomial logistic regression in one-shot learning is related to the norms of the weight vectors in multinomial logistic regression, and propose a novel loss term called underrepresented-classes promotion (UP) which effectively addresses the data imbalance problem in one-shot learning.
- Our experimental results show that the proposed UP term significantly helps improve the recognition coverage rate from 25.65% to 77.48% at the precision of 99% for one-shot classes, while still keep an overall top-1 accuracy of 99.8% for normal classes.

2. Related Work

2.1. Low-shot learning for face recognition

We abstract face recognition into two steps. The first step is face feature extraction, and the second step is to estimate the person's identity from the extracted face feature. Recently, with the quick development of deep convolutional neural network, the major focus in face recognition has been to learn a good face feature space, in which faces of the same person are close to each other, and faces of different persons are far away from each other. There have been steady progresses in this direction [1, 17, 16, 18, 15, 19, 20]. Moreover, many benchmark tasks for face recognition also focus on getting good face features. For example, the verification task with the LFW dataset [9] has become the de facto standard test to evaluate face feature.

We observe less effort in estimating the person's identity from his/her face feature. Many face identification tasks, e.g., MegaFace [10] or LFW [9] with the identification setup, are typically based on the similarity comparison between the images in the gallery set and the query set, which is essentially a K-nearest-neighborhood (KNN) method to estimate the persons' identity. In the most ideal case, if we have a perfect face feature extractor (inter-class distance is always larger than the intra-class distance), KNN method is good enough to estimate the persons' identity. Unfortunately, no one has a perfect face feature extractor for now.

For the large scale face recognition problem, KNN (based on a reasonably good, yet not perfect feature extractor) might not be the best solution. If we use all the face images for every person in the gallery, the complexity is usually too high for large scale recognition, and the gallery dataset needs to be very clean to ensure high precision. If we don't keep all the images per person, how to construct representer for each class is still an open problem. One straight forward way is to use the average of face features for all the face images of the same person to represent this person. However, this method may lead to poor performance.

We choose to investigate the multinomial logistic regression (MLR) to estimate the persons' identity from his/her face feature. One of the advantages of MLR is that after feature extraction, the computing complexity of estimating the persons' identity is linear to the number of persons, not the number of images in the gallery. Another advantage of MLR is that the weight vectors for each classes are estimated using the information from all the classes, while in the KNN setup, the query image only needs to be close enough to one local class to be recognized. Though MLR needs some time to train the classifier for all the persons to be recognized while KNN could keep a general face feature extractor fixed, in many real scenarios, a few images of the persons in the novel set are naturally available, and it worths the time updating or retraining a model if better performance is offered. In the most recent and largest face recognition challenge, MS-Celeb-1M [5], the best performance is achieved by using the MLR setup [25, 24].

One of the major challenge of using the MLR classifier is that the MLR classifier trained with standard cross entropy loss does not perform well on the novel set. We have not seen a lot effort in this direction except boosting the number of samples in the novel set. Especially for the deep learning-based feature extractor, this is still an open area. This is our major focus in this paper: we propose an underrepresented classes promotion (UP) term to improve the classifier performance on the novel set. Details are provided in the next section.

In the general image recognition domain, the recent low-shot learning work [6] also attracts a lot of attentions by its good performance. In this work, the authors propose to first shrink the features (especially for the misclassified examples) to penalize the difference between classifiers learned on large and small datasets, and then generate "hallucinating" examples to further transfer the variations from the base set to the novel set. Compared with the work in [6], we focus more on the classifier part to handle the imbalanced data problem. This is mainly because for our task, a good face representation model (especially generalization capability) is learned from a sufficiently large scale dataset, base set, whereas for general visual recognition, the representation model from the base set is not generalizable enough. Experimental results in section 4 demonstrate that the feature improvement method in [6] only has very limited impact on the final results for the low-shot face recognition. Good face feature also enables us to aggressively focus on the one-shot learning setting rather than the low-shot learning setting in [6].

2.2. Dataset

Base set – One of our contributions in this paper is that we publish a large scale, near noise-free training dataset for general face representation learning. In the literature, several datasets have been published to facilitate the research in the area. We summarize examples in Table 1.

As shown in Table 1, our training dataset is considerably larger than the publicly available datasets except for the MS-Celeb-1M dataset [4]. Note that our dataset emphasizes on different aspects compared with MS-Celeb-1M. MS-Celeb-1M targets at recognizing as many as possible celebrities in the one-million celebrity list so the celebrity coverage is important, and the noisy label for the less popular celebrity is inevitable [25, 24, 23]. Therefore, MS-Celeb-1M inspires work including data cleaning, training with noisy-labels, etc. The base set provided with this paper is mainly to train a robust, generalizable face feature extractor and isolate the problem of one-shot learning from feature

Dataset	Available	people	images
IJB-A [11]	public	500	5712
LFW [9, 8]	public	5K	13K
YFD [22]	public	1595	3425 videos
CelebFaces [17]	public	10K	202K
CASIA-WebFace [3]	public	10K	500K
Our Base set	**public**	20**K**	about 1.2**M**
MS-Celeb-1M [4]	public	100K	about 10M
Facebook	private	4K	4400K
Google	private	8M	100-200M

Table 1: Face Datasets

learning. For these reasons, in contrast to MS-Celeb-1M, we have created a smaller yet nearly noise-free version .

Moreover, for the convenience of feature evaluation, we do not include the celebrities in LFW in our 20K dataset. Thus researchers can directly leverage this dataset and evaluate performance on the LFW verification task. In our experiments, With very minimal effort, we learned a good face feature with face verification accuracy of 98.88% on LFW [9, 8].

Novel set – We provide 20 images for 1000 persons to test the performance on the novel set. Some examples are shown in Figure 1 and Figure 7 . For comparison, the LFW dataset [9], which is the de facto standard, has less than 100 persons having more than 20 images. The benchmark task in MegaFace [10] focuses on 80 identities for the query set to be recognized, though millions of images provided as distractors.

Our benchmark task evaluate models with a large number of persons in order to include large variations in age, race, gender, professions, etc. Moreover, our benchmark task evaluate models with many images per person in order to include variations in expressions, lighting, poses, etc. This is to evaluate the model's generalization ability within each person.

3. Methodology

Our method includes the following two phases. The first phase is *representation learning*. In this phase, we build face representation model using all the training images from the *base set*.

The second phase is *one-shot learning with underrepresented-classes promotion (UP)*. In this phase, we train a multiclass classifier to recognize the persons in both *base set* and *novel set* based on the representation model learned in phase one. We design the underrepresented-classes promotion (UP) technology to improve the recognition performance for the persons in the novel set.

3.1. Representation learning

In the representation learning phase, we train a $20,000$-class classifier using all the training images of the $20,000$ persons in the base set. As we have described in section 1, there are about 50-100 images per person in the base set. The wrong labels in the base set is very limited (less than 1% based on manual check). We save 5 images per person for testing and use the rest of these images for training. We crop and align face areas to generate the training data, with some examples shown in Figure 6. We release the aligned face images as well as face detection results so that our experimental results are easily reproducible.

Our face representation model is learned from predicting the $20,000$ classes. More specifically, we consider each person as one class and train a deep convolutional neural network (ConvNet) supervised by the softmax with the cross-entropy loss. We have tried different network structures and adopted the standard residual network with 34 layers [7] due to its good trade-off between prediction accuracy and model complexity. Feature extracted from the last pooling layer is used as the face representation.

As shown in the Table 3, the resnet-34 model leads to a result comparable to the state-of-the-art performance on the LFW verification task [9, 8], which demonstrates the value of our base set in terms of face representation learning. As all the persons in LFW have been excluded from the base set, this result indicates a good generalization ability of the learned face feature, making it possible to decouple the classifier learning problem from the feature learning. We will explore more options for feature learning in our future work.

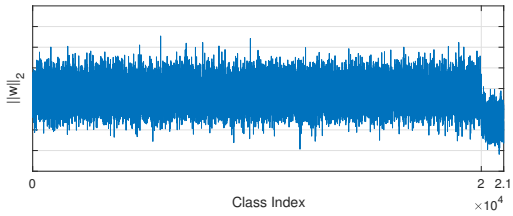
3.2. One-shot Learning with UP

In the one-shot learning phase, we train a $21,000$-class classifier using the training data from both the base set and the novel set, treating each person as one class. As we have discussed in the introduction section, in the novel set, there are $1,000$ persons (mutually exclusive from the base set). Each person in the novel set has only **one** image for training while 20 for testing.

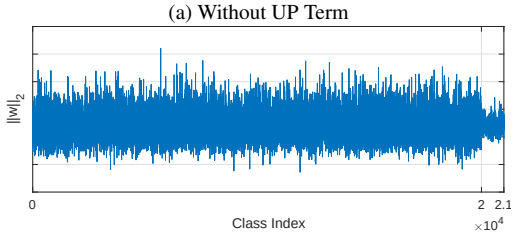
We build this multi-class classifier by using multinomial logistic regression based on the 34-layer residual network [7], which is the same network structure as the one we used in the feature learning phase. We first use the parameters of the network trained in phase one to initialize the network, and then further fine-tune the network in phase two.

3.2.1 Challenges of One-shot

We very briefly review the multinomial logistic regression with the standard cross entropy loss. The probability that



(a) Without UP Term



(b) With UP Term

Figure 2: Norm of the weight vector $\mathbf{w}$ with/without UP term in Eq. 6. The x-axis is the class index. The rightmost 1000 classes on the x-axis correspond to the persons in the novel set. As shown in the figure, without the UP term, $\|\mathbf{w}_k\|_2$ for the novel set is much smaller than that of the base set, while with the UP term, on average, $\|\mathbf{w}_k\|_2$ for the novel set tends to have similar values as that of the base set. This promotion introduces significant performance improvement, details presented in section 4.

the n^{th} sample x_n belongs to the k^{th} class is calculated as,

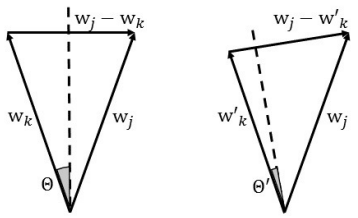
$$p_k(x_n) = \frac{\exp(\mathbf{w}_k^T \boldsymbol{\phi}(x_n))}{\sum_i \exp(\mathbf{w}_k^T \boldsymbol{\phi}(x_n))}\,, \tag{1}$$

where $\mathbf{w}_k$ is the weight vector for the k^{th} class, the subscript i is the class index, and $\boldsymbol{\phi}(\cdot)$ denotes the feature extractor for image x_n. Note that in all of our experiments, we always set the bias term $b_k = 0$. We conducted comprehensive experiments and empirically found that removing the bias term from the standard softmax layer in a convolutional deep neural network does not affect the performance. The cross entropy is used as the loss to guide the training.

$$\mathcal{L} = -\sum_n t_{k,n} \log p_k(x_n), \tag{2}$$

where $t_{k,n} \in \{0, 1\}$ is the ground truth label indicating whether x_n belongs to the k^{th} class.

Unfortunately, the loss function in Eq. 2 does not lead to a good performance for the persons in the novel set. As presented in section 4, for testing images in the novel set, the coverage at the precision of 99% is only 25.65%, while for testing images in the base set, the coverage is 100% at the precision of 99%. Moreover, in our experiments with the loss function Eq. 2, we found that the norms of the weight vectors for the novel classes are much smaller than the norms of the weight vectors for the base classes, with an example shown in Figure 2.



(a) $\|\mathbf{w}_k\|_2 = \|\mathbf{w}_j\|_2$ (b) $\|\mathbf{w}_k\|_2 < \|\mathbf{w}_j\|_2$

Figure 3: Relationship between the norm of $\mathbf{w}_k$ and the volume size of the partition for the k^{th} class. The dash line represents the hyper-plane (perpendicular to $\mathbf{w}_j - \mathbf{w}_k$) which separates the two adjacent classes. As shown, when the norm of $\mathbf{w}_k$ decreases, the k^{th} class tends to possess a smaller volume size in the feature space.

The low coverage for the novel classes is related to the small values of the norms of the weight vectors for the novel classes. Without loss of generality, we discuss the decision hyperplane between any two adjacent classes. We apply Eq. 1 to both the k^{th} class and the j^{th} class to determine the decision hyperplane between the two classes (note we don't have bias terms throughout our paper):

$$\frac{p_j(x)}{p_k(x)} = \frac{\exp(\mathbf{w}_j^T \boldsymbol{\phi}(x))}{\exp(\mathbf{w}_k^T \boldsymbol{\phi}(x))} = \exp[(\mathbf{w}_j - \mathbf{w}_k)^T \boldsymbol{\phi}(x)] \tag{3}$$

As shown in Figure 3, the hyperplane to separate two adjacent classes k and j is perpendicular to the vector $\mathbf{w}_j - \mathbf{w}_k$. When the norm of $\mathbf{w}_k$ gets decreased, this hyperplane is pushed towards the k^{th} class, and the volume for the k^{th} class also gets decreased. As this property holds for any two classes, we can clearly see the connection of the norm of a weight vector and the volume size of its corresponding partition space in the feature space.

Here we discuss the reason that the weight vectors for the novel classes have much smaller norms. If we generate a convex hull for the training samples of one class in the base set, typically, the volume of this convex hull is much larger than that of the convex hull of the samples for one class in the novel set. Furthermore, the weight vector $\mathbf{w}_k$ gets updated when

$$|p_k(x_n) - t_{k,n}| < \epsilon\,, \tag{4}$$

where ϵ is a very small positive number. This is because the gradient of Eq. 2 with respect to $\mathbf{w}_k$ is

$$\frac{\partial L}{\partial \mathbf{w}_k} = \sum_n (p_k(x_n) - t_{k,n}) \boldsymbol{\phi}(x_n)\,. \tag{5}$$

Therefore, with a larger convex hull claimed by more samples in the feature space, the base classes have larger chance

to update their weight vectors and tend to have larger weight vector norms to satisfy Eq. 4, compared with the novel classes.

3.2.2 Underrepresented Classes Promotion

In this sub-subsection, we propose a method to promote the underrepresented classes, a.k.a. the classes with limited number of (or only one) samples. Our method is based on a prior which we design to increase the volumes of the partitions corresponding to the novel classes in the feature space.

Based on the previous analysis, we introduce a new term to the loss function with the assumption that on average, the persons in the novel set and the persons in the base set should have similar volume sizes for their corresponding partitions in the feature space.

$$\mathcal{L}_{up} = \sum_n -t_{k,n} \log p_k(x_n) + \frac{1}{|C_n|} \sum_{k \in C_n} \| \|\mathbf{w}_k\|_2^2 - \alpha \|_2^2 \,, \tag{6}$$

where α is the average of the squared norms of weight vectors for the base classes,

$$\alpha = \frac{1}{|C_b|} \sum_{k \in C_b} \|\mathbf{w}_k\|_2^2. \tag{7}$$

We use C_b and C_n to denote the sets of the class indices for the *base set* and the *novel set*, respectively. As shown in Eq. 6, the average of the squared norms of the weight vectors in the *novel set* is promoted to the average of the squared norms of the weight vectors for the *base set*. We call this term underrepresented-classes promotion (**UP**) term.

For every mini-batch, we jointly optimize the cross entropy term and the UP loss term. The derivative we sent back for back propagation is the summation of the derivative of cross entropy and the derivative of the UP term. We keep the rest of the optimization the same as a regular deep convolutional neural network.

3.2.3 Alternative Methods

Adding extra terms of $\mathbf{w}_k$ to the cost function is essentially to inject prior knowledge to the system. Different assumptions or observations yield to different prior terms to the weighting vectors. Here we discuss several alternatives to the UP-prior.

One typical method to handle insufficient data problem for regression and classification problems is to shrink $\mathbf{w}_k$, [21, 2]. Here we choose the $L2$-norm option for optimization efficiency.

$$\mathcal{L}_{l2} = \sum_n -t_{k,n} \log p_k(x_n) + \sum_k \|\mathbf{w}_k\|_2^2 \,. \tag{8}$$

Another option is to encourage all the weight vectors to have similar or even the same norms. A similar idea has been proposed in [14] for the purpose of accelerating the training speed. We adopt the soft constraint on the squared norm of $\mathbf{w}$ here.

$$\mathcal{L}_{eq} = \sum_n -t_{k,n} \log p_k(x_n) + \sum_{k \in \{C_n \cup C_b\}} \| \|\mathbf{w}_k\|_2^2 - \beta \|_2^2 \,, \tag{9}$$

where

$$\beta = \frac{1}{|\{C_n \cup C_b\}|} \sum_{k \in \{C_n \cup C_b\}} \|\mathbf{w}_k\|_2^2. \tag{10}$$

Note the major difference between this cost function and the cost function in Eq. 6 is that, in Eq. 9, the values of the norms of all $\mathbf{w}_k$ get affected and pushed to the same value, while in Eq. 6, only the values of the norms of $\mathbf{w}_k$ for *novel set* classes get promoted. The performance of all these options is presented in Section 4.

4. Experimental Results

4.1. Toy Example

To better visualize and illustrate our idea, we construct a toy example based on the MNIST data set with a fixed feature representation. As shown in Figure 4 (a), we have ten classes (color-coded) in the feature space (each dot corresponds to one sample). In an ideal case, we have sufficient samples for each of the classes, as shown in (a).

Unfortunately, in many situations, we only have limited number of samples for certain classes. We simulate an extreme case, as shown in Figure 4 (b). For the left-most class, we only have one sample (colored in red). In this situation, the standard multinomial logistic regression will "ignore" this class, with the solution space shown in Figure 4 (c).

In order to address the above issue, we apply our underrepresented class promotion (UP) in Eq. 6 and get the result shown in Figure 4 (d). This is achieved by promoting the squared norm of the weight vector of the underrepresented class, as illustrated in Table 2.

4.2. One-shot Face Recognition

As we have described in section 3, we first train a general face representation model with the training images in the base set, and then train a multi-class classification model with the training images in both the base and novel sets. We list the experimental results in details in the following subsections.

4.2.1 Face Representation Learning

We use the LFW [9, 8] verification task to evaluate our face representation model trained using the training images in

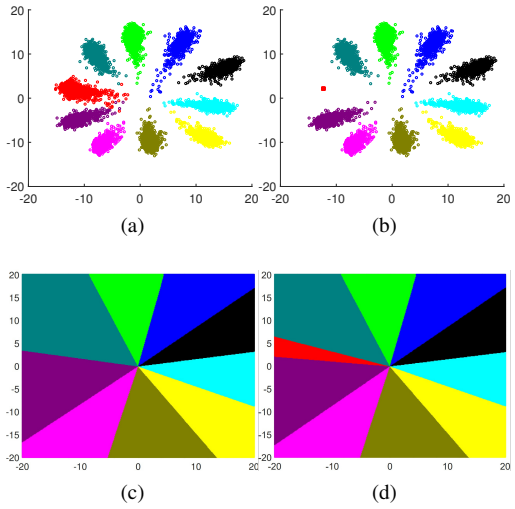


Figure 4: A toy example to illustrate our UP term, better viewed in color. (a) is an ideal case, where we have sufficient examples for all the classes. Each dot in the figure corresponds to a sample in the training dataset, while different colors represent different classes. (b) is the case we discuss in this paper, where for one class (leftmost, red), we have only one sample. (c) and (d) are the classification results by optimizing without/with our UP term in Eq. 6. As shown in (d), our UP term successfully claims a reasonable volume for the red class (leftmost) with only one example. Corresponding norms of the weight vectors are shown in Table 2.

	Without UP	With Up
$\frac{1}{9}\sum_{k\neq 9}\|\mathbf{w}_k\|_2$	2.29	2.36
$\|\mathbf{w}_k\|_2$, k = 9	1.75	**2.38**

Table 2: The norms of the weight vectors for the one-shot class and base classes with and without the underrepresented class promotion (UP). As shown in Figure 4 (b), there are 9 base classes, indexed from 0 to 8. The average of the norms of the weight vectors for these base classes is listed in the first row in this table. There is one underrepresented class (leftmost class in Figure 4) (b), indexed by 9. The norm of its weight vector effectively gets increased by the UP loss term.

the base set in phase one. The LFW verification task is the de facto standard for face feature evaluation. The task is to verify if a given face pair (in total 6000) belongs to the same person or not. The verification accuracy with different models are listed in Table 3.

As shown, with resnet-34 [7] and the training data in the base set, we can achieve a result comparable to the state of the art. Among the methods using single model and public data, we consider our model as achieving a cutting-edge performance. One advantage of our model is that we train our model using public data set and standard model structure, which makes our work easy to reproduce and flexible to extend for better performance. According to Table 3, we regard our model good enough to let us start to investigate the one-shot learning phase. We will leave the face feature representation improvement as our future work.

Methods	Dataset	Network	Accuracy
JB [1]	Public	–	96.33%
DeepID2,3 [16, 18]	Public	200	99.53%
FaceNet [15]	Private	1	99.63%
DeepFace[13]	Public	1	97.27%
Human	–	–	97.53%
Ours (AlexNet)	Public	1	96.97%
Ours (Resnet-34)	Public	1	98.88%

Table 3: Face Feature Evaluation with LFW verification.

4.2.2 One-shot Face Recognition

In phase two, we train a $21,000$-class classifier to recognize the persons in both the base set and the novel set. In the base set, there are $20,000$ persons, each of which having $50-100$ images for training and 5 for testing. In the novel set, there are 1000 persons, each having one image for training and 20 for testing. The experimental results in this paper were obtained with $100,000$ test images for the base set and $20,000$ test images for the novel set. In order to facilitate research in this direction, we release the labels for $20,000$ test images for the base set (out of the 100K base set test), and release labels for another $5,000$ test images for the novel set, to build a development set. We focus on the recognition performance in the novel set while monitoring the recognition performance in the base set to ensure that the performance improvement in the novel set does not harm the performance in the base set.

To recognize the test images for the persons in the novel set is a challenging task. The one training image per person was randomly preselected, and the selected image set includes images of low resolution, profile faces, and faces with occlusions. Some examples of the training images are shown in Figure 1 and Figure 7. As shown, the training images in the novel set show a large range of variations in gender, race, ethnicity, age, camera quality (or evening drawings), lighting, focus, pose, expressions, and many other parameters. Moreover, we applied de-duplication algorithms to ensure that the training image is visually different from the test images, and the test images can cover many different looks for a given person. Some examples can be seen in

Method	C@99%	C@99.9%
Fixed Feature	25.65%	0.89%
SGM [6]	27.23%	4.24%
Update Feature	26.09%	0.97%
Direct Train	15.25%	0.84%
Shrink Norm (Eq.8)	32.58%	2.11%
Equal Norm (Eq.9)	32.56%	5.18%
UP Term (Eq.6)	**77.48%**	**47.53%**

Table 4: Coverage at Precisions = 99% and 99.9% on the **novel set**. Please refer to subsection 4.2.2 for the detailed descriptions for all the methods. As shown in the table, our UP loss significantly improves the recall at precision 99% and 99.9% .

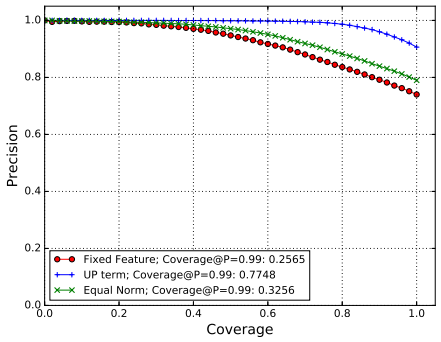


Figure 5: Precision-Coverage curves of different methods on the novel set with close-set identification protocol. We draw selected curves for better visualization. Please refer to Table 4 for detailed comparisons.

Figure 1, and more examples are shown in the supplementary materials.

The experimental results of our method and the alternative methods are listed in Table 4. We use coverage rate at precision 99% and 99.9% as our evaluation metrics since this is the major requirement for a real recognizer. The methods in the table are described as follows.

All the methods in Table 4 are based on a $21,000$-class classifier (trained with different methods). Note that we boost all the samples in the novel set for **100** times for all the methods , since the largest number of samples per person in the base set is about 100.

The "Fixed Feature" in Table 4 means that, in phase two, we do not update the feature extractor and only train the classifier in Eq. 2 with the feature extractor provided by phase one.

The SGM, known as squared gradient magnitude loss, is obtained by updating the feature extractor during phase one using the feature shrinking method as described in [6]. Compared with the "Fixed-Feature", SGM method introduces about 2% gain in recall when precision requirement is 99%, while 4% gain when precision requirement is 99.9%. The improvement for face recognition by feature shrinking in [6] is not as significant as that for general image. The reason might be that the face feature is already a good representation for faces and the representation learning is not a major bottleneck. Note that we did not apply the feature hallucinating method as proposed in [6] for fair comparison and to highlight the contribution of model learning, rather than data augmentation. To couple the feature hallucinating method (may need to be modified for face) is a good direction for the next step.

The "Update Feature" method in Table 4 means that we fine-tune the feature extractor simultaneously when we train the classifier in Eq. 2 in phase two. The feature updating does not change the recognizer's performance too much.

The rest three methods (shrink norm, equal norm, UP-method) in Table 4 are obtained by using the cost functions defined in Eq. 8, Eq. 9, and Eq. 6 as supervision signals for deep convolutional neural network in phase two, respectively, with the face feature updating option. As shown in the table, our UP term improves the coverage@precision=99% and coverage@precision=99.9% significantly.

The coverage at precision 99% on the base set obtained by using any classifier-based methods in Table 4 is 100%. The top-1 accuracy on the base set obtained by any of these classifier-based methods is $99.80 \pm 0.02\%$. Thus we do not report them separately in the table.

5. Conclusion and Future Work

In this paper, we have studied the problem of one-shot face recognition, by creating a benchmark dataset consisting of $20,000$ persons for face feature learning and $1,000$ persons for one-shot learning. We reveal that the deficiency of multinomial logistic regression in one-shot learning is related to the norms of the weight vectors in multinomial logistic regression, and propose a novel loss called underrepresented-classes promotion to effectively address the data imbalance problem in one-shot learning. The evaluation results on the benchmark dataset show that the new loss term brings a significant gain by improving the recognition coverage rate from 25.65% to 77.48% at the precision of 99% for one-shot classes, while still keeping an overall accuracy of 99.8% for normal classes.

In the future, we will continue this study for open domain face identification and see how the proposed UP term can help improve the recognition accuracy in more general scenarios. We are also interested in applying the UP prior on the ImageNet dataset under the same setting as in [6] and explore more options to improve low-shot learning in general visual recognition problems.

References

[1] *A Practical Transfer Learning Algorithm for Face Verification*. Proc. of Int'l Conf. on Computer Vision (ICCV), 2013.

[2] G. Cao, Y. Guo, and C. A. Bouman. High dimensional regression using the sparse matrix transform (SMT). In *Proc. of IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.*, pages 1870–1873, 2010.

[3] S. L. Dong Yi, Zhen Lei and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

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

[5] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: Challenge of recognizing one million celebrities in the real world. In *Electronic Imaging*, 2016.

[6] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. *arXiv preprint arXiv:1606.02819*, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[10] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. *ArXiv e-prints*, 2015.

[11] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. In *Science*, pages 1332–1338, 2015.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[14] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, 2016.

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, NIPS'14, pages 1988–1996. MIT Press, 2014.

[17] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[18] Y. Sun, X. Wang, and X. Tang. DeepID3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2014.

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2746–2754. IEEE, 2015.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[22] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[23] X. Wu, R. He, and Z. Sun. A lightened CNN for deep face representation. *CoRR*, abs/1511.02683, 2015.

[24] Y. Wu, J. Li, Y. Kong, and Y. Fu. Deep convolutional neural network with independent softmax for large scale face recognition. In *Proc. of ACM Int'l Conf. on Multimedia*, pages 1063–1067, 2016.

[25] H. Ye, W. Shao, H. Wang, J. Ma, L. Wang, Y. Zheng, and X. Xue. Face recognition via active annotation and learning. In *Proc. of ACM Int'l Conf. on Multimedia*, pages 1058–1062, 2016.

6. Image Visualization

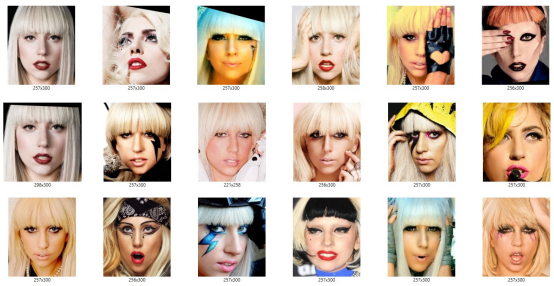


Figure 6: Example training images in the *base set* used for *representation learning*. These are a portion of the images we have for Lady Gaga, which covers a large diversity of her appearance. Note that we have cropped and aligned the face area to generate the training data.

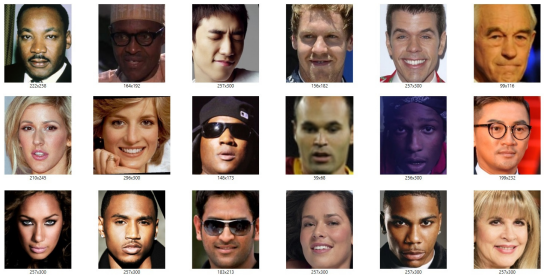


Figure 7: Example training images in the *novel set* used for *one-shot learning*. In this set, only **one** training image is provided for each person. Here we present 18 images for 18 different persons. As shown, there are large variations in gender, race, ethnicity, age, camera quality, lighting, focus, occlusion condition, pose, expressions, and many other aspects.