Know You at One Glance: A Compact Vector Representation for Low-Shot Learning

Yu Cheng*[1] Jian Zhao*[2] Zhecan Wang[3] Yan Xu[1] Karlekar Jayashree [1] Shengmei Shen [1] Jiashi Feng [2]
[1] Panasonic R&D Center Singapore [2] National University of Singapore [3] Franklin. W. Olin College of Engineering
CHEN0974@e.ntu.edu.sg zhaojian90@u.nus.edu zhecan.wang@students.olin.edu
{yan.xu, karlekar.jayashree, shengmei.shen}@sg.panasonic.com elefjia@nus.edu.sg

Abstract

*Low-shot face recognition is a very challenging yet important problem in computer vision. The feature representation of the gallery face sample is one key component in this problem. To this end, we propose an Enforced Softmax optimization approach built upon **C**onvolutional **N**eural **N**etworks (CNNs) to produce an effective and compact vector representation. The learned feature representation is very helpful to overcome the underlying multi-modality variations and remain the primary key features as close to the mean face of the identity as possible in the high-dimensional feature space, thus making the gallery basis more robust under various conditions, and improving the overall performance for low-shot learning. In particular, we sequentially leverage optimal dropout, selective attenuation, ℓ_2 normalization, and model-level optimization to enhance the standard Softmax objective function for to produce a more compact vectorized representation for low-shot learning. Comprehensive evaluations on the MNIST, **L**abeled **F**aces in the **W**ild (LFW), and the challenging MS-Celeb-1M Low-Shot Learning Face Recognition benchmark datasets clearly demonstrate the superiority of our proposed method over state-of-the-arts. By further introducing a heuristic voting strategy for robust multi-view combination, and our proposed method has won the Top-1 place in the MS-Celeb-1M Low-Shot Learning Challenge.*

1. Introduction

Recently, deep learning techniques have made great breakthroughs in many area both academically and industrially. In computer vision, advances of deep learning approaches have remarkably boosted the performance of face recognition. Several approaches claim to have achieved [21, 18] or even surpassed [14, 19] human performance on several benchmarks. The simple yet powerful structure of **C**onvolutional **N**eural **N**etworks (CNNs) is able to efficiently learn useful information from real-world images and capture the intrinsic connections beneath the big data. Recent works on CNNs mainly focus on network architectures [20, 6], non-linear activations[1, 23, 11], and objective function optimization [15, 22, 16].

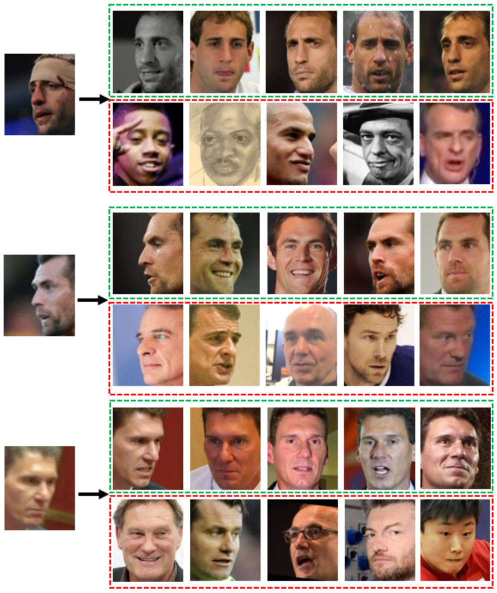

Figure 1: Matched (bounded with green boxes) and non-matched (bounded with red boxes) cases in the challenging low-shot learning problem. The left most column represents the gallery data with only 1 face image of each identity available for training, while the right 5 columns represent the probe (query) face images captured in different conditions (poses, illumination, resolution, *etc.*) with those of the gallery. Our model consistently give the correct recognition results for all challenging scenarios (faces with occlusion, drawings, and low-resolution). More details are presented in Sec. 4.3.

Theoretically, the CNNs can be regarded as a manifold

* indicate equal contributions. Yu Cheng, Jian Zhao, and Zhecan Wang were interns at Panasonic R&D Center Singapore during this work. Jian Zhao is the corresponding author.

learning scheme, which is able to approximate any function with a certain small error [2]. Therefore, CNNs are able to learn an ideal projection from the training data representing the real-world data distribution to an abstract vector space depending on the specific objective function. For instance, a deep model will learn a ℓ_2 distribution from distance-based objective function [16, 17] and a polar distribution based on matrix multiplication guidance, *e.g.*, Softmax Cross-Entropy scheme [23].

However, in practice, it is considerably difficult to develop a perfect model due to unavoidable limitations from modern algorithms (*e.g.*, gradient vanishing in back-propagation, and over fitting) and those from hardware (*e.g.*, memory and computational consumption for huge models). Thus, the CNNs usually learn a rough approximation instead where many samples cannot be clearly classified. Such hard cases of traditional methods usually result from robustless sparse representations with sensitivity to diverse variations, which becomes a huge barrier for the low-shot learning problem. This is very challenging for face recognition in the wild as the available training data for each identity is limited, as shown in Figure 1. Different from human beings who can effortlessly recognize the identities from very few face images by accurately capturing the intrinsic features, the data-driven deep learning algorithms still requires quite a few training data to achieve satisfactory performance. This problem becomes even exacerbated if the gallery face images is captured under extreme conditions such as poses, illumination, and resolution. Therefore, it becomes necessary for CNNs to learn a compact vector representation to minimize the effect from disturbance.

In this paper, we propose an enforcing scheme for the standard Softmax objective function by narrowing the decision boundaries of each class in order to guide the model to produce a more compact vector representation for effectively solving the challenging low-shot learning problem on face recognition. Through the compact vector representation learning, the deep features belonged to the same class are located closer to each other while those belonged to different classes are separated clearer. Such compactness results in less variation on irrelevant variables from facial poses, illumination, resolution, *etc*. Therefore, it becomes a good choice for low-shot learning problem due to the increased compactness and robustness of the gallery basis and query encoding. In particular, we sequentially leverage an optimal dropout scheme to overcome the gradient vanishing and over-fitting problems, a selective attenuation scheme for the last fully connected layer of CNNs to compact the intra-class distance and sparse the inter-class distance, a ℓ_2 normalization operation to balance the decision boundaries of majority and minority classes, and a model-level optimization to learn better feature vector distribution and dense multi-class clusters.

Comprehensive evaluations on the MNIST [12], **L**abeled **F**aces in the **W**ild (LFW) [8], and the challenging MS-Celeb-1M [4] Low-Shot Learning Face Recognition benchmark datasets clearly demonstrate the superiority of our proposed method over state-of-the-arts.

Moreover, we further propose a heuristic voting strategy for robust multi-view combination, and our proposed method has won the Top-1 place in the MS-Celeb-1M Low-Shot Learning Challenge[1].

Our contributions are summarized as follows.

- We propose a novel enforcing scheme for the standard Softmax objective function by narrowing the decision boundaries of each class in order to guide the model to produce a more compact vector representation for effectively solving the challenging low-shot learning problem on face recognition.
- We introduce an optimal dropout scheme to overcome the over-fitting problem; a selective attenuation scheme for the last fully connected layer of CNNs to compact the intra-class distance and sparse the inter-class distance; a ℓ_2 normalization operation to balance the decision boundaries of majority and minority classes, and a model-level optimization to learn better feature vector distribution and dense multi-class clusters.
- Comprehensive evaluations on the MNIST, LFW, and the challenging MS-Celeb-1M Low-Shot Learning Face Recognition benchmark datasets clearly demonstrate the superiority of our proposed method over state-of-the-arts. Moreover, we further propose a heuristic voting strategy for robust multi-view combination, and our proposed method has won the Top-1 place in the MS-Celeb-1M Low-Shot Learning Challenge.

2. Related Works

Recently, several works have been devoted on optimization of CNN objective functions to achieve better performance on specific tasks. Center loss [22] provides an approach to cluster the samples both in the perspective of angle and ℓ_2 distance by adding a ℓ_2 regularization term. With this approach, the learned deep features are more compact in the $\mathbb{R}^n$ space. Large Margin Softmax [15] proposes a loss function which narrows the area for each class. This loss function is able to compact the deep features of each class. However, similar to center loss [22], it is unable to untangle the unbalanced data problem in the low-shot learning problem.

For low-shot learning, due to the extremely limited training samples for Novel set identities, most methods are focusing on developing effective algorithms for feature expansion. In [5], the authors propose a feature generating approach by

[1] http://www.msceleb.org/leaderboard/c2

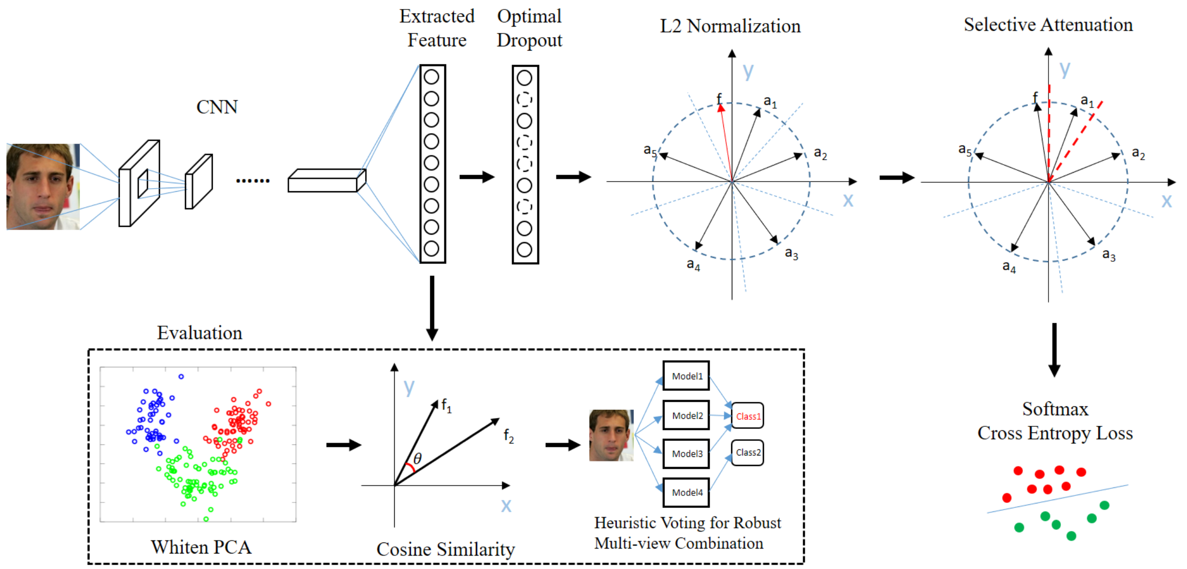

Figure 2: Enforcing scheme overview. The proposed Enforced Softmax scheme contains 4 main components: optimal dropout, selective attenuation, ℓ_2 normalization, and model-level optimization. Through this scheme, a model is enabled to learn a compact vector representation, which is beneficial for solving the low-shot learning problem on face recognition. Best viewed in color.

training a transferring **M**ulti-**L**ayer **P**erceptron (MLP) which takes the input of (z_1, z_2, x) and outputs x' by applying the $z_1 \rightarrow z_2$ transformation to x. This generating scheme is quite helpful to expand the Novel set. However, the generator is trained based on the assumption that the nearest two pairs have the same variation, which does not always hold in real-world scenario. This might potentially lead to wrong generation and could be a fetal error for certain classes, as the gallery vectors are wrongly biased.

The other trend of approaches to solve the low-shot learning problem is to generalize the decision boundary of each class. In [3], the authors proposed an **U**nderrepresented-classes **P**romotion (UP) term to achieve reasonable decision boundaries for identities of the Novel set during training. However, merely balancing the decision area is not enough for low-shot learning since there still exist quite a lot hard cases which could be easily misclassified with a high confidence score.

In this paper, we propose an enforcing scheme to compact the feature vector representation by sequentially leveraging optimal dropout, selective attenuation, ℓ_2 normalization and model-level optimization to enhance the standard Softmax objective function for guiding the model to produce a more compact vectorized representation. This method is effective for solving low-shot learning problems by normalizing their decision boundaries and separate us well with other related works.

3. Enforced Softmax

Our proposed Enforced Softmax scheme is able to 1) learn discriminative yet generative compact vector representations, and 2) boost the low-shot learning face recognition performance in presence of large multi-modality variances. As shown in Figure 2, the Enforced Softmax scheme sequentially leverages optimal dropout, selective attenuation, ℓ_2 normalization and model-level optimization to enhance the standard Softmax objective function for guiding the model to produce a more compact vectorized representation. This method is effective for solving low-shot learning problems by normalizing their decision boundaries. We now present each component in detail.

3.1. Compact Vector Representation Learning

Optimal Dropout A CNN model deployed for face recognition using feature retrieval strategy can be regarded as a feature extractor:

$$v(\vec{x}) = f(x), \tag{1}$$

where x denotes the network input (RGB face image), $f(\cdot)$ denotes the non-linear encoding function learned by a CNN model, $v(\vec{x})$ denotes the learned feature vector.

For any manifold, there exists an intrinsic dimensionality that conforms to the following definition, as firstly introduced in [24]:

Definition 3.1. Manifold A subset $\mathcal{M} \subset R^d$ is called a p-smooth $(p > 0)$ manifold with intrinsic dimensionality $m = m(\mathcal{M})$, if there exists a constant $c_P(\mathcal{M})$ such that

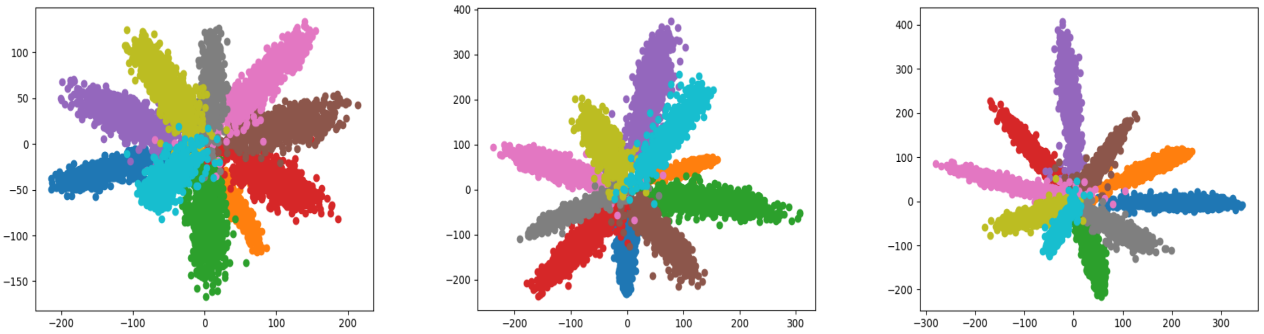

Figure 3: Visualized comparison of the learned feature vectors with the proposed enforcing scheme using different attenuation factors on MNIST [12]. The standard Softmax with attenuation factor of 1.0 (left) results in more sparse vector representation compared with those with attenuation factor of 0.9 (middle) and 0.7 (right). The feature vectors become more compact with the decrease of the attenuation factors. Best viewed in color.

given any $x \in \mathcal{M}$, there exists m vectors $v_1(x), ..., v_m(x) \in \mathbb{R}^d$, $\forall x' \in \mathcal{M} : inf_{\gamma \in \mathbb{R}^m} \left|\left| x' - x - \sum_{j=1}^{m} \gamma_j v_j(x) \right|\right| \leq c_P(\mathcal{M}) ||x' - x||^{1+p}$, where γ is a map of $x \in \mathbb{R}^d$ to $[\gamma_{v \in C}] \in \mathbb{R}_C$ such that $\sum_v \gamma_v(x) = 1$, C is a set of anchor points in d-dimensional space $\mathbb{R}^d$.

Thus, an approximation can be computed by maximum likelihood estimation [13]:

$$\begin{aligned} \hat{m_k}(X_i) =& [\frac{1}{k-1} \sum_{j=1}^{k-1} log(\frac{T_k(X_i)}{T_j(X_i)})]^{-1}, \\ \hat{m_k} =& \frac{1}{N} \sum_{j=1}^{k-1} \hat{m_k}(X_i). \end{aligned} \tag{2}$$

According to Def. 3.1 and Eqn. (2), qualitatively, a dataset intrinsically has a lower bound of feature dimensionality m to fully express all the data samples, and for face recognition or generic object classification, the required dimensionality will be lower since some information will not contribute to the overall performance.

Hence, based on optimal intrinsic dimensionality assumption, we propose a drop-out scheme for training CNN models. Since we aim to avoid the gradient vanishing and over-fitting problem while minimizing the information loss, the dropped vector is supposed to contain the same essential information as the compact counterparts do. Thus, the optimal dropout rate should reach its maximum while not hurting the accuracy. The dropout of the feature layer can be regarded as a layer assembling operation, which means we assemble different parameterized layers in a dropout style. This operation can efficiently improve the robustness of learned feature vectors. Moreover, it is also helpful on choosing the best **P**rincipal **C**omponent **A**nalysis (PCA) strategy, since the intrinsic information in different dimensions is highly correlated after an optimal dropout.

Selective Attenuation We then define every column vector of the weights of the last fully connected layer as an “anchor vector“ which represents the center of each class. Therefore, the decision boundary can be derived when two anchor vectors give the same prediction.

$$\vec{p} \cdot \vec{a_1} = \vec{p} \cdot \vec{a_2}, \tag{3}$$

where $\vec{a_1}$ and $\vec{a_2}$ denote the anchor vectors and $\vec{p}$ denotes the feature vector.

However, in such cases, the samples located closed to the decision boundary can be wrongly classified with a high probability. A simple yet effective solution is to compact the intra-class distance and sparse the inter-class distance of the feature vectors, through which the hard samples will be adjusted and located in the correct decision area.

Therefore, we propose to impose the following regularization term to the standard Softmax Cross-Entropy loss to optimize the whole network and learn the relevant parameters:

$$S_t := aS_t, \tag{4}$$

where a denotes the attenuation factor (margin) to control the intra-class distance.

Note that Eqn. (4) is only applied to the confidence scores of genuine samples. Since we have made modifications of the standard Softmax loss to a more complicated objective function where genuine samples are treated differently from imposter ones, the target manifold is changed to a more complex shape which is expected to contain more paddle points and local minimas. Optimizers will probably face difficulties on convergence when facing large number of classes. Thus, during the training process, the attenuation factor is firstly set to 1.0, so that the model is only trained with the standard Softmax loss. After convergence, the value is decreased step-by-step in order to gradually shift the manifold from the standard Softmax to our target enforced version.

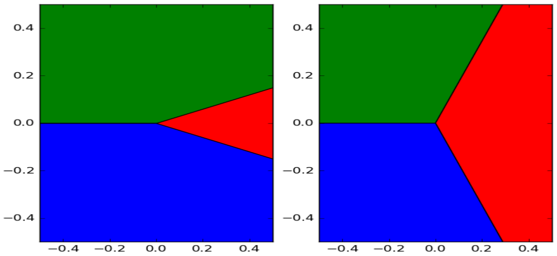

Figure 4: Illustration of decision boundaries before and after ℓ_2 normalization. The minor classes occupy a very small decision area due to the small number of training data (left). A more reasonable decision area can be generated after ℓ_2 normalization (right).

ℓ_2 Normalization For evaluation, we choose cosine similarity as the confidence score measurement to make it as simple as possible to advance the time efficiency of the proposed approach,

$$Scr = \frac{\vec{a} \cdot \vec{b}}{||a||\,||b||}. \tag{5}$$

Note that the final confidence score is relevant to the inner product of two normalized vectors, which is different from what we did during the training process. The decision boundaries during evaluation is identical for each feature vector, but different during training due to various of anchor vectors, as described in Eq. (3). For unbalanced datasets, especially the low-shot dataset, where identities of the Novel set are provided with very few training images, the decision areas for minorities will be compressed to extremely small, as shown in Figure 4. Thus, the minor classes make no contribution to training, and the decision areas for major classes will become sparse. Therefore, we propose a ℓ_2 normalization pipeline as the substitution of the last fully connected layer in the standard CNN model. The standard Softmax loss is then changed to Eqn. (6) and the normalization pipeline is illustrated in Figure 5.

$$P_i = \frac{e^{\hat{f} \cdot \hat{a_i}}}{\sum_j e^{\hat{f} \cdot \hat{a_j}}}. \tag{6}$$

In this normalization pipeline, we can alternatively choose to enable either one of the normalization operations or both. Since normalizations on features does not bring any changes to decision boundary, we focus on the normalization operation on anchor vectors (the weight matrix), as it is expected to give a more uniform feature distribution. For consistence with the cosine evaluation as in Eqn. (5), we normalize both anchor vectors and feature during training.

However, considering an extreme condition that every anchor vector is perpendicular to each other after convergence, the prediction score after Softmax will become:

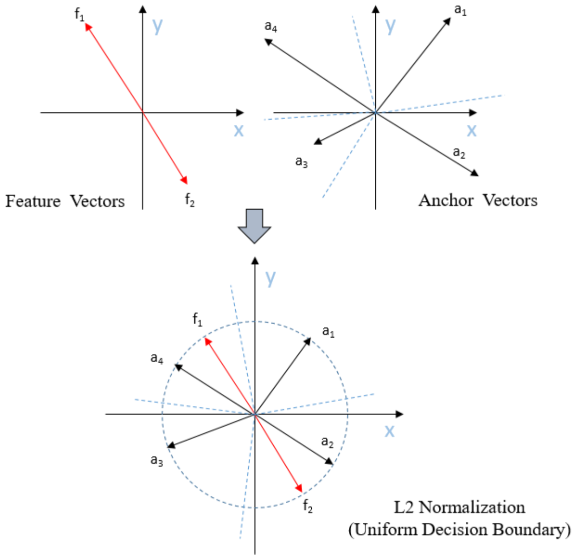

Figure 5: Illustration of ℓ_2 normalization scheme for the last fully connected layer. It is deployed to normalize the decision boundaries of different classes.

$$P = \begin{cases} \frac{e}{e+\sum_{n-1} 1} & \text{(Genuine sample)}, \\ \frac{1}{e+\sum_{n-1} 1} & \text{(Imposter sample)}, \end{cases} \tag{7}$$

where n denotes the number of total classes.

When n is large, it is unreasonable that there still exists a very large gradient from back-propagation after convergence. In order to address this issue, we multiply a scaling factor s before Softmax to modify the predicted scores of genuine samples:

$$P = \frac{e^s}{e^s + \sum_{n-1} 1}. \tag{8}$$

Thus, our final normalized Softmax objective function is:

$$P_i = \frac{e^{s\hat{f} \cdot \hat{a_i}}}{\sum_j e^{s\hat{f} \cdot \hat{a_j}}}. \tag{9}$$

Model-Level Optimization As discussed in previous sections, since the Softmax loss guides the feature vector to a weighted cosine clustering distribution, it is necessary to modify the structure of the feature layer of the CNN model.

CNNs usually contain some activation functions directly or indirectly connected with the feature layer. Modern CNN models mostly adopt **Re**ctified **L**inear **U**nits (ReLU)+Average pooling scheme. The feature vectors are restrained in the $\mathbb{R}^{+n}$ space and all negative activations are canceled by force. This cancellation leads to a gradient path for anchor vectors to distribute on the negative areas and form obtuse angles between feature vectors. Thus, it is necessary to remove the ReLU activation to facilitate the anchor vectors converging to the centers of classes by spreading the feature vectors over the total space. As the feature vectors are concentrated to the centers of classes under the Enforced Softmax, the spreading of feature vectors would enable CNN models with stable and fast convergence and help improve the overall performance on low-shot learning problems.

Moreover, we further propose to change the average pooling layer to a 7×7 depthwise convolution layer as it is expected to learn a better pooling scheme. We also remove the activation function of the feature layer for a better convergence.

3.2. Heuristic Voting for Robust Multi-View Combination

A single model is difficult to produce satisfactory results due to the fact that hard cases appear frequently in low-shot learning problems. We further propose a heuristic voting strategy at the score level for robust multi-view combination of multiple CNN models.

We first normalize all confidence scores of each trained model to 0-1, and we assign the model with the best performance as the main model while we regard the others as auxiliary models. Then, we sum up the confidence scores of the auxiliary models to that of the main model if they have the same prediction. We then divide the total testing data into multiple splits, sorted by the more reliable order of confidence scores. Then, we take the lowest group of our predictions as the hard cases as they are laying on the edge of decision boundaries of each model. Hard cases are mostly the face images from the identities whose gallery are quite different from the query during testing. Thus, we replace the predictions in the last split with the results from the gallery to achieve better performance on identities of the Novel set, and add the score by 1.0 because we strongly believe that the hard cases are face images from identities of the Novel set. The overall pipeline is illustrated in Figure 6. This heuristic voting strategy effectively and efficiently improves the overall performance via this hard case mining process.

4. Experiments

We verify the effectiveness of the proposed method on MNIST [12], LFW [8], and MS-Celeb-1M [4] low-shot learning face recognition benchmark datasets.

4.1. MNIST

We first train a toy model on the MNIST dataset [12], to evaluate the compactness of the learned feature vectors. The details of the network architecture is provided in Table 1. The model is trained to distinguish 10 different handwritten digits with various attenuation factors, and evaluated with the inter-class and intra-class cosine similarity based on Eqn. (5). Results are visualized in Figure 3 and listed in Table 2.

With the decreasing of the attenuation factor, the intra-class cosine similarities become larger and the inter-class similarities become smaller. Thus, the model has learned a more compact vector representation with the proposed Enforced Softmax strategy, which is significantly helpful to achieve better performance in feature retrieval with cosine similarity evaluation.

Operation	Kernel/Padding	Output	Activation
Convolution	$5 \times 5/3$	32	ReLU
Convolution	$5 \times 5/2$	64	ReLU
Convolution	$3 \times 3/1$	128	ReLU
Fully Connected	N/A	2	None

Table 1: Toy model architecture for MNIST [12].

Attenuation Factor	Inter-Class	Intra-Class
1.0 (Standard Softmax)	0.054	0.864
0.9	0.11	0.90
0.8	0.107	0.927
0.7	0.132	0.943

Table 2: Comparison of cosine similarity for inter-class and intra-class with different attenuation factors on MNIST [12]. The increase of intra-class similarity ensures more compact feature vector representation.

4.2. LFW

In order to measure the reliability (*i.e.*, the model should give a high recall at low fall-out) of the learned feature vectors, we then perform evaluation on the Open Testing set of LFW [8]. The Open set is composed of 596 gallery face images and $10,090$ probe face images, where $9,494$ of them are distractors. The evaluation protocols are determined by **T**rue **P**ositive **R**ate (TPR)@**F**alse **P**ositive **R**ate (FPR)=0.01 and Top-1 accuracy.

We conducted several experiments with different network architectures to study the effect of our proposed enforcing scheme on face recognition task. The experiments are performed based on 3 different magnitudes of training data and 3 different network architectures. All results consistently show that our proposed enforcing strategy gives better performance compared with the standard Softmax Cross-Entropy loss, as shown in Table 3.

4.3. MS-Celeb-1M Low-Shot Learning

The MS-Celeb-1M [4] Low-Shot Learning Challenge provides a new benchmark dataset for face recognition, which contains two subsets: Base set and Novel set. Base set contains 20k identities with 50 to 70 images for each. Novel set contains 1k identities with only 1 image for each. The evaluation protocol is determined by Coverage@Precision=0.99 on Novel set. During evaluation, the predictions are sorted by the confidence scores predicted by deep models in descending order. Then we take the Top-x images, where 99% of them are correctly recognized, to compute the coverage: x divided by the number of all testing images. For validity

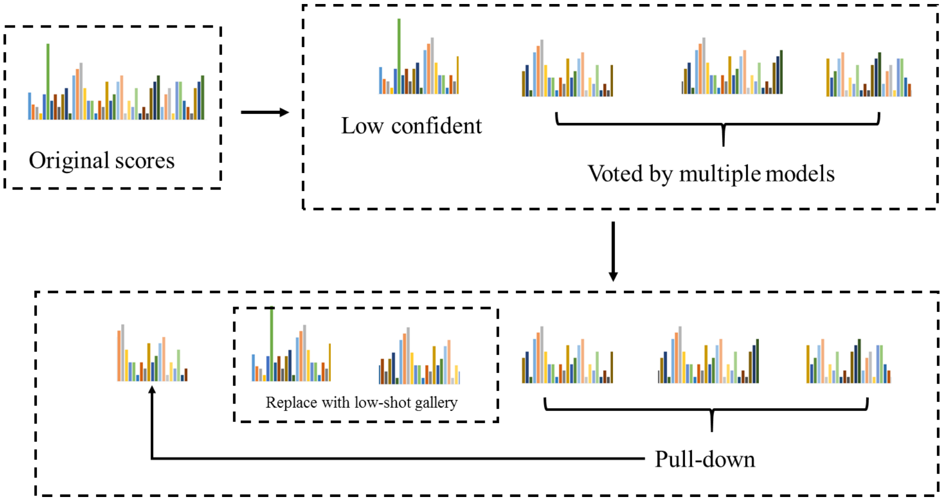

Figure 6: Heuristic voting for robust multi-view combination overview. We first redistribute the confidence scores by the number of deep models. Then, we replace the low confidence scores with the results from the 1k gallery comparison. We then perform a pull-down operation to the samples where the predictions of the 21k gallery are different from those of the 1k gallery. Such samples are regarded as the Base set data.

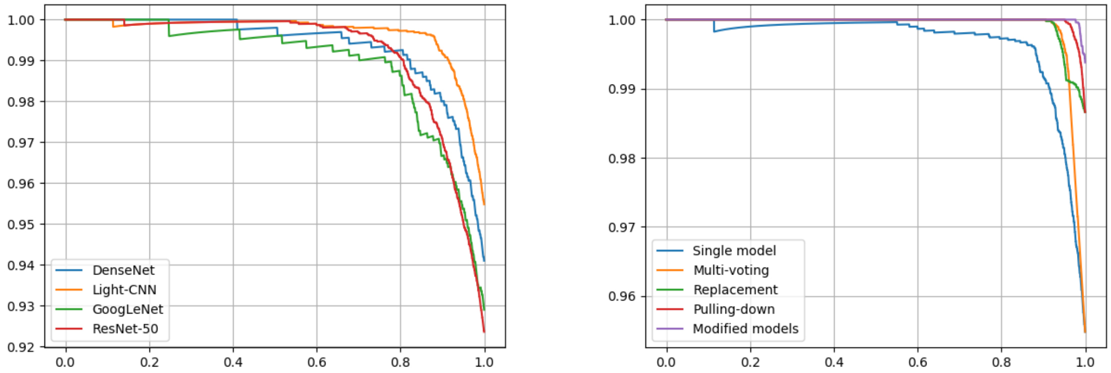

Figure 7: Coverage@Precision curves on MS-Celeb-1M Low-Shot Learning Challenge Development set. The performance of separate model (left) is satisfactory on Development set but varies dramatically due to different network architectures (views). The proposed heuristic voting strategy combines the strengths of each model with compensation to each other. Thus, it provides better performance (right). Best viewed in color.

Network	Training ID	TPR@FPR=0.01	Top-1	Scheme
ResNet-18 [6]	10k	0.201	-	None
ResNet-18 [6]	10k	0.343	-	At
ResNet-50 [6]	10k	0.315	-	None
ResNet-50 [6]	10k	**0.437**	-	At
ResNet-50 [6]	50k	0.686	-	None
ResNet-50 [6]	50k	0.758	-	At
ResNet-50 [6]	50k	0.857	-	L2
ResNet-50 [6]	50k	**0.896**	0.9908	At+L2
ResNet-50 [6]	80k	0.824	-	At
ResNet-50 [6]	80k	**0.903**	0.9913	At+L2
Light-CNN-29 [23]	80k	0.889	-	None
Light-CNN-29 [23]	80k	**0.929**	0.993	At
Light-CNN-29 [23]	80k	0.917	-	At+L2

Table 3: Comparison of TPR@FPR=0.01 and Top-1 accuracy on LFW [8] Open set verification protocol. At and ℓ_2 denote selective attenuation and ℓ_2 normalization, respectively.

of the testing results, the performance of Top-1 accuracy on Base set is also monitored and required to be better than 99%.

Since the magnitude of the Base set is not enough for training a well generalized CNN model as the feature extractor for effectively solving the low-shot learning problem on face recognition in the wild, we construct an extended training dataset using the face images of the identities from MS-Celeb-1M Challenge 1[2]. In particular, we filtered out the identities overlapped with the 1k identities of the Novel set according to the MIDs, and we only keep the identities with more than 20 face images to construct a well balanced training dataset. The final constructed training dataset contains 80k identities in total where 60k of them are from our extension and the other 20k are from the original Base set. We also build an additional Testing set using the data from challenge 1 as the Validation set in our early work to tune the hyper parameters. The Validation set comprises 5 images

[2] http://www.msceleb.org/celeb1m/1m

for each identity from the Base set and 20 images for each identity from the Novel set.

We trained 4 models with different architectures and then predict the final recognition results using our proposed heuristic voting strategy for robust multi-view combination. The 4 models are Light-CNN-29 [23], DenseNet [7], ResNet-50 [6] and GoogLeNet [20] with Bach Normalization [10], respectively. For evaluation, we random crop and flip the face images of the 1k identities of the Novel set up to 42 times and compute the corresponding mean encoding to construct a more generalised basis feature vector representation for gallery retrieval.

We first evaluate the performance of separate model, as shown in Table 4 and Figure 7 (left). With the proposed enforcing scheme, all models achieve satisfactory performance and the best model achieves 91.5% when Coverage@precision=0.99, outperforming the state-of-the-art by 14.02%.

We then evaluate the performance of combined models, as shown in Table 5 and Figure 7 (right). Firstly, the feature vectors are compressed by Whiten PCA to our optimal dimensionality (*i.e.* 512 by cross-validation). Then we perform the proposed heuristic voting strategy step-by-step. For MS-Celeb-1M Challenge 2, as the evaluation protocol separately tests the Top-1 accuracy on the Base set and Coverage@Precision=0.99 on the Novel set, it is necessary to divide the testing data into the Base set data and the Novel set data. Motivated by this, we "pull down" the confidence score where the current result is different from the prediction of the 1k gallery. Hereby we have divided the whole Testing set into 6 splits, where the Base set images are separated from Novel set images to the negative side. This hard case mining strategy is proved to be a good solution for unbalanced data.

As can be seen, every step gives an improvement and finally we achieve 99.56% when Coverage@precision=0.99. After applying Joint Bayes [9] metric learning strategy to DenseNet and GoogLeNet by cross-validation, we finally achieve 100% on the Development set (outperforming the state-of-the-art by 22.52%) and 99.01% on the Testing set when Coverage@precision=0.99 while keeping 99.74% Top-1 accuracy on the Base set, which significantly outperforms other state-of-the-arts. This results have won the Top-1 place in the MS-Celeb-1M Low-Shot Learning Challenge (Challenge 2). Please refer to MS-Celeb-1M Low-Shot Learning Challenge official leaderboard[3] for fully detailed results.

Model	Validation	Development
Guo *et al.* [3]	-	77.48%
GoogLeNet [20]	75.20%	76.34%
ResNet-50 [6]	80.30%	80.74%
DenseNet [7]	82.00%	82.30%
Light-CNN-29 [23]	90.30%	91.50%

Table 4: Comparison of Coverage@Precision=0.99 of separate model on MS-Celeb-1M [4] Low-Shot Learning Challenge Validation and Development set.

Step	Validation	Development	Testing
Guo *et al.* [3]	-	77.48%	-
Original	91%	91.5%	-
Voting	95.7%	96.26%	-
Replacement	97.84%	98.34%	-
Pulling Down	99.1%	99.56%	98.57%
Optimal Combination	**99.63**%	**100**%	**99.01**%

Table 5: Comparison of Coverage@Precision=0.99 of each combination step on MS-Celeb-1M [4] Low-Shot Learning Challenge Validation, Development and Testing sets. The proposed heuristic voting strategy significantly improves the overall performance from 91.5% to 99.56%. With the DenseNet and GoogLeNet incorporated with Joint Bayes metric learning strategy, our optimal combination performance (highlighted in bold) achieves 100% on the Development set and 99.01% on the Testing set.

5. Conclusion

In this paper, we propose a novel enforcing scheme for the standard Softmax objective function by narrowing the decision boundaries of each class in order to guide the model to produce a more compact vector representation for effectively solving the challenging low-shot learning problem on face recognition. Our framework can be easily extended to other generic object classification tasks. Comprehensive evaluations on the MNIST, LFW, and the challenging MS-Celeb-1M Low-Shot Learning Face Recognition benchmark datasets clearly demonstrate the superiority of our proposed method over state-of-the-arts. Moreover, we further propose a heuristic voting strategy for robust multi-view combination, and our proposed method has won the Top-1 place in the MS-Celeb-1M Low-Shot Learning Challenge.

Acknowledgement

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248, and Shenzhen Qianhai DeepvisionX Technology Co.,Ltd, China.

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

We would like to thank Junliang Xing (Institute of Automation, Chinese Academy of Sciences), Lin Xiong (Panasonic), Sugiri Pranata (Panasonic), Yongping Zhai (National University of Defense Technology) for helpful discussions.

[3] http://www.msceleb.org/leaderboard/c2

References

[1] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[2] B. C. Csáji. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48, 2001.

[3] Y. Guo and L. Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017.

[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 *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

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

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[8] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[9] S. Ioffe. Probabilistic linear discriminant analysis. *Computer Vision–ECCV 2006*, pages 531–542, 2006.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[11] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *arXiv preprint arXiv:1706.02515*, 2017.

[12] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[13] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.

[14] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[15] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516, 2016.

[16] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[18] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[19] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[20] 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.

[21] 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.

[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[23] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015.

[24] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *Advances in neural information processing systems*, pages 2223–2231, 2009.