High Performance Large Scale Face Recognition with Multi-Cognition Softmax and Feature Retrieval

Yan Xu*[1] Yu Cheng*[1] Jian Zhao[2] Zhecan Wang[3] Lin Xiong[1] Karlekar Jayashree[1]
Hajime Tamura[4] Tomoyuki Kagaya[4] Sugiri Pranata[1] Shengmei Shen[1] Jiashi Feng[2] Junliang Xing[5]
[1] Panasonic R&D Center Singapore [2] National University of Singapore
[3] Franklin. W. Olin College of Engineering [4] Panasonic Corporation
[5] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Abstract

*In this paper, we introduce our solution to the Challenge-1 of the MS-Celeb-1M challenges which aims to recognize one million celebrities. To solve this large scale face recognition problem, a **M**ulti-**C**ognition **S**oftmax **M**odel (MCSM) is proposed to distribute training data to several cognition units by a data shuffling strategy. Here we introduce one cognition unit as a group of independent softmax models, which is designed to increase the diversity of the one softmax model to boost the performance for models ensemble. Meanwhile, a template-based **F**eature **R**etrieval (FR) module is adopted to improve the performance of MCSM by a specific voting scheme. Moreover, a one-shot learning method is applied on collected extra 600K identities due to each identity has one image only. Finally, testing images with lower score from MCSM and FR are assigned new labels with higher score by merging one-shot learning results. Extensive experiments on the MS-Celeb-1M testing set demonstrate the superiority of the proposed method. Our solution ranks the first place in both two settings of the final evaluation and outperforms other teams by a large margin.*

1. Introduction

In the last decade, deep neural networks, especially the **D**eep **C**onvolutional **N**eural **N**etworks (**DCNNs**), have greatly boosted the performance with high accuracy and robustness of the face recognition task [15, 12, 13, 19, 11, 10]. However, most of existing face recognition methods mainly focus on finding whether two face images are from the same person, *i.e*., the face verification problem, rather than recognizing it by outputting a specific name directly. Another issue is that current public face dataset [6, 16, 20, 14] are far from to be sufficient to build face recognition systems, which limits the development of academic research and industrial applications. To address these issues, Guo *et al*. [2] propose a large scale knowledge database with one million celebrities where each of them is linked to a unique entity key and provide a benchmark for large scale face recognition task.

Typically, there are two general kinds of methods to recognize from face images. One is model-based method. More specifically, it models this problem as a classification problem and considers each celebrity as a class. To solve large scale face recognition via this method, Li *et al*. [9] design a multi-view deep representation learning to obtain discriminative features as input of a classifier. Wu *et al*. [18] propose an independent softmax model (ISM) to handle large scale classification problem. Instead of training a classier to directly predict all the classes at the same time, they train several ISMs to predict probabilities for a part of classes. By applying this strategy, the scale of classification problem of each group is reduced. Meanwhile, it is not necessary to require deeper CNN architectures, more training time costs and GPU resources. However, there is no correlation among different ISMs and they do not share global labels information while only focus on a part of classes. The conflict will happen when some less discriminate classes appear in improper ISMs. Given a testing image with its true label from a less discriminate ISM, the final predicted result may not be correct due to the maximum probability among all ISMs is not from the correct ISM model. More details will be discussed in Section 3.

Template-based method is the second category. First, a gallery set (template set) with multiple images of each targeted identity is constructed. Then, for a given image in the query set, the most similar image and its class in the gallery set are retrieved, and the predicted class is assigned to the given query image. It is noteworthy that the template-based method is very convenient for adding/removing entries in the gallery set when the gallery set is not very large, because

* indicate equal contributions. Yu Cheng, Jian Zhao, and Zhecan Wang were interns at Panasonic R&D Center Singapore during this work. Yan Xu is the corresponding author: yan.xu@sg.panasonic.com

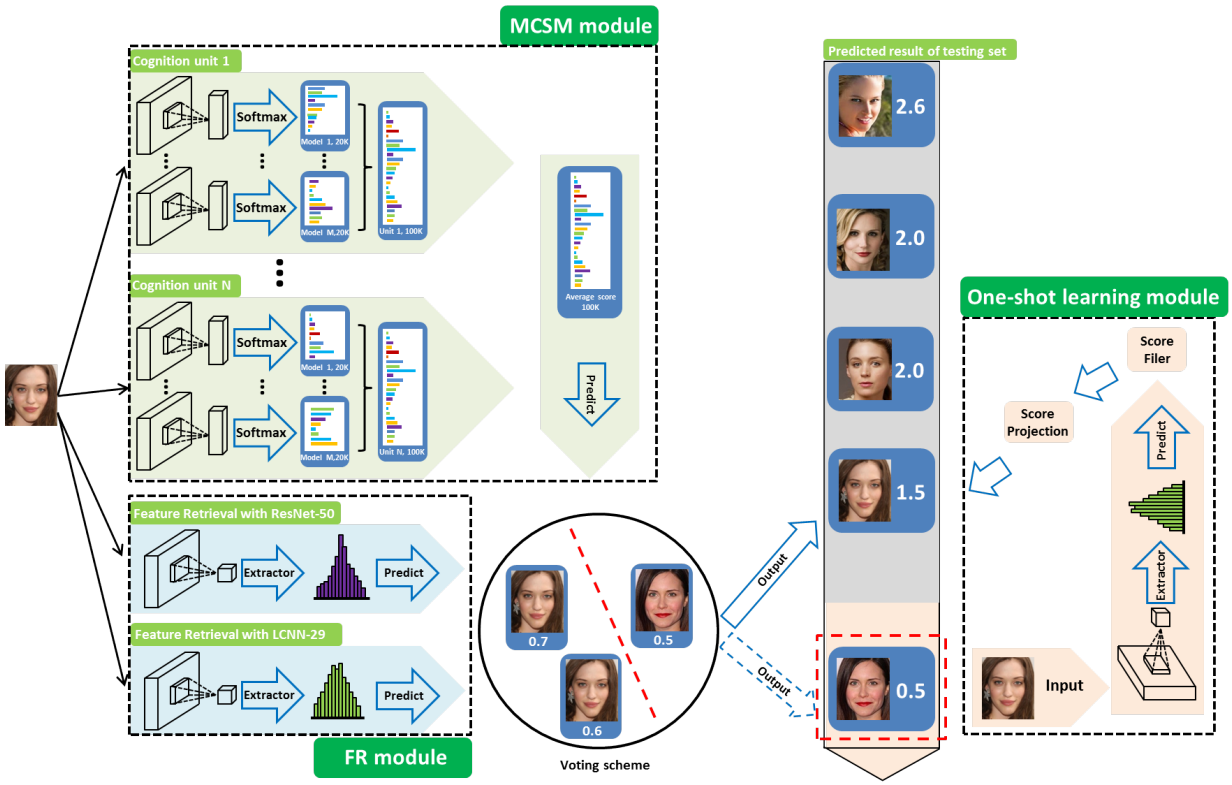

Figure 1: Pipeline of our proposed method for recognizing one million celebrities. This method includes three modules: MCSM module, Feature Retrieval module, and one-shot learning module. For each testing image, the predicted results from the MCSM module and the FR module are selected by a voting scheme. Then, all the scores of whole testing set are sorted from maximum to minimum with their predicted labels. Finally, a one-shot learning method with extra data is applied on testing images with lower confidence score after MCSM and FR. Best viewed in color.

the feature of a targeted face image can be extracted off line. Besides, the template-based method can be viewed as one solution for one-shot learning problem [1, 3, 7] when the training data is extremely limited, *i.e.* one identity has one image only, because it can perform well by computing similarity between gallery set and query set based on a perfect feature extractor. However, with the increasing number of identities in the gallery set, it is difficult to build a complicated index to shorten the retrieval time. Moreover, the accuracy of the template-based method highly relies on the annotation accuracy in the gallery set.

In this paper we propose a novel framework with multi-cognition softmax model and feature retrieval module to address the large scale face recognition problem shown in Figure 1. First, the provided training data is cleaned, since we observe that some images are assigned with the wrong labels and the data distribution is unbalanced. To clean the noisy data, three steps are presented to construct a 100K cleaned training database with not only high quality but also unbroken identity set. Second, we apply the cleaned training data on our proposed MCSM which is viewed as a model-based method. In order to train large scale data with limited time effectively, the whole training data is distributed to several cognition units by a data shuffling scheme and results from each cognition unit are merged by a score-level average and max-max strategy. The targeted training data of each cognition is divided into several small subsets by applying an independent softmax model. Third, a feature retrieval strategy viewed as template-based method is combined with MCSM by a specific voting scheme. To construct a huge gallery set and shorten the retrieval time, we average the features which are extracted by DCNN from all the images of each targeted identity, it means each targeted identity is fully represented by a low dimensional feature. Moreover, we collect around 600K aligned identities as extra data, then a one-shot leaning is applied on them due to each identity has one image only. Finally, testing images with lower confidence score from MCSM and FR are assigned new labels with higher score by merging one-shot learning results.

To summarize, in this paper we have made the following three contributions:

- We propose a novel framework with multi-cognition softmax model and feature retrieval module to handle large scale face recognition problem.

- We deploy an efficient data cleaning method with three steps, which can construct a cleaned training database with high quality while not throwing away any identity.
- We conduct the comprehensive experiments on MS-Celeb-1M to evaluate our method. It shows that our proposed framework can perform well on large scale face recognition problem. On the MS-Celeb-1M challenge, we obtain excellent results and have 79.1% and 87.5% Coverage@Precision=95%, which ranks the first place in both random set and hard set[1].

2. Data Processing

Constructing a large scale face dataset requires lot of time and efforts. Some datasets are massive but they are private and cannot be downloaded, *i.e*. Google [12], Baidu [10]. Based on this observation, Guo *et al*. [2] construct one million celebrities database and release 99892 celebrities as original training data. After carefully analysing the released dataset, we observe that there are lots of noises in it. For example, some images belong to one celebrity while those are included in other celebrities. Some images are very blurry and even clearly not human faces. Also, the distribution of original training data is unbalanced. Some "rich" celebrities have many samples, and a lot of celebrities are "poor" with a few samples per person (see Figure 2). This long-tail characteristic of data distribution is evaluated by Zhou *et al*. [22] and shows that large amounts training data improve the face recognition system's performance and celebrities with only a few samples do not help to boost the recognition accuracy.

Another observation is that the provided training set in MS-Celeb-1M only covers 75% of celebrities in the testing set, which means the upper bound of recognition recall rate does not exceed 75%. MS doesn't provide samples for the remaining celebrities but allows the participants to collect them as extra data to exceed 75% coverage.

Based on the two observations, we propose a efficient method to clean the provided training data. Moreover, we collect around 600K celebrities as extra data to exceed the upper bound of recognition recall rate.

2.1. Data cleaning for provided training data

In order to pre-process provided original training data, we define "Cleaned image" and "Cleaned celebrity" shown as below:

- *Cleaned image*: the image belongs to one celebrity rather than other celebrities.
- *Cleaned celebrity*: all the images of this celebrity are cleaned images, and number of cleaned images is more than 10.

[1] http://www.msceleb.org/leaderboard/iccvworkshop-c1

Table 1: Information of original training data and cleaned training data

	#. of classes	#. of images
Original data	99892	8456240
Cleaned data	100000	5084127

Based on the two definitions, we design below three steps for data cleaning of original data:

Step1: A semantic bootstrapping method is adopted to clean the original training data, which is inspired by [17]. We train a 99892-way softmax classifier using ResNet-50 [5] till a satisfied training accuracy. During testing phase, for each image of one specific celebrity, if its predicted probability P_0 is upper than a given threshold T_0, which is set to 0.7, and its predicted label is the same as ground truth, we accept this image as "Cleaned image". Then, we will retain those celebrities which are not satisfied "Cleaned celebrity" for Step 2.

Step 2: A feature-based clustering method is proposed to further clean around 30K celebrities after Step 1. Based on our observation, the top 5 ranked images from one celebrity are very reliable, we adopt them as the gallery set, and the remaining images of the same celebrity are viewed as the query set. Moreover, in order to represent one face image by a dense feature, a modified LCNN-29 model [17] trained on cleaned 70K celebrities via Step 1 is deployed to extract features for all the gallery and query images. Then we fuse deep features of top 5 gallery images for each celebrity. Meanwhile, the cosine similarity between gallery features and query features are computed. If similarity score is upper than a predefined threshold T_1, which is set to 0.5, we will collect this query image as "Cleaned image". Finally, those celebrities which are not satisfied "Cleaned celebrity" will be passed to Step 3.

Step 3: We manually collect those passed celebrities from Step 2 via Internet. After step 2, there are still about 5K celebrities needed to be cleaned further. We believe these 5K celebrities are the most difficult identities to be cleaned among the original noisy training data. No matter what deep neural network model or method is utilized, they cannot satisfy "Cleaned image" and "Clean celebrity". Based on this observation, we manually collect these 5K celebrities in terms of the specific celebrity's name via Google or Bing. Moreover, in order to build up a 100K training dataset exactly, we add some celebrities from MID-Name list[2].

By applying three steps above, we construct a 100K cleaned training database with not only high quality but also unbroken identity set.

2.2. Data collection for extra training data

To collect extra data as much as possible within a limited time and maintain high quality, for each celebrity, we only

[2] http://www.msceleb.org/download/list

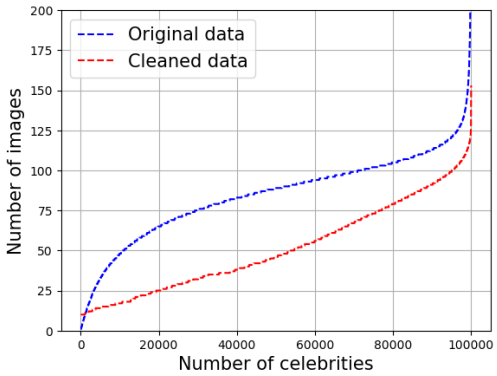

Figure 2: Distribution of original training data and cleaned training data.

download one image by Google based on its specific name provided by MID-Name list. With our best efforts, we collect around 600K data as extra data. Then MTCNN [21] as a powerful face detector is deployed to detect the face with a bounding box and 5 facial points. For those images which cannot be handled by MTCNN, OpenCV detector, and PA-CNN facial localization method [4] are utilized. Based on 5 facial points, we crop the face patch from the downloaded image. In our experiments, the face patches we extracted are similar with the aligned images provide by MS, see Figure 3.

Finally, we collect around 600K aligned image as extra data, then a one-shot learning method is applied on them which will be discussed in Section 6.

3. Multi-Cognition Softmax Model (MCSM)

How to train the large scale data within a limited time effectively? A model-based **M**ulti-**C**ognition **S**oftmax **M**odel (MCSM) is proposed in this paper. By virtue of this module, the large scale training data is distributed to several cognition units by a data shuffling strategy and the results of each cognition are merged by a score-level average and max-max method. Furthermore, the targeted training data of each cognition unit is divided into several small subsets by applying an independent softmax model.

In this section, first, we introduce SM and ISM, respectively. Then, we investigate the conflict that may occur in the experiments. At last, we elaborate that MCSM could obtain higher accuracy due to overcome the conflict from ISMs.

3.1. Softmax Model (SM)

SM is aimed to recognize one million celebrities by training a classifier that has the ability to predict large scale classes. For a given image $x \in \mathbb{R}^{c\times w\times h}$, the probability of i^{th} class is computed by SM as:

$$P(C_i|x) = \mathcal{F}(f_i(x)), \tag{1}$$

where $\mathcal{F}()$ is the softmax function, $f_i(x)$ is the i^{th} dimensional output, which can be expressed as:

Figure 3: Visualization of some samples from extra data. We collect each extra celebrity per image via Google with its specific MID name.

$$f_i(x) = \mathbf{a}_i \cdot \mathbf{v}(x), \tag{2}$$

where $\mathbf{a}_i$ is the i^{th} column vector of the last fully connected layer's weight $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_n]$, and $\mathbf{v}(x)$ is the feature vector of the input image x , which is extracted by the convolutional neural network.

For the purpose of approximation, we compute the intrinsic dimensionality for fully expression of the targeted dataset by maximum likelihood estimation [8].

$$\begin{aligned} \hat{m_k}(X_i) =& [\frac{1}{k-1}\sum_{j=1}^{k-1}\log(\frac{T_k(X_i)}{T_j(X_i)})]^{-1} \\ m_k =& \frac{1}{N}\sum_{i=1}^{N}\hat{m_k}(X_i). \end{aligned} \tag{3}$$

where $T_k(X_i)$ represents the k^{th} nearest sample of X_i, m_k stands for the dimensionality under k^{th} nearest estimation.

As the increasing number of images and classes in the targeted dataset, intrinsic dimensionality fully expressed information will be increased. Therefore, the targeted mapping function $\mathbf{v}(x)$ will be more complex. As the complexity of representation increases, deeper convolutional layers are required to approximate. Furthermore, the risk of misclassification may be increased especially when classifying harder samples. Moreover, it is obvious that huge time and memory costs are impractical when a SM is trained.

3.2. Independent Softmax Model (ISM)

Instead of training a classifier to directly predict all the classes, Wu *et al.* [18] train several ISMs to predict probabilities for a part of classes and the final results are computed by a max-max scheme. There is an assumption [18]: if the test image does not belong to any identity of the trained model, the predicted score under all training identities is considerably low. Particularly, this holds true for most cases.

Softmax function for any sample x of i^{th} class is formulated as:

$$\mathcal{F}(f_i(x)) = \frac{e^{f_i(x)}}{\sum_{j=1}^{N} e^{f_j(x)}}, \tag{4}$$

where N is the number of classes, $f(x)$ is the output of the last fully connected layer. Then the maximum score of each ISM can be computed as:

$$\mathcal{F}_{\max}(f(x)) = \frac{1}{\sum_{j=1}^{N} e^{f_j(x) - f_{\max}(x)}}. \tag{5}$$

Consider a situation that the maximal three f_i values of two ISMs (model A and model B) are given 8, 10, and 12 (label of 12 is ground truth and from model A), respectively, and others are negative values. Due to exponential operations, those negative values are considerable small, they can be neglected. Furthermore, if 12 and 10 locate in model A, and 8 belongs to model B, the maximum probability for the two models are:

$$\begin{cases} \mathcal{F}^A_{\max} = \frac{1}{e^{12-12}+e^{10-12}}, \\ \mathcal{F}^B_{\max} = \frac{1}{e^{8-8}}. \end{cases} \tag{6}$$

Obviously, the assumption fails in this situation, because the highest score of the two models is from model B rather than model A. The assumption becomes true when 12 is in one ISM while 8, 10 locate in another ISM model. Therefore, the conflict heavily depends on the location of less discriminate classes from ISMs.

3.3. Multi-Cognition Softmax Model

Although ISM can be easily trained by several models within a limited time, it throws away correlation among the ISMs and appears conflict when some less discriminate classes belongs to improper ISMs. The SM could solve the conflict problem while it requires more complex architectures and heavy computational resources. Inspired by the advantages and limitations of both the SM and ISM, we propose the MCSM to boost the performance of the large scale face recognition task.

The conflict between several similar identities could be solved when they locate in the different models properly, and the face recognition system will keep more robust when multiple groups of models are trained with different views of data. By applying shuffling overall training data, multiple groups are correlated, which is similar to the idea of increasing individual diversity as did in bagging strategy. In this paper, the group of models is called cognition unit. Mathematically, each cognition unit can deal with large scale recognition problem as:

$$\mathbf{A}_k \cdot \mathbf{V}(x) \approx f_{i,k}(x) = \mathbf{a}_{i,k} \cdot \mathbf{v}_i(x). \tag{7}$$

where i, k stand for the k^{th} identity number in the i^{th} model, $\mathbf{A}, \mathbf{V}$ and $\mathbf{a}, \mathbf{v}$ represents column vector of last fully connected layer and feature vector from one softmax model and independent softmax model, respectively.

When the number of cognition increases, in order to perform as accurately as one SM and have less conflict among cognition units, averaging the scores of cognition unites is applied, and then the eventual prediction is more similar to one SM due to the probability of conflict reduced exponentially. Therefore we take multi-cognitions with shuffling training data for more accurate approximation:

$$\mathbf{A}_k \cdot \mathbf{V}(x) \approx \frac{1}{m} \sum_{p=1}^{m} \mathbf{a}_{i,k,p} \cdot \mathbf{v}_{i,p}(x). \tag{8}$$

where p is the p^{th} cognition unit among total m cognition units of MCSM.

The advantages of MCSM are obvious, we are able to build a system that comprises several models which are trained on small amount of data. This divide-and-conquer approach requires less time cost and computational resources for each model. Meanwhile, high performance is obtained by MCSM.

4. Feature Retrieval (FR)

For the large scale face recognition task, besides the mode-based method, another main stream is the template-based feature matching proposal. In our approach, the deep convolutional neural network model viewed as a feature extractor to extract facial features, and then pass the facial features to the verification task.

The feature extractor is trained by adding a weight matrix followed by feature layer. The training process is performed as:

$$P(C_i|x) = \mathcal{F}(\mathbf{a}_i \cdot \mathbf{v}(x)), \tag{9}$$

where $\mathbf{a}_i$ represents the i^{th} column vector which is called "anchor vector". Under this training scheme, the deep model learns a probability:

$$P(C_i|x) = \mathcal{F}(||\mathbf{a}_i||_2 \, ||\mathbf{v}(x)||_2 \cos(\mathbf{a}_i \cdot \mathbf{v}(x))), \tag{10}$$

Here, softmax is viewed as the normalization function in Eq. (10). Therefore, the probability of image x belonging to class C_i is directly related to its cosine similarity with i^{th} anchor vector $\mathbf{a}_i$.

So far the abstract features of facial images can be represented via deep model. After being trained with large amount of identities with facial images, the model will spontaneously learn an angle-dependent feature distribution. Therefore, the high dimensional images are mapped into a low dimensional cosine space (hyper spherical surface):

$$\mathbb{R}^h \longmapsto \mathbb{S}^l. \tag{11}$$

In testing phase, we first construct feature-based templates. Assuming we have m samples of i^{th} template, then the anchor vector $\mathbf{b}_i$ can be computed by:

$$\mathbf{b}_i = \frac{1}{m} \sum_{j=1}^{m} \frac{\mathbf{v}(x_{i,j})}{||\mathbf{v}(x_{i,j})||_2}, \tag{12}$$

where $\mathbf{v}(x_{i,j})$ is the j^{th} image's feature of the i^{th} template.

Then, for given a testing image y, the similarity score S_i is simply computed by inner product of i^{th} anchor vector $\mathbf{b}_i$ and its feature vector $\mathbf{v}(y)$:

$$S_i = \mathbf{b}_i^T \cdot \mathbf{v}(y). \tag{13}$$

5. Voting Scheme

A recognition system with good performance should have high recall ratio at low fall-out ratio when outputting final result, especially when several independent results are obtained by different methods, *e.g*. MCSM, FR with ResNet-50, FR with LCNN-29. How can we obtain higher performance after combining different models while let the system to cover the coverage as much as we can? It requires us to not only improve the top-1 accuracy of the ensemble models, but also ensure correct predictions for the samples with large score. Therefore, we propose a multi-perspective voting scheme to enforce the predicted results with high score.

First, a model with best performance is selected to be the primary model while others as auxiliary models. The predicted scores are initially set to the same value as the main model's prediction. Then, these scores are increased by a if the predicted results of all auxiliary models are the same as the primary model's. Next, if the predicted results of all the auxiliary models are the same but different from the primary model, we replace the predicted label and score with auxiliary models' prediction and their averaging score plus an offset p, as we believe it would probably be a wrong prediction from main model. In our experiments, we set a, p as 1.6 and 0.8, respectively. Detail process is described in Algorithm 1.

Algorithm 1 Voting scheme for different models

Input: Testing set $T_i, i \in (1, 2...n)$; Predicted score set and predicted label set of T: $S_j, L_j, j \in (1, 2...m)$
Output: Final predicted score and result S_r, L_r

1: $S_r \leftarrow S_m, L_r \leftarrow L_m$
2: **for** $t = 1, 2, ..., n$ **do**
3: **if** $L_1^{T_t} = L_2^{T_t} = ... = L_{m-1}^{T_t} = L_r^{T_t}$ **then** $S_r^{T_t} \leftarrow S_r^{T_t} + a$
4: **if** $L_1^{T_t} = L_2^{T_t} = ... = L_{m-1}^{T_t} \neq L_r^{T_t}$ **then** $S_r^{T_t} \leftarrow \frac{1}{m-1}\sum_{j=1}^{m-1} S_j^{T_t} + p, L_r^{T_t} \leftarrow L_1^{T_t}$
5: **end for**

By means of the voting scheme, the samples with higher confidence score are predicted correctly with high probability, whereas the samples with lower scores are predicted correctly with low probability. Moreover, the voting scheme requires high top-1 accuracy of each model to ensure that the predictions with high score are more reliable.

6. One-shot Learning for Extra Data

As extra data, we design a template-based method of one-shot learning for the extra 600K data since each celebrity contains one image only.

First, the 600K data is viewed as the gallery set and development set or final testing data as query set. Second, a LCNN-29 model from FR module, which learns a compact and discriminate facial feature by narrowing the decision boundaries of each class during the training process, is deployed to extract deep features for all the gallery and query images. To further boost the performance, during the feature extraction phase, we apply multi-patch testing [11, 10], *i.e*. 25-crops with flipping, on each image, then the deep features are normalized by using $L2$-norm. Third, we compute the cosine similarity between gallery and query features.

We only apply one-shot learning method on testing images where confidence score is less than a high threshold T_h from MCSM and FR. After getting the predicted confidence score from one-shot learning method, those images with scores are less than a low threshold T_l are filtered out. The high threshold T_h and low threshold T_l are set experimentally to 2.2 and 0.6, respectively. Finally, the selected scores are projected to softmax domain as following:

$$S_2 = a - b \times e^{c \times S_1}. \tag{14}$$

where S_1 is the score from cosine similarity. S_2 is the projected score of softmax and a, b, c are 2, 5, -6, respectively. By deploying the above strategy, a testing image with lower confidence score from MCSM and FR is assigned to a new label with higher score by merging results of one-shot leaning method.

7. Experiments

In this section, first we introduce the dataset and evaluation protocol for MS-Celeb-1M Challenge-1. Then, extensive experiments with different modules are presented. Last, we show our submitted result in the final testing set of Challenge-1.

7.1. Dataset and evaluation protocol

MS provides around 100K celebrities as training data and defines two different evaluation sets.The one is development set provided with label, the other is testing set without label. All participants should tune their models or parameters to fit the development set as much as possible. After that, based on best models, participants should submit the result of testing set to MS. Meanwhile, MS also divides each evaluation set (development set or testing set) into two subsets, random set and hard set, respectively. The purpose of the former is to reveal how many celebrities are truly covered by the models to be tested. The latter aims to evaluate the generalization ability and the robustness of the model on complex situations.

Table 2: Compared results between original data and cleaned data. dev1 and dev2 indicate hard set and random set, respectively

Training data	Model	C@P=95%		C@P=99%	
		dev1	dev2	dev1	dev2
Original data	1	6.2	4.7	5.0	5.0
	1,2,3,4,5	24.0	50.0	6.0	5.0
Cleaned data	1	14.1	18.7	11.0	10.0
	1,2,3,4,5	58.1	76.0	43.3	40.3

Moreover, only about 75% classes of the evaluation sets are covered in training set. That means the upper bound of recognition recall ratio is 75% only, if extra data is not collected.

In order to measure the recall ration of a face recognition system, a coverage with precision protocol is proposed as: $precision = C/M$, $coverage = M/N$. N denotes the number of images in the evaluation set, C indicates the number of images are recognized correctly among the recognized M images. For matching the real scenarios, the evaluation protocol only returns the recall ratio at a given precision 95% and 99% for random set and hard set.

7.2. Experiments with Data cleaning

Our proposed data cleaning strategy aims to clean noisy training data while maintain the same training data scale. In this section, we conduct experiments to show the effectiveness of data cleaning. ResNet-50 as our backbone deep is deployed to train the ISMs with original noisy data and cleaned data, respectively. We train 5 ISMs covered the whole 100K training data. Each ISM is assigned exactly 20K celebrities to the last fully connected layer. The results of coverage with 95% precision (C@P=95%) and 99% precision (C@P=99%) are shown in Table 2. From the table, we observe that data cleaning strategy not only increases the coverage at 95% precision but also obtains the improvement at 99% precision. On one hand, it demonstrates the necessity of our data cleaning strategy, especially for large scale noisy data. On the other hand, although noisy images are filtered out, the class of celebrity does not be removed. It is the fundamental stone to train deep neural networks for recognition of large scale celebrities.

7.3. Experiments with MCSM

In this section, 5 units of cognition with independent softmax models are evaluated and each unit contains 5 ISMs. Specifically, the first 4 cognition units share the same model architecture, which is ResNet-50. Nevertheless, the last one unit is changed to LCNN-29. The final results are combined by a score-level average and max-max strategy among different cognition units. The results with increasing the number of cognition units are shown in Table 3 and Figure 4. It is obvious the performances are improved on both dev1 set and dev2 set when more units are added. After 5 units of cognition with 25 models are combined totally, we achieve the coverage 64.6% of dev1 and 78.4% of dev2 at precision 95%. There are 9.6% and 7.4% coverage improvement compared with single unit result, respectively. In particular, when the unit 5 is added, the coverage at precision 95% of dev2 is increased from 78.0% to 78.4%, while it is decreased from 65.4% to 64.6% of dev1. We argue that the performance of dev1 is saturated because dev1 set is obvious harder than former. We believe it is the reason of the limited samples of development set.

Table 3: Results of MCSM with different number of cognition units. Each cognition unit contains 5 ISMs

Cognition units	C@P=95%		C@P=99%	
	dev1	dev2	dev1	dev2
1	55.0	71.8	39.8	25
1,2	60.6	75.8	45.2	25.6
1,2,3	63.4	76.4	49.2	23.4
1,2,3,4	65.4	78.0	53.0	32.6
1,2,3,4,5	64.6	78.4	52.0	32.6

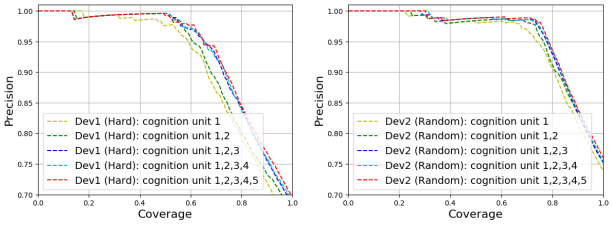

Figure 4: Results of MCSM with different number of cognition units

In a conclusion, experiments demonstrate that MCSM could improve the performance significantly compared with ISM. Furthermore, MCSM requires less training time and computational resources.

7.4. Experiments with Feature Retrieval

In this section, we perform experiments to verify the effectiveness of feature retrieval module. The results are shown in Table 4 and Figure 5. LCNN-29 and ResNet-50 obtain high coverage at precision 95% for both dev1 set and dev2 set, while low coverage at precision 99%. Furthermore, although feature retrieval obtains high Top-1 accuracy, it cannot guarantee correct labels for the predicted results with a large confidence score, which is the limitation of feature retrieval for large scale classification problem. Only relying on feature retrieval is not robust when the system requires high precision. Therefore, it is our inspiration that if we combined feature retrieval with MCSM whether the performance can be improved or not. Based on our experiments, it is worth noting that fusion of MCSM and FR by a specific voting scheme can overcome the limitation of feature retrieval

Table 4: Results of feature retrieval (FR)

Method		C@P=95%		C@P=99%	
		dev1	dev2	dev1	dev2
FR	LCNN-29	69.0	78.8	5.0	8.8
	ResNet-50	56.4	76.4	10.2	28.8
MCSM		64.6	78.4	51.2	61.0
MCSM+FR		71.8	79.8	51.0	61

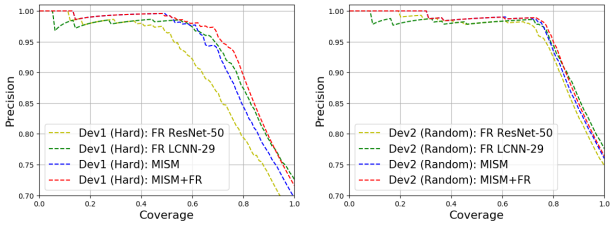

Figure 5: Results of feature retrieval (FR)

method and maintain the advantages of MCSM method.

The last row of Table 4 demonstrates that the combination result, when MCSM and FR are fused, we not only achieve higher coverage at precision 95% but also maintain the similar coverage at precision 99%. In other words, it indicates that the combination of MCSM and feature retrieval by a voting strategy has quite well ability to recognize large scale celebrities with large coverage and high accuracy at the same time.

7.5. Experiments with One-shot learning

In this section, first, the experiments of one-shot learning method with different testing views are performed. Then, we show the result of ensemble all modules (MCSM+FR+One-shot) on the development set. Specifically, to evaluate the performance of one-shot learning method, we compare the difference between center crop and 25-crops with random and flip setting in Table 5. From the results, 25-crops obtains better performance than center crop on Top-1 accuracy protocol. It is indeed because that there is only one image for each identity on the training and testing data. Multi-patch strategy increase more information for one-shot learning problem.

All results of different methods on development set are viewed in Table 6 and Figure 6. By applying one-shot learning method on extra data, the performance is improved significantly from 71.8% to 83.2% and from 79.8% to 90.2% on dev1 set and dev2 set at precision 95%, respectively. Last but not least, our Top-1 accuracy of dev1 set and dev2 set are 79.8% and 85.8%, respectively, which beyond the upper bound of recognition recall ratio (75%).

7.6. Evaluation on testing data

In this section, the results of this year among different teams on the final 50K test images are compared and shown

Table 5: Top-1 accuracy results of one-shot learning with multi-patch testing

One-shot learning	Top-1 accuracy	
	dev1	dev2
Center-crop	54.6	63.0
25-crop	60.4	66.2

Table 6: Results of different methods on development set

Method	C@P=95%		C@P=99%	
	dev1	dev2	dev1	dev2
Multi-View [9]	40.8	50.6	28.0	21.8
ISM [18]	50.2	76.6	34.6	25.6
MCSM	64.6	78.4	51.2	61.0
MCSM+FR	71.8	79.8	51.0	61.0
MCSM+FR+One-shot	83.2	90.2	51.0	61.0

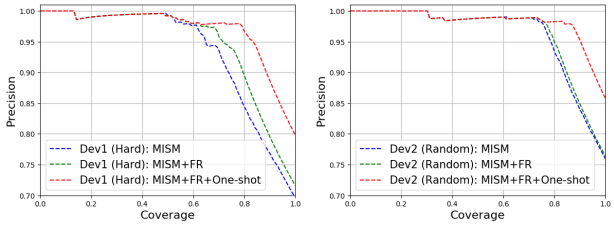

Figure 6: Results of different methods on development set

Table 7: Compared results with state-of-arts on testing set

Teams	C@P=95%	
	dev1	dev2
SmileLab	62.1	79.2
Turtle	73.0	86.2
Ours	79.1	87.5

in Table 7. It is obvious that we obtain 79.1% coverage at precision 95% for dev1 and 87.2% coverage at precision 95% for dev2, which outperforms other teams by a large margin.

8. Conclusions

In this paper, we have proposed a novel method by fusing multiple cognition units and feature retrieval module to solve large face recognition problem. The MCSM module divides a single classifier into several small classifiers while sharing global labels information by a shuffling data strategy. Meanwhile, a feature retrieval module is proposed to further boost the performance. With the help of collected extra data, a one-shot learning method is applied on testing images with lower confidence score from MCSM and FR. The top performance of MS-Celeb-1M challenge demonstrates the superiority of our method.

References

[1] Y. Guo and L. Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017. 1

[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*, pages 87–102. Springer, 2016. 1, 3

[3] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. *arXiv preprint arXiv:1606.02819*, 2016. 1

[4] K. He and X. Xue. Facial landmark localization by part-aware deep convolutional network. In *Pacific Rim Conference on Multimedia*, pages 22–31, 2016. 4

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[6] 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. 1

[7] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 1

[8] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 777–784, 2005. 4

[9] J. Li, J. Zhao, F. Zhao, H. Liu, J. Li, S. Shen, J. Feng, and T. Sim. Robust face recognition with deep multi-view representation learning. In *ACM International Conference on Multimedia*, pages 1068–1072, 2016. 1, 8

[10] 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. 1, 3, 6

[11] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015. 1, 6

[12] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1, 3

[13] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[14] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014. 1

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1

[16] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011. 1

[17] 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. 3

[18] Y. Wu, J. Li, Y. Kong, and Y. Fu. Deep convolutional neural network with independent softmax for large scale face recognition. In *ACM International Conference on Multimedia*, pages 1063–1067, 2016. 1, 4, 8

[19] L. Xiong, J. Karlekar, J. Zhao, J. Feng, S. Pranata, and S. Shen. A good practice towards top performance of face recognition: Transferred deep feature fusion. *arXiv preprint arXiv:1704.00438*, 2017. 1

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 4

[22] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015. 3