FEATURE LEARNING FOR ONE-SHOT FACE RECOGNITION

Lingxiao Wang, Yali Li, Shengjin Wang

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ABSTRACT

One-shot face recognition is a challenging open problem which requires recognizing novel identities from only one gallery face. One-shot classes are squeezed and neglected in the feature space for classification due to data imbalance. Moreover, training samples deficience is a major obstacle to intra-class clustering. In this paper, we propose a novel framework based on CNN of *balancing regularizer* and *shifting center regeneration* which regulates norms of weight vector into same scale and adjusts clustering center to deal with deficient training data. Comprehensive evaluations on MS-celeb-1M low-shot face dataset demonstrate that our methods improve one-shot face recognition notablely which achieve 88.78% coverage at precision=0.99 using restricted data without hybrid classifiers or multi-model. Moreover, experiments on LFW prove that CNN model trained with proposed methods can obtain more discriminative and compact feature representations. Since there are many identities that have only few training samples available online, our methods have great significance for improving data utilization and strengthening feature representation for face recognition.

Index Terms— One-shot Learning, Face Recognition, Feature Learning

1. INTRODUCTION

In recent years, Convolutional Neural Network(CNN) has achieved great breakthroughs and boosted the state-of-the-art in many vision tasks[1, 2]. In face recognition field, methods[3, 4, 5] based on CNN have achieved or even surpassed human performance on several benchmarks. CNNs of considerable depth and width utilize the information of large scale training data in real world to learn the intrinsic pattern for each class supervised by label signals. Although different kinds of loss functions like Softmax loss[3], center loss[6], A-softmax[7], triplet loss[8] and so on have been applied for learning face representations, barriers still exist in one-shot face recognition[9]. Unlike human beings who can recognize identities from very few gallery faces, deep learning methods require a large number of training data. However, many identities only have one or few training samples available to be collected. This imbalance of training data is a huge challenge for building large-scale face recognizer. To study this problem, Guo *et al*. [10] provided a benchmark dataset of $21,000$ persons which is a subset of MS-Celeb-1M[11]. It is divided into two sets, $20,000$ persons in base set with tens of training samples and $1,000$ persons in novel set with only one sample. Details of dataset are shown in Table 1. It mainly focuses on the Precision-Coverage performance on the novel set, while also monitors the performance on the base set.

Several works have been devoted to one-shot face recognition. Guo *et al*. [10] proposed a UP term to achieve reasonable decision area and improve the classifier performance during CNN training. It leads to significant enlightenment but just focusing on classifier is not enough for one-shot learning. Method in [12] adopted hybrid classifiers of a CNN and a Nearest Neighbor(NN) model to handle two kinds of training sets and its final predictions are the fusion of CNN and NN results. Its main contribution of classifier fusion relys on threshold choosing and is hard to generalize. Authors in [13] proposed an enforcing scheme to obtain compact feature representation by enforced softmax and heuristic voting. Although this approach achieves a high coverage using external training data, its voting strategies involving multi-model combination are over complicated and limited in close-set problem. Method in [14] introduced a hard example mining method. It maintains a list of the most similar identities for each identity and generates better mini-batches. It furtherly trains the last classifier but there is a huge gap of performances between development and test set. Since it is impractical to preset all the unseen test identities in training, deeply learned features are adopted for face verification and identification instead of using predicted labels. In this way, one-shot learning issues have great importance to general deep feature learning because it can take full advantage of open online data where many celebrities have very few images for training. Thus, the trend of approaches above does not pay enough attention to feature learning and is limited in close-set protocol.

It is revealed that novel classes occupy less feature space due to small norms of weight vectors of classifier. More-

This work was supported by the National Natural Science Foundation of China under Grant No. 61771288 and the state key development program in 13th Five-Year under Grant No. 2016YFB0801301.

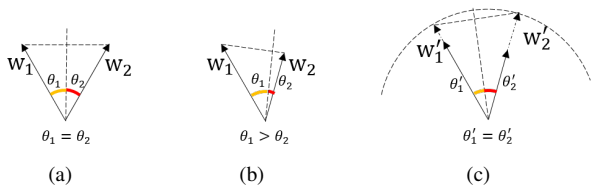

Fig. 1. Norms of $\mathbf{w}_k$ influence classification space of class k. We regulate $\mathbf{w}_1$ and $\mathbf{w}_2$ into same scale so that class 1 and 2 have similar volume size in feature space.

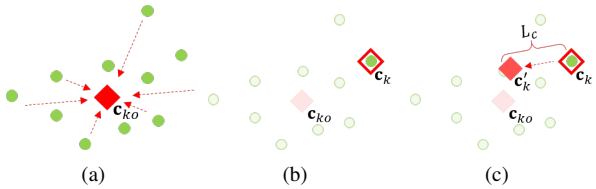

Fig. 2. (a) Ideally, center $\mathbf{c}_{ko}$ is averaged by the features of the corresponding classes based on mini-batch. (b) When a class has only one sample, its center $\mathbf{c}_k$ is decided only by this sample. (c) We shift $\mathbf{c}_k$ a little to $\mathbf{c}'_k$, then center loss L_c is computed based on the new center.

over, lack of training sample hinders intra-class clustering and makes no contribution to center loss. To this end, we propose methods of *balancing regularizer* and *shifting center regeneration* which regulate norms of weight vector into same scale and adjust the center loss computing strategy of deficient training data. Comprehensive Experiments on MS-Celeb-1M low-shot face dataset and LFW[15] demonstrate that proposed methods can improve coverage performance greatly on one-shot face recognition and guide models to produce discriminative and compact feature representations.

2. PROPOSED METHOD

In this section, *balancing regularizer* is first introduced to unfold the feature space of the one-shot classes and balance weight vectors of all classes. Then *shifting center regeneration* is designed to assist novel classes to cluster effectively during learning process. These proposed novel learning strategies can produce more compact and distinguishable feature representation for one-shot face recognition.

2.1. One-shot Learning with Balancing Regularizer

As decribed in [10], base sets are defined as the classes that have many images for training. In the novel set, each identity has only one image for training and many for test. The main sticking point of one-shot problem is to use the imbalanced training data and obtain representative features. The standard cross entropy loss which guides the training can be written as:

$$L = -\sum_i p_k(\mathbf{x}_i) \log p_k(\mathbf{x}_i), \tag{1}$$

where $p_k(\mathbf{x}_i)$ is the softmax probablity of sample $\mathbf{x_i}$ belonging to class k,

$$p_k(\mathbf{x}_i) = \frac{\exp(\mathbf{w}_k^T \mathbf{x}_i + \mathbf{b}_k)}{\Sigma \exp(\mathbf{w}_k^T \mathbf{x}_i + \mathbf{b}_k)}. \tag{2}$$

We set bias term $\mathbf{b}_k = 0$ and $\mathbf{w}_k$ is the weight vector for class k of the last fully-connected layer. However, this standard loss function cannot handle one-shot face recognition problem partially due to a critical fact found by Guo *et al*. [10] that the norms of weight vectors for novel classes are usually much smaller than the ones for base classes. Novel weight vectors cannot get trained effectively due to deficient training samples. This leads to novel classes being compressed in feaure space as shown in Fig 1(a) and 1(b). When the norm of $\mathbf{w}_k$ is smaller than others, its corresponding class occupies less volume size in feature space. Thus, a great measure of samples are wrongly classified in test phase.

Based on above fact and analysis, we introduce a balancing regularizer term to training. Instead of directly apply L2-normalization to weight vector of the last fully-connected layer like [7, 13] which actually has little help for compressed classes in training, we add a regularizer for weights as an explicit supervising term. It unfolds squeezed space of novel classes and balance norms of weights of all classes(Fig 1(c)),

$$L_b = \frac{1}{|K|} \sum_{k \in K} \| \|\mathbf{w}_k\|_2^2 - \beta \|_2^2, \tag{3}$$

where β is a constant term that constrains each weight vector with a roughly same length. Regulated by balancing regularizer, two kinds of classes occupy similar volume in feature space.

Balancing regularizer has a similar format to underrepresented classes promotion(UP) in [10] which is written as:

$$L_{up} = \frac{1}{|C_n|} \sum_{k \in C_n} \| \|\mathbf{w}_k\|_2^2 - \alpha \|_2^2, \ \alpha = \frac{1}{|C_b|} \sum_{k \in C_b} \|\mathbf{w}_k\|_2^2, \tag{4}$$

where C_b and C_n denote the base set and novel set respectively. Equations 3 and 4 demonstrate that these two methods balance norms of weights of all classes in the similar way, but balancing regularizer has simpler form and less computational complexity than UP term.

2.2. Shifting Center Regeneration

Center loss has been proved effective in face recognition problems[6]. With joint supervision of softmax loss and center loss, a CNN can obtain deep features that have both inter-class discrimination and intra-class compactness. However, there is an obstacle presented to the classes with only

Set	Classes	Train. Imgs/AVG	Dev. Imgs/AVG
Base	20,000	1155175/58	20,000/1
Novel	1,000	1,000/1	5,000/5

Table 1. Information of training and development datasets.

one training sample that their cluster centers rely merely on this one sample during training. In other words, the center of a novel class is no other than its one and only training sample for each iteration. Since training samples are randomly selected, they are possibly on the periphery of the cluster in real feature space shown in Fig 2(a) and 2(b). Center losses of those classes are invalid and of no help to supervise the compactness. To this end, we propose shifting center regeneration illustrated in Fig 2(c) to mend center loss under the one-shot learning circumstance. Since a novel class is lack of training samples to update valid center loss, a possible virtual "center" is reconstructed near the present central position at each iteration. By regenerating a virtual new cluster center using our method, intra-class center losses no longer rely only on one sample and make a positive contribution to the feature learning. Specifically, the position of present cluster center $\mathbf{c}_k$ is shifted randomly and a new "center" $\mathbf{c}'_k$ is regenerated here,

$$\mathbf{c}'_k = \mathbf{c}_k + \Delta_{\mathbf{c}_k}, \tag{5}$$

where $\Delta_{\mathbf{c}_k}$ is a random vector following a truncated normal distribution. To avoid inter-class space overlapping, the range of center shifting must be limited which requires the angle between $\mathbf{c}_k$ and $\mathbf{c}'_k$ to be small. In our work, it's set around $1°$. Note that shifting center regenration is only applied to the class that has one training sample.

In this way, the modified center loss and its gradients can be computed as below:

$$L_c = \frac{1}{2}\sum_{y_i \in C_b} \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2 + \frac{1}{2}\sum_{y_i \in C_n} \|\mathbf{x}_i - \mathbf{c}'_{y_i}\|_2^2, \tag{6}$$

and

$$\begin{aligned} -\frac{\partial L_c}{\partial \mathbf{x}_i} = &-(\delta(y_i \in C_b) \cdot (\mathbf{x}_i - \mathbf{c}_{y_i}) \\ &+\delta(y_i \in C_n) \cdot (\mathbf{x}_i - \mathbf{c}_{y'_i})), \end{aligned} \tag{7}$$

where $\delta(\cdot) \in \{0, 1\}$ is an indicator function and $\mathbf{c}_k$ is updated as the original way in [6]. If $\mathbf{c}_k (k \in C_n)$ is computed only in a mini-batch, it can be directly derivated from Eq 5 and Eq 7 that $-\frac{\partial L_c}{\partial \mathbf{x}_i}$ has the same direction with $\Delta_{\mathbf{c}_k}$. Thus, randomly shifted $\mathbf{c}'_k$ also expand the volume that one class takes up in feature space. Since low-shot problems which provide a few training samples for novel classes have similar characteristics with one-shot problems, shifting center regeneration can be generalized to low-shot learning.

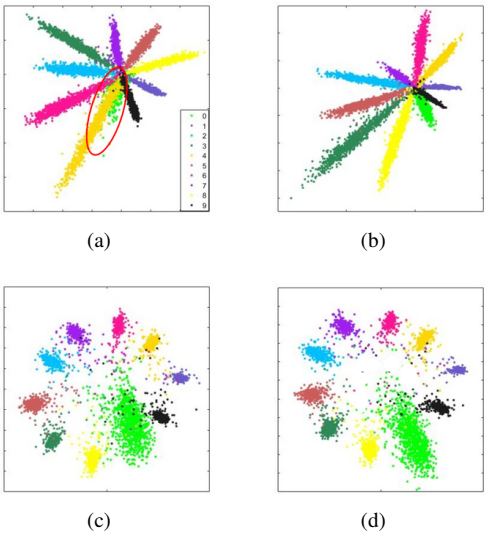

Fig. 3. A toy example to illustrate effectiveness of our method. Class **0** has one training sample represented by bright green. (a) Novel class is not distinguishable with softmax. (b) Balancing regularizer extend the decision space. (c) CNN learn more compact feature by center loss but class **0** has more classification mixtures due to lack of valid center loss back-propagation. (d) By regenrating shifting center, we obtain discriminative and compact feature representation.

We apply both balancing regularizer and shifting center regeneration to one-shot face recognition problems. Although training samples are imbalanced and limited, these two novel methods remarkably improve discrimination and compactness of feature learning by unfolding space of compressed classes and shifting cluster center.

3. EXPERIMENTS

3.1. A Toy Example

To better illustrate our methods, we conduct experiments on MNIST dataset and simulate one-shot learning by retaining only one training sample for one certain class. As shown in Fig 3, we visualize ten classes in feature space and one dot represents one sample in test set. Class **0** (in bright green and marked with a red circle) has only one training sample which is replicated by 1,000 times in training.

Supervised by the standard softmax, novel class **0** is not distinguishable from other classes shown in Fig 3(a) due to limited training samples. By applying balancing regularizer in Eq 3, novel class gets better representation shown in Fig 3(b). To reduce intra-class variance, Fig 3(c) and Fig 3(d) introduce center loss and Fig 3(d) is obtained with added shifting center regeneration. As shown in above, our methods of balancing regularizer and shifting center regeneration are effective to one-shot feature learning.

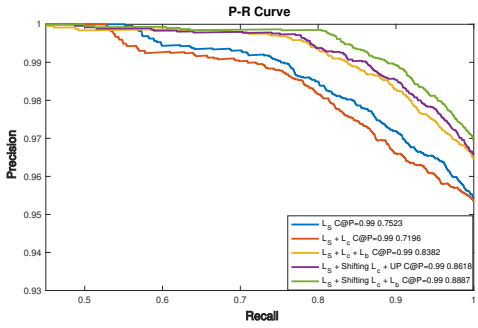

Fig. 4. P-R curves on MS-celeb-1M low-shot dataset.

3.2. Experimental Setup

Since the test set is unlabeled, we evaluate our methods on MS-Celeb-1M low shot face development set[10] which has 20,000 base classes and 1,000 novel classes. **Only** the provided training data are used for training our models. Since the key point of one-shot face recognition problem is the performance on novel classes, recognition coverage at a given precision 99% is evaluated. For N images in the measurement set, precision and coverage(recall) are defined as: $precision = C/M, coverage = M/N$, where C is the correct number of recognized M images at a given threshold.

We use ResNet-34[16] which has both good performance and moderate size as the feature learning network. A classifier of 20,000 classes is pre-trained using the base set. Then we finetune on the same network with both base and novel set. For all our one-shot learning methods, samples in the novel set are boosted and replicated by 100 times. Scores and classification results output by classifier are directly adopted-because of close-set protocol. Also, test samples are computed cosine metric with the only gallery image in the novel training set to get the maximum class index. We promote the confidence scores of test samples whose results are identical with the maximum cosine distance index.

3.3. One-shot Face Recognition

We compare coverage rate of our method and other baselines at precision 99% and 99.9% to reflect capability of methods on one-shot face recognition. All the shown methods have top-1 accuracy more than 98% on the base set so results on novel set has no sacrifice of performance on base set. Table 2 shows that our method using shifting center regeneration and balancing regularizer (denoted as Shifting L_c and L_b) leads to a huge improvement over softmax or center loss. Fig 4 illustrates Precision-Recall curves for better visualization. Tabel 3 shows the comparision with other methods on development set[10]. It is noteworthy that our approach achieves the performance comparable to the state-of-the-art **without** external data, multi-model or hybrid classifiers.

Although it is trained on one-shot dataset, our single model achieves a performance comparable to the state-of-the-art on LFW verification task. Table 4 shows that feature representation ability can be improved by introducing our novel methods into training on imbalanced data. There is a great potentiality of generalizing our methods to improve data utilization and overcome the low-shot problem.

Method	C@99%	C@99.9%
L_S	75.23%	56.64%
$L_S + L_c$	71.96%	52.97%
$L_S + L_c + L_b$	83.82%	48.84%
L_S+ Shifting L_c+ UP	86.18%	58.50%
L_S+ Shifting $L_c + L_b$	**88.87%**	**62.20%**

Table 2. Coverage at Precision=99% and 99.9% on MS-celeb-1M low-shot dataset.

Method	External Data	Hybrid models / classifiers	C@99%
NUS[13]	Yes	Yes	100%
SmileLab[12]	No	Yes	92.78%
SGM[17]	No	No	27.23%
UP[10]	No	No	77.48%
DM[14]	No	No	88.32%
Ours	No	No	88.87%

Table 3. Comparison on development set. Our approach achieves the performance comparable to the state-of-the-art **without** external data, multi-model or hybrid classifiers. Results in [14] with additional classifier training are not listed.

Method	Dataset	Accuracy
Deepface[18]	Public	97.27%
FaceNet[8]	Private	**99.63%**
$L_S + L_c$	Imbalanced Public	98.05%
$L_S + L_c + L_b$	Imbalanced Public	98.53%
L_S+ Shifting $L_c + L_b$	Imbalanced Public	**98.78%**

Table 4. Verification accuracy on LFW.

4. CONCLUSION

In this paper, we propose *balancing regularizer* and *shifting center regeneration* to one-shot face recognition. The key idea is balancing feature space of all classes and adjusting clustering to deal with deficient training data. Experiments demonstrate that proposed methods improve the performance of one-shot face recognition significantly and also produce discriminative and compact feature representation.

5. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[4] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1988–1996, 2014.

[5] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[6] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[7] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1.

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[9] B. M. Lake, R Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332, 2015.

[10] Yandong Guo and Lei Zhang, "One-shot face recognition by promoting underrepresented classes," *arXiv preprint arXiv:1707.05574*, 2017.

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 87–102.

[12] Yue Wu, Hongfu Liu, and Yun Fu, "Low-shot face recognition with hybrid classifiers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1933–1939.

[13] Jian Zhao, Yu Cheng, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, Jiashi Feng, Jian Zhao, Yu Cheng, and Zhecan Wang, "Know you at one glance: A compact vector representation for low-shot learning," in *International Conference on Computer Vision*, 2017.

[14] Evgeny Smirnov, Aleksandr Melnikov, Sergey Novoselov, Eugene Luckyanets, and Galina Lavrentyeva, "Doppelganger mining for face representation learning," in *International Conference on Computer Vision*, 2017.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] Bharath Hariharan and Ross Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy*, 2017.

[18] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, pp. 41.1–41.12.