MarginDistillation: Distillation for Face Recognition Neural Networks with Margin-Based Softmax

Svitov David, Alyamkin Sergey

***Abstract*—The usage of convolutional neural networks (CNNs) in conjunction with the margin-based softmax approach demonstrates the state-of-the-art performance for the face recognition problem. Recently, lightweight neural network models trained with the margin-based softmax have been introduced for the face identification task for edge devices. In this paper, we propose a distillation method for lightweight neural network architectures that outperforms other known methods for the face recognition task on LFW, AgeDB-30 and Megaface datasets. The idea of the proposed method is to use class centers from the teacher network for the student network. Then the student network is trained to get the same angles between the class centers and face embeddings predicted by the teacher network.

***Keywords*—ArcFace, distillation, face recognition, margin-based softmax.

I. Introduction

THE development of edge devices has sparked significant interest in lightweight face recognition access systems. This type of solution is based on optimized neural network architectures for mobile devices. A typical example of such network is MobileFaceNet [1] designed specifically for the face recognition on devices with the low computing power. The usage of margin-based softmax approach [2]–[4] in the training procedure helps to obtain the state-of-the-art performance for face recognition tasks.

Despite the fact that fast and compact mobile network architectures give lower face recognition accuracy than the full-size ones, in some applications such as biometric access systems, it nevertheless plays a critical role. Distillation is a method that helps to achieve the highest accuracy for mobile neural network architectures where the knowledge is transferred from a large teacher network to a small student network. In this paper we propose a novel distillation method called *MarginDistillation* to reduce the gap between teacher and student networks during the distillation process.

The idea of the proposed method is to copy class centers from a teacher network to a student network and freeze class centers for the whole distillation procedure where the student network is trained to get angles between given class centers and face embeddings the same as in the teacher network. It allows the student network to better reproduce the results of teacher network trained with the margin-based loss function.

The main contributions of our work are:

- We have proposed a novel method for the distillation of neural networks trained with the margin-based softmax.
- The proposed method allows reducing a gap between the teacher and student networks for face recognition problem. The accuracy of the mobile face recognition neural network achieved with our method exceeds other known distillation methods on different datasets: LFW [5], AgeDB-30 [6] and MageFace [7] dataset.
- In the presented work we made direct comparison of different distillation methods. The code for implemented methods and comparison experiments is available on the github.

II. Related Works

A. Margin-Based Softmax

There are several variations of the margin-based softmax used for training of neural networks for the face recognition problem. They include Cosface [4], Sphereface [3] and ArcFace [2] approaches which all can be described by the general formula:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(cos(\theta_{y_i} m_1 + m_2) - m_3)}}{e^{s(cos(\theta_{y_i} m_1 + m_2) - m_3))} + \sum_{j=1, j\neq y_i}^{n} e^{s\, cos\theta_j}}. \tag{1}$$

The listed methods are obtained from (1) by substitution of parameters. Sphereface: $m_1 = 4, m_2 = m_3 = 0$; Cosface: $m_1 = 1, m_2 = 0, m_3 = 0.35$; Arcface: $m_1 = 1, m_2 = 0.5, m_3 = 0$. The ArcFace approach for the face recognition task demonstrates the state-of-the-art performance on LFW, AgeDB-30 and MegaFace datasets.

B. Distillation

Knowledge distillation from a teacher network to a student network was proposed by Hinton *et al.* [8]. It is an approach for training a small student neural network by transferring knowledge from a large teacher network. The key idea of distillation proposed by Hinton is to transfer the knowledge about smoothed probability distribution of the output layer from the teacher network to the student network.

Some researchers continue to develop the idea of using a smoothed probability distribution as labels for training a student network. For example, Fukuda *et al.* [9] proposed an approach to distil an ensemble of neural networks into a single student network. Sau & Balasubramanian [10] proposed a regularization method that allows training a student network with a noisy teacher. Furlanello *et al.* [11] trained the student network parametrized identically to the teacher network.

Svitov David and Alyamkin Sergey are with Expasoft LLC, Novosibirsk, Russia. (e-mail: d.svitov@expasoft.tech; s.alyamkin@expasoft.com).

Svitov David is under postgraduate in the Institute of Automation and Electrometry of the Siberian Branch of the Russian Academy of Sciences.

Another approach to the knowledge transfer is distillation of hidden layers. Huang *et al.* [12] trained the student network to reproduce the distribution of weights on the hidden layers of the teacher network. Romero *et al.* [13] used the outputs of intermediate layers of teacher and student network in distillation procedure to regularize training. Chen *et al.* [14] used preservation of local object relationships for regularization. In their work, L_2 distances between feature vectors of the student network are minimized depending on the distance between the corresponding vectors in the teacher network. Wonpyo *et al.* [15] proposed relational knowledge distillation that penalizes structural differences in the samples relations.

For training lightweight face recognition neural networks with the margin-based softmax, the following distillation methods are used: Triplet distillation [16], Angular distillation [17] and Margin Based Knowledge Distillation [18].

In the *Triplet distillation* approach, the student neural network is trained with a triplet loss function and margin. Where the margin is calculated based on the distances between the anchor and negative and the anchor and positive examples predicted by the teacher network.

Angular distillation approach minimizes the angle between the teacher and student embedding vectors for each sample.

In the *Margin Based Knowledge Distillation* it is proposed to distil the knowledge via smooth probability distribution obtained from (1) via dividing by the temperature value T.

III. Proposed Approach

A. Teacher and Student Networks

In our approach, the ResNet100 [19] architecture was chosen as a teacher network. It has a large number of parameters and helps to achieve a high accuracy on face recognition tasks. The novel lightweight architecture called MobileFaceNet(ReLU) [1] was used as the student network. In our experiments we made one modification of MobileFaceNet architecture: the dimension of the embedding vector was increased to 512 to make it compatible with ResNet-100's embeddings. Table I shows comparison of parameters for the teacher and student networks.

TABLE I
Parameters of the Considered Networks

	ResNet100	MobileFaceNet
FLOPs / 10^9	24.2	0.44
Size / MB	261.2	5.3
Number of parameters / 10^6	52.56	1.19
Time / ms	401 ± 25.7	42.2 ± 5.48

Network run time was measured for $112 \times 112 \times 3$ input images on a machine with the processor: Intel Xeon(R) CPU E3-1270 v3 @ 3.50GHz × 8.

B. Margin Distillation

Let $x_{s_i} \in R^D$ denote the feature vector of student network for the sample with number i, $x_{t_i} \in R^D$ denotes the feature vector of teacher network for the same sample. We will denote the weight matrices of the last layer of student and teacher networks respectively by $W_s \in R^{D \times n}$ and $W_t \in R^{D \times n}$. The column with the index j corresponding to the center of the class y_i will be denoted by $W_{s_j} \in R^D$ and $W_{t_j} \in R^D$ for the student and teacher networks.

Methods based on adding the margin m to the $softmax$ function normalize the weight matrix and sample vectors by 1: $||W_j|| = 1$ and $||x_i|| = 1$. This normalization allows considering the output of the $logit$ layer as the cosine of the angles between the sample vectors and corresponding class centers: ${W_j}^T x_i = ||W_j|| \cdot ||x_i|| cos(\theta_j) = cos(\theta_j)$. We will consider ArcFace [2] as a special case of the margin-based softmax approach since it gives the best performance among the margin-based methods:

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\, cos\theta_j}}. \tag{2}$$

In ArcFace, the margin m is fixed at 0.5. We propose to distil the knowledge from the teacher network by calculating the margin values m for each sample i. The proposed distillation method contains two key ideas:

- Class centers found by the teacher network are used for the student network: $W_s = W_t$. Since the class centers are learning values, a deeper network is able to learn more optimal position of classes on the hypersphere.
- The calculated margin values m_i are used for distillation. They explicitly control the distance between vectors x_{s_i} and corresponding class centers W_{s_j}: larger m_i leads to the stronger attraction of the vector x_{s_i} to the class center. It is proposed to calculate m_i based on the information from the teacher.

The intuition behind the proposed method is to pull feature vectors and class center closer to each other for the student network when these vectors are close for the teacher network. It allows the knowledge transfer from the teacher to the student to be more efficient, because the student network focuses on samples with more confident predictions while paying less attention to samples with the low confidence.

Margins m_i are calculated based on the angle between $x_{t_i} \in R^D$ and $W_{t_j} \in R^D$. In other words, margin values are calculated based on the angle between the center of the class y_i and the sample vector i in the teacher network. Margin m for the sample with the index i is calculated similarly to the triplet distillation margin:

$$m_i = \frac{m_{max} - m_{min}}{a_{max}} a_i + m_{min}, \tag{3}$$

$$a_i = \frac{W_{t_j}^T x_{t_i}}{||W_{t_j}|| \cdot ||x_{t_i}||}, \tag{4}$$

where we fix $m_{max} = 0.5$ and $m_{min} = 0.2$ - maximum and minimum margin values. And a_{max} takes the value of the largest angle a in the mini-batch.

Our approach to the distillation allows transmitting the information about relative vectors position on the hypersphere without imposing the strict limitation on student feature vectors.

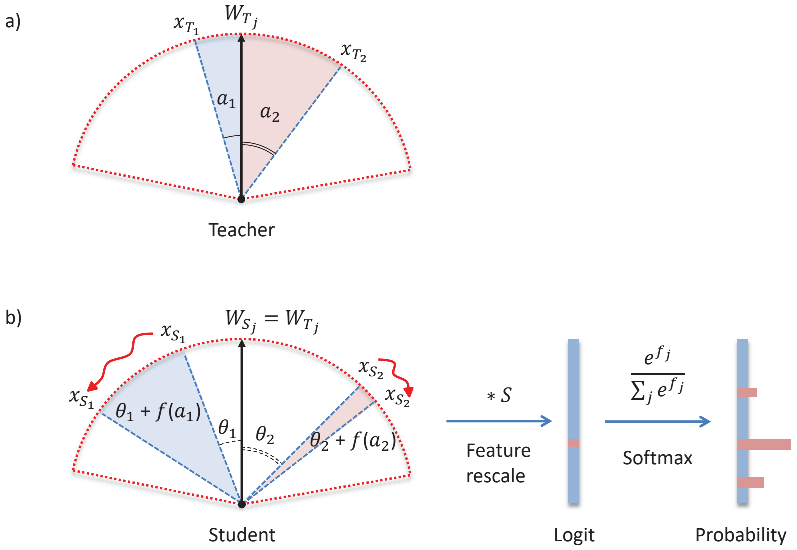

Fig. 1 a) Using (4) from the center of W_{t_j} class and the sample vector x_{t_i} of the teacher network, the angle a_i is calculated. b) The angle a_i is obtained from the teacher network and is used to calculate the margin $f(a_i)$ according to (3). The smaller is the angle between the sample vector and the center of the corresponding class in the teacher network, the greater is the margin when training the student network. The margin calculated this way is used as m in ArcFace (2). The larger is the margin, the stronger the student network pulls the feature vector towards the center of the class

IV. Experiments

A. Implementation Details

Pre-processing. We used MTCNN method [20] for detection and alignment of a face in an image. For training procedure MS1MV2 [2] was used. It is semi-automatic cleaned MS-Celeb-1M [21] dataset proposed by ArcFace authors.

After making the face alignment using the key points obtained by MTCNN, the images were cropped to 112 × 112 pixels. The pixel values were normalized to the range [-1, 1].

Training. In order to get a teacher network, the ResNet100 was trained with the ArcFace loss function. All distillation methods were compared in the same scenario where the knowledge was transferred from the trained ResNet100 to MobileFaceNet(ReLU). We used the following setup for distillation by our approach: mini-batch size was 512, learning rate was set to 0.1 and decreased 10 times by 100'000, 160'000 and 220'000 iterations. Training was performed by the SGD algorithm with a momentum of 0.9 and weight decay of $5e{-}4$. The values of maximum and minimum possible margins were fixed as: $m_{max} = 0.5$ and $m_{min} = 0.2$. Scale factor s was set to 64 as in ArcFace. For training MarginDistillation, a modification of the official implementation of ArcFace on MXNet was used.

In order to compare MarginDistillation with other distillation methods for margin-based softmax proposed previously, we imlemented the Triplet distillation [16], Angular distillation for feature direction [17] and Margin Based Knowledge Distillation [18] on MXNet. These methods were trained with the parameters recommended in the corresponding papers. The source code can be found in the article repository.

Evaluation. *LFW, AgeDB-30:* The considered datasets are widely used in the evaluation tasks for facial verification algorithms. They contain about 3000 positive and 3000 negative pairs of examples. At the testing stage, the trained network was used to obtain a feature vector for a pre-processed image of the face and its horizontal flip copy. Both vectors were then concatenated. The resulting vector was used for verification. The accuracy was measured as the percentage of correctly verified pairs of examples.

MegaFace: It is the most representative and challenging open testing protocol for the face recognition task. MegaFace includes 1 million facial images for 690'000 people, as a sample for the formation of distractors, and 100'000 for 530 people from the FaceScrub [22] dataset for identification. The measured metric is the top-1 accuracy for identification with 1 million distractors.

B. Experimental Results

TABLE II
Verification Accuracy at LFW and AgeDB-30

Architecture	Training method	LFW %	AgeDB-30 %
ResNet100 (teacher)	ArcFace [2]	99.76	98.21
MobileFaceNet (student)	ArcFace [2]	99.51	96.13
MobileFaceNet	triplet distillation L2 [16]	99.56	96.23
MobileFaceNet	triplet distillation cos [16]	99.55	95.60
MobileFaceNet	margin based with T=4 [18]	99.41	96.01
MobileFaceNet	angular distillation [17]	99.55	96.01
MobileFaceNet	MarginDistillation (our)	**99.61**	**96.55**

The experiments used the version of MobileFaceNet with ReLU.

As shown in Table II, the trained teacher network reaches 99.76% on LFW and 98.21% on AgeDB-30. The student network trained with ArcFace reaches 99.51% on LFW and 96.13% on AgeDB-30. Our approach gives the best results on AgeDB-30 reaching 96.55% and on LFW reaching 99.61%.

TABLE III
IDENTIFICATION ACCURACY USING MEGAFACE PROTOCOL WITH 1 MILLION DISTRACTORS

Architecture	Training method	MegaFace acc. %
ResNet100 (teacher)	ArcFace [2]	98.35
MobileFaceNet (student)	ArcFace [2]	90.62
MobileFaceNet	triplet distillation L2 [16]	89.10
MobileFaceNet	triplet distillation cos [16]	86.52
MobileFaceNet	margin based with T=4 [18]	90.77
MobileFaceNet	angular distillation [17]	90.73
MobileFaceNet	MarginDistillation (our)	**91.70**

The experiments used the version of MobileFaceNet with ReLU.

Since many algorithms show high accuracy on LFW dataset, it cannot be used to conclude which algorithm is suitable for usage in real life scenarios. MegaFace dataset is more challenging, which includes a much larger number of people and images. On the MegaFace dataset the teacher network reaches 98.35%. The student network trained with ArcFace reaches 90.62%. As shown in Table III, our method demonstrates the best accuracy of 91.70%. In Triplet distillation methods, some draw-down of accuracy was noted, although these methods demonstrated good results on LFW.

C. Ablation Study

For more detailed analysis of the proposed method, we conducted an ablation study where we measured an impact of each novelty proposed. In Table IV we examined the following aspects:

- Initialization by teacher - initialization of student class centers by teacher class centers. The last fully-connected layer of the student is initialized by weights of the teacher network: $W_s = W_t$.
- Usage of m_i instead of m - the usage of computed m_i for distillation instead of constant $m = 0.5$.
- Centers freezing - making student class centers untrainable.

The main increase of accuracy is observed when using the teacher centers for the student without the ability to modify them - 0.9%. Trained centers of student classes in all cases lead to degradation of performance on LFW dataset. So, in order to achieve the highest accuracy during the knowledge transfer from a teacher to student network, class center initialization plays a key role. The usage of adaptive margins m_i during distillation gives additional increase of accuracy.

TABLE IV
ABLATION STUDY OF THE PROPOSED METHOD ON LFW

Initialization by teacher	Using m_i instead of m	Centers freezing	LFW %
✓	✓	✓	**99.61**
✗	✓	✗	99.43
✓	✗	✓	99.60
✓	✓	✗	98.31
✓	✗	✗	99.55

V. CONCLUSION

A network distillation approach for the face recognition was introduced: it is MarginDistillation. We demonstrated the effectiveness of usage of class centers from the teacher network and teacher dependent margin to distil networks with the margin-based softmax. The proposed method was compared with other distillation methods and it demonstrated superior performance on LFW, AgeDB-30 and challenging Megaface datasets. Implementation of our method is available at: https://github.com/david-svitov/margindistillation

REFERENCES

[1] S. Chen, Y. Liu, X. Gao, and Z. Han, “Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices,” in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[3] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[5] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,” 2008.

[6] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “Agedb: the first manually collected, in-the-wild age database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–59.

[7] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, “The megaface benchmark: 1 million faces for recognition at scale,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.

[8] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[9] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, “Efficient knowledge distillation from an ensemble of teachers.” in *Interspeech*, 2017, pp. 3697–3701.

[10] B. B. Sau and V. N. Balasubramanian, “Deep model compression: Distilling knowledge from noisy teachers,” *arXiv preprint arXiv:1610.09650*, 2016.

[11] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born again neural networks,” *arXiv preprint arXiv:1805.04770*, 2018.

[12] Z. Huang and N. Wang, “Like what you like: Knowledge distill via neuron selectivity transfer,” *arXiv preprint arXiv:1707.01219*, 2017.

[13] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” *arXiv preprint arXiv:1412.6550*, 2014.

[14] H. Chen, Y. Wang, C. Xu, C. Xu, and D. Tao, “Learning student networks via feature embedding,” *arXiv preprint arXiv:1812.06597*, 2018.

[15] W. Park, D. Kim, Y. Lu, and M. Cho, “Relational knowledge distillation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[16] Y. Feng, H. Wang, R. Hu, and D. T. Yi, “Triplet distillation for deep face recognition,” *arXiv preprint arXiv:1905.04457*, 2019.

[17] C. N. Duong, K. Luu, K. G. Quach, and N. Le, “Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks,” *arXiv preprint arXiv:1905.10620*, 2019.

[18] D. Nekhaev, S. Milyaev, and I. Laptev, “Margin based knowledge distillation for mobile face recognition,” in *Twelfth International Conference on Machine Vision (ICMV 2019)*, vol. 11433. International Society for Optics and Photonics, 2020, p. 114330O.

[19] 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*, 2016, pp. 770–778.

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[21] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition,” in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[22] H.-W. Ng and S. Winkler, “A data-driven approach to cleaning large face datasets,” in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 343–347.