Marginal Loss for Deep Face Recognition

Jiankang Deng
Imperial College London
UK
j.deng16@imperial.ac.uk

Yuxiang Zhou
Imperial College London
UK
yuxiang.zhou10@imperial.ac.uk

Stefanos Zafeiriou
Imperial College London
UK
s.zafeiriou@imperial.ac.uk

Abstract

Convolutional neural networks have significantly boosted the performance of face recognition in recent years due to its high capacity in learning discriminative features. In order to enhance the discriminative power of the deeply learned features, we propose a new supervision signal named marginal loss for deep face recognition. Specifically, the marginal loss simultaneously minimises the intra-class variances as well as maximises the inter-class distances by focusing on the marginal samples. With the joint supervision of softmax loss and marginal loss, we can easily train a robust CNNs to obtain more discriminative deep features. Extensive experiments on several relevant face recognition benchmarks, Labelled Faces in the Wild (LFW), YouTube Faces (YTF), Cross-Age Celebrity Dataset (CACD), Age Database (AgeDB) and MegaFace Challenge, prove the effectiveness of the proposed marginal loss.

1. Introduction

Face representation through the deep convolutional network embedding is considered the state-of-the-art method for face verification, face clustering, and recognition [24, 18, 17]. The deep convolutional network is responsible for mapping the face image, typically after a pose normalisation step, into an embedding feature vector such that features of the same person have a small distance while features of different individuals have a considerable distance.

The various face recognition approaches by deep convolutional network embedding differ along three primary attributes. The first attribute is the training data employed to train the model. The identity number of public available training data, such as VGG-Face [17], CAISA-WebFace [30], MS-Celeb-1M [7], MegaFace [12], ranges from several thousand to half million. Although MS-Celeb-1M and MegaFace have a significant number of identities, they suffer from annotation noises [29] and long tail distribution [31]. By comparison, private training data of Google [18] even has several million identities. The second attribute is the network architecture. High capacity deep convolutional networks such as ResNet [8, 9, 27, 31] and Inception-ResNet [21] usually obtains better performance compared to VGG network [19, 17] and Google Inception V1 network [22, 18]. The third attribute is the design of loss function. The contrastive loss [20] and the triplet loss [18] utilise pair training strategy. The contrastive loss function consists of positive pairs and negative pairs. The gradients of the loss function pull together positive pairs and push apart negative pairs. Triplet loss minimises the distance between an anchor and a positive sample and maximises the distance between the anchor and a negative sample from a different identity. In [24] and [17], a classification layer is trained over a set of known identities. The feature vector is then taken from an intermediate layer of the network and used to generalise recognition beyond the set of identities used in training. The training procedure of the contrastive loss [20] and the triplet loss [18] is very tricky due to the selection of effective training samples. The classification-based methods [24, 17] suffer from massive GPU memory consumption on the classification layer when the identity number increases to million level, and prefer balanced and sufficient training data for each identity.

In this paper, we employ the large scale public available training data (MS-Celeb-1M) and the high capacity ResNet network structure. We propose a new supervision signal named marginal loss to enhance the discriminative power of the deeply learned features. Specifically, the marginal loss simultaneously minimises the intra-class variances as well as maximises the inter-class distances by focusing on the marginal samples. With the joint supervision of softmax loss and marginal loss, we can easily train a robust CNNs to obtain more discriminative deep features.

Our main contributions are summarised as follows.

1. We propose a new loss function (called marginal loss) to minimise the intra-class variances as well as maximise the inter-class distances of the deep features. With the joint supervision of the marginal loss and the

[0]https://github.com/davidsandberg/facenet

softmax loss, the highly discriminative features can be obtained for robust face recognition, as supported by our experimental results. We show that the proposed loss function is very easy to implement in the CNNs and can be directly optimised by the standard SGD.

2. Extensive experiments are conducted to prove the effectiveness of the proposed method on the public datasets. We verify the excellent performance of our new approach on Labelled Faces in the Wild (LFW) [10] and YouTube Faces (YTF) datasets [28]. The proposed method is robust under age variations and obtains state-of-the-art results on Cross-Age Celebrity Dataset (CACD) [3] and Age Database (AgeDB) [15]. The proposed marginal loss also achieves state-of-the-art results on MegaFace Challenge [12], which is the largest public face database with one million faces for recognition.

2. Related Work

To enhance the discriminative power of the deep features, Wen *et al*. [27] add a new supervision signal, called centre loss, to softmax loss for face recognition task. Specifically, the centre loss simultaneously learns a feature centre for each identity and penalises the distances between the deep features of examples and their corresponding feature centres. With the joint supervision of softmax loss and centre loss, this method can easily obtain inter-class dispersion and intra-class compactness. However, the on-line centre for each identity doubles the memory consumption of the last CNN layer.

To alleviate long tail distribution of the real-world face recognition training data, Zhang *et al*. [31] propose a new loss function called range loss to effectively utilise the entire long-tailed data in the training process. The optimisation objective of range loss is the k greatest ranges harmonic mean values in one class and the shortest inter-class distance within one batch. Both the range value and the centre value are calculated based on groups of samples, which alleviates unbalanced comparison times on the long tail training data. However, it is the softmax loss that most need uniform distribution across the classes, and the ability to increase inter-class differences within one mini-batch is limited because there are only four identities within each mini-batch as described in the original paper.

Tadmor *et al*. [23] point out that solving the classification-based training is easier than the comparative training, but the classification-based method requires a large training set because of the significant increase in the number of parameters due to the large classification layer. To speed up the convergence rate of stochastic gradient descent method, Tadmor *et al*. propose a multibatch method, which first generates features for a mini-batch of k face images and then constructs an unbiased estimate of the full gradient by relying on all $k^2 - k$ pairs from the mini-batch. The objective function requires features of positive pairs to be below a global threshold while features of negative pairs should be above the global threshold. Compared to the standard gradient estimator that relies on random $k/2$ pairs and has a variance of order $1/k$, the multibatch method is bounded by $O(1/k^2)$, both in theory and in practice.

3. Marginal Loss

The most widely used classification loss function, softmax loss, is presented as follows:

$$L_s = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the deep feature of the i-th samples, belonging to the y_i-th identity. The feature dimension d is set as 512 in this paper following [27, 31]. $W_j \in \mathbb{R}^d$ denotes the j-th column of the weights $W \in \mathbb{R}^{d \times n}$ in the last fully connected layer and $b \in \mathbb{R}^n$ is the bias term. The batch size and the identity number is m and n, respectively.

In order to enhance the discriminative power of the deeply learnt features, we propose the marginal loss function to minimise the intra-class variations and keep inter-classes distances within the batch. When x_i and x_j are from the same class, their distance $\|x_i - x_j\|_2^2$ should be close up to the threshold θ. By comparison, when x_i and x_j are from the different class, their distance $\|x_i - x_j\|_2^2$ should be farther than the threshold θ.

$$L_m = \frac{1}{m^2 - m} \sum_{i,j,i \neq j}^{m} \left(\xi - y_{ij} \left(\theta - \left\| \frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|} \right\|_2^2 \right) \right)_+, \tag{2}$$

where $y_{ij} \in \{\pm 1\}$ indicates whether the faces x_i, and x_j are from the same class or not, $(u)_+ := max(u, 0)$ is the marginal (hinge) loss [5], θ is the threshold to distinguish whether the faces are from the same person or not, and ξ is the error margin besides the classification hyperplane.

In Figure 1, we give a toy example to illustrate the proposed marginal loss. As we can see, the farthest intra-class samples and the nearest inter-class samples are selected to compute the loss, which can decrease the intra-class variances as well as keep inter-class distances. Due to the limitation of the GPU memory, the number of different identities within one batch is set as 16 in this paper. If we randomly select 16 individuals and 16 images per individual, the marginal loss looks like the matrix shown in Figure2(a). The gray-scale value indicates the loss value. The diagonal blocks are the penalty on intra-class variations, and the off-diagonal values are the penalty on inter-class confusion. As we can see, the inter-class loss is not as obvious as the intra-

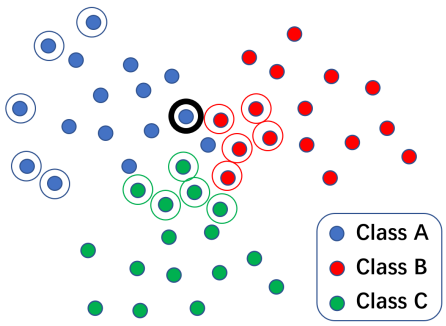


Figure 1. A simulated 2-D feature distribution graph in one batch. We only show three classes in this batch. For a particular sample from class A (marked by the black circle), the farthest intra-class samples (marked by blue circles) and the nearest inter-class samples (marked by green and red circles) are selected to compute the marginal loss. The farthest intra-class samples and the nearest inter-class samples are both called marginal samples, which are distributed around the decision boundary. The objective of marginal loss is to decrease the intra-class variances as well as keep inter-class distances.

class loss, and the marginal loss degenerates into a similar formulation to only control intra-class variances as the centre loss [27]. Since similar persons hardly get together randomly, we calculate the feature centre for each identity in an off-line way and re-rank the reading sequence of the training data. For each step, we randomly select one identity and its 15 nearest neighbour identities according to the off-the-shelf feature centres, which can increase the probability of the effective inter-class marginal loss, as is shown in Figure 2(b). The identity feature centres are updated after each identity has been selected for training. Compared to the range loss [31], the proposed marginal loss is also calculated based on groups of samples, but easier to implement.

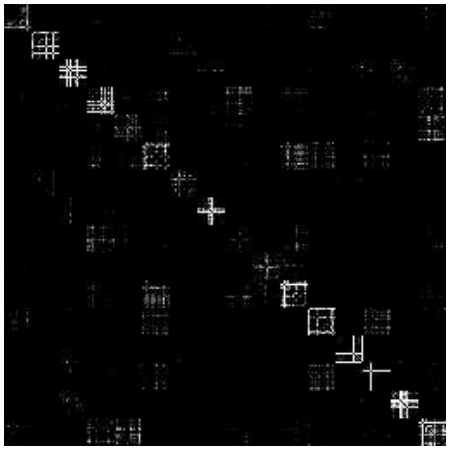
We adopt the joint supervision of softmax loss and marginal loss to train the CNNs for discriminative feature learning.

$$L = L_s + \lambda L_m \tag{3}$$

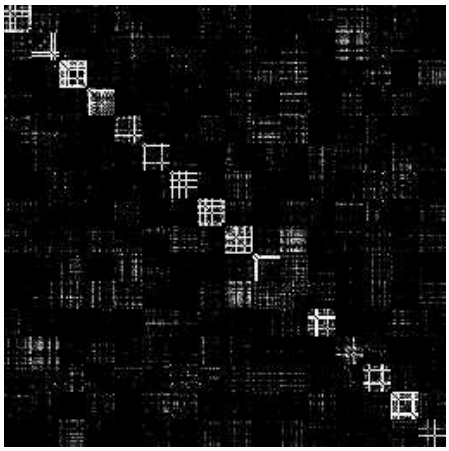
where λ is used for balancing the two loss functions. Clearly, the CNNs supervised by joint softmax and marginal loss are trainable and can be optimised by standard SGD. Compared to the multibatch method [23], the softmax loss provides separable features from a global view, and prevent the marginal loss degrading to zeros [27].

4. Experiments

Extensive experiments results are reported in this section on current popular and important face recognition benchmarks, including Labelled Faces in the Wild (LFW) [10], YouTube Faces (YTF) [28], Cross-Age Celebrity Dataset (CACD) [3], Age Database (AgeDB) [15]and MegaFace Challenge [12]. Our approach achieves state-of-the-art performance on LFW and YTF with significantly less training data (4M instead of 200M as Google FaceNet), while we outperform other algorithms on CACD and AgeDB by a notable amount. The proposed marginal loss also ranks 3rd under the large training data protocol on MegaFace Challenge, the largest face identification and verification benchmark.



(a) Random



(b) Nearest Neighbour

Figure 2. Marginal loss matrix by random identity selection and nearest neighbour selection according to feature centres. The nearest neighbour selection can increase the probability of the effective inter-class marginal loss.

4.1. Experiment Settings

Training data. We use the MS-Celeb-1M [7] dataset as our training data. To get a high-quality training data, we re-rank all face images of each identity by their distances to the identity centre [27]. For a particular identity, the face image whose feature vector is far from the identitys feature centre (normalised distance > 1.5) will be automatically removed. We further manually remove the obvious noisy data from

the farthest face images for each identity. After removing the images of identities appearing in testing datasets, we finally obtain a dataset which contains about 4M images of 82k unique identities.

Data prepossessing. To balance the training data for each identity, we employ faces synthesis method proposed in [14] to augment the training data. The generated face images introduce new intra-class facial appearance variations, including pose, shape and expression. The preferred choices for augmentation are faces closest to the identity centre, because augmentation on marginal faces may amplify the noise. After data augmentation, there are at least 100 face images for each identity. We use five facial landmarks (eye centres, nose tip and mouth corners) for similarity transformation to normalise the face images [27]. The faces are cropped to 112×96 RGB images. Following a previous convention, each pixel (ranged between $[0, 255]$) in RGB images is normalised by subtracting 127.5 then dividing by 128.

Network Settings. We implement the proposed marginal loss method in Tensorflow [2] with our modifications on multi-GPU implementation to hold more samples within one batch. We employ the ResNet structure with 27 convolutional layers [27, 31], but add batch normalisation layers [11]. We set 256 as the batch size, with 16 identities in one batch and 16 images per identities [23]. The learning rate is started from 0.01, and divided by 10 at the 80K, 160K, 240K iterations. Total iteration is 480K. The threshold θ and the error margin ξ in Equation 3 are set to 1.2 and 0.3, respectively. The in Equation 3 is set to 1.

Test settings. The deep features ($512d$) are taken from the output of the fully connected layer. The score is computed by the Cosine Distance of two features. Nearest neighbour and threshold comparison are used for both identification and verification tasks. Note that, we only use a single model for all the testing.

4.2. Experiments on the LFW and YTF datasets

We evaluate the proposed marginal loss model on two famous face recognition benchmarks, LFW (image) and YTF (video) datasets, under unconstrained environments. LFW dataset [10] contains $13,233$ web-collected images from 5749 different identities, with large variations in pose, expression and illuminations. Following the standard protocol of *unrestricted with labelled outside data*, we test on $6,000$ face pairs and report the experiment results in Table 1. YTF dataset [28] consists of $3,425$ videos of $1,595$ different people. The clip durations vary from 48 frames to $6,070$ frames, with an average length of 181.3 frames. We also follow the *unrestricted with labelled outside data* protocol and report the results on $5,000$ video pairs in Table 1.

In Table 1, we compare the proposed marginal loss method against many existing state-of-the-art models, including DeepID [20], VGG Face [17], Deep Face [24], Fusion [25], FaceNet [18], Baidu [13], Center Loss [27], Range Loss [31], Multibatch [23], Aug [14], and our baseline softmax loss. From the results, we can see that the joint softmax and marginal loss outperforms the softmax loss by a significant margin (from 98.87% to 99.48% in LFW and from 94.16% to 95.98% in YTF), which indicates that the proposed marginal loss can enhance the discriminative power of the deeply learned face features. Our method even obtains similar performance with FaceNet [18], which is trained on a largest private dataset with several million identities.

Methods	Images	LFW (%)	YTF (%)
DeepID [20]		99.47	93.20
VGG Face [17]	2.6M	98.95	97.30
Deep Face [24]	4M	97.35	91.40
Fusion [25]	500M	98.37	
FaceNet [18]	200M	99.63	95.10
Baidu [13]	1.3M	99.13	
Center Loss [27]	0.7M	99.28	94.9
Range Loss [31]	1.5M	99.52	93.70
Multibatch [23]	2.6M	98.8	
Aug [14]	0.5M	98.06	
Softmax Loss	4M	98.87	94.16
Marginal Loss	4M	99.48	95.98

Table 1. Verification performance of different methods on LFW and YTF datasets

4.3. Experiments on the CACD and AgeDB datasets

The CACD dataset is a face dataset for age-invariant face recognition, containing 163,446 images from 2,000 celebrities with labelled ages. It includes varying illumination, pose variation, and makeup to simulate practical scenario. However, the entire CACD dataset contains some incorrectly labelled samples and some duplicate images. Following the state-of-the-art configuration [3], we test the proposed method on a subset of CACD [3], CACD-VS, which consists of 4000 image pairs (2000 positive pairs and 2000 negative pairs) and have been carefully annotated.

We follow the ten-fold cross-validation rule to compute the face verification rate and compare our result with the existing methods in this dataset. As can be seen from Table 2, the proposed marginal loss significantly outperforms all the published results [4, 6, 3, 26] on this dataset, even surpassing the human-level performance with a clear margin. Even though our method has not explicitly trained on age-related data, the marginal loss is, to some extent, robust against age variations.

To further observe the influence of age variations on face recognition, we give the recognition results on the AgeDB.

Methods	Acc (%)
High-Dimensional LBP [4]	81.6
Hidden Factor Analysis [6]	84.4
Cross-Age Reference Coding [3]	87.6
LF-CNNs [26]	98.5
Human Average	85.7
Human Voting	94.2
Centre Loss [27]	97.475
Marginal Loss	98.95

Table 2. Verification performance of different methods on CACD

Methods	5 Yr	10 Yr	20 Yr	30 Yr
VGG Face [17]	93.15	92.18	89.15	85.08
Centre Loss [27]	95.93	95.15	93.07	90.72
Marginal Loss	98.12	97.95	97.15	95.75

Table 3. Verification performance (%) of different methods under different year gaps (5 years, 10 years, 20 years, and 30 years) on AgeDB

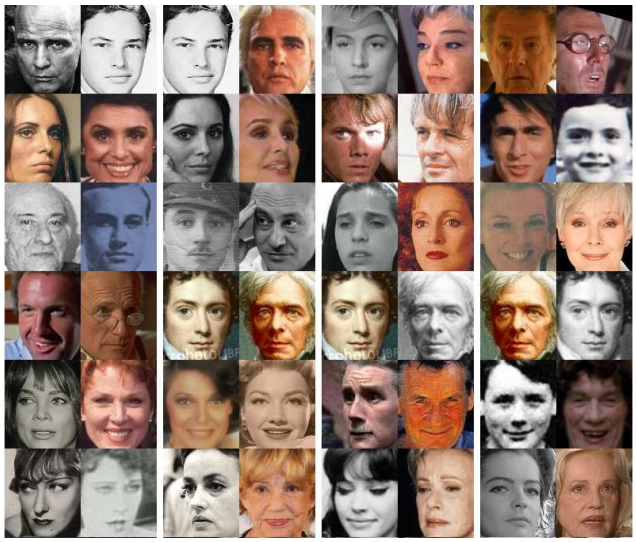


Figure 3. Positive sample pairs from AgeDB with the gap of 30 years. Facial appearances undergo dramatical changes in this time span.

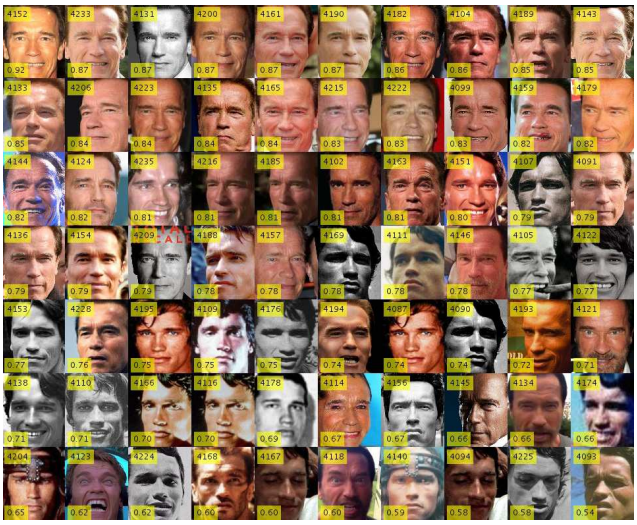
The experiments were conducted on a subset of the final version of AgeDB, as AgeDB was further extended by the time it became publicly available [15]. AgeDB is an in-the-wild dataset with large variations in pose, expression, illuminations, and age. AgeDB contains $12,240$ images of 440 distinct subjects, such as actors, actresses, writers, scientists, and politicians. Every image is annotated with respect to the identity, age and gender attribute. The minimum and maximum ages are 3 and 101, respectively. The average age range for each subject is 49 years. There are four groups of test data with different year gaps (5 years, 10 years, 20 years and 30 years, respectively). Each group has ten split of face images, and each split contains 300 positive examples and 300 negative examples. The face verification evaluation metric is just the same with LFW. In Table 3, we compare the proposed marginal loss to the baseline methods, VGG Face [17] and Centre Loss [27]. The proposed marginal loss significantly outperforms the baseline methods and can keep the high accuracy level even on the subset of 20 years gap. When the year gap increases, all of the methods experience a noticeable performance drop, which indicates that age variation is still very challenging especially when the year gap is larger than 30 years. Figure 3 shows some positive pairs from the subset of 30 years gap. As the face appearance changes dramatically, face verification on this dataset is very challenging.

4.4. Experiments on the Mega-face Challenge

MegaFace datasets [12] are recently released as the largest testing benchmark, which aims at evaluating the performance of face recognition algorithms at the million scale of distractors. MegaFace datasets include gallery set and probe set. The gallery set, a subset of Flickr photos from Yahoo, consists of more than one million images from 690K different individuals. The probe sets are two existing databases: Facescrub [16] and FGNet [1]. Facescrub is a publicly available dataset that containing 100K photos of 530 unique individuals, in which $55,742$ images are males, and $52,076$ images are females. FGNet is a face ageing dataset, with 1002 images from 82 identities. Each identity has multiple face images at different ages (ranging from 1 to 69).

In Figure 4, we give two example probe identities from FaceScrub and FGNet. We put the Cosine Distance between each face and the identity feature centre on the left bottom. There is cropped version of FaceScrub images provided by MegaFace, but our method uses a different crop size (112×96). As a result, we crop and rescale faces from the original images in FaceScrub. We also put the face ID number on the top left of each face image in order to refine the noisy faces. (We use the same visualisation method to manually refine our training data MS-Celeb-1M [7]. Annotators can learn the identity from the first several face images and remove the noisy and misaligned faces from the end.) Compared to FaceScrub face images, the faces from FG-Net exhibit larger appearance variations due to the long age span, especially for the childhood face images. The intra-class variance dramatically increases the difficulty when using the FG-Net faces as the probe images.

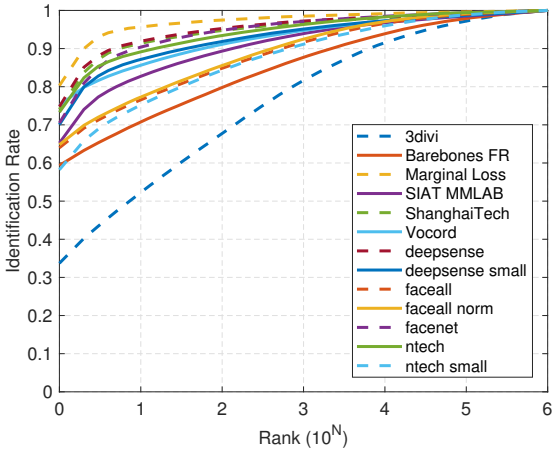
There are two testing scenarios (identification and verification) under two protocols (large or small training set). The training set is defined as large if it contains more than 0.5M images and 20K subjects. Our training data belongs



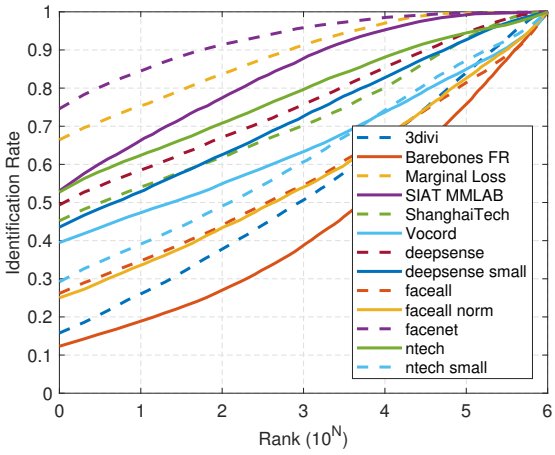
(a) FaceScrub

(b) FGNet

Figure 4. Example probe face images from FaceScrub dataset and FGNet dataset. Compared to FaceScrub face images, the faces from FG-Net exhibit larger appearance variations due to the long age span, especially for the childhood face images. The intra-class variance dramatically increases the difficulty when using the FG-Net faces as the probe images.



(a) FaceScrub



(b) FGNet

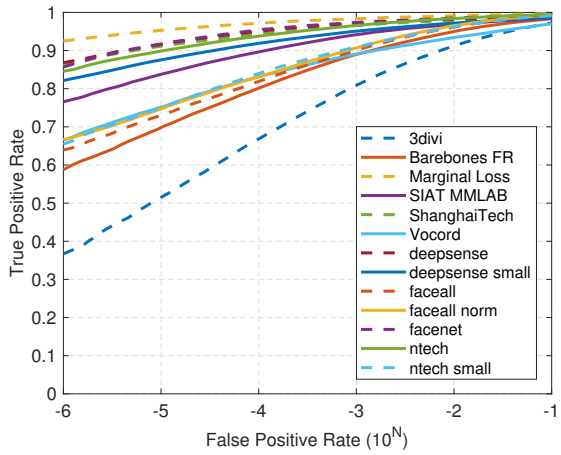
Figure 5. CMC curves of dierent methods with 1M distractors on Set 1. The results of other methods are provided by MegaFace team.

to the large training set, so we follow the protocol of large training set.

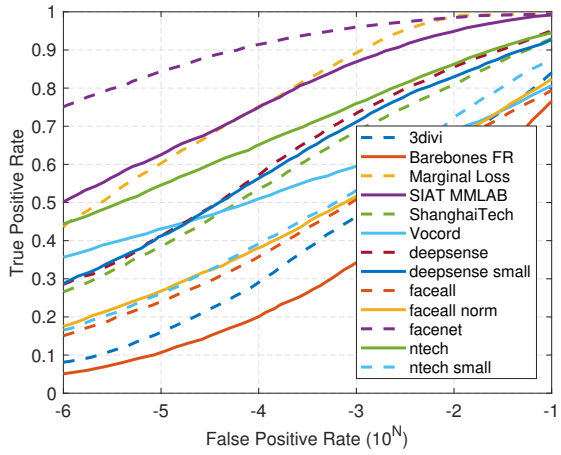
Face Identification. Face identification aims at matching a given probe image to the ones with the same person in the gallery. The MegaFace gallery contains different scale of distractors, from 10 to 1 million, leading to increasing challenge during testing. In face identification experiments, we present the results by Cumulative Match Characteristics (CMC) curves. It reveals the probability that a correct gallery image is ranked on top-K. The results are shown in Figure 5. When using FaceScrub as the probe set, the proposed marginal loss obtains the state-of-the-art result among the most recent public results. Taking FGNet as the probe set, Google FaceNet obtains the best result by a large margin, which indicates Google FaceNet is very robust under age variations.

Face Verification. For face verification, the algorithm should decide a given pair of images is the same person or not. 4 billion negative pairs between the probe and gallery datasets are produced in the MegaFace dataset. We compute the True Accept Rate (TAR) and False Accept Rate (FAR) and plot the Receiver Operating Characteristic (ROC) curves of different methods in Figure 6. When using FaceScrub as the probe set, the proposed marginal loss still obtains the state-of-the-art result among the most recent public results. Taking FGNet as the probe set, the proposed marginal loss witnesses a dramatical performance drop because our training data contains fewer age variations compared to Google's 500M private training data.

To meet the practical high precision demand, face recognition models should achieve high performance against millions of distractors. In this case, only Rank-1 identification rate with at least 1M distractors and verification rate at a



(a) FaceScrub



(b) FGNet

Figure 6. ROC curves of different methods with 1M distractors on Set 1. The results of other methods are provided by MegaFace team.

low false accept rate (*e.g.*,10^6) are very meaningful [12]. We include the top 3 methods from the latest MeagaFace Challenge1 leadboard under Large and Small protocol in Table 4 and Table 5.

From these results we have the following observations. First, large-scale training data usually perform better when taking FaceScrub and FGNet as the probe set. Second, FGNet dataset is more challenging than FaceScrub dataset. Most of the methods experience a dramatical performance drop, except for the Google FaceNet. Taking FaceScrub as the probe set, the proposed marginal loss ranks 3rd on both face identification and verification tasks under the large protocol. Our method even outperforms the Google FaceNet by 9.782% on face identification and 6.167% on face verification, respectively, which confirms the advantage of the proposed marginal loss. Taking FGNet as the probe set, our method still ranks 3rd on face identification task. Due to the limitation of age variation within our training data, the performance drop is inevitable.

Methods	protocol	Id (%)	Ver(%)
YouTu Lab	Large	**83.290**	91.340
DeepSense V2	Large	81.298	**95.993**
Vocord deepVo1.2	Large	80.258	77.143
Google FaceNet v8	Large	70.496	86.473
GRCCV	Small	77.677	74.887
SphereFace	Small	75.766	90.045
DeepSense	Small	70.983	82.851
Centre Loss [27]	Small	65.234	76.516
Marginal Loss	Large	80.278	92.640

Table 4. FaceScrub results: identification rates and verification TAR at 10^6 FAR of different methods on MegaFace.

Methods	protocol	Id (%)	Ver(%)
Google FaceNet v8	Large	**74.594**	**75.550**
SIATMMLAB	Large	71.247	67.954
DeepSense V2	Large	63.632	56.767
SIAT MMLAB	Small	55.304	50.144
SphereFace	Small	47.555	40.094
DeepSense	Small	43.540	29.610
Marginal Loss	Large	66.432	43.703

Table 5. FGNet results: identification rates and verification TAR at 10^6 FAR of different methods on MegaFace.

5. Conclusion

In this paper, we have proposed a new loss function, referred to as marginal loss. The marginal loss can minimise the intra-class variances as well as maximise the inter-class distances. By combining the marginal loss with the softmax loss to jointly supervise the learning of CNN, the discriminative power of the deeply learned features can be highly enhanced for robust face recognition. We utilise MS-Celeb-1M, a public available large-scale training data, to train the CNN model. Extensive experiments on several large scale face benchmarks, such as LFW, YTF, CACD, AgeDB, and MegaFace, have convincingly demonstrated the effectiveness of the proposed approach. Due to the limitation of the age variance in our training data, the proposed method experiences a performance drop when there is large year gap. In the future, we will try to propose an age-invariant method to alleviate this problem.

Acknowledgement Stefanos Zafeiriou was partially funded by EPSRC project EP/N007743/1 (FACER2VM), as well as by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520 (TeSLA).

References

[1] Fg-net aging database, www-prima.inrialpes.fr/fgnet/.2002. 5

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 4

[3] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer, 2014. 2, 3, 4, 5

[4] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3025–3032, 2013. 4, 5

[5] C. Gentile and M. K. Warmuth. Linear hinge loss and average margin. In *Advances in neural information processing systems*, volume 11, pages 225–231, 1998. 2

[6] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2872–2879, 2013. 4, 5

[7] 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, 5

[8] 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. 1

[9] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 1

[10] 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. 2, 3, 4

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[12] 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*, pages 4873–4882, 2016. 1, 2, 3, 5, 7

[13] 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. 4

[14] I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016. 4

[15] 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 Workshop*, 2017. 2, 3, 5

[16] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 343–347. IEEE, 2014. 5

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015. 1, 4, 5

[18] 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. 1, 4

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[20] 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. 1, 4

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 1

[22] 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. 1

[23] O. Tadmor, T. Rosenwein, S. Shalev-Shwartz, Y. Wexler, and A. Shashua. Learning a metric embedding for face recognition using the multibatch method. In *Advances In Neural Information Processing Systems*, pages 1388–1389, 2016. 2, 3, 4

[24] 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. 1, 4

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2746–2754, 2015. 4

[26] Y. Wen, Z. Li, and Y. Qiao. Latent factor guided convolutional neural networks for age-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4901, 2016. 4, 5

[27] 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. 1, 2, 3, 4, 5, 7

[28] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011. 2, 3, 4

[29] 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. 1

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

[31] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao. Range loss for deep face recognition with long-tail. *arXiv preprint arXiv:1611.08976*, 2016. 1, 2, 3, 4