More Information Supervised Probabilistic Deep Face Embedding Learning

Ying Huang [*1] **Shangfeng Qiu** [*1] **Wenwei Zhang** [1] **Xianghui Luo** [1] **Jinzhuo Wang** [2]

Abstract

Researches using margin based comparison loss demonstrate the effectiveness of penalizing the distance between face feature and their corresponding class centers. Despite their popularity and excellent performance, they do not explicitly encourage the generic embedding learning for an open set recognition problem. In this paper, we analyse margin based softmax loss in probability view. With this perspective, we propose two general principles: 1) monotonically decreasing and 2) margin probability penalty, for designing new margin loss functions. Unlike methods optimized with single comparison metric, we provide a new perspective to treat open set face recognition as a problem of information transmission. And the generalization capability for face embedding is gained with more clean information. An auto-encoder architecture called Linear-Auto-TS-Encoder(LATSE) is proposed to corroborate this finding. Extensive experiments on several benchmarks demonstrate that LATSE help face embedding to gain more generalization capability and it boost the single model performance with open training dataset to more than 99% on MegaFace test.

1. Introduction

Face recognition performance has gained dramatic improvement in recent years. Margin based loss functions(Schroff et al., 2015; Liu et al., 2017b) play an important role in this process. The most common solutions treat face recognition as a classification problem. These works utilize deep convolutional network, such as VGGNet (Simonyan & Zisserman, 2014) or ResNet (He et al., 2016), to transfer the landmark aligned face images to their corresponding class label. And comparison loss function(Hadsell et al., 2006) is employed in the learning process, such as softmax cross entropy loss. Recent researches demonstrate that we can gain more discriminative face embedding by explicitly adding margin to loss. FaceNet (Schroff et al., 2015) adds margin by penalizing the distance between face feature and their corresponding class centers in euclidean space (triplet loss). While SphereFace (Liu et al., 2017b) penalizes this distance in hyperspherical angle space (angular margin softmax loss).

Figure 1. Examples of generated frontal face from different identity. In rows top to bottom: original input face image, the generated frontal face image from the corresponding feature, the absolute difference between input image and the generated one.

However, all these methods strengthen the embedding learning process from the perspective of comparison. So these recognition algorithms optimize single discriminative ability metric. At the same time, they do not give a general guidance for new margin loss function designing.

While robust face embedding learning is an open set problem(Geng et al., 2018; Boult et al., 2019). We can not involve images of all possible human identities in the training dataset. So we seek to solve some primary questions: *How to gain more discriminative power with appropriate margin definition? Whether discriminative ability equals generative capability in open set situation?*

Following these questions, we firstly explore the effect of margin in softmax cross entropy loss (Liu et al., 2016; Wang et al., 2018b; Deng et al., 2019a) from probabilistic view. Through detailed analysis, this work proposes two principles to guide the definition of new margin loss function. One is monotonically decreasing. Let θ be the angle between face feature and their class center. The definition of transformation function, which transforms θ to probability value,

[*]Equal contribution [1]Guangzhou Huya Technology Co., Ltd, Guangzhou, Guangdong, China [2]Department of Engineering Science, University of Oxford, Oxford, Oxfordshire, United Kingdom. Correspondence to: Ying Huang <huanying@huya.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

should monotonically decrease in the domain of θ. The other is margin probability penalty. New margin loss should guarantee a non-negative probability penalty following this principle.

And we find single comparison metric may not direct the model to gain optimally generalized face embedding. In contrast to these single-metric-based methods, we investigate a different way. In our work, we treat face recognition as an information transmission problem, which can be supervised by class label. Different from the classification task which only strengthens the discriminative power, information transmission can be seen as a multi-objective optimization problem. One optimization objective is comparison-metric-based discriminative embedding learning. While the other objective is that how much accurate information is passed to the face feature.

Base on this motivation, we built the learning framework as an auto-encoder (Hinton & Salakhutdinov, 2006) architecture, which can be trained with a teacher student (Hinton et al., 2015; Liu et al., 2017a) learning strategy. This Linear-Auto-TS-Encoder(LATSE) architecture can guarantee the proper information pass through the whole network. Different parts in framework promise distinct effects for the embedding learning. In the decode part is a generative network (Goodfellow et al., 2014; Zhao et al., 2016) which transfers face embedding to the corresponding individual frontal face image. This generative model is supervised by pixel level image label. The generated frontal faces are shown in Figure 1. At the encode part, we employ a deep ResNet to learn face embedding. A new definition of margin loss following the proposed principle plays the role as comparison loss. The goal of teacher network is filtering noise and ensuring cleaner information for face embedding learning. The effects of these parts are discussed in details below. We conduct extensive experiments on large scale face recognition benchmarks. The experimental results verify our findings and the effectiveness of the proposed architecture.

2. Related Work

Margin Based Comparison Loss. Learning face embedding with comparison loss can be divided into two main streams. One stream methods directly obtain face embedding from the raw image through comparing match/non-match pairs, such as triplet loss (Schroff et al., 2015). The other methods first train a multi-class CNN using margin-based softmax loss function, such as large-margin softmax loss (Liu et al., 2016) or arccos loss (Deng et al., 2019a). Then the feature layer before softmax is used as face embedding. Recent works mainly focus on how to adjust margin or other hyper parameters. Liu (Liu et al., 2019) thinks that the margin between unbalance face classes should be adaptively learned during the training process. RegularFace (Zhao et al., 2019) is proposed which explicitly penalize the angle between an identity and its nearest neighbor in order to increase the inter-class separability. AdaCos (Zhang et al., 2019) is proposed to tune the margin and scale parameter automatically to strengthen the discriminative power for face embedding. However, the single optimization metric, which enhances the intra-class compactness and inter-class dispersion, in all these methods may ignore the nature of open set recognition problem.

Generative Model. Generative network learning usually starts with a latent code from noise distribution. Then the model maps this code to a generative sample, such as probabilistic GAN (Goodfellow et al., 2014) and Energy Based GAN (Zhao et al., 2016). In Cycle GAN (Zhu et al., 2017), a cycle-consistent loss is proposed to learn from unpaired images. However, few work explores whether a generative model can enhance the embedding generalization ability in the open set recognition situation.

Teacher Student (TS) Learning Strategy. Noises are inevitable in large-scale datasets and heavily affect the performance of face recognition algorithms(Wang et al., 2018a). It is expensive to get extremely clean large-scale datasets. So how to learn with noises plays a significant role for model training. MentorNet (Jiang et al., 2017) utilizes a pre-trained teacher network to drop and update corrupted labels for the student network in learning process. De-Coupling (Malach & Shalev-Shwartz, 2017) alleviates influence of noises through updating the parameters only when the predictions from two classfiers are same. Co-Mining (Wang et al., 2019) simultaneously traines two peer networks to redistribute the labels of raw data. Despite these methods have made efforts to reduce the corruption of noisy labels. There are still drawbacks to be improved, such as careful design for face recognition with large number of classes, less resource consumption and easier training process.

Multi-Modal Supervision For CNN Models. CNN for Image recognition is generally trained with image level semantic labels (Krizhevsky et al., 2012). While several researches demonstrates that multi-modal supervision can strengthen the generative ability of the learned models in different types of tasks. MTCNN (Zhang et al., 2016) and RetinaFace (Deng et al., 2019b) find that face detection performance can be boosted by using face landmarks along with bounding boxes label. Mask-rcnn (He et al., 2017) boosts the performance for the task of instance segmentation, bounding-box object detection, and person keypoint detection by utilizing object level bounding boxes and pixel level semantic labels together. Depth and time related information is used in Depth-Patch-Net (Atoum et al., 2017) and AuxNet (Liu et al., 2018b) to improve the accuracy for the task of anti-face spoofing. However, there is few work to explore whether face recognition performance can be

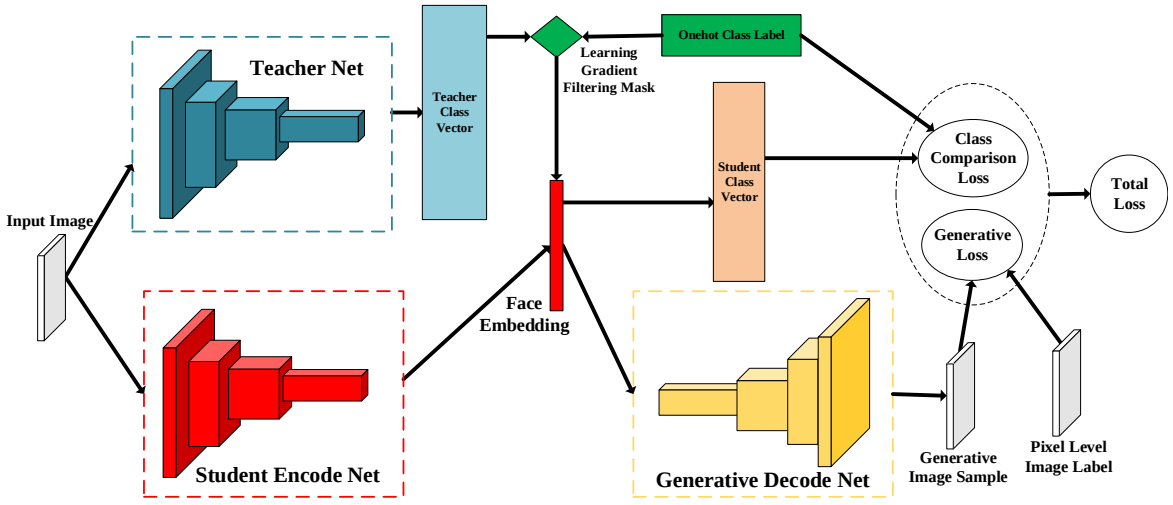

Figure 2. The proposed Linear-Auto-TS-Encoder(LATSE) Architecture. The red bottom part is the student encode net. While the yellow part is generative decode net. The top blue part is teacher network. A linear comparison loss and a pixel level generative loss are employed for final face embedding learning.

improved by extra supervision. In this work, we leverage pixel level face image as extra supervision to improve face recognition accuracy.

3. Learning Face Feature Embedding with Multi-Modal Supervision

The intuition behind this work is more orthogonal information lead to less uncertainty. And we treat face recognition as a problem of information transmission. Following this guide, we build our learning architecture as an auto-encoder framework which is trained with a teacher student learning strategy. This architecture is illustrated in Figure 2. Although there are three key components in the proposed algorithm. The model can be trained end-to-end from scratch data. In the following paragraphs, we will first illustrate the probability theory behind margin comparison loss. Then details for multi-level supervision are gave. After that is teacher student learning strategy and whole network training algorithm.

3.1. Margin Softmax Loss in Probability View

There is a line of research to add margin in softmax cross entropy loss. Previous works illustrate their theory by giving us a geometric interpretation (Liu et al., 2016; Deng et al., 2019a). In this paper, we try to explain the effect of margin in probability view. At the same time, we summarize two principles which can be followed when new margin loss is needed.

Firstly, we starts from the most widely used softmax loss function. Let x_i represent the extracted feature for a face image from identity y_t. W_{y_t} and W_j are the model weights in the fully-connected classifier layer for identity y_t and y_j. b_{y_t}, b_j are the biases. N is the total image number and K is the total category number in the training set. Then the softmax loss function can be presented as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_t}^T x_i + b_{y_t}}}{\sum_{j=1}^{K} e^{W_j^T x_i + b_j}} \tag{1}$$

This loss is consisted of two components and each part plays a distinct role in model optimization. One is the softmax function which transforms the predicted value from fully-connected layer to the probability of corresponding class. The other is a cross entropy loss to measure the difference between predicted probability and the given label distribution. Then we can decouple the softmax loss in Equation 1 to these parts. For an input image x, the softmax function computing the probability for the target label y_t can be presented as:

$$P_{softmax}(x_i, y_t) = \frac{e^{W_{y_t}^T x_i + b_{y_t}}}{\sum_{j=1}^{K} e^{W_j^T x_i + b_j}} \tag{2}$$

Let $q(x)$ represent the real data distribution and $p(x)$ be the model predicted probability distribution. Then the cross entropy loss computes the mutual information $H(p, q)$ between them, which can be presented as:

$$CELoss = H(p, q) = -\sum_{x} q(x)\log p(x) \tag{3}$$

Following the researches for margin softmax loss (Liu et al., 2017b; Deng et al., 2019a), we do not modify the format

of cross entropy loss. And we adjust margin in softmax function. For simplicity, we fix the bias term $b_j = 0$ as in (Liu et al., 2016), normalize the feature $\|x_i\| = 1$ and classifier layer weight $\|W_j\| = 1$. Then the predicted value from fully-connected layer $W_j^T x_i + b_j$ is simplified as $\|x_i\|\|W_j\| \cos\theta$. After that, the output is multiplied with a scale parameter s. At this time, softmax function is formulated as:

$$P_\theta(x_i, y_t) = \frac{e^{s(\|x_i\|\|W_{y_t}\| \cos\theta_{y_t})}}{\sum_{j=1}^{K} e^{s(\|x_i\|\|W_j\| \cos\theta_j)}} = \frac{e^{s\cos\theta_{y_t}}}{\sum_{j=1}^{K} e^{s\cos\theta_j}} \tag{4}$$

where θ_j is the angle between the predicted face feature and its normalized class center. In previous works (Liu et al., 2018a; Chen et al., 2019), the angle θ is demonstrated to be correlated to visual semantic. Decreasing this angle can boost the discriminative ability for the learned model. Margin based softmax methods try to multiple (m_1) or add positive margin value (m_2 or m_3) in the target label, which is expressed in format:

$$P_m(x_i, y_t) = \frac{e^{s(\cos(m_1\theta_{y_t}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{y_t}+m_2)-m_3)} + \sum_{j\neq y_t}^{K} e^{s\cos\theta_j}} \tag{5}$$

Rethinking the effect of softmax function, it normalizes the predicted value $W_j^T x_i + b_j$ from the fully-connected layer and transfers this value to the probability of the corresponding class. Then we can define a general function F to normalize the predicted value:

$$F(x_i, y_t) = \frac{\phi(x_i, W_{y_t})}{\sum_{j=1}^{K} \phi(x_i, W_j)} \tag{6}$$

Where $\phi(x, W) >= 0$ should always hold for any input pair (x, W). At the same time, the definition of this generalized function F must satisfy probability law. While in softmax function, it employs $\phi(x, W) = e^{W^T x}$ as the definition of ϕ. By adding margin m in this normalization function F, we can get F_m :

$$F_m(x_i, y_t) = \frac{\chi(x_i, W_{y_t}, m)}{\chi(x_i, W_{y_t}, m) + \sum_{j\neq y_t}^{K} \phi(x_i, W_j)} \tag{7}$$

Where χ and ϕ are general functions to transform fully-connected layer output to a non-negative value.

The first principle for the definition of χ and ϕ is non-negative and to penalize probability when margin is added. This principle is formulated as:

$$\phi(x, W) \geq \chi(x, W, m) \geq 0 \tag{8}$$

When we have non-negative margin value m_1 and m_2, χ should be monotonically decreasing when the margin increases.

$$\chi(x, W, m_1) \leq \chi(x, W, m_2), m_1 \geq m_2 \geq 0 \tag{9}$$

Margin based loss decreases the probability of target label when the model predicts same angle between feature and class center, expressed as $F_m \leq F$.

Another principle for function χ and ϕ is to make them monotonic decrease in the domain of θ ($x_i W_j^T = \cos\theta_j$ when $\|x_i\| = \|W_j\| = 1$). If the model tries to get the same probability under a margin based function, it must step forward to decrease the angle between feature and class center. This constraint makes the learned model gain more discriminative power by learning small angle θ between similar inputs.

Following these proposed principles, we can generalize the margin based probability function F_m to any new formats. In this paper, we test our method with a linear definition under these principles, specialize $\chi = e^{h(\theta)}$ and $h(\theta) = s(-a\theta + b)$. We set $\phi = e^{g(\theta)}$ and $g(\theta) = s\cos\theta$. This specialization case can be formulated as:

$$P_{linear}(x_i, y_t) = \frac{e^{s(-a\theta_{y_t}+b)}}{e^{s(-a\theta_{y_t}+b)} + \sum_{j\neq y_t}^{K} e^{s\cos\theta_j}} \tag{10}$$

Finally, the $DLoss$ is computed by cross entropy loss function between model prediction P_{linear} and label probability q. This is defined as:

$$DLoss(x) = -\sum_x q(x) \log P_{linear}(x) \tag{11}$$

Visualize the effect of margin by target logit curve. Different kinds of margin try to decrease the predicted probability for ground truth by penalizing the target predicted value, namely increase the cross entropy loss value under the same angle θ. Compared to the origin softmax function, we can view their relations in Figure 3. Despite the similar goal for different margin losses, the proposed principles will lead us to define a better normalization function. They provide a guidance when problem occurs in model training. From target logit curve in the domain of θ, we can find SphereFace(Liu et al., 2017b) has a nonlinear penalty value

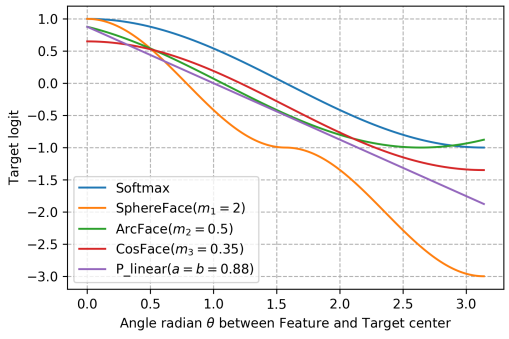

Figure 3. Target logit curves for different loss function

which increases along with angle θ. This leads divergence at the beginning of training when the initial angle θ is large. Arcface(Deng et al., 2019a) is not monotonic decreasing in the domain of θ, so it will shrink margin instead of magnifying it when angle is too large. CosFace(Wang et al., 2018b) is too smooth in the domain of $\theta \in (0, 0.5)$, meaning it will have little loss decay if the angle θ becomes smaller in this area. This makes it hard to shrink angle at the final stage of training. By contrast, the proposed linear format stably increases the target predicted value along with angle decreasing. This setting reduces the difficulty for the model to learn small angle θ between face embedding and its class center.

3.2. Pixel Level Supervision for Face Generative Net

Owing to single classification optimization metric (intra-class compactness and inter-class dispersion), margin based comparison loss encounters the bottleneck to gain more generalization ability for face embedding learning. We propose whether model can transfer more information to the face feature as another metric, which also should be optimized in the embedding learning process. So a generative network is added to realize this purpose with pixel level supervision. This generative part will try to restore the frontal face image from the embedding. Therefore the learned face embedding need to keep as much information as possible. Although we train this generative part with pixel level supervision. It is gained without cost of extra human labeling expect the identity class label for the whole image. The loss to supervise the generative model is consisted of two components. One is $L1Loss$ for frontal face regression. The other is $SimLoss$ correlated to structural similarity index (Wang et al., 2004), which compares the similarity between generative sample and the input image. Given an face image x^c for identity c. The generated face is x^c_{gen}. And the label for generative model y^c_{img} is the momentum mean image from c, which can be expressed as:

$$y^c_{img} = (1 - momentum) * x^c + momentum * y^c_{img} \quad (12)$$

Then We gain the definition of $SimLoss$:

$$SimLoss(x^c_{gen}, y^c_{img}) = \frac{1 - SSIM(x^c_{gen}, y^c_{img})}{2} \quad (13)$$

Next the total generative loss $GLoss$ is expressed as:

$$GLoss(x^c_{gen}, y^c_{img}) = L1Loss + SimLoss \quad (14)$$

The $L1loss$ suggest the model to transfer more information to embedding. While $SimLoss$ make the generated image to resemble the frontal face of the input identity.

3.3. Teacher Student Learning Strategy

Nowadays, most of large-scale datasets are obtained by utilizing search engine. Researchers apply automatic or semi-automatic approaches to clean the identity label for these datasets. This process leads to two types of noise : 1) label mess, which means images from same person may be marked with different labels. 2) distractor, which means some labeled examples are not part of any class within the dataset. Both types of noise will pollute the generalization ability of the learned face embedding. Despite current methods (Jiang et al., 2017; Malach & Shalev-Shwartz, 2017; Wang et al., 2019) have provided various strategies for model training in noisy dataset. Some of them are designed for binary classification and others may make training process too complicated. To solve these deficiencies and take into account the character of face embedding, we propose a teacher student learning strategy to filter label noise real-time in training process.

This strategy uses only the correct predictions gained by the pre-trained teacher model to update the student network, which guarantees training samples sufficiently clean with the prior knowledge of the teacher network. Recent researches (Yu et al., 2019; Han et al., 2018) mention the memorization effects of deep neural networks, which argue that networks would first memorize training data of clean labels. So we train a teacher network in the large-scale dataset to make the teacher network memorize data of clean labels as much as possible and then fix its parameters in next process. Then, the teacher network will filter noise for the student network, which would further strengthen representational ability of the face embedding. The filtering strategy of the teacher network is optional, which provides more flexibility. Besides, there is no need to simultaneously train the teacher and student model. This setting simplifies the procedure of training and achieves better results in consumption of training time and resource than other methods.

Intuitively, the more correct knowledge gained from teacher, student can form more distinct cognition about the problem. Equally like that in the training process, the face embedding learned by the student network would possess better discriminability and generalization with the clean data under the help of the auxiliary teacher network. Besides, in the view of information transmission, the teacher net plays role as high reliability signal channel between face image data distribution and identity label.

3.4. Whole network training

For the whole network training. First we gain a teacher model on training data. Then the parameters from teacher network are fixed. Next we estimate the student network and generative model on same dataset with $Loss$ formulated as:

$$Loss(x) = DLoss(x) + GLoss(x) \quad (15)$$

Finally the parameters from student model are saved for testing. The whole training process can be illustated by

Algorithm 1 Linear-Auto-TS-Encoder Learning algorithm

Input: face images X, face class label Y
First train $teacherNet$ by
for $i = 1$ **to** $maxiter$ **do**
  Learn $teacherNet$ with X and Y using $DLoss$
end for
Then fix parameters of $teacherNet$
Next Estimate $studentNet$, $genNet$ with X and Y by
for $i = 1$ **to** $maxiter$ **do**
  $prob_{yteacher} = teacherNet.forward(x_i)$
  $em_{student} = studentNet.forward(x_i)$
  $prob_{student} = FCLMSoftmax.forward(em_{student})$
  $DLoss = CELoss(prob_{student}, y_i)$
  $xgen_i = genNet.forward(em_{student})$
  $GLoss = GLossFun(xgen_i, x_i)$
  $Loss = DLoss + GLoss$
  $gradGLoss = genNet.backward(GLoss)$
  $gradDLoss = FCLMSoftmax.backward(DLoss)$
  $gradFromLoss = gradDLoss + gradGLoss$
  if $y_i \in topk(prob_{yteacher}, k)$ **then**
    $gradForEm = filter(gradFromLoss)$
  end if
  $studentNet.backward(gradForEm)$
end for
Finally save $studentNet$ parameters for testing

Algorithm1.

4. Experiments

4.1. Implementation Details

Datasets. We employed CASIA (Yi et al., 2014) as small training set and MS1MV2 (Guo et al., 2016) or MS1M-RetinaV (Deng et al., 2019a) from ArcFace (Deng et al., 2019a) as large training dataset. We compared the performance for both face verification and identification tasks on several benchmark datasets, including Labelled Faces in the Wild (LFW) (Huang et al., 2008), Celebrities in Frontal Profile (CFP-FP) (Sengupta et al., 2016), Age Database (AgeDB-30) (Moschoglou et al., 2017), Cross-Age LFW (CALFW) (Zheng et al., 2017), Cross-Pose LFW (CPLFW) (Zheng & Deng, 2018) and Megaface (Nech & Kemelmacher-Shlizerman, 2017).

Experimental Settings. The proposed method were implemented with MXNet (Chen et al., 2015). We reduced memory cost with the help of memonger (Chen et al., 2016). The data preprocessing step followed paper of margin based softmax loss (Liu et al., 2017b; Wang et al., 2018b; Deng et al., 2019a). The detected face were aligned by five facial key points and resized to fix dimension ($112 \times 112 \times 3$) as the network input.

ResNet (He et al., 2016) with depth of 34, 50, 100 and 124 were employed as the encode part in the auto-encoder, followed by a structure of Batch Normalization (Ioffe & Szegedy, 2015), Dropout, Fully-Connected layer and Batch Normalization to get the final 512 dimension face embedding. A reversed ResNet-18 with deconvolution layer (Noh et al., 2015) to up-sample feature map was adopted as the generative decode part. A same setting with the encode part was employed for teacher network. The parameters in teacher network were fixed during the student network training process.

For hyperparameter setting, we adopted $s = 64$ as the normalized scale in spherical manifold by following (Liu et al., 2017b). Models were trained in eight NVIDIA Tesla V100 GPUs(16GB) with total batch size 768. The learning rate started from 0.1 and was divide by 10 at 10K, 16K, 20K, 22k iterations. We set weight decay to $5e^{-4}$ and momentum to 0.9. At test time, we only computed the 512 dimension feature for each normalized face from the student network and compared feature cosine angle value as the similarity score between different face images.

Table 1. Face verification results (%) with different margin loss.(models trained on CASIA, ResNet50)

Loss Functions	LFW	CFP-FP	AgedDB-30
ArcFace (Deng et al., 2019a)	99.53	95.56	**95.15**
SphereFace (Liu et al., 2017b)	99.42	-	-
CosFace (Wang et al., 2018b)	99.51	95.44	94.56
LinearFace($a = 0.88, b = 0.88$)	**99.55**	**97.01**	94.66
LinearFace($a = 0.88, b = 1$)	99.48	96.81	95.05
LinearFace($a = 1, b = 1$)	99.48	96.80	94.40

4.2. Ablation Study of proposed key components

The performance with different margin loss. Firstly, we explored the effect of different margin losses. For fair comparison, the proposed LinearFace method was trained following the setting in (Deng et al., 2019a) with ResNet50 as embedding feature backbone on CASIA dataset. The verification results on LFW, CFP-FP and AgeDB-30 were reported to compare their performance in Table 1. From the result, we observed better accuracy on LFW and CFP-FP by employing the linear margin function compared to other margin softmax losses, which demonstrated the effectiveness of the proposed principle in Section 3.1.

The effectiveness of Teacher Student Strategy. To show the effectiveness of teacher student learning strategy, we conducted numerous validity experiments. All experiments employed MS1MV2 as training dataset and Arcface loss as training loss to guarantee fairness of comparison. We set the model performance from (Deng et al., 2019a) as baseline.

Table 2. Face verification results (%) with Teacher Student learning strategy

Method	LFW	CALFW	CFP-FP
resnet34	99.65	95.85	92.12
resnet34(k=1)	**99.76**	96.03	97.23
resnet34(k=3)	**99.76**	**96.05**	**97.27**
resnet50	99.80	95.80	92.74
resnet50(k=1)	**99.81**	95.93	98.25
resnet50(k=3)	**99.81**	**95.95**	**98.30**
resnet100	99.77	**96.10**	98.27
resnet100(k=1)	**99.83**	**96.10**	98.64
resnet100(k=3)	**99.83**	**96.10**	**98.71**

Table 3. Megaface performance (%) with Teacher Student learning strategy

Method	Identification	Verification
resnet34	96.09	96.72
resnet34(k=1)	97.61	98.03
resnet34(k=3)	**97.74**	**98.03**
resnet50	97.26	97.62
resnet50(k=1)	98.26	98.48
resnet50(k=3)	**98.33**	**98.55**
resnet100	98.35	98.6
resnet100(k=1)	98.56	98.8
resnet100(k=3)	**98.70**	**98.83**

Figure 4. Samples selected by teacher network. Examples from same row were marked with same label. The leftmost column: data of true label. The other columns: data of noisy labels. The student network actually avoid pollution of noisy labels with the selection of teacher network.

Table 4. The effect with different proposed component

Component	CALFW	CPLFW	AgedDB-30
LinearFace	93.3	89.08	94.66
LinearFace+TS	93.73	89.75	95.05
LinearFace+Gen	94.3	90.13	**95.91**
LinearFace+TS+Gen	**94.58**	**90.33**	**95.91**

The performance for student network was explored with setting different k values of teacher network. We assume the image label was correct if it was included in teacher network predicted top k probabilities. Only when image label fell in teacher's predicted top k classes, the gradient would pass to student network for updating its parameters. As shown in Table 2 , our teacher student learning strategy could be applied effectively with various network depth(34,50 and 100). The strategy not only obtained better results in several benchmarks, especially got more than 6% promotion in CFP-FP when depth was 50. But also outperformed baseline for large-scale evaluation set. Results in Table 3 verified that our strategy strengthen the generalization ability of face embedding. Besides, we made visualization of clean training samples provided by teacher network during student training process in Figure 4. Obviously, the student network could get cleaner samples in large-scale training dataset with the help of the teacher network.

The advantage of proposed architecture. Next we compared four architectures with different components equipped in them to show the advantages of the proposed architecture. The face verification performance on different benchmark datasets is shown in Table 4. LinearFace represents an architecture equipped with only the proposed margin loss. TS is short for the teacher student learning strategy and Gen is short for the generative decode part. The row LinearFace+TS+Gen showed the performance of the proposed architecture in this paper. All these model were trained on CASIA dataset with ResNet50 as backbone. The results showed that the proposed architecture can effectively boost the performance for face verification.

Visualize samples from generative model. Resent researches illustrated why margin based algorithm works for face embedding learning by giving an geometric interpretation. But the perceptual intuition what the feature learned through the embedding network was missing. With the help of the auto-encoder architecture, we have a chance to visualize what different embeddings look like directly. In Figure 5, we showed the generative samples for same identity. We employed different embedding models from early, middle and last stages of training as the encode model to get the face image feature. Then a pre-trained generative decode net was employed to restore these samples from features. The figure showed that along with the training process stepping forward, more details were preserved in the restore frontal face image. This suggested that the embedding model has gained more discriminative power. And the model increased the capacity to retain information.

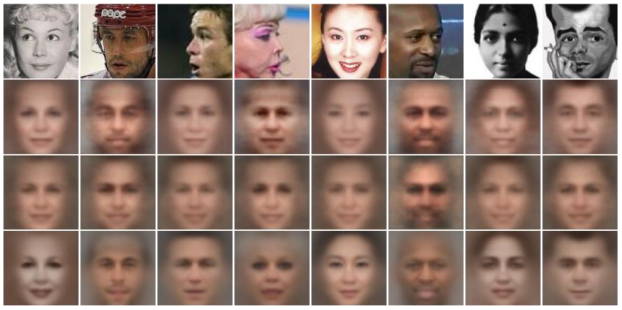

Figure 5. Examples of generated frontal face for same identity from different embedding models of various training stage. In rows top to bottom: early stage embedding model, middle stage embedding model and last stage embedding model. More details reserved in the generated frontal face images from latter stage embedding model.

Table 5. Face verification results (%) for different algorithms

Method	LFW	CALFW	CPLFW
Human-Individual	97.27	82.32	81.21
Human-Fusion	**99.85**	86.50	85.24
CenterLoss (Wen et al., 2016)	98.75	85.48	77.48
SphereFace (Liu et al., 2017b)	99.27	90.30	81.40
VGGFace2 (Cao et al., 2018)	99.43	90.37	84.00
ArcFace (Deng et al., 2019a)	**99.82**	95.45	92.08
Proposed LATSE	**99.82**	**96.20**	**93.48**

4.3. Evaluation Comparison

Results for face Verification on LFW, CALFW, CPLFW. LFW (Huang et al., 2008) dataset is the most widely used benchmark for unconstrained face verification on images. Recent algorithms(Deng et al., 2019a; Wang et al., 2018c) gain nearly saturation performance on it. Instead of only comparing performance on LFW, we also employed CALFW and CPLFW datasets, which involved higher pose and age variations for same identities from LFW, to evaluate the performance for face verification. The experiment results were reported in Table 5. From the results, we could find that several algorithms gained better performance than human-individual. By comparing results from different algorithms, the results shown that although both Arcface and our proposed LATSE gain similar performance on LFW test, our method improved 1% accuracy by average on CALFW and CPLFW benchmarks. This showed the proposed method generalized better when more human pose and age variations were involved.

Results for MegaFace test. The MegaFace (Kemelmacher-Shlizerman et al., 2016) dataset includes 1,000,000 images of 690,000 different identities as the distractors in the gallery set and 100,000 photos of 530 unique celebrity from FaceScrub as the probe set. There are two testing scenarios, one for face identification and the other is face verification. The testing process is conducted under two different training protocols (large or small training set, while training set is defined as large if there is more than 500,000 images in it). Following the setting in ArcFace (Deng et al., 2019a), we employed CASIA as training dataset for protocol of small, and use MS1M-RetinaV dataset as the training set for protocol large. For MegaFace test data inference, we used the cleaned version from ArcFace to test the proposed method, which is noted as 'r' in the Table 6. Both the identification and verification accuracy results were reported to compare the performance for large scale face recognition. Our method improved both accuracy under small and large training protocols. The results showed that the proposed architecture gain more generalization capability for large scale face recognition in an open set situation. We gained more than 99% accuracy on MegaFace test without private dataset.

Table 6. Megaface results (%) for different algorithm

Method	Identification	Verification
Softmax (Liu et al., 2017b)	54.85	65.92
TripletLoss (Schroff et al., 2015)	64.79	78.32
SphereFace (Liu et al., 2017b)	72.73	85.56
SphereFace+(Liu et al., 2018a)	73.03	-
CosFace (Wang et al., 2018b)	77.11	89.88
ArcFace(CASIA) (Deng et al., 2019a)	77.50	92.34
ArcFace,r(CASIA) (Deng et al., 2019a)	91.75	93.69
Proposed LATSE,r(CASIA)	**94.74**	**95.38**
TripletLoss (Schroff et al., 2015)	70.49	86.47
CosFace (Wang et al., 2018b)	82.72	96.65
ArcFace,r (Deng et al., 2019a)	98.35	98.48
SV-AM-Softmax,r (Wang et al., 2018c)	98.82	99.03
Proposed LATSE,r	**99.14**	**99.19**

5. Concluding Remarks

In this paper, we illustrate how the margin based loss methods work for face embedding learning in a perspective of probability. Two principles are proposed for new margin loss function designing. At the same time, in view of comparison metric encounters the bottleneck to gain more generalization ability, we propose to regard the open set face recognition as a problem of information transmission. Based on this intuition, we propose an auto-encoder architecture trained with a teacher student learning strategy, which increased the generalization ability for the face embedding. The extensive experimental results on several benchmarks show clear advantages of the proposed method.

References

Atoum, Y., Liu, Y., Jourabloo, A., and Liu, X. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 319–328. IEEE, 2017.

Boult, T., Cruz, S., Dhamija, A., Gunther, M., Henrydoss, J., and Scheirer, W. Learning and the unknown: Surveying steps toward open world recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9801–9807, 2019.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition*, pp. 67–74. IEEE, 2018.

Chen, B., Liu, W., Garg, A., Yu, Z., Shrivastava, A., Kautz, J., and Anandkumar, A. Angular visual hardness, 2019.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

Chen, T., Xu, B., Zhang, C., and Guestrin, C. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

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

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019b.

Geng, C., jun Huang, S., and Chen, S. Recent advances in open set recognition: A survey, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pp. 87–102. Springer, 2016.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1735–1742. IEEE, 2006.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8527–8537. 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969. IEEE, 2017.

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

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

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

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

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels, 2017.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., and Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4873–4882, 2016.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Liu, H., Zhu, X., Lei, Z., and Li, S. Z. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11947–11956, 2019.

Liu, W., Wen, Y., Yu, Z., and Yang, M. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 2, pp. 7, 2016.

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., Rehg, J. M., and Song, L. Iterative machine teaching. In *Proceedings of the International Conference on Machine Learning*, pp. 2149–2158. JMLR. org, 2017a.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 212–220, 2017b.

Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., and Song, L. Learning towards minimum hyperspherical energy. In *Advances in Neural Information Processing Systems*, pp. 6222–6233, 2018a.

Liu, Y., Jourabloo, A., and Liu, X. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 389–398. IEEE, 2018b.

Malach, E. and Shalev-Shwartz, S. Decoupling "when to update" from "how to update", 2017.

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

Nech, A. and Kemelmacher-Shlizerman, I. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7044–7053, 2017.

Noh, H., Hong, S., and Han, B. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

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

Sengupta, S., Chen, J.-C., Castillo, C., Patel, V. M., Chellappa, R., and Jacobs, D. W. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9. IEEE, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2014.

Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., and Change Loy, C. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision*, pp. 765–780, 2018a.

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

Wang, X., Wang, S., Zhang, S., Fu, T., Shi, H., and Mei, T. Support vector guided softmax loss for face recognition. *arXiv preprint arXiv:1812.11317*, 2018c.

Wang, X., Wang, S., Wang, J., Shi, H., and Mei, T. Comining: Deep face recognition with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9358–9367, 2019.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

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

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does disagreement help generalization against label corruption?, 2019.

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

Zhang, X., Zhao, R., Qiao, Y., Wang, X., and Li, H. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10823–10832, 2019.

Zhao, J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Zhao, K., Xu, J., and Cheng, M.-M. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1136–1144, 2019.

Zheng, T. and Deng, W. Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, pp. 18–01, 2018.

Zheng, T., Deng, W., and Hu, J. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.