ArcFace: Additive Angular Margin Loss for Deep Face Recognition

Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou

CVPR 2019

https://arxiv.org/abs/1801.07698

Presenter: Minho Park

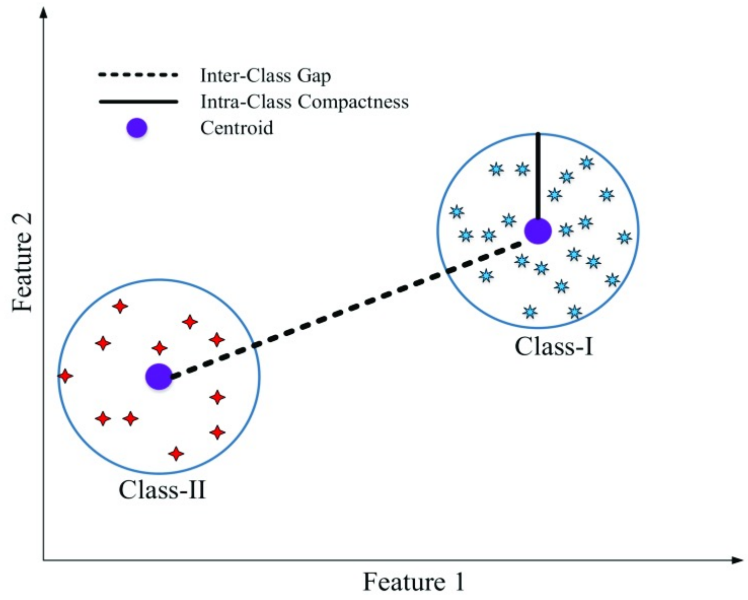
Metric learning

- Metric learning aims to learn a similarity (distance) function.
- Traditional metric learning usually learns a matrix $\boldsymbol{A}$ for a distance metric $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\boldsymbol{A}} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \boldsymbol{A} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$ upon the given features $\boldsymbol{x}_1, \boldsymbol{x}_2$.
- Recently, prevailing deep metric learning usually uses neural networks to automatically learn discriminative features $\boldsymbol{x}_1, \boldsymbol{x}_2$ followed by a simple distance metric such as Euclidean distance $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$.
- E.g. Contrastive loss, triplet loss

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Classification

- Inter-class separability & Intra-class compactness



The concept of compactness and separability.

Ahmed, Saeed & Lee, YoungDoo & Hyun, Seung & Koo, Insoo. (2018). Covert Cyber Assault Detection in Smart Grid Networks Utilizing Feature Selection and Euclidean Distance-Based Machine Learning. Applied Sciences. 8. 772. 10.3390/app8050772.

Centre loss

- To achieve intra-class compactness

$$
\begin{aligned}
\mathcal{L} &= \mathcal{L}_S + \lambda \mathcal{L}_C \\
&= \mathcal{L}_S + \lambda \frac{1}{2} \sum_i \left\| \boldsymbol{x}_i - \boldsymbol{c}_{y_i} \right\|_2^2
\end{aligned}
$$

Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." European conference on computer vision. Springer, Cham, 2016.

Centre loss

- To achieve intra-class compactness

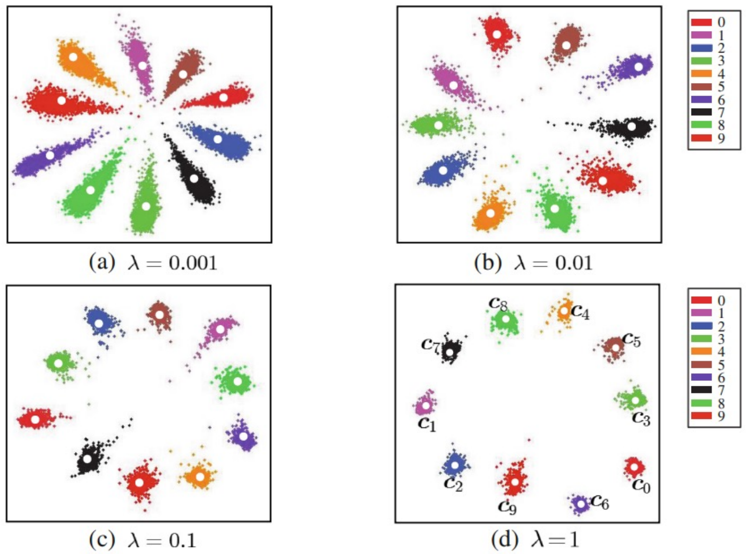


Fig. 3. The distribution of deeply learned features under the joint supervision of softmax loss and center loss.

Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." European conference on computer vision. Springer, Cham, 2016.

Deep Hypersphere Embedding

- Map feature to a unit sphere. ($W_j^T \boldsymbol{x} = \|W_j\| \|\boldsymbol{x}\| \cos \theta_j$)
- WHY?

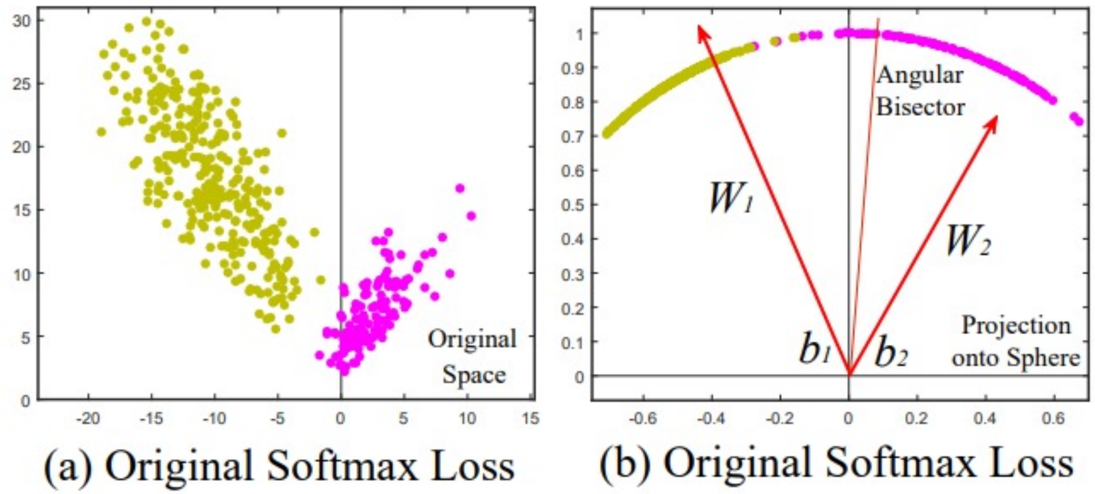


Figure 2: Comparison among softmax loss, modified softmax loss and A-Softmax loss.

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

SphereFace

- Normalized version of Softmax Loss (NSL)

$$\mathcal{L}_{A-softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}$$

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Geometry interpretation

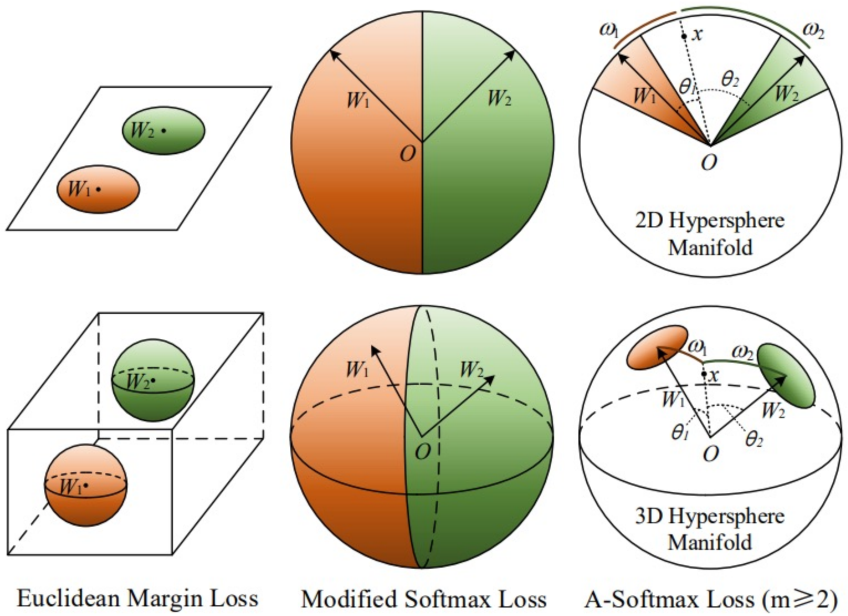


Figure 3: Geometry interpretation.

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Feature distribution

- Results

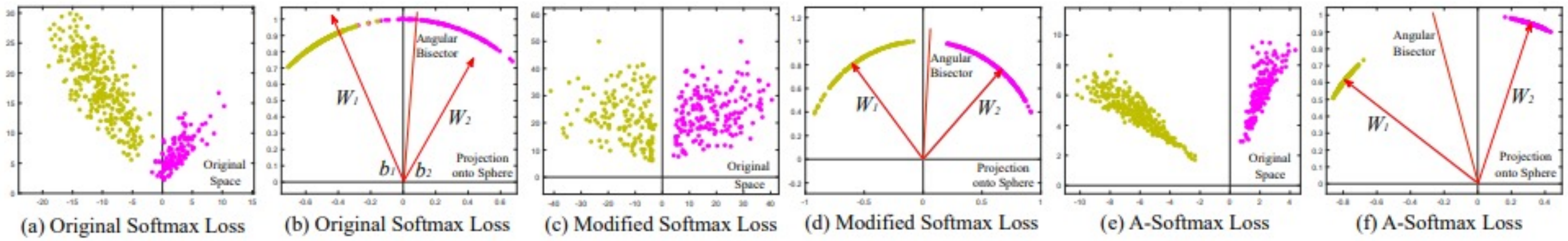


Figure 2: Comparison among softmax loss, modified softmax loss and A-Softmax loss.

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

SphereFace

- Angular softmax loss (A-Softmax)

$$\mathcal{L}_{A-softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{s\cos m_1 \theta_{y_i}}}{e^{s\cos m_1 \theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos \theta_j}}$$

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

CosFace

- Large Margin Cosine Loss (LMCL)

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

Wang, Hao, et al. "Cosface: Large margin cosine loss for deep face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

ArcFace

- Additive angular margin

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

ArcFace

- Additive angular margin

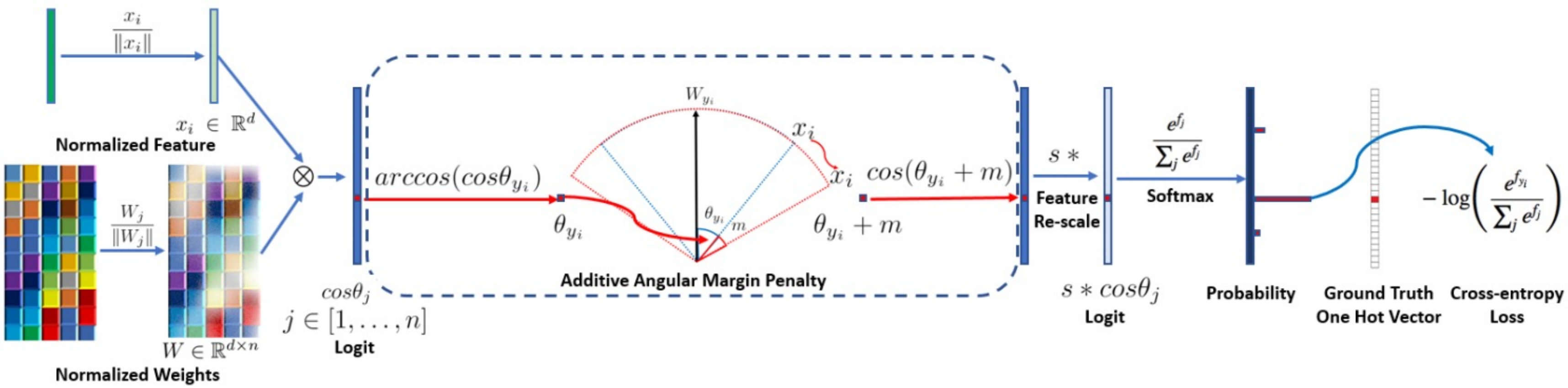
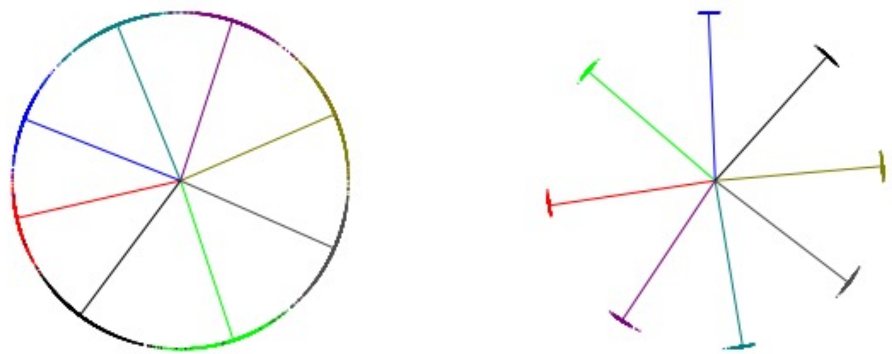


Figure 2. Training a DCNN for face recognition supervised by the ArcFace loss.

Toy examples



(a) Softmax (b) ArcFace

Figure 3. Toy examples under the softmax and ArcFace loss on 8 identities with 2D features.

Comparison with SphereFace and CosFace

- Combining all of the margin penalties.

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

Liu, Weiyang, et al. "Sphereface: Deep hypersphere embedding for face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Numerical Similarity

- From the view of numerical analysis, different margin penalties, no matter add on the angle or cosine space, all enforce the intra-class compactness and inter-class diversity by penalising the target logit.

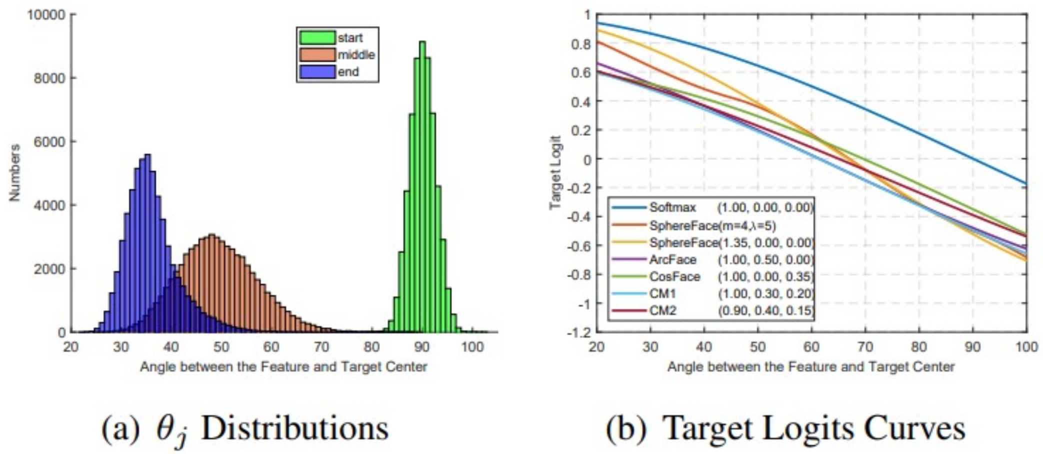


Figure 4. Target logit analysis.

Geometric Difference

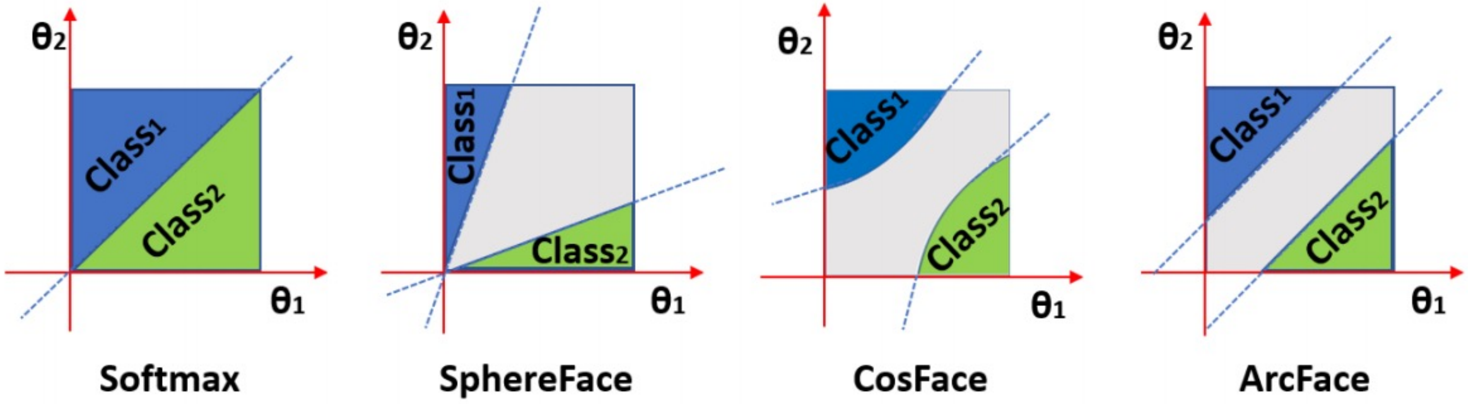


Figure 5. Decision margins of different loss functions under binary classification case.

Comparison with Other Losses

- Intra-Loss

$$\mathcal{L}_{intra} = \mathcal{L}_S + \frac{1}{\pi N} \sum_{i=1}^{N} \theta_{y_i}$$

Comparison with Other Losses

- Inter-Loss

$$\mathcal{L}_{inter} = \mathcal{L}_S - \frac{1}{\pi N(n-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq y_i}^{n} \operatorname{arccos}\left(W_{y_i}^T, W_j\right)$$

Comparison with Other Losses

- Triplet-Loss

$$\arccos\left(x_i^{pos}, x_i\right) + m \leq \arccos\left(x_i^{neg}, x_i\right)$$

Experiments

- 10 Datasets

Datasets	#Identity	#Image/Video
CASIA [43]	10K	0.5M
VGGFace2 [6]	9.1K	3.3M
MS1MV2	85K	5.8M
MS1M-DeepGlint [2]	87K	3.9M
Asian-DeepGlint [2]	94 K	2.83M
LFW [13]	5,749	13,233
CFP-FP [30]	500	7,000
AgeDB-30 [22]	568	16,488
CPLFW [48]	5,749	11,652
CALFW [49]	5,749	12,174
YTF [40]	1,595	3,425
MegaFace [15]	530 (P)	1M (G)
IJB-B [39]	1,845	76.8K
IJB-C [21]	3,531	148.8K
Trillion-Pairs [2]	5,749 (P)	1.58M (G)
iQIYI-VID [20]	4,934	172,835

Table 1. Face datasets for training and testing.
"(P)" and "(G)" refer to the probe and gallery set, respectively.

Results

1. **W_j is nearly synchronised with embedding feature centre for ArcFace (14.29°)**, but there is an obvious deviation (44.26°) between W_j and the embedding feature centre for Norm-Softmax. Therefore, **the angles between W_j cannot absolutely represent the inter-class discrepancy on training data.** Alternatively, the **embedding feature centres** calculated by the trained network **are more representative.**

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

Table 3. The angle statistics under different losses ([CASIA, ResNet50, loss*]). Each column denotes one particular loss. **"W-EC"** refers to the **mean of angles between W_j and the corresponding embedding feature centre. "W-Inter"** refers to the **mean of minimum angles between W_j's. "Intra1"** and **"Intra2"** refer to the **mean of angles between x_i and the embedding feature centre** on CASIA and LFW, respectively. "Inter1" and "Inter2" refer to the **mean of minimum angles between embedding feature centres** on CASIA and LFW, respectively.

Results

2. Intra-Loss can effectively compress intra-class variations but also brings in smaller interclass angles.
3. Inter-Loss can slightly increase inter-class discrepancy on both W (directly) and the embedding network (indirectly), but also raises intra-class angles.

Results

4. ArcFace already has very good intra-class compactness and inter-class discrepancy.
5. Triplet-Loss has similar intraclass compactness but inferior inter-class discrepancy compared to ArcFace.

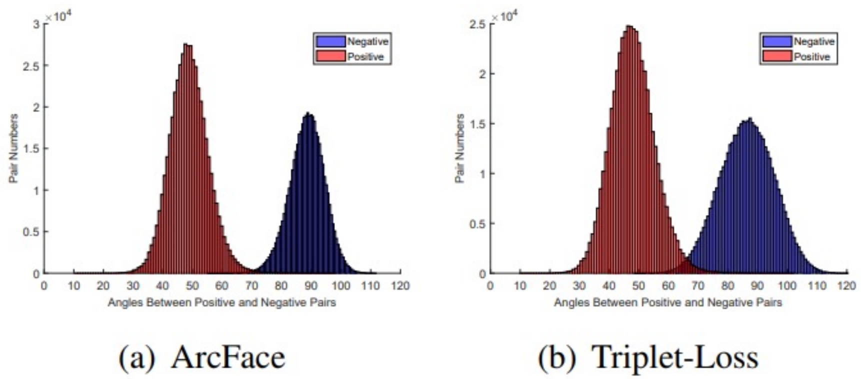


Figure 6. Angle distributions of all positive pairs and random negative pairs (~ 0.5M) from LFW.

Results

Loss Functions	LFW	CFP-FP	AgeDB-30
ArcFace (0.4)	99.53	95.41	94.98
ArcFace (0.45)	99.46	95.47	94.93
ArcFace (0.5)	**99.53**	**95.56**	**95.15**
ArcFace (0.55)	99.41	95.32	95.05
SphereFace [18]	99.42	-	-
SphereFace (1.35)	99.11	94.38	91.70
CosFace [37]	99.33	-	-
CosFace (0.35)	99.51	95.44	94.56
CM1 (1, 0.3, 0.2)	99.48	95.12	94.38
CM2 (0.9, 0.4, 0.15)	99.50	95.24	94.86
Softmax	99.08	94.39	92.33
Norm-Softmax (NS)	98.56	89.79	88.72
NS+Intra	98.75	93.81	90.92
NS+Inter	98.68	90.67	89.50
NS+Intra+Inter	98.73	94.00	91.41
Triplet (0.35)	98.98	91.90	89.98
ArcFace+Intra	99.45	95.37	94.73
ArcFace+Inter	99.43	95.25	94.55
ArcFace+Intra+Inter	99.43	95.42	95.10
ArcFace+Triplet	99.50	95.51	94.40

Table 2. Verification results (%) of different loss functions ([CASIA, ResNet50, loss*]).

Method	#Image	LFW	YTF
DeepID [32]	0.2M	99.47	93.20
Deep Face [33]	4.4M	97.35	91.4
VGG Face [24]	2.6M	98.95	97.30
FaceNet [29]	200M	99.63	95.10
Baidu [16]	1.3M	99.13	-
Center Loss [38]	0.7M	99.28	94.9
Range Loss [46]	5M	99.52	93.70
Marginal Loss [9]	3.8M	99.48	95.98
SphereFace [18]	0.5M	99.42	95.0
SphereFace+ [17]	0.5M	99.47	-
CosFace [37]	5M	99.73	97.6
MS1MV2, R100, ArcFace	5.8M	**99.83**	**98.02**

Table 4. Verification performance (%) of different methods on LFW and YTF.

Why Face Recognition needs ArcFace loss?

Face Recognition

- Definition: Face recognition is the problem of **identifying and verifying people** in a photograph **by their face.**
- Process
  1. **Face Detection.** Locate one or more faces in the image and mark with a bounding box.
  2. **Face Alignment.** Normalize the face to be consistent with the database, such as geometry and photometrics.
  3. **Feature Extraction.** Extract features from the face that can be used for the recognition task.
  4. **Face Recognition.** Perform matching of the face against one or more known faces in a prepared database.

https://machinelearningmastery.com/introduction-to-deep-learning-for-face-recognition/

Datasets

- #Identity = 1K ~ 94K

Datasets	#Identity	#Image/Video
CASIA [43]	10K	0.5M
VGGFace2 [6]	9.1K	3.3M
MS1MV2	85K	5.8M
MS1M-DeepGlint [2]	87K	3.9M
Asian-DeepGlint [2]	94 K	2.83M
LFW [13]	5,749	13,233
CFP-FP [30]	500	7,000
AgeDB-30 [22]	568	16,488
CPLFW [48]	5,749	11,652
CALFW [49]	5,749	12,174
YTF [40]	1,595	3,425
MegaFace [15]	530 (P)	1M (G)
IJB-B [39]	1,845	76.8K
IJB-C [21]	3,531	148.8K
Trillion-Pairs [2]	5,749 (P)	1.58M (G)
iQIYI-VID [20]	4,934	172,835

Table 1. Face datasets for training and testing.