UNIVERSITY OF SURREY

Angular SparseMax for Face Recognition

Chi Ho Chan and Josef Kittler

Centre for Vision, Speech and Signal Processing, University of Surrey,

United Kingdom

Introduction

Both probabilistic loss and metric learning loss methods can attain excellent performance in face recognition.

Metric learning loss methods: usually suffer from high computational cost. To mitigate this problem, they require carefully designed sample mining strategies. Unfortunately, the performance is very sensitive to these strategies.

Probabilistic loss methods: SoftMax activation is widely used to map a score vector to a posterior probability distribution. The most appealing property of Softmax is that it is simple to evaluate and differentiate, and it can be turned into a (convex) negative log-likelihood loss function by taking the logarithm of its output

SoftMax

Face representation trained on the SoftMax loss exhibits an inherently good separation between classes but the intraclass variations may be very poor.

The limitation of the SoftMax activation is that the resulting probability distribution always has a full support. In other words, the probability of the face embedding belonging to every training subject is never zero, although it may be very small.

Thus, most variants of SoftMax loss functions directly employ a margin on the angular score vector, z. These variants, summarized in Tab I, use the class label information to increase the margin to improve the discriminatory power.

TABLE I

SURVEY OF MARGIN-BASED ANGULAR SCORES FOR TRAINING USING SOFTMAX LOSS

Ref.	Intra-class score	Inter-class score
[4]	$cos(m\theta_{\mathbf{y}=\mathbf{1}})$	$cos(\theta_{\mathbf{y}=\mathbf{0}})$
[2]	$cos(\theta_{\mathbf{y}=\mathbf{1}}+m)$	$cos(\theta_{\mathbf{y}=\mathbf{0}})$
[6], [7]	$cos(\theta_{\mathbf{y}=\mathbf{1}})-m$	$cos(\theta_{\mathbf{y}=\mathbf{0}})$
[8]	$cos(\theta_{\mathbf{y}=\mathbf{1}}+m)$	$cos(\theta_{\mathbf{y}=\mathbf{0}})$, if $cos(\theta_{\mathbf{y}=\mathbf{1}}+m) \geq cos(\theta_{\mathbf{y}=\mathbf{0}})$ $t(cos(s(\theta_{\mathbf{y}=\mathbf{0}}))+1)-1$, otherwise.

SparseMax

$$\hat{\mathbf{f}}_{\Omega}(\mathbf{z}) = \underset{\mathbf{p} \in \triangle^d}{\operatorname{argmin}}(\|\mathbf{p} - \mathbf{z}\|^2) \qquad (4)$$

Algorithm 1: Computing Sparsemax

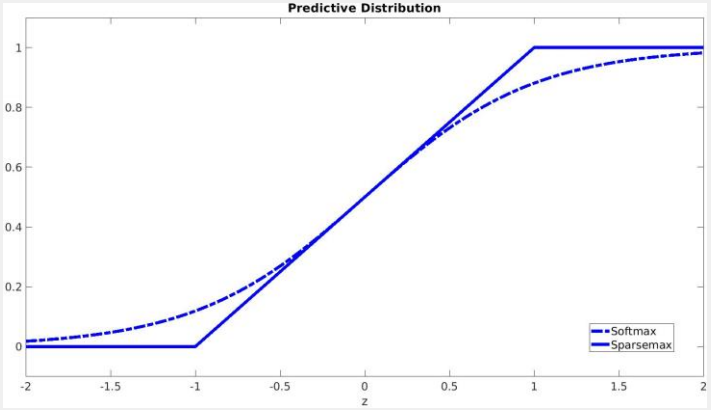
Data: $\mathbf{z}$

Sort the components of $\mathbf{z}$ as $z_1 \geq \cdots \geq z_C$

Find $\rho := \max\{1 \leq i \leq C \quad | \quad 1 + iz_i > \sum_{j \leq i} z_j\}$

Define $\tau = \frac{(\sum_{j \leq \rho} z_j) - 1}{\rho}$

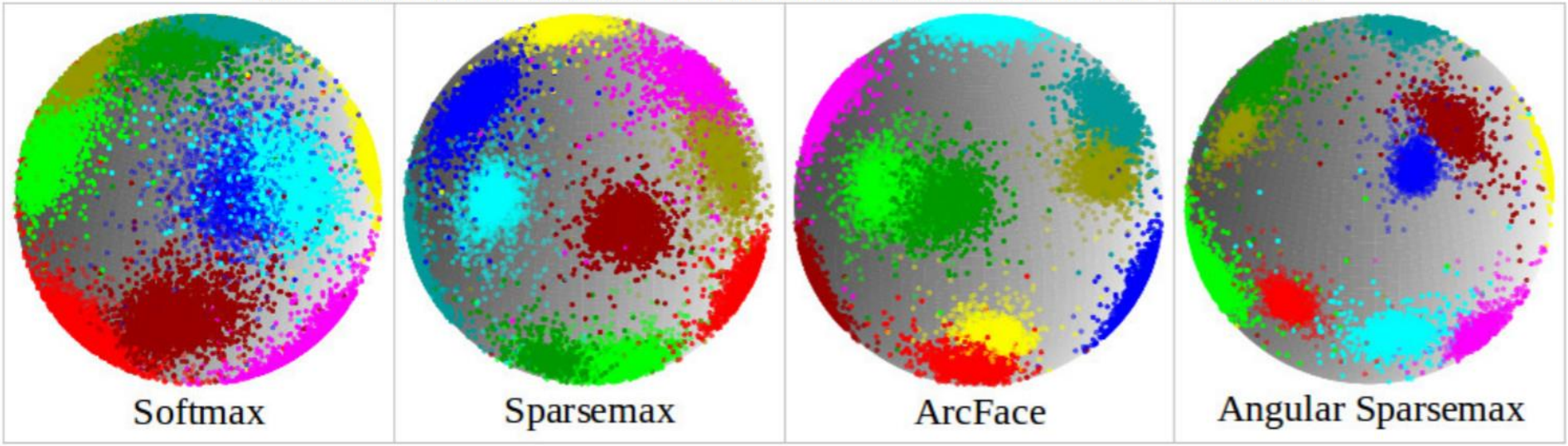
Result: $\mathbf{p} = \max(\mathbf{z} - \tau \mathbf{1}, 0)$



Angular SparseMax

$$\mathbf{z} = s \cdot cos(\boldsymbol{\phi}) \quad | \quad \boldsymbol{\phi} \in \mathcal{R}^C, \phi_i := \begin{cases} \theta_i + m & y_i = 1, \\ \theta_i & y_i = 0. \end{cases} \tag{9}$$

Fig. 2. Feature distribution visualization of several loss functions. Different colors denote different classes.



Exploratory Experiment

1)Effect of Temperature, s: The best performance is achieved when the temperature is set to 1.9. Therefore, we choose this temperature for training on the CASIA dataset in the following experiments.

2)Effect of Additive Angular Margin, m: Figure 4 plots the result of angular SparseMax for different margins. The best rank 1 recognition rate and verification rate are achieved at margin equal to 0.2. This finding is used by the proposed method to compare it with the state of the art algorithms when the model is trained on the CASIA dataset.

Fig. 3. System Performance on MegaFace with 1M distractors against temperature. Note: The Verification Rate is measured at 1e-6 False acceptance Rate and Recognition Rate is at Rank 1.

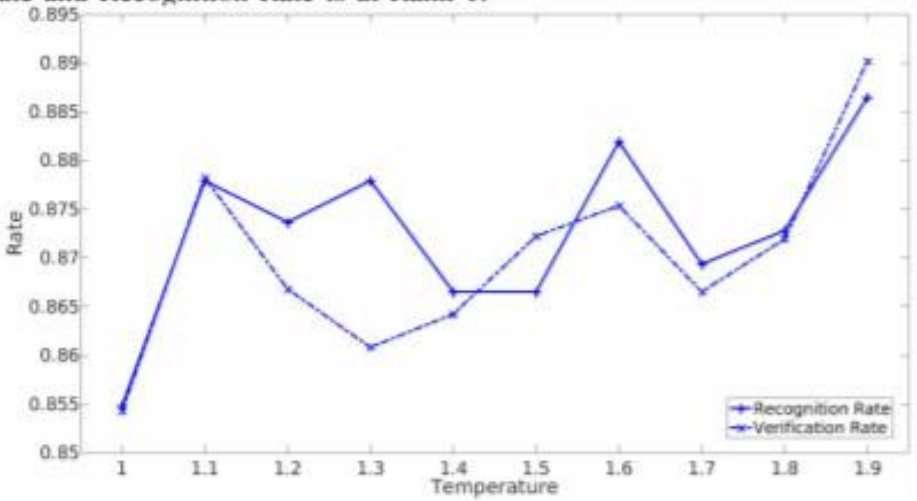
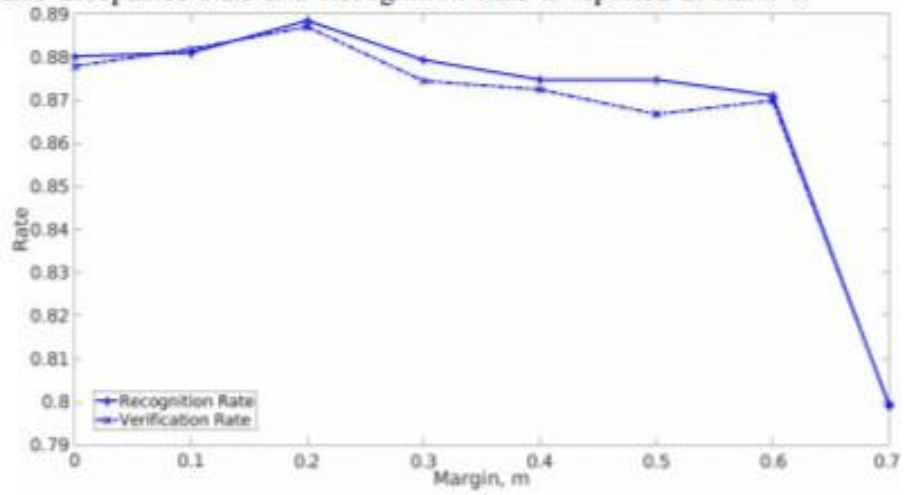


Fig. 4. System Performance on MegaFace with 1M distractors against Additive Angular Margin. Note: The Verification Rate is measured at 1e-6 False acceptance Rate and Recognition Rate is reported at Rank 1.



Benchmarks

TABLE III

PERFORMANCE OF ANGULAR SPARSEMAX AND OTHER SAMPLING-BASED SOFTMAX VARIANTS ON LFW, CFP-FP, AGEDB-30, AND MEGAFACE CHALLENGE1 IDENTIFICATION (MR) AND VERIFICATION (1M) WITH 1M DISTRACTORS. "MR" REFERS TO THE RANK-1 FACE IDENTIFICATION ACCURACY WITH 1M DISTRACTORS, AND "MV" REFERS TO THE FACE VERIFICATION TAR AT 10^6. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND BEST RESULTS ARE UNDERLINED.

Methods	Arch.	LFW	CFP-FP	AGEDB	MR	MV
CASIA Dataset						
SphereFace [3]	ResNet-64	99.2				
ArcFace [4]	ResNet-50	**99.53**	**95.56**	**95.15**	**91.75**	**93.69**
ArcFace [26]	MobileFacenet	99.28		93.05		88.09
Softmax s=64.	MobileFacenet	98.85	92.86	90.18	75.76	71.88
ArcFace s=64., m=0.42	MobileFacenet	99.28	94.20	93.05	88.11	88.08
Sparsemax s=1.9	MobileFacenet	99.32	93.34	92.82	88.02	87.79
Angular Sparsemax s=1.9, m=0.2	MobileFacenet	99.17	93.64	92.43	88.84	88.71
MS-Celeb-1M Dataset						
Softmax [30]	ResNet-50	99.30	87.23	94.48	91.25	
SphereFace [30]	ResNet-50	99.59	91.37	96.62	96.04	
ArcFace [30]	ResNet-50	99.68	92.26	97.23	96.97	
Intra D + Linter Arc [30]	ResNet-50	**99.73**	93.07	**97.30**	97.02	
ArcFace [26]	MobileFacenet	99.55				92.59
ArcFace s=64., m=0.5	MobileFacenet	99.67	96.13	96.53	**97.40**	**97.83**
Sparsemax s=1.	MobileFacenet	99.67	96.67	96.67	97.26	97.66
Angular Sparsemax s=1.9, m=0.1	MobileFacenet	99.65	**97.17**	97.03	**97.40**	97.60

Conclusions

- We opt for an alternative predictor function - the SparseMax. We describe how it maps the score vector to a sparse probability distribution.
- Using SparseMax as a baseline, we developed Angular SparseMax which adds the additive angular margin into the score vector to further improve the discriminative power of the embedding feature.
- The proposed loss function is experimentally validated. In terms of performance, it compares well with the state of the art Arcface loss.