UNIVERSITY OF SURREY

Angular Sparsemax for Face Recognition

Chi Ho Chan, Josef Kittler
CVSSP, University of Surrey, Guildford, Surrey, GU2 7XH, UK.
(chiho.chan, j.kittler)@surrey.ac.uk

Abstract

The Softmax prediction function is widely used to train Deep Convolutional Neural Networks (DCNNs) for large-scale face recognition and other applications. The limitation of the softmax activation is that the resulting probability distribution always has a full support. This full support leads to larger intraclass variations. In this paper, we formulate a novel loss function, called Angular Sparsemax for face recognition. The proposed loss function promotes sparseness of the hypotheses prediction function similar to Sparsemax [1] with Fenchel-Young regularisation. By introducing an additive angular margin on the score vector, the discriminatory power of the face embedding is further improved. The proposed loss function is experimentally validated on severaldatabases in terms of recognition accuracy. Its performance compares well with the state of the art Arcface loss.

Background

Face representation trained on the softmax loss exhibits an inherently good separation between classes but the intraclass variations may be very poor. 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, y to increase the margin to improve the discriminatory power.

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, x, belonging to every training subject is never zero, although it may be very small.

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

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

SparseMax

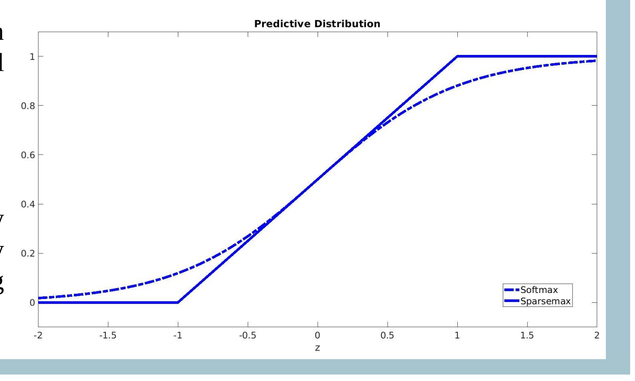
A prediction function is to map an angular score vector to a posterior probability. This mapping problem presented in Eq. 1, can be solved by maximizing an affinity term, subject to a confidence constraint $\Omega(\mathbf{p})$.

$$\hat{\mathbf{f}}_{\Omega}(\mathbf{z}) \in \underset{\mathbf{p} \in dom(\Omega)}{\operatorname{argmax}} \left(\mathbf{z}^{\intercal}\mathbf{p} - \Omega(\mathbf{p})\right) \quad (1)$$

If Ω function is Shannon Entropy, the prediction function is SoftMax. If Ω is Gini Index, it will become SparseMax, presented in Eq 4.

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

SparseMax tends to yield sparse probability distribution as the resulting distribution is likely to assign exactly zero probability to low-scoring choices.

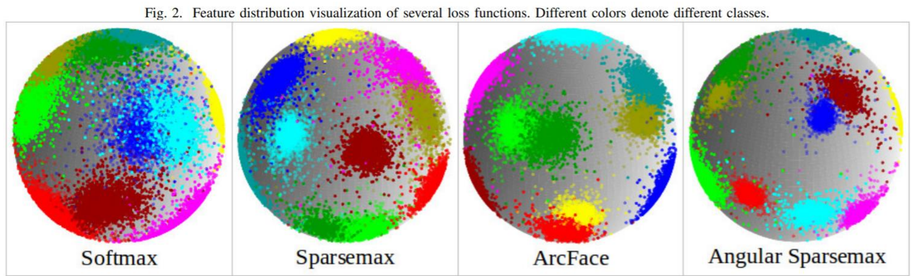


Angular SparseMax

To further improve the discriminatory power of face representation, we employ SparseMax on the marginal angular scores. Because of the limited resource, the additive angular margin, presented in Eq 9, is employed.

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

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



Exploratory Experiments

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.

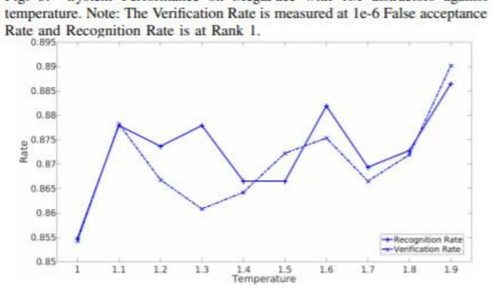
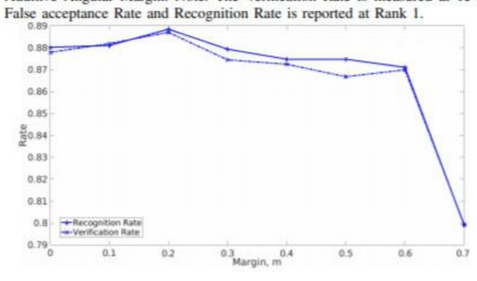


Fig. 4. System Performance on MegaFace with 1M distractors against Additive Angular Margin. Note: The Verification Rate is measured at 1e-6 False acceptance Rate and Recognition Rate is reported at Rank 1.



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

In this paper, 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.

Acknowledgements

This work was supported by the EPSRC Programme Grant (FACER2VM) EP/N007743/1.