Range Loss for Deep Face Recognition with Long-Tailed Training Data

Xiao Zhang[1,2] Zhiyuan Fang[1,3] Yandong Wen[4] Zhifeng Li[5] Yu Qiao[1,6]

[1] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, China

[2] Tianjin University [3] Southern University of Science and Technology [4] Carnegie Mellon University

[5] Tencent AI Lab, China [6] The Chinese University of Hong Kong, Hong Kong SAR, China

ICCV17 International Conference on Computer Vision 2017

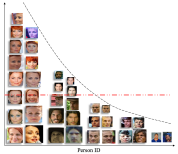
Introduction

➢ Motivation:

- Abundant training data and well-designed training strategies are indispensable for effective deep face models. However, many large scale face datasets exhibit long-tail distribution where a small number of entities have large number of face images while a large number of persons only have very few face samples (long tail).

➢ Goal:

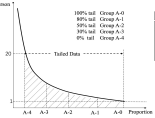
- Explore the negative effect of long-tailed training data.
- Analyze deep feature vectors.
- Design a new loss to relieve the effect of long-tailed training data.



Shortcomings of Long-Tailed Data

➢ Explorations on Long-Tailed training data:

- In this set, there are 700k images for roughly 10k identities, and 1 million images for the remaining 90k identities. We further divide the dataset into several groups according to the proportions of tailed data the table.



Groups	Num of Identities	Images	Tail Ratio
A-0	99,891	1,687,691	100.00%
A-1	81,913	1,620,526	80.00%
A-2	54,946	1,396,414	50.00%
A-3	26,967	1,010,735	30.00%
A-4	10,000	699,832	0.00%

Groups	Acc. on LFW	Acc. on YTF
A-0 (100% tail)	97.87%	92.0%
A-1 (80% tail)	98.03%	92.2%
A-2 (50% tail)	98.25%	92.9%
A-3 (30% tail)	97.18%	91.7%
A-4 (0% tail)	95.97%	89.4%

- We train the VGG-Face with cross-entropy loss and examine their performances on LFW and YTF's tasks. Training with A-1 and A-2 leads to better performance than A-0, even A-0 has more samples. On the other hand, removing too much tailed data like A-3 and A-4 will drop the performance.

➢ Explorations with contrastive[1], triplet[2] and Center Losses[3]

Groups	Contrastive Loss		Triplet Loss		Center Loss	
	Acc. on LFW	Acc. on YTF	Acc. on LFW	Acc. on YTF	Acc. on LFW	Acc. on YTF
A-0 (with long-tail)	**98.35%**	**92.7%**	**98.10%**	**92.3%**	**98.23%**	**92.4%**
A-1 (cut 20% tail)	98.45%	93.1%	98.13%	92.3%	98.50%	92.7%
A-2 (cut 50% tail)	**98.47%**	**93.3%**	**98.40%**	**93.2%**	**98.57%**	**93.2%**
A-3 (cut 70% tail)	96.23%	91.1%	97.87%	91.7%	97.85%	92.0%
A-4 (cut 100% tail)	95.97%	89.4%	97.33%	91.1%	97.33%	91.1%

Reference

[1] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification, **NIPS**, 2014

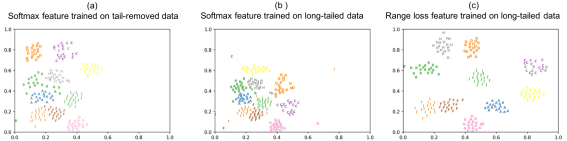
[2] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering, **CVPR**, 2015

[3] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition, **ECCV,** 2016

Analysis of Deep Feature Vectors

➢ Visualization of Deep Feature Vectors

- Well-trained CNN can map the input face image to feature vectors with rich identity information. For recognition tasks, we expect CNN model to output similar deep feature vectors for same persons and far apart vectors for different persons. Each face image is mapped to a 4096-dimensional feature by VGG-Nets trained with above loss functions. Since it is hard to analyze high dimension vectors, we use t-SNE to transform these vectors into 2-D vectors, as shown in these figures.



➢ Statistics of Deep Feature Vectors

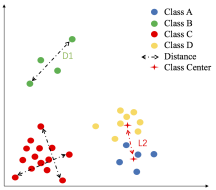
Model	Loss	Data	Intra-Class Evaluation			Inter-Class Evaluation		
			SD	EM	Kurtosis	SD	EM	Kurtosis
VggNet	Softmax	A-2	56.9768	77.9928	-1.3590	285.6691	342.0850	-1.4122
		A-0	22.8880	33.3013	0.3803	71.5123	88.4179	-1.9884
	Contras.	A-2	26.3160	36.4437	-1.4130	122.3764	150.9392	-1.7051
		A-0	22.8497	31.5918	-1.0697	109.9323	134.6600	-1.6276
	Triplet	A-2	26.0807	36.7853	-0.9714	113.1263	134.5524	-1.0984
		A-0	23.0050	31.9569	-0.9448	106.7124	129.6840	-1.4492
	Center	A-2	18.9627	25.7436	-0.9558	180.4223	136.4760	-1.2578
		A-0	15.2288	18.9850	-0.5807	118.6627	92.2623	-1.3320
	Range	A-2	42.2881	63.4769	**-1.3308**	125.4162	153.2813	-1.3468
		A-0	27.2868	39.5968	**-1.4060**	208.1743	249.8967	-1.0050

- The intra-class and inter-class statistics expose differences between long-tail model and cut-tail model. Here SD is standard deviation and EM is the average Euclidean metric. Kurtosis describes the 4th order statistics of feature distribution. Infrequent extreme deviated vectors lead to high kurtosis. We expect a low kurtosis because infrequent extreme deviation is harmful for face recognition task.

Range Loss

➢ The Range Loss

- This paper addresses this challenge by proposing range loss to handle imbalanced data. This new loss can help to reduce kurtosis (infrequent extreme deviation) while enlarge the inter-class distances.



$$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_R = -\sum_{i=1}^{M}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \lambda\mathcal{L}_R$$

$$\mathcal{L}_R = \alpha\mathcal{L}_{R_{intra}} + \beta\mathcal{L}_{R_{inter}}$$

$$\mathcal{L}_{R_{intra}} = \sum_{i\subseteq I}\mathcal{L}^i_{R_{intra}} = \sum_{i\subseteq I}\frac{k}{\sum_{j=1}^{k}\frac{1}{\mathcal{D}_j}}$$

$$\mathcal{L}_{R_{inter}} = \max(M - \mathcal{D}_{Center}, 0)$$
$$= \max(M - \|\bar{x}_{\mathcal{Q}} - \bar{x}_{\mathcal{R}}\|_2^2, 0)$$

Where $\mathcal{L}_R$ is range loss, $\mathcal{L}_{R_{intra}}$ is the intra-class part of range loss and $\mathcal{L}_{R_{inter}}$ is the inter-class part.

Experiment Results

➢ Performances on LFW and YTF sets

- For comparison, we trained CNN models under the supervision of softmax loss only, contrastive loss, joint triplet loss, center loss, and range loss, respectively (the last four are jointly used with softmax loss).

Groups	Contrastive Loss		Triplet Loss		Center Loss		**Range Loss**	
	LFW	YTF	LFW	YTF	LFW	YTF	LFW	YTF
A-0 (100% tail)	98.35%	92.7%	98.10%	92.3%	98.22%	92.4%	**98.63%**	**93.5%**
A-2 (50%)	98.47%	93.3%	98.40%	93.2%	98.57%	93.2%	**98.45%**	**93.2%**

- When trained with long tailed dataset A-0, range loss clearly outperforms baseline model with softmax loss, from 97.87% to 98.63% in LFW and 92.00% to 93.50% in YTF. Contrastive loss and triplet loss with full tailed data leads to lower accuracy, while our range loss can effectively e xploit the tailed data part to enhance the training, with accuracy increase by 0.18% in LFW and 0.3% in YTF from A-2 to A-0. Moreover, range loss can prevent kurtosis rising and extend the inter-class distance with long tailed data from the statistics

➢ Comparison with state-of-the-art methods

- To further examine the ability of range loss, we utilize a 30-layer residual CNN. The intention of this experiment is to examine the potential ability and generalization of rang loss with deeper networks and cleaner data.

Methods	LFW	YTF
DeepID-2+ [26]	99.47%	93.20%
FaceNet [23]	99.63%	95.10%
Baidu [16]	99.13%	-
Deep FR [20]	98.95%	97.30%
DeepFace [29]	97.35%	91.40%
Center Loss [31]	99.28%	94.90%
Softmax Loss	98.27%	93.10%
Range Loss	**99.52%**	**93.70%**

- Range loss performs well with stronger nets like ResNets for long tail training. We train ResNets on long-tailed data with and without range loss, and test these two models on BLUFER dataset. ResNet with range loss obtained accuracies of 92.10% in verification (FAR=0.1%) and 63.69% in Open-Set Identification (Rank=1, FAR=1%) while version without range loss gets 90.03% and 60.02%, respectively.

Conclusions

- Long-Tailed training data do have negative effect on the generalization ability.
- Features extracted from long-tailed model have higher kurtosis and more infrequent extreme deviation on their distribution.
- Range loss can effectively relieve the effect of long-tailed training data.