CELEB-500K: A LARGE TRAINING DATASET FOR FACE RECOGNITION

Jiajiong Cao, Yingming Li, and Zhongfei Zhang*

College of Information Science & Electronic Engineering, Zhejiang University, China
{jiajiong, yingming, zhongfei}@zju.edu.cn

ABSTRACT

In this paper, we propose a large training dataset named Celeb-500K for face recognition, which contains 50M images from 500K persons. To better facilitate academic research, we clean Celeb-500K to obtain Celeb-500K-2R, which contains 25M aligned face images from 365K persons. Based on the developed dataset, we achieve state-of-the-art face recognition performance and reveal two important observations on face recognition study. First, metric learning methods have limited performance gain when the training dataset contains a large number of identities. Second, in order to develop an efficient training dataset, the number of identities is more important than the average image number of each identity from the perspective of face recognition performance. Extensive experimental results show the superiority of Celeb-500K and provide a strong support to the two observations.

***Index Terms*—** face recognition, face dataset, large scale, convolutional neural networks

1. INTRODUCTION

In this paper, we propose a large training dataset named Celeb-500K for deep learning [1, 2, 3, 4, 5, 6] based large scale face recognition [7, 8, 9, 10, 11]. The training dataset consists of 50M images from 500K persons. Our paper focuses on addressing the following two issues in face recognition. First, according to Table 1, there are large gaps on dataset scale between publicly available datasets and private datasets. For example, CelebFace [12] has only $1/800$ identities and $1/500$ images of the Google dataset. Therefore, compared with industrial applications, the academic research community can only resort to smaller scaled datasets resulting in typically biased conclusions. Thus the efficacy of the proposed methods on larger training datasets needs further verification. For example, many metric learning methods including Contrastive Loss [12], Center Loss [13] and Triplet Loss [14] have greatly improved the face recognition performance of models trained on smaller public datasets such as CelebFace and CASIA-WebFace, but their efficiency on larger scale datasets needs further investigation.

*Corresponding author

Table 1. Recent face recognition training datasets.

Dataset	Available	#People	#Images
YFD [15]	public	1595	3425 videos
VGGFace [16]	public	2600	2.6M
VGGFace2 [10]	public	9131	3.3M
CelebFaces [12]	public	10K	202K
CASIA-WebFace [17]	public	10K	500K
MS-Celeb-1M [18]	public	100K	10M
Celeb-500K	**public**	**500K**	**50M**
Facebook [18]	private	4K	4.4M
Google [18]	private	8M	$100-200$M

The second is about the importance of intra- and inter-identity variations when a dataset is large enough. The former focuses on the Average Image Number of Each Identity (**AINEI**) and the latter on overall Identity Number (**IN**). For example, VGGFace [16] and VGGFace2 [10] pay more attention to AINEI. Every identity has more than 330 and $1,000$ images in VGGFace2 and VGGFace on average, respectively. On the other hand, MS-Celeb-1M chooses to increase IN as shown in Table1. However, there are still not enough efforts on studying which way is more efficient when the dataset is large enough, from the perspective of face recognition performance.

Consequently, Celeb-500K is developed to cope with the above problems. Since the original images of Celeb-500K are downloaded from the Internet, the original images are not directly suitable for model training and there are many noisy labels. Therefore, we first perform basic processing including face detection, alignment and affine warp to get standard face images. Then, we use a two-step bootstrapping strategy to clean the dataset. Finally, we obtain a cleaned version called Celeb-500K-2R containing 25M aligned face images from 356K celebrities. Extensive experiments are conducted on the developed dataset. Based on Celeb-500K-2R along with a simple baseline, we achieve state-of-the-art face recognition accuracy and face verification rate on LFW. Furthermore, Celeb-500K-2R helps achieve much higher verification

Fig. 1. Data development procedure. Celebrity list is referred to [18]. We adopt MTCNN [19] for face detection and alignment.

rate over all the other public datasets at an extremely small false alarm rate.

According to our experimental results, we have interesting observations. First, out of our expectation, the previous metric learning methods including CenterLoss [13] and TripletLoss [14] have little influence on models trained on Celeb-500K-2R. Second, when IN is fixed, the increase of AINEI has very limited impact on face recognition performance. On the contrary, the model performance increases dramatically when we increase IN but fix the number of total images. These important observations are critically useful in guiding the further investigation. In summary, the contributions of our work are as follows:

- We propose a large face recognition training dataset Celeb-500K that consists of 50 million images from 500 thousand persons and a cleaned version called Celeb-500K-2R containing 25 million aligned faces from 356 thousand celebrities.
- We show that many previous metric learning methods do not have performance gains as would be expected on the training dataset with a large identity number.
- We show that inter-identity variation plays a more important role than intra-identity variation when the dataset scale is large. This can be a useful principle for developing a valid dataset.

2. DATASET DEVELOPMENT

As illustrated in Figure 1, the dataset is developed in several steps. In particular, we first refer to the celebrity list from [18], which consists of 1M celebrity names. Then, we search each name on search engines to obtain 100 image urls for each person. After downloading all the images of the first 500K people, we get a dataset consisting of 50M images named Celeb-500K. To help facilitate the standard model training, we further perform face detection, alignment and affine warp on the original images. Specifically, we adopt MTCNN [19], which is a unified multi-stage cascade CNN model, for face detection and alignment. For affine warp, we fix the aligned image size as 128×112 $(h \times w)$ and the five landmark points as follows: left eye pupil $(x : 38, y : 55)$, right eye pupil $(x : 73, y : 55)$, tip of nose $(x : 56, y : 75)$, left mouth corner $(x : 42, y : 95)$ and right mouth corner $(x : 71, y : 95)$.

After all these processes, we obtain a dataset that contains 35M aligned face images from 500K celebrities. The decrease of the number of images is mainly due to the false detection of MTCNN.

3. BOOTSTRAPPING LABEL CLEANING

Since the images are downloaded from search engines without manually labeling, the labels usually contain much noise. This can significantly influence the model performance [20]. Therefore, we conduct a two-step bootstrapping strategy to automatically clean the dataset. In particular, we first pretrain the model on CelebFace [12], which contains 202599 manually labeled images from 10177 persons. Second, we finetune the model on 356K identities with the 35M face images and use the trained model to predict labels for the training images. Then, we select the images whose predicted label is the same as the ground truth and whose probability is larger than 0.7. After the first step bootstrapping, a new dataset called Celeb-500K-1R is obtained, which consists of 23M face images from 356K celebrities.

Table 2. The performance on 500K image pairs of LFW. The performance gain of bootstrapping strategy is significant.

Dataset	AUC	FAR=0.1%	FAR=0.01%
CelebFaces	65.73%	64.78%	46.37%
Celeb-500K	88.05%	92.72%	87.91%
Celeb-500K-1R	95.63%	96.38%	94.75%
Celeb-500K-2R	**98.29%**	**98.71%**	**95.28%**

Then, the model is further finetuned on Celeb-500K-1R and the bootstrapping cleaning is performed for the second time. In particular, the 35M face images are relabeled with the trained model and the images whose predicted label is the same as the ground truth and whose probability is bigger than 0.7 are accepted. Finally, we obtain Celeb-500K-2R, which contains 25M face images from 356K celebrities.

To evaluate the performance of the refined datasets, we randomly select 500K face pairs from LFW and report AUC (Area Under Curve) of PR (Precision and Recall) curves as

well as VR@FAR (Verification Rate at False Alarm Rate) of ROC (Receiver Operating Characteristic) curves in Tabel 2. The two-step bootstrapping cleaning strategy brings significant performance gains to the models. On the other hand, the models trained on our datasets outperform the model pre-trained on CelebFace by **40-50 percentage** at 0.01% FAR. Note that the images used in Section 4 are all from Celeb-500K-2R. For comprehensive performance analysis, we also select 500K images from 10K persons from Celeb-500K-2R, called Celeb-10K.

4. EXPERIMENT

4.1. Experiment setup

Network Architecture: We adopt ResNet-20 [4] as our baseline model. Besides SoftmaxLoss, CenterLoss [13] or TripletLoss [14] is also employed for some models. In particular, we train CenterLoss with SoftmaxLoss together but we train TripletLoss alone after the model converges on SoftmaxLoss. All the training images are cropped to 112×96 $(h \times w)$ from 128×112 $(h \times w)$ before fed into networks.

Training Process: In Section 4.2, we first pre-train the baselines on Celeb-10K for convergence on larger IN. Then we finetune them on Celeb-500K-2R or MS-Celeb-1M. In Section 4.3 and Section 4.4, we first pre-train the baselines on CelebFace. Then we finetune them on Celeb-500K-2R or MS-Celeb-1M. For pre-training, we set the initial learning rate at 0.02 and decrease it two times to 0.0002 during training. For finetuning, we set the initial learning rate at 0.002, which is decreased two times to 0.00002 during training and the learning rate of the last fully-connected (fc) layer is ten times of those of other layers. All the training process is performed on Caffe [21] with a mini-batch SGD algorithm.

4.2. Comparison with other datasets

To verify the efficacy of the proposed Celeb-500K-2R, we train our baselines on CASIA-WebFace [17], CelebFace [12], MS-Celeb-1M [18] as well as our datasets. Note that we use the cleaned version of MS-Celeb-1M from [20] for fair comparisons. We only report results of other datasets trained on SoftmaxLoss with CenterLoss since this configuration often outperforms those with SoftmasLoss alone and with SotmasLoss and TripletLoss together.

As shown in Table 3, the models trained on Celeb-500K-2R outperform the models trained on other datasets by a large margin. To be specific, the accuracy of Celeb-500K-2R is **0.8 percentage** higher and VR@FAR=0 is more than **35 percentage** higher than those with the models trained on CASIA-WebFace and CelebFace. Further, the accuracy of Celeb-500K-2R is still significantly higher and VR@FAR=0 is more than **1.5 percentage** higher than those with MS-Celeb-1M which contains around 5M images from 70K persons.

Table 3. Comparisons with other state-of-the-art methods and different training datasets on the 6K LFW pairs [22].

Method	Acc on LFW	VR@FAR=0
VGG [16]	97.27%	52.40%
CenterLoss [13]	98.70%	61.40%
CASIA-WebFace	98.40%	60.95%
CelebFace	98.68%	63.75%
MS-Celeb-1M	99.25%	96.72%
Celeb-10K	97.50%	45.02%
Celeb-10K+CenterLoss	98.45%	60.83%
Celeb-10K+TripletLoss	98.50%	60.08%
Celeb-500K-2R	99.33%	98.27%
Celeb-500K-2R+CenterLoss	**99.37%**	98.30%
Celeb-500K-2R+TripletLoss	99.25%	**98.35%**

Since the performance on the 6K LFW pairs [22] is saturated, we report the VR@FAR=10^{-5} of different methods on randomly selected 500K pairs of LFW in Table 4, which shows larger performance gains of Celeb-500K-2R. Specifically, the verification rate of Celeb-500K-2R is around **5 percentage** higher than that of MS-Celeb-1M and more than **45 percentage** higher than those of CASIA-WebFace and CelebFace. Therefore, Celeb-500K-2R is very suitable for practical applications, which require a high verification rate at an extremely low false alarm rate.

Table 4. Verification rates at a 10^{-5} false alarm rate of different methods on the randomly selected 500K LFW pairs.

Method	VR@FAR=10^{-5}
VGG [16]	21.07%
CenterLoss [13]	35.89%
CASIA	36.23%
CelebFace	39.58%
MS-Celeb-1M	79.01%
Celeb-10K	25.38%
Celeb-10K+CenterLoss	36.42%
Celeb-10K+TripletLoss	37.85%
Celeb-500K-2R	83.58%
Celeb-500K-2R+CenterLoss	**84.26%**
Celeb-500K-2R+TripletLoss	83.79%

4.3. Limitation of current metric learning methods

Many metric learning methods have been proposed to improve the performance of face recognition. Their positive influence is significant on CASIA-WebFace and CelebFace.

However, their efficacy on larger datasets is not verified. Consequently, we train baselines with different metric learning methods on both Celeb-10K and Celeb-500K-2R to investigate this issue. As illustrated in Table 3 and Table 4, metric learning methods do significantly improve the performance for Celeb-10K, while they have little performance gain when we train the models on Celeb-500K-2R.

The limitation of metric learning methods lies on their working mechanism. In particular, metric learning methods usually work as a regularizer on the second last fc layer, while SoftmaxLoss works on the last fc layer which maps feature vector to an one-hot classifier vector. When the dataset has several thousand identities such as CelebFace and Celeb-10K, the number of parameters of the second last layer is of the same order of that of the last layer. Thus the metric losses can significantly regularize the model. On the other hand, Celeb-500K-2R contains 356K identities, enlarging the number of parameters of the last fc layer to be around 30 times of that of the second last layer. This makes the regularization of metric losses negligible. Therefore, more effective metric learning methods need to be developed aiming at the last fc layer.

4.4. IN vs AINEI

To study the impact of IN and AINEI on face recognition, we conduct two experiments. First, baselines are trained on the subsets of Celeb-500K-2R with the same IN but different AINEI. As shown in Table 5, the increase of AINEI improves the performance by 4.66 percentage when IN is fixed at 10K, but improvement disappears when IN increases to 100K and 356K.

Table 5. Results on LFW using models trained on subsets of Celeb-500K-2R consisting of different AINEI but fixed IN.

#Images	#Identities	Acc on LFW	VR@FAR=0
50K	10K	97.50%	55.11%
100K	10K	98.20%	57.67%
500K	10K	**98.45%**	**59.77%**
500K	100K	99.25%	96.58%
1M	100K	**99.33%**	96.70%
5M	100K	99.25%	**96.72%**
5M	356K	**99.37%**	98.35%
10M	356K	99.33%	**98.42%**
25M	356K	**99.37%**	98.40%

Second, we train the baselines on datasets with different IN but fix the number of total images, to get rid of the influence of the total dataset size. Therefore, when we increase IN of a dataset, AINEI does not keep unchanged but decreases. Though AINEI decreases, the performance gain of VR@FAR=0 is still $37 - 39$ percentage high as shown in Table 6, after increasing IN from 10K to 100K. This performance gain is around eight times of that promoted by AINEI increase (4.66 percentage) in Table 5. However, when IN is further increased to 356K, the model does not converge, because of too small AINEI.

The results of Table 5 and Table 6 show the importance of keeping a large IN or inter-identity variation when developing a large scale training dataset. The experimental results indicate that images from different persons contain more information than those from the same person. On the other hand, AINEI or intra-person variation should not be too small considering for model convergence. According to our experience, a most appropriate training dataset for a high face recognition performance should consist of as a large IN as possible while a $10 - 20$ AINEI is sufficient.

Table 6. Results on LFW using models trained on subsets of Celeb-500K-2R consisting of different IN but fixed number of images.

#Images	#Identities	Acc on LFW	VR@FAR=0
500K	10K	98.20%	57.67%
500K	100K	**99.25%**	**96.58%**
500K	356K	-	-
1M	10K	98.45%	59.77%
1M	100K	**99.25%**	**96.58%**
1M	356K	-	-

5. CONCLUSION

In this paper, we propose a training dataset for face recognition named Celeb-500K, which contains 50M images from 500K celebrities. After standardizing the images and cleaning the labels, we obtain Celeb-500K-2R, which contains 25M aligned face images from 356K celebrities. We achieve the state-of-the-art performance and beat all the other public datasets based on this cleaned dataset along with a simple baseline network. Further, we reveal two important observations for future study on face recognition. First, the current metric learning methods do not have the expected performance gains on dataset with a large IN. Second, for a large scale face recognition, the increase of intra-identity variation brings a much more significant performance gain than simply increasing intra-identity variation.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61702448, 61672456), the Key R&D Program of Zhejiang Province (No. 2018C03042), and the Fundamental Research Funds for the Central Universities (No. 2017QNA5008, 2017FZA5007).

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[7] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[8] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," *CoRR*, vol. abs/1704.01719, 2017.

[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," *arXiv preprint arXiv:1710.08092*, 2017.

[11] Emily M Hand and Rama Chellappa, "Attributes for improved attributes: A multi-task network for attribute classification," *arXiv preprint arXiv:1604.07360*, 2016.

[12] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," *Advances in neural information processing systems*, pp. 1988–1996, 2014.

[13] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015.

[15] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[16] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015, vol. 1, p. 6.

[17] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[18] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," August 2016.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[20] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light cnn for deep face representation with noisy labels," *arXiv preprint arXiv:1511.02683*, 2015.

[21] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[22] Gary B Huang, Manu Ramesh, Tamara Berg, and Erick Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *Technical Report 07-49, University of Massachusetts, Amherst*, 2007.