MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices

29 Apr 2018 (cs.CV)

https://arxiv.org/abs/1804.07573v2

Abstract

- 100万未満のパラメタを利用したモバイルCNNモデルの提案
- モバイル機器組み込みの高精度リアルタイム認識に特化
- 既存の弱点の分析と、その克服
- MobileNetV2よりも高精度かつ2倍以上の速度を実現
- 携帯電話で0.18秒
- 顔認証のState-of-the-Art

I.INTRODUCTION

- 顔認証はデバイスアンロック、Appログイン、モバイル決済などで使用される重要な技術
- オフラインで使用されることもある→リソスに限りがある
- 小型かつ高精度が望ましいが、深くて大きいCNNはデカイ

I.INTRODUCTION

TABLE II
PERFORMANCE COMPARISON AMONG MOBILE MODELS TRAINED ON CASIA-WEBFACE

Network	LFW Acc.	AgeDB-30 Acc.	Params	Speed (CPU)
MobileNetV1	98.63%	88.95%	3.2M	60ms
ShuffleNet (1 ×, g = 3)	98.70%	89.27%	**0.83M**	27ms
MobileNetV2	98.58%	88.81%	2.1M	49ms
MobileNetV2-GDConv	98.88%	90.67%	2.1M	50ms
MobileFaceNet	**99.28%**	**93.05%**	**0.99M**	24ms
MobileFaceNet (112 × 96)	99.18%	92.96%	0.99M	21ms
MobileFaceNet (96 × 96)	99.08%	92.63%	0.99M	**18ms**
MobileFaceNet-M	99.18%	92.67%	0.92M	24ms
MobileFaceNet-S	99.00%	92.48%	**0.84M**	23ms
MobileFaceNet (ReLU)	99.15%	92.83%	0.98M	23ms
MobileFaceNet (expansion factor×2)	99.10%	92.81%	1.1M	27ms

In the last column, we report actual inference time in milliseconds (ms) on a Qualcomm Snapdragon 820 CPU of a mobile phone with 4 threads (using NCNN [30] inference framework).

- 大きなモデルは、モバイルや組み込みには適していない

TABLE IV
FACE VERIFICATION EVALUATION ON MEGAFACE CHALLENGE1

Method	Protocol	VR @ FAR10^{-6}
SIAT MMLAB [34]	small	76.72%
DeepSense-Small	small	82.85%
SphereFace-Small [20]	small	90.04%
Beijing FaceAll V2	small	77.60%
CosFace (3-patch) [22]	small	**92.22%**
MobileFaceNet	small	85.76%
MobileFaceNet (R)	small	88.09%
MobileFaceNet	large	90.16%
MobileFaceNet (R)	large	92.59%
Google-FaceNet v8 [18]	large	86.47%
SIATMMLAB Tencent Vision	large	87.27%
DeepSense V2	large	95.99%
Vocord-deepVo V3	large	94.96%
CosFace (3-patch) [22]	large	97.96%
iBUG_DeepInsight (ArcFace [5])	large	**98.48%**

"VR" refers to face verification TAR (True Accepted Rate) under 10^{-6} FAR (False Accepted Rate). MobileFaceNet (R) are evaluated on the refined version of MegaFace dataset (c.f. [5]).

I.INTRODUCTION

The major contributions

1. 最後のconv層の後、pooling層ではなくglobal depthwise convolutionを用いる
2. モバイル用の顔の特徴クラスを設計
3. MobileFaceNetsが顔認証のための最先端モバイルCNNと比較して大幅に効率がよいことを実験を通して示した

II.RELATEWORKS

- 視覚認識の最近のアキテクチャ[1,2,3,9]
- １から訓練できる小さなネットワク[9]

  ←AlexNet on ImageNetレベルの精度で1/50のパラメタ量
- MobileNet[1]は深さ方向に分離可能なconv層を用いることで軽量化した
- MobileNetV2 [3]は、線形のボトルネックを持つ逆行列構造に基づいて、モバイルモデルの効率を向上させる
- 顔認証の高精度軽量アキテクチャはない。

II.RELATEWORKS

- 軽量モデルを得る方法にknowledge distillation(知識蒸留)[16]がある
- これは事前訓練されたネットワクを圧縮する
- 今回は使わない

III.APPROACH

- 顔認証のモバイルネットワクの弱点を克服する
- モバイルデバイス上で高精度リアルタイム顔照合CNNモデルへのapproachについて説明もする

- 結果の再現性維持に、ArcFaceロスを使用[5]
  公開デタセット上で学習

A.　顔認証のモバイルネットワクの弱点

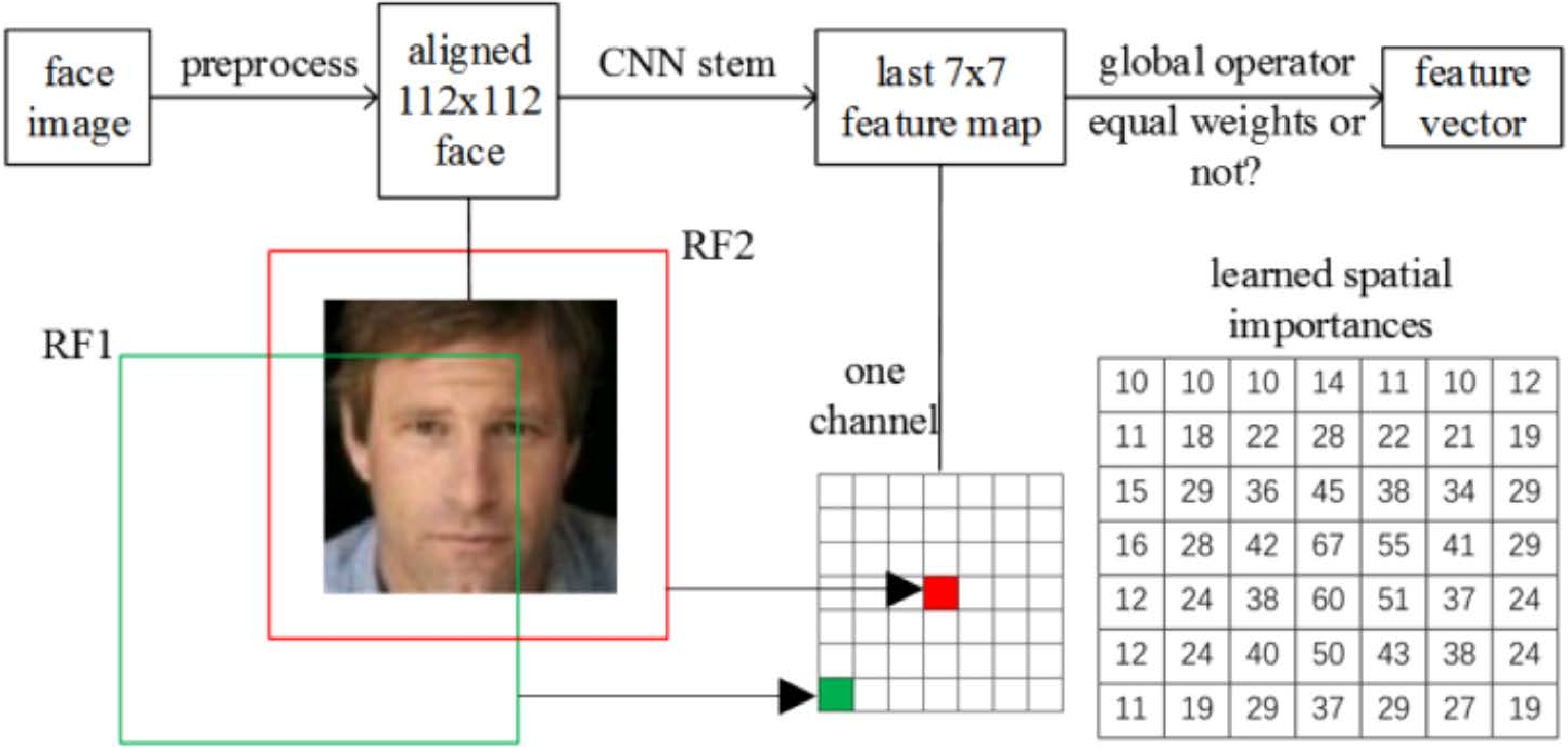
- 視覚認識における最新のモバイルネットワクには、global平均pooling層(GAPool層)がある
- GAPool層があると、それがないCNNよりも精度が低いことが観察されている
- しかし、この現象の理論的分析は行われていない
- 受容場理論[19]でこの現象について簡単な分析を行う

A.　顔認証のモバイルネットワークの弱点

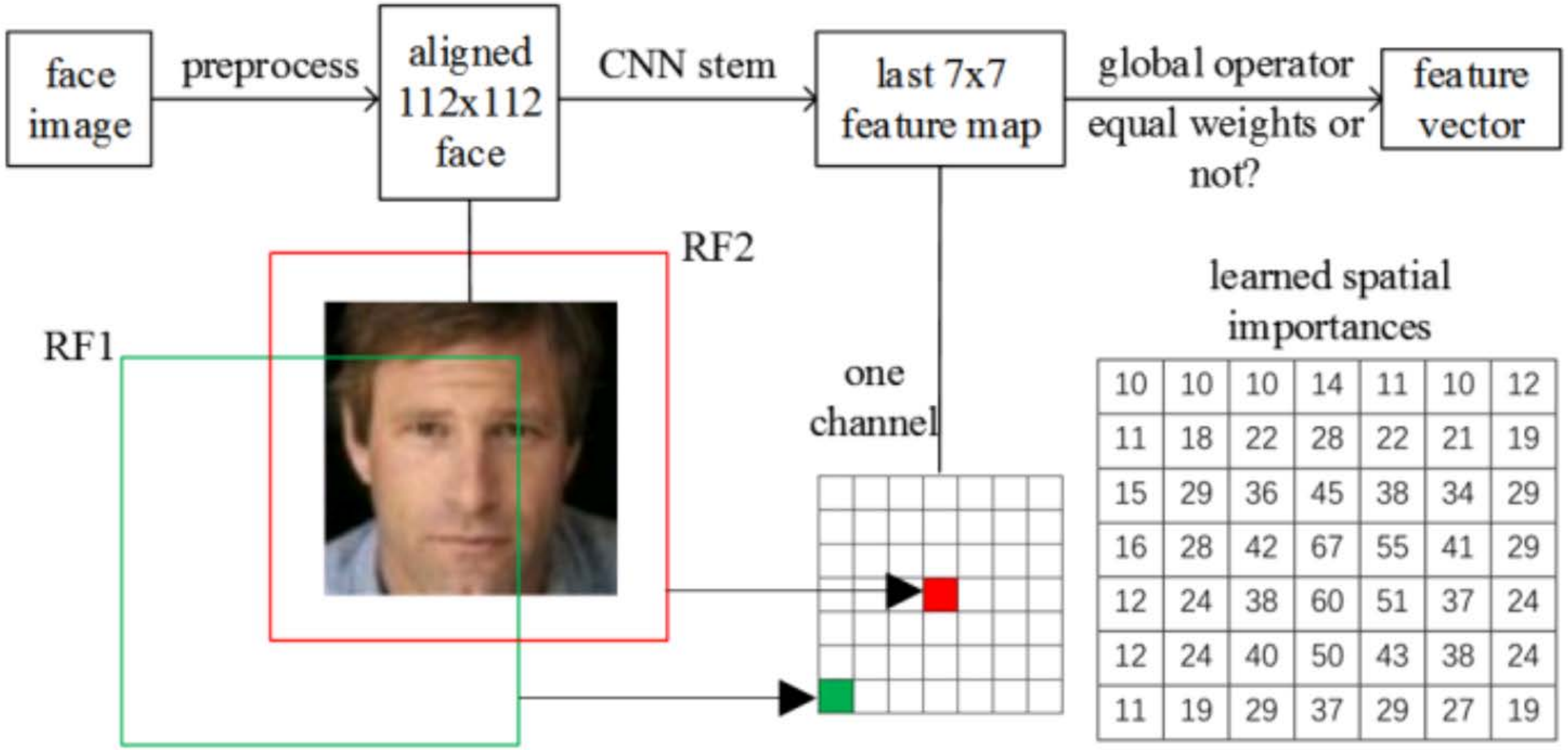
- 一般的に顔認証では
  - preprocessing(前処理)
  - 学習済みモデルでの特徴量抽出
  - 特徴を用いた顔の照合

  がある。
- 顔を検出し、顔ランドマクに基づいて類似度で画像を並べる。
- 112×112のRGB画像の各画素を127.5を128で除算することで正規化
- CNNを埋め込んだ顔特徴は、各整列された顔を特徴ベクトルに写像する

A. 顔認証のモバイルネットワークの弱点



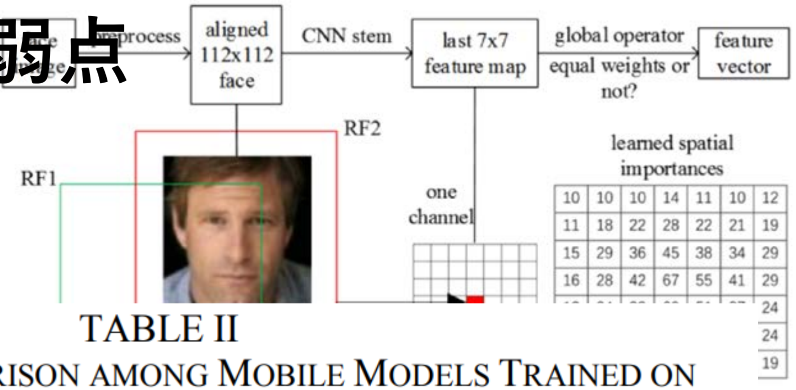
A.　顔認証のモバイルネットワークの弱点



10	10	10	14	11	10	12
11	18	22	28	22	21	19
15	29	36	45	38	34	29
16	28	42	67	55	41	29
12	24	38	60	51	37	24
12	24	40	50	43	38	24
11	19	29	37	29	27	19

- 角の受容野の中心は入力画像の隅にある(緑)
- 中央部の受容野の中心は入力画像の中心(赤)
- 中心にある画素ほど出力に影響を与える
- 出力上の受容野内の衝突分布はほぼガウス分布

A. 顔認証のモバイルネットワークの弱点



- MobileNetV2にはFMap-endは大きすぎる
- 特徴ベクトルとして直接使えない
- GAPool層を特徴ベクトルとして使うのは自然だが、検証精度が悪い

- もうひとつの自然な選択としてGAPool層を接続されたレイヤにし、Fmap-endをコンパクトな特徴ベクトルに投影する方法がある
- これによってモデル全体に多数のパラメタが追加される

TABLE II
PERFORMANCE COMPARISON AMONG MOBILE MODELS TRAINED ON CASIA-WEBFACE

Network	LFW Acc.	AgeDB-30 Acc.	Params	Speed (CPU)
MobileNetV1	98.63%	88.95%	3.2M	60ms
ShuffleNet (1×, g = 3)	98.70%	89.27%	**0.83M**	27ms
MobileNetV2	98.58%	88.81%	2.1M	49ms
MobileNetV2-GDConv	98.88%	90.67%	2.1M	50ms
MobileFaceNet	**99.28%**	**93.05%**	**0.99M**	24ms
MobileFaceNet (112 ×96)	99.18%	92.96%	0.99M	21ms
MobileFaceNet (96 ×96)	99.08%	92.63%	0.99M	**18ms**
MobileFaceNet-M	99.18%	92.67%	0.92M	24ms
MobileFaceNet-S	99.00%	92.48%	**0.84M**	23ms
MobileFaceNet (ReLU)	99.15%	92.83%	0.98M	23ms
MobileFaceNet (expansion factor×2)	99.10%	92.81%	1.1M	27ms

B.　グローバル深度畳み込み

- 重要度の異なるFMap-endを扱うために、GAPool層をグロバル深度畳み込み層（GDConv）で置き換える
- GDConv={入力サイズ=0, パッド= 0, ストライド= 1}

$$G_m = \Sigma_{i,j} K_{i,j,m} \cdot F_{i,j,m} \qquad (1)$$

$$W \cdot \mathrm{H} \cdot M \qquad (2)$$

- F:　入力特徴マップ　HxWxM
- K:　conv kernel　　HxWxM
- G:　出力　　　　　　1x1xM
- GDConvのあるMobileNetVとないNetでは、あるほうがLFWとAgeDBで大幅に精度が向上する
- GDConv層は、MobileFaceNetsにとって効率的な構造

III.APPROACH

C.　MobileFaceNetアキテクチャ

- MobileNetV2よりもずっと小さく
- 非線形としてPReLUを使用

1. ネットワクの始めに高速ダウンサンプリング戦略
2. 最後のいくつかのConv層で早期の次元削減戦略
3. 特徴出力層

- としてGDConvの後に1x1Conv層

TABLE I
MOBILEFACENET ARCHITECTURE FOR FEATURE EMBEDDING

Input	Operator	t	c	n	s
$112^2 \times 3$	conv3x3	-	64	1	2
$56^2 \times 64$	depthwise conv3x3	-	64	1	1
$56^2 \times 64$	bottleneck	2	64	5	2
$28^2 \times 64$	bottleneck	4	128	1	2
$14^2 \times 128$	bottleneck	2	128	6	1
$14^2 \times 128$	bottleneck	4	128	1	2
$7^2 \times 128$	bottleneck	2	128	2	1
$7^2 \times 128$	conv1x1	-	512	1	1
$7^2 \times 512$	linear GDConv7x7	-	512	1	1
$1^2 \times 512$	linear conv1x1	-	128	1	1

IV.EXPERIMENTS

- MobileFaceNetモデルとベスラインモデルの学習設定についての説明
- 複数の最新の顔認証モデルとの性能比較します。

IV.EXPERIMENTS

A. Training settings and accuracy comparison on LFW and AgeDB

- MobileNetV1、ShuffleNet、MobileNetV2を使用
- Stride = 2の設定が非常に精度が低い
- 最初の畳み込みレイヤではストライド= 1
- すべてのモデルはCASIA-Webfaceセットと、ArcFace lossで訓練済み
- モデルを最適化にモメンタムSDG(0.9)を使用
- bs=512
- lr=0.1　~
- 60Kイテレション

B. Evaluation on MegaFace Challenge1

TABLE IV

FACE VERIFICATION EVALUATION ON MEGAFACE CHALLENGE1

Method	Protocol	VR @ FAR10^{-6}
SIAT MMLAB [34]	small	76.72%
DeepSense-Small	small	82.85%
SphereFace-Small [20]	small	90.04%
Beijing FaceAll V2	small	77.60%
CosFace (3-patch) [22]	small	**92.22%**
MobileFaceNet	small	85.76%
MobileFaceNet (R)	small	88.09%
MobileFaceNet	large	90.16%
MobileFaceNet (R)	large	92.59%
Google-FaceNet v8 [18]	large	86.47%
SIATMMLAB Tencent Vision	large	87.27%
DeepSense V2	large	95.99%
Vocord-deepVo V3	large	94.96%
CosFace (3-patch) [22]	large	97.96%
iBUG_DeepInsight (ArcFace [5])	large	**98.48%**

"VR" refers to face verification TAR (True Accepted Rate) under 10^{-6} FAR (False Accepted Rate). MobileFaceNet (R) are evaluated on the refined version of MegaFace dataset (c.f. [5]).

V.CONCLUSION

- MobileFaceNetの提案
- モバイル端末上でのリアルタイム顔認識
- モバイルCNNと比較して効率が良くなった

REFERENCES

References

[1] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," CoRR, abs/1704.04861, 2017.

[2] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," CoRR, abs/1707.01083, 2017.

[3] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," CoRR, abs/1801.04381, 2018.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," arXiv preprint arXiv: 1607.08221, 2016.

[5] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," arXiv preprint arXiv: 1801.07698, 2018.

[6] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Workshop on Faces in "Real-Life" Images: Detection, Alignment, and Recognition, 2007.

[7] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard, "The megaface benchmark: 1 million faces for recognition at scale," In Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. "Agedb: The first manually collected in-the-wild age database," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2017.

[9] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In Advances in neural information processing systems, pages 1097–1105, 2012.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," In IEEE Conference on

[24] Wenjie Luo, Yujia Li, Raquel Urtasun, Richard Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," In Advances in Neural Information Processing Systems 29 (NIPS 2016), pages 4898-4906, 2016.

[25] F. Chollet. "Xception: Deep learning with depthwise separable convolutions," arXiv preprint arXiv:1610.02357, 2016.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint arXiv:1411.7923, 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," In Proceedings of the IEEE international conference on computer vision, pages 1026–1034, 2015.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," In International Conference on Machine Learning, pages 448–456, 2015.

[29] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, Dmitry Kalenichenko, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference." arXiv preprint arXiv: 1712.05877, 2017.

[30] "NCNN: a high-performance neural network inference framework optimized for the mobile platform," url: https://github.com/Tencent/ncnn, the version in Apr 20, 2018.

[31] Yaniv Taigman, Ming Yang, Marc' Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," In Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al, "Deep face recognition," In BMVC, volume 1, page 6, 2015.

[33] Yi Sun, Xiao gang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," In Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," In European Conference on Computer Vision (ECCV), pages 499–515, 2016.

[35] Weihong Deng, Binghui Chen, Yuke Fang, Jiani Hu, "Deep Correlation Feature Learning for Face Verification in the Wild," IEEE Signal Processing Letters, PP.99:1-1, 2017.

[36] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," In IEEE International Conference on Image Processing (ICIP),