ShuffleFaceNet: A Lightweight Face Architecture for Efficient and Highly-Accurate Face Recognition

Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Miguel Nicolás-Díaz
Advanced Technologies Application Center (CENATAV)
7A ♯21406 Siboney, Playa, P.C. 12200, Havana, Cuba
{ymartinez,hmendez,mnicolas}@cenatav.co.cu

Luis S. Luevano, Leonardo Chang, Miguel Gonzalez-Mendoza
Tecnologico de Monterrey, School of Engineering and Science,
México
luisluevano@outlook.com, {lchang,mgonza}@tec.mx

Abstract

The recent success of convolutional neural networks has led to the development of a variety of new effective and efficient architectures. However, few of them have been designed for the specific case of face recognition. Inspired on the state-of-the-art ShuffleNetV2 model, a lightweight face architecture is presented in this paper. The proposal, named ShuffleFaceNet, introduces significant modifications in order to improve face recognition accuracy. First, the Global Average Pooling layer is replaced by a Global Depth-wise Convolution layer, and Parametric Rectified Linear Unit is used as a non-linear activation function. Under the same experimental conditions, ShuffleFaceNet achieves significantly superior accuracy than the original ShuffleNetV2, maintaining the same speed and compact storage. In addition, extensive experiments conducted on three challenging benchmark face datasets, show that our proposal improves not only state-of-the-art lightweight models but also very deep face recognition models.

1. Introduction

Deep neural networks (DNNs) have recently achieved a series of breakthroughs in many computer vision tasks, including unconstrained face recognition [33]. However, modern highly-accurate face recognition methods are usually built upon very deep convolutional neural networks (CNNs) [6, 26, 39] which means that they comprise a long sequence of convolutional layers. As a result, the models require a high amount of computational resources such as large memory and powerful GPUs in order to achieve high-performance results. For example, the widely used VGG-Face network [26] involves 138 million parameters, more than 500 MB memory size and over 15G floating-point operations (FLOPs) to classify a single 224 $\times$ 224 image. Therefore, this kind of models is usually unable to deploy on real-time applications or computationally limited platforms such as robots, smart phones and, mobile devices.

In order to overcome these limitations, recent developments have focused on building small and efficient neural networks without significantly deteriorating their performance. Some approaches have tried to compress or accelerate pre-trained networks by using techniques such as pruning [9], knowledge distillation [11], low-rank approximation [41] and quantization [16]. In the last few years, developing lightweight deep neural networks is one of the most promising solutions to obtain better speed-accuracy trade-off [14, 40, 19, 27]. SqueezeNet [14], MobileNets [27] and ShuffleNets [40, 19] are among the most popular ones for common visual recognition tasks, showing impressive results. However, just a few works have proposed accurate lightweight architectures specifically designed for face recognition [2, 4, 7, 36, 37], so this topic deserves further attention.

In this paper, we propose a new lightweight architecture named ShuffleFaceNet, that extends the extremely efficient network ShuffleNetV2 [19] to the domain of face recognition. Firstly, we replace the Global Average Pooling layer for a Global Depth-wise Convolution layer in order to obtain a more discriminative face representation. Secondly, we use Parametric Rectified Linear Unit (PReLU) as a non-linear activation function due to its accuracy improvement over the Rectified Linear Unit (ReLU) function. As a result, we designed four ShuffleFaceNet models with different complexity levels that use less than 4.5 millions parameters,

with a maximum computational complexity and model size of 1.05G FLOPs and 18 MB, respectively. The experiments conducted on images and videos benchmark datasets show that our proposal improves both state-of-the-art lightweight and very deep CNNs for face recognition. The major contributions of this work are summarized as follows:

- We ensure not only speed and compact storage space, but also significant improvements on face recognition accuracy, by using a Global Depth-wise Convolution layer to output a discriminative feature vector and PReLU as a non-linear activation function.

- We design an efficient and accurate lightweight face architecture, with four different complexity levels. The resulting ShuffleFaceNet models are less than 20 MB of size and have an actual inference CPU time of about 37 ms, which is suitable for deploying on real-time applications, as well as, mobile and embedded devices.

- We demonstrate that the proposal achieves better performance than state-of-the-art lightweight CNNs as well as very deep CNNs, on two popular face recognition benchmarks and on a recently released lightweight face recognition challenge.

The paper is organized as follows. Section 2 reviews the existing lightweight CNNs for face recognition. Section 3 introduces the lightweight ShuffleFaceNet architecture proposed for face recognition. Experimental results are given in Section 4, and finally, we conclude in Section 5.

2. Related Work

Deploying efficient and lightweight deep face recognition architectures remains a challenge for real-world applications. A light CNN framework was proposed in [37] to obtain a compact and low dimensional face representation. For this, the authors introduce an alternative of ReLU function, named Max-Feature-Map (MFM), to suppress the low-activation neurons in each convolutional layer. Also, small convolution filters, Network in Network layers, and Residual Blocks are used to reduce parameter space and improve performance. Three architectures for light CNNs were evaluated, showing better performance in terms of speed and storage space compared with state-of-the-art big face models. However, their most accurate architecture (light CNN-29) has 12.6 million parameters and about 3.9G FLOPs, which make it not-so-suitable for mobile and embedded platforms. An alternative to spatial convolutions, named shift operation was presented in [36], which requires zero FLOPs and zero parameters and can be easily and efficiently implemented. To demonstrate the shift operation's effectiveness in the face recognition task, the authors take the original FaceNet [28] and propose a new ShiftFaceNet model that reduces the parameter size by 35×, with at most 2% drop of accuracy in above three verification benchmarks. LMobileNetE [4] is an improved version of MobileNet [27] that achieves comparable face verification accuracy with the lowest running time, but the model size is 112MB, which is a larger-sized model rather than lightweight.

Recently, MobileFaceNets [2] were introduced for high-accuracy and real-time face verification on mobile and embedded devices. Experiments on the Labeled Faces in the Wild (LFW) database showed that MobileFaceNets achieve an accuracy similar to that of state-of-the-art large size models, with faster inference speed. However, in the case of the MegaFace dataset, accuracy decreases slightly. MobiFace [7] is another lightweight CNN designed for face recognition on mobile devices, which adopts fast downsampling and bottleneck residual block with the expansion layers and achieves high performance with 99.7% on LFW database and 91.3% on MegaFace database.

Regardless of the few works focused on highly efficient and lightweight architectures specifically designed for face recognition, there are other lightweight CNNs [14, 40, 19] that have shown excellent performance in image classification tasks and deserve further attention in face recognition. For example, SqueezeNet [14] is a very small CNN architecture that achieves AlexNet-level accuracy on ImageNet with 50× fewer parameters. Its success is given by three main strategies: first, replace 3×3 filters with 1×1 filters, which has 9× fewer parameters; second, decrease the number of input channels to 3×3 filters and finally, downsample late in the network so that convolution layers have large activation maps. Zhang et al. [40] proposed an extremely computational efficient CNN architecture named ShuffleNet, which utilizes point-wise group convolution and channel shuffle operations. Compared with MobileNet, it achieves superior performance by a significant margin on ImageNet classification. ShuffleNetV2 [19] was inspired by ShuffleNet, but considering some practical aspects in its design to get a more efficient network architecture while maintaining high levels of accuracy. For this, a simple operator, named channel split, was introduced, allowing to maintain a large number and equally wide channels with neither dense convolution nor too many groups.

3. ShuffleFaceNet Architecture

In this section, we detail the lightweight ShuffleFaceNet architecture designed for face recognition. The proposal is inspired by the state-of-the-art network ShuffleNetV2 [19], but adding some strategies aimed at improving its robustness on this task.

Most of the deep networks designed for image classification, including ShuffleNetV2, use the output of the Global

Name	Kernel/Stride	Output Size	Output Channels			
			$0.5\times$	$1\times$	$1.5\times$	$2\times$
Image	-	112×112	3	3	3	3
Conv1	$3 \times 3/2$	56×56	24	24	24	24
Stage2	-	28×28	48	116	176	244
Stage3	-	14×14	96	232	352	488
Stage4	-	7×7	192	464	704	976
Conv5	$1 \times 1/1$	7×7	1024	1024	1024	2048
GDConv	$7\times7/1$	1×1	1024	1024	1024	1024
LinearConv	$1\times1/1$	1×1	128	128	128	128

Table 1. ShuffleFaceNet architecture for four different levels of complexity.

Average Pooling (GAP) layer as a feature vector in the embedding process. However, in the case of face recognition, this strategy has shown to be less accurate [37, 2, 4]. This is due to GAP layer treats each unit of the output feature map equally, which is not consistent with the theory that different kinds of units bring more or less discriminative information for extracting a face feature vector. Using a Fully Connected (FC) layer instead, allows us to learn different weights to these units and project the information embedded into a compact face feature vector. Nevertheless, the FC layer ended up having a large number of weights, which not only requires more computational power but also increases the model size. Recently, a Global Depth-wise Convolution (GDC) layer was used in [2] to treat different units of the output feature map with different importance, showing be an efficient structure for face recognition. In this work, we replace the GAP layer of ShuffleNetV2 with a GDC layer.

On the other hand, ShuffleNetV2 model is based on the ReLU activation function [24] that offers usually high dimensional and sparse features. To alleviate this problem, several activation functions have been proposed [37, 10, 20, 38]. We choose PReLU [10] as the non-linearity rather than ReLU which has shown to be better for face recognition [1, 2, 7] since it allows negative responses that in turn improves the network performance.

In addition, we use a fast downsampling strategy at the beginning of our network and a linear 1×1 convolution layer following a GDC layer, as the feature output layer. Consequently, a compact 128-dimensional face representation is obtained.

The detailed structure of the proposed ShuffleFaceNet architecture is shown in Table 1, the number of channels in each block is scaled to generate networks of different complexities, denoted as $0.5\times$, $1\times$, $1.5\times$ and $2\times$. The building blocks in Stages 2-4 consist of DenseNet blocks [12].

4. Experimental evaluation

In this section, we assess the performance of our lightweight ShuffleFaceNet architecture from the aspects of accuracy, speed, and model size. The proposal is compared with several state-of-the-art models on three benchmark datasets for face recognition.

4.1. Training and Network settings

We use as training set the cleaned MS1M dataset [8], which contains 5.1 million face images of 93K identities. We take Labeled Faces in the Wild (LFW) [13], Celebrities in Frontal Profile (CFP) [29] and Age Database (AgeDB) [23] as the validation sets. All face images are re-aligned to the size of 112×112 by using the RetinaFace detector [5], and each pixel (ranged between [0; 255]) in RGB images is normalized by subtracting 127.5 then divided by 128. All feature embedding dimensions are set to 128.

We set the batch size as low as 256 and train models on two Nvidia GeForce GTX 1080Ti (11GB) GPUs. The learning rate starts from 0.1, and it is divided by 10 at the 100K, 140K, 160K iterations. The total iteration step is set as 200K. We used a Stochastic Gradient Descent optimizer, setting the momentum at 0.9 and weight decay at 5e-4. The parameter initialization for convolution is Xavier with random sampling from a Gaussian normal distribution. All experiments are implemented on the MxNet framework [3].

We trained ShuffleFaceNet with four complexity levels and different loss functions such as SoftMax, CosFace [32] and ArcFace [4]. Table 2 shows the verification accuracy obtained on the LFW, CFP-FP, and AgeDB datasets for each model, as well as the number of parameters, the model sizes, and the complexity in terms of FLOPs.

As we can see, in the case of LFW, the performance improvement for each ShuffleFaceNet with a different complexity level is not significant, since this dataset is almost saturated. In contrast, both CosFace and ArcFace outperform SoftMax, especially under large pose and age variations. In addition, we find that CosFace and ArcFace perform very similar; however, this last one is slightly better.

As expected, the higher the complexity, the higher the accuracy. However, there is not a significant difference between levels $1.5\times$ and $2\times$. Thus, in order to get a better

Method	Complexity FLOPs	# Params. (M)	Model size (MB)	Loss Function	Accuracy LFW	Accuracy CFP-FP	Accuracy AgeDB
ShuffleFaceNet 0.5×	66.9M	0.5	1.9	SoftMax	96.87 ± 1.2	88.60 ± 1.4	78.33 ± 2.0
				CosFace	**99.23 ± 0.5**	**92.59 ± 1.4**	**93.22 ± 1.4**
				ArcFace	99.07 ± 0.5	91.87 ± 1.5	92.45 ± 1.7
ShuffleFaceNet 1×	275.8M	1.4	5.6	SoftMax	96.91 ± 0.8	90.42 ± 2.0	80.40 ± 1.5
				CosFace	99.42 ± 0.3	95.07 ± 0.7	95.13 ± 1.0
				ArcFace	**99.45 ± 0.4**	**96.04 ± 0.9**	**96.33 ± 0.7**
ShuffleFaceNet 1.5×	577.5M	2.6	10.5	SoftMax	96.37 ± 0.9	90.57 ± 1.0	80.37 ± 2.7
				CosFace	99.62 ± 0.2	96.79 ± 0.6	96.75 ± 0.7
				ArcFace	**99.67 ± 0.3**	**97.26 ± 0.7**	**97.32 ± 0.8**
ShuffleFaceNet 2×	1.05G	4.5	18	SoftMax	97.15 ± 0.7	91.57 ± 1.0	81.95 ± 2.9
				CosFace	99.58 ± 0.3	97.33 ± 0.6	97.08 ± 0.9
				ArcFace	**99.62 ± 0.4**	**97.56 ± 0.6**	**97.28 ± 0.8**

Table 2. Verification results (%) on LFW, CFP-FP and AgeDB databases for different loss functions.

Method	LFW Accuracy	CFP-FP Accuracy	AgeDB Accuracy	Complexity (FLOPs)	#Params. (M)	Model Size (MB)	GPU Speed (ms)
ShuffleNetV2* 1.5×	99.52 ± 0.4	96.21 ± 1.1	94.78 ± 1.1	577.3M	2.5	10.1	0.77
ShuffleFaceNet 1.5×	99.67 ± 0.3	97.26 ± 0.7	97.32 ± 0.8	577.5M	2.6	10.5	0.77

Table 3. Performance comparison between our ShuffleFaceNet and the original ShuffleNetV2 on LFW, CFP-FP and AgeDB databases.

speed-accuracy trade-off, we have decided to use in the remaining experiments, the ShuffleFaceNet 1.5× trained with ArcFace loss function.

4.1.1 Comparison with ShuffleNetV2 architecture

We compare our ShuffleFaceNet 1.5× with the original ShuffleNetV2 1.5× [19] in order to show the advantages of the proposal for the case of face recognition. For a fair comparison, ShuffleNetV2 1.5× is trained from scratch by ArcFace loss function under the same training setting as our ShuffleFaceNet. In the rest of this paper we will refer to the resulting model as ShuffleNetV2* 1.5×.

Table 3 presents the verification accuracy of the tested models on LFW, CFP-FP and AgeDB datasets. In addition, it shows the number of parameters, the model size and the inference time. It can be seen that the proposal outperforms the original ShuffleNetV2 model on the three datasets. On the other hand, although the model size and number of parameters increase a little bit, the inference time remains the same. This means that the guidelines considered in the design of the efficient ShuffleNetV2 were maintained [19].

4.2. Performance assessment

In order to evaluate the effectiveness of ShuffleFaceNet 1.5×, we conducted several experiments on two popular benchmarks as well as on a recently released Lightweight Face Recognition Challenge. Note that, we do not re-train or fine-tune the ShuffleFaceNet model on any training set of the testing database. Thus, we directly extract the features of the ShuffleFaceNet learned on the cleaned MS1M dataset described previously, and perform the comparison of these features by a metric.

4.2.1 MegaFace Challenge 1 on FaceScrub

Method	Rank-1	VR@FAR=10^{-6}
Vocord-DeepVo1	75.1	67.3
Deepsense-large	74.8	87.8
CenterLoss [34]	65.2	76.5
FaceNet [28]	70.5	86.5
CosFace [32]	82.7	96.7
ResNet50-ArcFace [4]	77.5	92.3
ResNet100-ArcFace [4]	81.0	97.0
Light CNN-4 [37]	60.2	62.3
Light CNN-9 [37]	67.1	77.5
Light CNN-29 [37]	73.5	84.7
MobileFaceNet [2]	-	90.2
ShuffleNetV2* 1.5×	69.6	84.1
ShuffleFaceNet 1.5×	77.4	93.0

Table 4. Performance evaluation on the MegaFace Challenge 1 using FaceScrub as test set. Rank-1 refers to face identification accuracy (%) at first position with 1 million distractors, and VR (%) corresponds to the verification TAR for a FAR value of 10^{-6}.

The MegaFace database [17] is one of the largest publicly available testing benchmarks for evaluating the performance of face recognition algorithms at the million scale of distractors. MegaFace includes a gallery set and a probe set. The gallery set consists of a subset of Flickr photos from Yahoo [31], containing more than one million images from 690K different individuals. The probe sets are two existing databases: FaceScrub and FGNet. In this work, we use FaceScrub [25] as the probe set that contains 100K photos of 530 unique individuals.

Table 4 summarizes the results obtained by the proposed ShuffleFaceNet $1.5\times$ and the original ShuffleNetV2* $1.5\times$ compared with state-of-the-art methods reported for both the identification and verification tasks. True Acceptance Rate (TAR) under False Acceptance Rate (FAR) of 10^{-6} is used to report the verification results, while the Rank-1 face accuracy is employed to the case of identification. Since our training set has more than 0.5 million images, it is regarded as large.

As we can see, on this large database, ShuffleFaceNet outperforms ShuffleNetV2* by almost 9% for both Rank-1 and VR@FAR=10^{-6} evaluation measures and shows its superiority with respect to the rest of light models such as MobileFaceNet and Light CNNs -4,-9 and -29. The obtained results are even better than existing very deep models such as Vocord and Deepsense, which are provided as baseline methods on the benchmark [17]. Other very deep models which provide better results need higher computational resources and storage space, as we will analyze later. For example, the LResNet100E model, that obtains the best results in this benchmark database, has a size of 250 MB.

4.2.2 Evaluation on YouTube Face database

The YouTube Faces (YTF) database [35] is a large video dataset for unconstrained face recognition in videos. It contains 3,425 videos of 1,595 subjects with significant variations on expression, illumination, pose, resolution, and background. An average of 2.15 videos are available for each subject. The average length of a video is 181.3 frames. For the standard protocol of the YTF database, a pair-matching benchmark corresponding to 5000 video pairs is provided. Specifically, these pairs are divided into ten splits, each one containing 250 positive pairs and 250 negative ones.

For each YTF video, we selected the 50 most frontal frames and compute the corresponding face descriptors. Finally, the video is represented by the average of the 50 face descriptors, and cosine similarity is used for comparison.

We compare the performance of our SuffleFaceNet $1.5\times$ with ShuffleNetV2* $1.5\times$ and state-of-the-art methods reported on this database. Table 5 presents the obtained verification results, in terms of the three metrics that are usually considered to report the verification results: the mean accuracy, the area under the curve (AUC) and the equal error rate (EER). It can be seen that the proposal achieves state-of-the-art results and also, in this case, outperforms the original ShuffleNetV2* for the three used metrics.

Method	Accuracy	AUC	EER
LBinVF^2 [21]	83.3	93.2	14.6
DeepFace-single [30]	91.4	96.3	8.6
ShiftFaceNet [36]	90.1	96.1	-
NAN [39]	95.7	98.8	-
FaceNet [28]	95.1	-	-
Light CNN-29 [37]	95.5	-	-
SphereFace [18]	95.0	-	-
VGG-Face [26]	97.3	-	-
CosFace [32]	97.6	-	-
CenterLoss [34]	94.9	-	-
TBE-CNN [6]	95.0	-	-
ShuffleNetV2* $1.5\times$	93.3	97.7	7.0
ShuffleFaceNet $1.5\times$	95.7	98.2	5.3

Table 5. Verification results (%) on YouTube Face database.

Recently, new relevant evaluation protocols were proposed for the YTF database: REP-YTF [22], containing open/closed-set identification for both video-to-video and video-to-image comparisons. Under these protocols, the YTF database is divided into ten random trials of training and test sets. On each trial, for both open and closed-set identification protocols, three different configurations of the test set are obtained by using the openness values: 0.2, 0.5, and 0.9, resulting in different gallery sizes. Performance metrics are computed and averaged over the ten random trials, and the standard deviation is also reported.

We compare ShuffleFaceNet $1.5\times$ and ShuffleNetV2* $1.5\times$ with the three best-performing methods reported on the REP-YTF by using open and closed-set identification protocols. Based on the results obtained on [22], we choose the LDA metric learning to perform the comparison. For the open-set identification protocol, Table 6 shows the mean Detection and Identification Rate (DIR) at rank-1 and the corresponding standard deviation over the 10 trials for a False Acceptance Rate (FAR) of 1% in both video-to-video and video-to-image scenarios. In the case of the closed-set identification protocol, the recognition rates at rank-1 are presented in Table 7. As we can see, under this protocol, the differences between ShuffleFaceNet and ShuffleNetV2* are also in general about 10% and both of them outperform baseline methods by a great margin. For example, in the video-to-image open set scenario, ShuffleFaceNet outperforms the accuracy of the popular VGG-Face model in more than 40%.

Method	video-to-video			video-to-image		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
LBinVF^2 + LDA	10.05 ± 2.1	8.57 ± 0.8	8.18 ± 1.0	6.58 ± 1.5	4.78 ± 0.8	4.53 ± 0.5
VGG-Face + JB	22.83 ± 3.6	18.16 ± 1.8	16.28 ± 1.5	17.33 ± 2.9	14.20 ± 2.4	13.14 ± 1.1
Dlib + LDA	25.97 ± 3.0	20.12 ± 1.2	17.99 ± 1.5	16.62 ± 4.2	14.26 ± 1.7	11.41 ± 1.0
ShuffleNetV2* $1.5\times$ + LDA	51.63 ± 4.7	45.70 ± 3.5	42.90 ± 3.2	49.43 ± 3.6	43.06 ± 1.7	39.86 ± 2.4
ShuffleFaceNet $1.5\times$ + LDA	59.79 ± 5.3	54.21 ± 3.3	51.44 ± 3.8	64.33 ± 3.5	59.64 ± 2.0	57.57 ± 3.0

Table 6. Performance comparison on YouTube Face database for REP-YTF open-set identification protocol in video-to-video and video-to-image scenarios. The results are reported as the mean DIR (%) at rank-1 and FAR = 1%.

Method	video-to-video			video-to-image		
	Op (0.2)	Op (0.5)	Op (0.9)	Op (0.2)	Op (0.5)	Op (0.9)
LBinVF^2 + LDA	37.70 ± 3.5	30.93 ± 1.6	29.85 ± 1.3	31.55 ± 3.8	24.46 ± 1.7	24.01 ± 1.4
VGG-Face + LDA	60.60 ± 3.2	54.59 ± 1.6	52.33 ± 1.3	53.87 ± 3.3	47.49 ± 1.8	45.02 ± 1.1
Dlib + LDA	71.91 ± 2.0	66.53 ± 1.7	64.18 ± 0.8	56.78 ± 3.4	50.90 ± 2.7	48.33 ± 2.1
ShuffleNetV2* $1.5\times$ + LDA	82.65 ± 2.4	79.22 ± 1.4	78.03 ± 1.2	76.88 ± 2.4	72.38 ± 1.4	71.20 ± 1.2
ShuffleFaceNet $1.5\times$ + LDA	86.83 ± 2.1	85.52 ± 1.1	84.61 ± 1.0	84.40 ± 2.1	81.89 ± 1.1	80.36 ± 1.0

Table 7. Identification rates at rank-1 on YouTube Face database for the REP-YTF closed-set identification protocol in video-to-video and video-to-image scenarios.

4.2.3 Lightweight Face Recognition Challenge

We conduct an additional experiment on the recently released ICCV 2019 Lightweight Face Recognition (LFR) Challenge [15]. Specifically, we participated in the Track 1 which requires float32 solutions with a computational complexity less than 1G FLOPs, a model size up to 20 MB and feature dimension up to 512. The selected ShuffleFaceNet $1.5\times$ model, based on the validation results in section 4.1, fulfill the requirements of this track. This model is evaluated and compared with the provided baseline MobileFaceNet in the two testing sets: large-scale images and large-scale videos.

The Trillion-Pairs dataset (deepglint-light) is used as the large-scale image test set, which contains about 274K face images from 5.7K identities from celebrities in the LFW name list and 1.58M face images from Flickr as distractors. In the case of the large-scale video test set, the iQIYI-VID test set (iQIYI-light) is used, that includes 200K videos of 10K identities. All test images were preprocessed to the size of 112×112 similar to the training data.

Table 8 shows the results obtained on the LFR Challenge in terms of True Positive Rate at a given False Acceptance Rate. Compared to the MobileFaceNet baseline, it can be seen the improvement of our ShuffleFaceNet architecture on the image test set, while the performance is very close to that obtained by the baseline model on the video dataset.

Method	deepglint-light	iQIYI-light
MobileFaceNet-baseline	64.69	47.19
ShuffleFaceNet $1.5\times$	75.31	44.55

Table 8. Results on the LFR Challenge reported in terms of TPR@FAR=10^{-8} and TPR@FAR=10^{-4} for the large-scale image (deepglint-light) and video (iQIYI-light) test sets, respectively.

4.3. Performance on Speed and Storage Space

In this section, we aim at evaluating the speed and storage space of the proposed architecture, in order to demonstrate its feasibility in real-time applications or computationally limited platforms.

We first compare the actual speed of our ShuffleFaceNets models with the ShuffleNetV2 $1.5\times$ and the MobileFaceNet model provided on the LFR Challenge [15]. The efficiency measurements were obtained by performing inference over 12,000 images from the Labeled Faces in the Wild dataset. We measured the inference time per image of these models on four different devices including CPU Intel i7-7700HQ (Mobile processor), Nvidia Quadro P2000, Nvidia GeForce GTX 1050Ti Mobile and Nvidia GeForce GTX 1660Ti. The obtained results are shown in Table 9. For our ShuffleFaceNet architecture, similar to the accuracy behavior, as complexity increases the inference time increases. Taking as reference the ShuffleFaceNet $1.5\times$, that was selected as the one with the better speed-accuracy trade-off, we can see that it is clearly faster than ShuffleNetV2 $1.5\times$ and MobileFaceNet models, especially on CPU, being MobileFaceNet the slowest. Moreover, on the different GPUs, the inference time of all evaluated models decrease considerably compared to GPU's. However, ShuffleFaceNet $1.5\times$ is faster than all the counterparts.

Method	Speed (miliseconds)			
	Mobile Intel i7	Quadro P200	1050 Ti	1660 Ti
MobileFaceNet [15]	62.4	5.5	7.3	3.3
ShuffleNetV2* 1.5×	33.0	5.3	12.3	2.8
ShuffleFaceNet 0.5×	12.0	1.4	2.0	0.7
ShuffleFaceNet 1×	23.8	3.0	3.3	1.2
ShuffleFaceNet 1.5×	29.1	4.7	4.7	1.9
ShuffleFaceNet 2×	37.5	6.4	6.9	2.4

Table 9. Comparison of inference time in different devices.

Table 10 presents the computational complexity in terms of FLOPs, the number of parameters and the model size of our proposed architecture taking as reference ShuffleFaceNet 1.5×, compared with several state-of-the-art face deep models that have been tested in previous sections. It can be seen that the proposal is one of the lightest from the evaluated models. This, together with its fast inference speed discussed above, make our ShuffleFaceNet 1.5× highly suitable for real-time or computationally limited face recognition applications.

Method	Complexity (FLOPs)	#Param. (M)	Model size
FaceNet [28]	1.6B	7.5	30
VGG-Face [26]	15G	138	526
Light CNN-4 [37]	1.5G	4.1	26
Light CNN-9 [37]	1.0G	5.6	32
Light CNN-29 [37]	3.9G	12.6	125
MobileFaceNet [2]	439.8M	1.0	4.0
MobileFaceNet [15]	933.3M	2.0	8.2
ShuffleFaceNet 1.5×	577.5M	2.6	10.5

Table 10. Storage space and complexity comparison of ShuffleFaceNet 1.5× with some state-of-the-art face recognition models.

5. Conclusion

In this paper, we have developed a lightweight convolution neural architecture named ShuffleFaceNet, to learn robust features for face recognition. Among the four complexity levels that were considered, the ShuffleFaceNet 1.5× model exhibits the best speed-accuracy trade-off. The extensive experiments conducted on different face recognition benchmarks show that the proposal achieves state-of-the-art results, maintaining extreme efficiency. Particularly, regarding the complexity of the related architecture, the size and the speed on common CPUs are consistently better. The ShuffleFaceNet 1.5× has a model size of about 10MB and an inference CPU time of 29 ms, which support its practical value for real-time and mobile face recognition applications.

References

[1] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 14

[2] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In J. Zhou, Y. Wang, Z. Sun, Z. Jia, J. Feng, S. Shan, K. Ubul, and Z. Guo, editors, *Biometric Recognition*, pages 428–438, 2018. 12, 13, 14, 15, 18

[3] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *CoRR*, abs/1512.01274, 2015. 14

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 12, 13, 14, 15

[5] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 14

[6] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 12, 16

[7] C. N. Duong, K. G. Quach, N. Le, N. Nguyen, and K. Luu. Mobiface: A lightweight deep learning face recognition on mobile devices. *arXiv preprint arXiv:1811.11080*, 2018. 12, 13, 14

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *CoRR*, abs/1607.08221, 2016. 14

[9] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 12

[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015. 14

[11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 12

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 14

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments, 2007. 14

[14] 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. 12, 13

[15] IBUG. Lightweight face recognition challenge. In *IEEE International Conference on Computer Vision*, 2019. 17, 18

[16] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training

of neural networks for efficient integer-arithmetic-only inference. *arXiv preprint arXiv:1712.05877*, 2017. 12

[17] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 16

[18] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 16

[19] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. *arXiv preprint arXiv:1807.11164*, 2018. 12, 13, 15

[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. 14

[21] Y. Martínez-Díaz, N. Hernandez, R. J. Biscay, L. Chang, H. Mendez-Vazquez, and L. E. Sucar. On fisher vector encoding of binary features for video face recognition. *Journal of Visual Communication and Image Representation*, 51:155–161, 2018. 16

[22] Y. Martínez-Díaz, H. Méndez-Vázquez, L. López-Avila, L. Chang, L. Enrique Sucar, and M. Tistarelli. Toward more realistic face recognition evaluation protocols for the youtube faces database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 413–421, 2018. 16

[23] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. pages 1997–2005, 07 2017. 14

[24] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress. 14

[25] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014. 16

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, pages 1–12, 2015. 12, 16, 18

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 12, 13

[28] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 13, 15, 16, 18

[29] S. Sengupta, J. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016. 14

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Closing the gap to human-level performance in face verification. deepface. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 5, page 6, 2014. 16

[31] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 16

[32] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 14, 15, 16

[33] M. Wang and W. Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018. 12

[34] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 15, 16

[35] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011. 16

[36] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. *arXiv preprint arXiv:1711.08141*, 2017. 12, 13, 16

[37] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 12, 13, 14, 15, 16, 18

[38] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015. 14

[39] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), In press*, 2017. 12, 16

[40] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *arXiv preprint arXiv:1707.01083*, 2017. 12, 13

[41] X. Zhang, J. Zou, K. He, and J. Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):1943–1955, 2016. 12