Exploring the Potential for Real-Time Vision Transformer-Level Precision on Face Recognition Scenarios through Binarization on Embedded Systems

Luis S. Luevano
Inria
University of Rennes, Rennes, 35042, France
luis-santiago.luevano-garcia@inria.fr

Miguel González-Mendoza
Tecnologico de Monterrey
Av. Eugenio Garza Sada 2501, N.L., 64700, Mexico
mgonza@tec.mx

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

Abstract

Efficient face recognition is a key challenge in Computer Vision, prompting the exploration of strategies that balance accuracy and computational resources. In this work, we conduct a thorough comparative analysis of three modern approaches for face recognition tasks: Lightweight Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Binarized Neural Networks (BNNs). We examine novel architectures for their design principles and efficiency tactics. We also explore ViT-based approaches with the potential for global context comprehension using self-attention mechanisms. To enhance efficiency, we investigate BNNs, utilizing binary weights and quantization techniques to reduce memory and computation requirements. Through a systematic evaluation and comparison, we identify the strengths and limitations of these approaches. Building over these insights, we suggest concrete refinements that could be present in BNN designs for harnessing the recent progress of these three modern avenues. We specifically focus on efficient and accurate face recognition scenarios being available on extremely hardware-constrained platforms, such as the Nvidia Jetson Nano.

1. Introduction

Face Recognition (FR) technology has gained widespread importance in various applications [16], from enhancing security systems to improving user experience in personal devices. In the realm of deep learning, recent advancements have opened up intriguing possibilities for combining different techniques to enhance face recognition systems. Specifically, the integration of Efficient Convolutional Neural Networks (CNNs) [8] and Vision Transformers (ViTs) [14], complemented by attention mechanisms, binarization [5], and quantization [19], offers a pragmatic approach to achieving both efficiency and accuracy in real-time applications.

Quantization techniques improve efficiency with 16 and 8-bit length computations, now being a standard for achieving high efficiency without seriously compromising accuracy in various Computer Vision applications. These types of approaches can also strike a balance between effectiveness and privacy, mitigating gradient leakage in machine learning computations, paramount in today's AI-powered landscape for sensitive applications [9]. Extreme quantization, also known as Binarization [5, 6] for 1-bit neural network computations, heavily reduces computing requirements by using mostly binary weights and intermediate map representations. This aggressive quantization offers a promising avenue for implementing extremely efficient face recognition solutions on the edge for real-world applications. Binary Neural Networks (BNNs) [6] have shown progress on quantization schemes [30] and architecture designs [2, 3, 17] to alleviate the penalties present in binarization scenarios, particularly information loss, computation stability, and the struggle to achieve real-time performance on affordable embedded hardware. This leaves Binarized FR scenarios remaining to be explored.

Further studies are needed to advance the state of the art with the potential of harnessing the efficiency benefits of Binarization with the higher accuracy of CNNs [8], and ViTs on FR scenarios [15], with current analysis limiting to 6 and 8-bit representations [13]. Studying the practical aspects of binarization could bring real-time efficient models to heavily constrained ARM-based platforms, such as

the Nvidia Jetson Nano. As such, in this work, we aim to contribute by exploring the limits of practical extreme quantization and architecture designs for FR, within the scope of bringing this application to a hybrid architecture accuracy in real-time for an embedded platform.

2. Lightweight CNNs for Face Recognition

In this section, we review of state-of-the-art Lightweight CNNs tailored for FR. The results are shown in Table 1.

MobileFaceNet In [4] the authors incorporate a Squeeze-And-Excitation component in its initial configuration, along with inverted bottleneck arrangements. It substitutes the conventional Global Average Pooling (GAP) layer with a Global DepthWise Convolution (GDC) in the embedding block, resulting in a 128-D embedding output, while employing PReLU [28] activations. The main resource for reducing computations is the employment of 1×1 convolutions for reducing the computational costs, while using strided DepthWise convolutions to reduce the feature map size while maintaining more of the input signal.

ShuffleFaceNet The ShuffleFaceNet [20] architecture introduces a rapid 2-stride downsampling technique at the initial convolutional layer. It substitutes the conventional GAP layer with a GDC layer and adopts the PReLU activation as well as MobileFaceNet. The feature representation is condensed to a concise 128 embedding after the GDC layer. The approach gradually diminishes intermediate representations while progressively expanding the channel size, reaching a 7×7×1024 configuration before the GDC layer. This method predominantly utilizes DepthWise operators in its block design. Here, the channel dimension is split, followed by convolutions applied only to a subset of input channels using pointwise and separable convolutions. Then, the feature map is concatenated and channel shuffling is employed, facilitating later-stage convolutions for additional channels. The downsampling block uses 2-strided DepthWise convolutions, 1×1 pointwise convolutions, and channel shuffling.

VarGFaceNet In this method [29], a consistent number of channels is utilized, while adjusting the quantity of groups in its variable group convolutional layer. This approach also features the PReLU activation, introduces a Squeeze-And-Excitation block in their modular design, with a pointwise convolutional layer before Fully Connected layers, outputting a 512-D embedding output. Achieving top performance involves Knowledge Distillation with a ResNet100 teacher network with 24GFLOPs of computational complexity. Although VarGFaceNet outperforms MobileFaceNet in accuracy for the LFR challenge [8] verification dataset, its standalone student network demands higher computational resources at 1.02GFLOPS, winning the LFR Challenge at ICCV2019.

Method	FLOPs	Params	LFW	CFP-FP	AGEDB30
VarGFaceNet	1.02G	4.9M	99.73%	97.67%	97.5%
MobileFaceNet-1.5	933.3M	2.0M	97.33%	**99.71%**	97.56%
ShuffleFaceNet-1.5	577.5M	2.6M	97.38%	97.25%	97.31%
ShuffleFaceNet-2.0	1.05G	4.5M	99.62%	97.56%	97.28%
GhostFaceNetV2-1	272.25M	6.88M	**99.86%**	99.33%	**98.62%**
EdgeFace[1]	153.99M	1.77M	99.68%	94.46%	95.72%
MobileFaceFormer[2]	**130.08M**	**1.38M**	99.60%	96.79%	97.69%

Table 1: Lightweight CNN and efficient Hybrid method performance on popular face verification benchmarks, trained on the MS1M-V3 [11] dataset. Compiled from [17, 1, 15]. [1] EdgeFace was trained on WebFace-260M [31] subsets instead of MS1M-V3 [11] under 2 Million parameters. [2] MobileFaceFormer was trained on CASIA-WebFace, FLOPs estimated as 2×MAdds reported on the original paper.

GhostFaceNet In this proposal [1], the authors leveraged GhostNetV2 [27], which incorporates long-range DFC attention branches, as their foundation. They introduced modifications to the GDC layer to generate a distinctive embedding vector, with PReLU as activation. Within the GDC layer, a 7×7 convolutional layer is followed by batch normalization, with a pointwise layer to reduce dimensionality. Further layers include flattening, batch normalization, and a linear activation function, with 512-dimensional embedding vector. In the Squeeze-And-Excitation (SE) module, the FC layers were replaced with DepthWise convolution layers, for enhancing discriminability in the SE modules.

3. Vision Transformers (Hybrid) methods

In this section, we discuss ViT-based architectures. Results for these methods for FR are also shown in Table 1. Figure 1 shows efficient self-attention designs.

Vision Transformer (ViT) In the original ViT paper [14], the authors propose to mimic attention mechanisms present in NLP models such as self-attention with image patches, imitating tokens present in text representation. This is processed with a transformer encoder with a Multilayer Perceptron (MLP) head, similar to other sequence models. They reported performance gains on image recognition and object detection tasks with interpretable results with the attention maps. However, the most glaring limitation is the self-attention mapping calculation with a $O(n^2)$ complexity, as it is defined for all the correspondences with the patch tokens in the Key matrix.

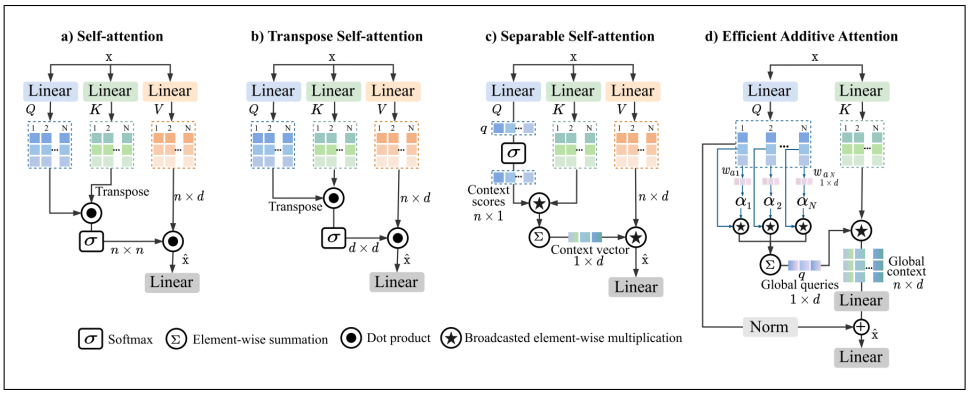

Figure 1: Efficient self-attention designs, taken from [25]. **(a)** ViT [14] Self-Attention, **(b)** EdgeFace Transposed Self-Attention [10], **(c)** MobileViT V2 Separable Self-attention [22], and **(d)** SwiftFormer's [25] Efficient Additive Attention.

MobileViTs In MobileViT(V1) [21], the authors employ strided 3×3 standard convolutions, akin to MobileNetV2 [24] with the Swish [23] activation function. The MobileViT block is designed with three MobileNetV2 (MV2) blocks, used for downsampling. The feature map undergoes an unfolding process before engaging with the transformer layer, followed by a re-folding step, subsequently merged with long-range attention branches. Remarkably, spatial properties akin to regular convolutional layers are preserved within the transformer encoding layer, as highlighted by the authors.

In MobileViT V2 [22], the authors also proposed a Separable Self-attention module, which alleviates the taxing self-attention process with computing single-vector softmax attention layers for the Query and Key matrices and broadcasted element-wise multiplications with the Value matrix. This reduces the process to a linear $O(n)$ complexity.

EdgeFace This method [10] is tailored for FR, based on the EdgeNeXT [18] hybrid architecture, and submitted for the EFaR 2023 Efficient Face Recognition Challenge [15] at IJCB 2023. This solution was evaluated as the best overall result across diverse benchmarks, under 2 Million parameters. The authors propose to use a standard embedding encoding block (head) comprised of a GAP layer, Normalization, Dropout, and a novel Low Rank Linear (LoRaLin) to compute a 512-Dimensional embedding. The LoRaLin layer helps to reduce computational costs while maintaining high accuracy performance and is used to replace regular Linear layers. LoRaLin factorizes a feature map into two separate layers depending on a rank-ratio hyperparameter for balancing computational performance. The Gaussian Error Linear Unit (GELU) [12] activation function is employed, as well as the transposed self-attention Query matrix to the channel dimension for further reducing operations to a linear order with respect to the number of tokens.

SwiftFormer In this approach [25], the authors propose to use an Efficient Additive Addition in the self-attention module. This makes the query matrix comprised of learnable vectors which are then pooled and added to produce the global context. This helps to use the Key matrix only for learning the global context and leave out the Value matrix entirely from the calculations. The rest of the architecture is comprised of convolutional encoders using 3×3 DepthWise convolutions, normalization, a pointwise convolution, and the GELU [12] activation, with another pointwise convolution and a residual connection at the end. The encoder uses the local representation encoding with a DepthWise convolution and a pointwise convolution before the Efficient Additive Attention module with spatial downsizing.

4. Binary Neural Networks tested on FR scenarios

Here, we explore BNN architectures and their evaluation in the FR context. We report efficiency on the Nvidia Jetson Nano platform and accuracy results on Table 2.

BinaryDenseNet In this approach [3], the authors avoid complex quantization schemes. Instead, they propose specialized design principles for BNNs, emphasizing a rich information flow, eliminating bottleneck structures, and proposing new block designs different from full-precision CNNs. Changes to the DenseNet bottleneck block involve replacing a convolutional layer with two layers with a reduced spatial size, while the DenseNet transition block sees an increased reduction rate for a lighter depth-channel representation. This design enhances efficiency by removing pointwise convolutions, reducing channels in residual connections, and introducing extra shortcuts not present in the original DenseNet architecture. Representation downsampling is streamlined through a block with max pooling, the

Method	1-bit MACs	32-bit MACs	Mem. (MB)	Params	Time (img/sec)	FPS	LFW	CFP-FP	AGEDB-30
BinaryDenseNet45 [3, 17]	1.59G	59.7M	4.2	13.1M	6.2s	0.16	**99.28%**	**92.88%**	**91.03%**
BinaryDenseNet37 [3, 17]	1.12G	48.9M	2.6	8.0M	4.4s	0.22	99.17%	92.59%	90.72%
BinaryDenseNet28 [3, 17]	893.2M	49.99M	1.8	4.5M	3.7s	0.27	99.17%	92.11%	90.72%
QuickNet [2, 17]	431.7M	6.8M	2.21	12.7M	1.8s	0.55	98.97%	92.00%	89.00%
QuicNet-small [2]	258.3M	**3.4M**	2.0	12.1M	1.1s	0.86	98.55%	90.65%	88.00%
BinaryFaceNet [17]	**184.9M**	3.8M	**1.3**	**506K**	**0.16s**	**6.25**	95.07%	77.93%	75.12%

Table 2: Efficiency and face verification accuracy results tested on popular face recognition benchmarks, expanded from [17]. The inference runtime efficiency was tested on a single ARM core of an Nvidia Jetson Nano.

QuickNet This method [2] also leverages Quantized DepthWise Convolutions, ReLU activations, Global Average Polling, and a 512-D face descriptor. The representation is spatially downsampled and with a low-dimensity channel expansion, ultimately expanding to 512 channels. Merging max pooling and Quantized DepthWise Convolutions provides an efficient reduction. Unlike Lightweight CNNs, QuickNet avoids depth channel fluctuations seen in MobileFaceNet and VarGFaceNet. The transition block utilizes a DepthWise convolution for downsampling, with ReLU activation at a floating-point level.

BinaryFaceNet In this method [17], the authors heavily reduce computation requirements by reducing the information degradation of heavy quantization by adding most of the computations using Grouped Convolutions in the Squeeze-And-Excitation block present in the first layers, while streamlining the rest of the feature extraction blocks with Quantized Separable Convolutions and aggregating separate skip connections at different points of the network. The authors also adopt the PReLU activations present in Lightweight CNN approaches for FR with the DoReFa [30] quantizer. This proposal achieves a remarkable efficiency, with 506K parameters only in the architecture design.

5. Methodology and results

In this work, we show results using state-of-the-art training methodologies, such as training from scratch using the SGD optimizer using the ArcFace loss function [7]. In all the reported results, training was performed using the MS1M-V3 dataset [11], except for EdgeFace in Table 1. For including QuickNet-small, we used the same training settings as [17]. We note that GhostFaceNet performs the best in the AGEDB-30 benchmark, usually the harder scenario between the three although with 3× more parameters than EdgeFace. For the efficiency part, BinaryFaceNet shows a remarkable efficiency performance in comparison with the rest of the selected BNNs, with less of 13% of verification accuracy penalty on the AGEDB-30 benchmark [8] against QuickNet-small. Furthermore, this design is the only one enabling face verification in real-time (6.25 FPS) in this extremely constrained scenario using the Larq Compute Engine [2] for Binarization.

6. Discussion and Conclusion

As shown in this research work, the potential avenues for studying and improving BNN methods for FR scenarios in this context include:

- Self-attention mechanisms on BNNs: the efficiency improvements present in separate [22] and additive [25] attention present an opportunity for implementing such mechanisms in BNN blocks and potentially coupling them with quantization capabilities.
- Knowledge Distillation: this training methodology [26, 29] could drastically improve FR performance across different scenarios by using a Transformer architecture as teacher and a design such as BinaryFaceNet as student, for highly-efficient and accurate FR.
- Privacy-preservation analysis on Quantized approaches for FR: with more focus on privacy-preserving computations for mitigating gradient leakage and protecting user privacy, binarization has proved to be a useful tool in general-purpose scenarios [9] but still further privacy analyses can be performed and are paramount for the adoption of mainstream AI technology for FR scenarios.

In this work, we explored the avenues of extremely efficient and highly-accurate Face Recognition, delving deeply into the architecture designs and evaluating their performance in FR and efficient hardware scenarios. Furthermore, we provided a discussion on leveraging the strengths of these approaches together and proposed research directions for real-world implementations of BNN technology in heavily constrained embedded scenarios.

Acknowledgement

This work was partially funded by the SOTERIA H2020 project. SOTERIA received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No101018342. This content reflects only the author's view. The European Agency is not responsible for any use that may be made of the information it contains.

The authors would like to thank the financial support from Tecnologico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # E120 - EIC-GI06 - B-T3 - D

References

[1] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446, 2023. 2

[2] Tom Bannink, Arash Bakhtiari, Adam Hillier, Lukas Geiger, Tim de Bruin, Leon Overweel, Jelmer Neeven, and Koen Helwegen. Larq compute engine: Design, benchmark, and deploy state-of-the-art binarized neural networks, 2020. 1, 4

[3] Joseph Bethge, Haojin Yang, Marvin Bornstein, and Christoph Meinel. Binarydensenet: Developing an architecture for binary neural networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1951–1960, 2019. 1, 3, 4

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. *Lecture Notes in Computer Science*, pages 428–438, 2018. 2

[5] Matthieu Courbariaux and Yoshua Bengio. Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1. *CoRR*, abs/1602.02830, 2016. 1

[6] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. *CoRR*, abs/1511.00363, 2015. 1

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 4

[8] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 1, 2, 4

[9] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Towards privacy aware deep learning for embedded systems. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 520–529, New York, NY, USA, 2022. Association for Computing Machinery. 1, 4

[10] Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Ketan Kotwal, and Sebastien Marcel. Edgeface: Efficient face recognition model for edge devices, 2023. 3

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, page 87–102, Cham, 2016. Springer International Publishing. 2, 4

[12] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. 3

[13] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[14] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 1, 2, 3

[15] Jan Niklas Kolf, Fadi Boutros, Jurek Elliesen, Markus Theuerkauf, Naser Damer, Mohamad Alansari, Oussama Abdul Hay, Sara Alansari, Sajid Javed, Naoufel Werghi, Klemen Grm, Vitomir Štruc, Fernando Alonso-Fernandez, Kevin Hernandez Diaz, Josef Bigun, Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Ketan Kotwal, Sébastien Marcel, Iurii Medvedev, Bo Jin, Diogo Nunes, Ahmad Hassanpour, Pankaj Khatiwada, Aafan Ahmad Toor, and Bian Yang. Efar 2023: Efficient face recognition competition, 2023. 1, 2, 3

[16] Stan Z. Li and Anil K. Jain. *Handbook of Face Recognition*. Springer Publishing Company, Incorporated, 2nd edition, 2011. 1

[17] Luis S. Luevano. *Closing the gap on affordable real-time very low resolution face recognition for automated video surveillance*. PhD thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey, 2022. https://hdl.handle.net/11285/650411. 1, 2, 4

[18] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *International Workshop on Computational Aspects of Deep Learning at 17th European Conference on Computer Vision (CADL2022)*. Springer, 2022. 3

[19] Brais Martinez, Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Training binary neural networks with real-to-binary convolutions, 2020. 1

[20] Yoanna Martinez-Diaz, Luis S. Luevano, Heydi Mendez-Vazquez, Miguel Nicolas-Diaz, Leonardo Chang, and Miguel Gonzalez-Mendoza. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[21] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2021. 3

[22] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *Transactions on Machine Learning Research*, 2022. 3, 4

[23] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017. 3

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[25] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. 2023. 3, 4

[26] David Svitov and Sergey Alyamkin. Margindistillation: distillation for margin-based softmax, 2020. 4

[27] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Chao Xu, and Yunhe Wang. Ghostnetv2: Enhance cheap operation with long-range attention. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[28] Ludovic Trottier, Philippe Giguere, and Brahim Chaib-draa. Parametric exponential linear unit for deep convolutional neural networks. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017. 2

[29] Mengjia Yan, Mengao Zhao, Zining Xu, Qian Zhang, Guoli Wang, and Zhizhong Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2, 4

[30] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients, 2018. 1, 4

[31] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2