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

Luis S. Luevano, Miguel González-Mendoza, Yoanna Martínez-Díaz, Heydi Méndez-Vázquez

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

Luis S. Luevano[1], Miguel González-Mendoza[2], Yoanna Martínez-Díaz[3], Heydi Méndez-Vázquez[3]

[1] Inria, Rennes, France
[2]Tecnologico de Monterrey, Nuevo León, Mexico
[3]Advanced Technologies Application Center (CENATAV), Havana, Cuba

ICCV23 PARIS

Introduction

- **Context:** Efficient Face Recognition (FR) 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 FR: Lightweight Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Binary Neural Networks (BNNs).

Motivation: To **enhance efficiency** for real-time scenarios on embedded devices, we investigate BNNs, utilizing binary weights and quantization techniques to reduce memory and computation requirements. We also explore Lightweight CNN and ViT-based approaches on FR scenarios, with the potential for global context comprehension using self-attention mechanisms for **improving accuracy**.

Goal: To 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 FR scenarios being available on **extremely hardware-constrained platforms**, such as the **Nvidia Jetson Nano**.

On Quantization and Binarization

- **State of the art Quantization**: Quantization analyses in the state of the art are limited to 16 and 8-bit computations for FR applications.
- **Challenges on Binarization**: Extreme quantization, also known as Binarization, using 1-bit neural networks, offers potential efficiency benefits but faces challenges including information loss, computation stability, and achieving real-time performance on affordable hardware.
- **Recent efforts using Binarization on FR**: Newer methods, such as BinaryFaceNet, show a remarkable efficiency performance with less than 13% of verification accuracy penalty on the AGEDB-30 benchmark against the state of the art. This design is the only one enabling face verification in real-time (6.25 FPS) on a single ARM core of an Nvidia Jetson Nano using the Larq Compute Engine framework.

Vision Transformer Efficiency vs Accuracy on ImageNet

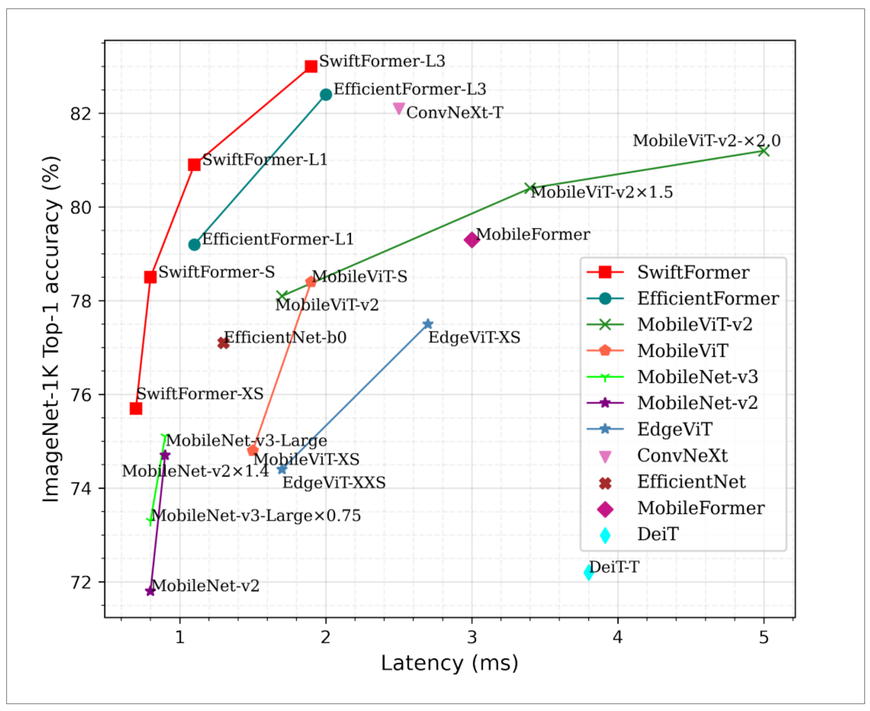

Figure: Modern ViT and Lightweight CNN designs benchmarked on ImageNet against inference time latency, taken from the SwiftFormer paper. We note that SwiftFormer reports the best accuracy-to-efficiency trade-off, showing potential for FR applications.

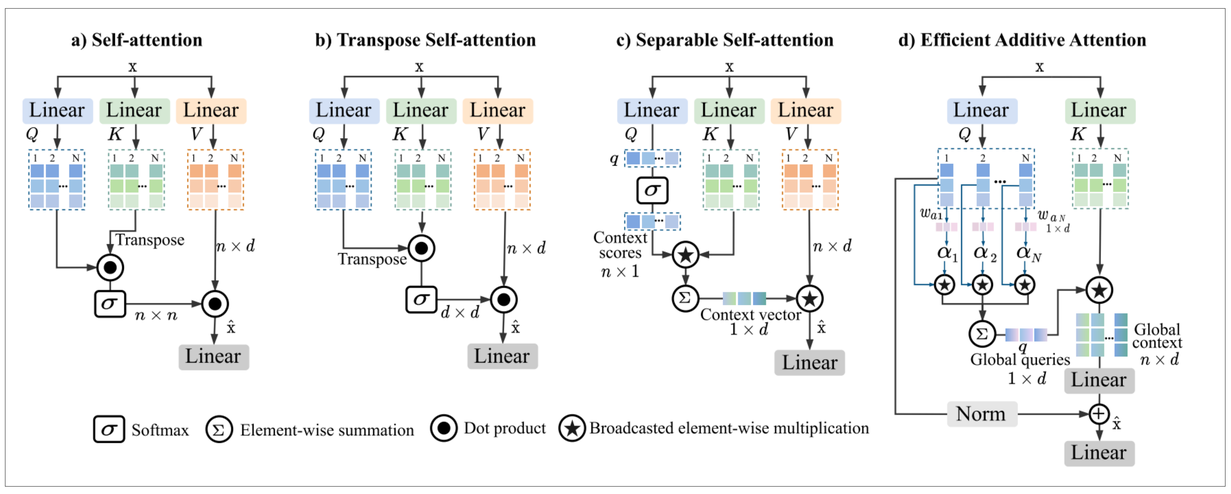

Figure: Efficient self-attention designs, taken from the SwiftFormer paper. **(a)** ViT Self-Attention, **(b)** EdgeFace Transposed Self-Attention, **(c)** MobileViT V2 Separable Self-attention, and **(d)** SwiftFormer's Efficient Additive Attention.

FR-specific methods with Lightweight CNNs & ViTs

Method	Type	FLOPs	Params	LFW	CFP-FP	AGEDB-30
VarGFaceNet	L.CNN	1.02G	4.9M	99.73%	97.67%	97.5%
MobileFaceNet-1.5	L.CNN	933.3M	2.0M	97.33%	**99.71%**	97.56%
ShuffleFaceNet-1.5	L.CNN	577.5M	2.6M	97.38%	97.25%	97.31%
ShuffleFaceNet-2.0	L.CNN	1.05G	4.5M	99.62%	97.56%	97.28%
GhostFaceNetV2-1	L.CNN	272.25M	6.88M	**99.86%**	99.33%	**98.62%**
EdgeFace[1]	Hybrid	153.99M	1.77M	99.68%	94.46%	95.72%
MobileFaceFormer[2]	Hybrid	**130.08M**	**1.38M**	99.60%	96.79%	97.69%

Table: Lightweight CNN and efficient Hybrid method performance on popular face verification benchmarks, trained on the MS1M-V3 dataset.
[1]EdgeFace was trained on WebFace-260M.
[2] MobileFaceFormer was trained on CASIA-WebFace, FLOPs estimated as 2×MAdds reported on the original paper.

BNN Accuracy on FR and Efficiency Evaluations

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

Table: Efficiency and face verification accuracy results tested on popular face recognition benchmarks. The inference runtime efficiency was tested on a single ARM core of an Nvidia Jetson Nano. All BNNs were trained on the MS1M-V3 dataset.

Research directions

- **Further ViT-based architecture analysis on FR**: Newer ViT-based methods, such as SwiftFormer, have the potential for improving results on FR benchmarks, based on their reported results on ImageNet performance, while also enhancing efficiency.
- **Self-attention mechanisms on BNNs**: the efficiency improvements present in separate and additive attention present an opportunity for implementing such mechanisms in BNN blocks and potentially coupling them with quantization capabilities.
- **Knowledge Distillation**: this training methodology 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 but further privacy analyses can be performed and are paramount for the adoption of mainstream AI technology for FR scenarios.

More on Face Recognition through Binarization

Acknowledgements

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