Energy-aware Retinaface: A Power Efficient Edge-Computing SOC for Face Detector in 40nm

1st Miao Sun
State Key Laboratory of ASIC and System
School of Microelectronics, Fudan University
Shanghai, China
18112020006@fudan.edu.cn

2nd Yingjie Cao
TiMESiNTELLi Inc.
Shanghai, China
stan.cao@timesintelli.com

3rd Patrick Yin Chiang
State Key Laboratory of ASIC and System
School of Microelectronics, Fudan University
Shanghai, China
pchiang@fudan.edu.cn

2021 IEEE 14th International Conference on ASIC (ASICON) | 978-1-6654-3867-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASICON52560.2021.9620286

Abstract—In this work, an energy-awaring face detector is implemented in 40nm technology SoC. Based on the art-of- state face detector, a highest accuracy retinaface detector (91.4% average precision) on the WIDER FACE dataset is quantized in the int8 domain. For this neural network, an 8-bit CNN accelerator in a hybrid SOC architecture is designed to achieve an end-to-end face detector. The entire detector runs at 15fps with 66.67mw power per frame. Furthermore, redundant layers in this CNN are analyzed based on this performance. For different sizes of face, some calculations can be reduced with no loss brought to results. To address this improvement, this network is divided into three branches according to different sizes of faces in a single input image. Besides, a simple two-layer classifier is trained to determine the calculation graph in the current run and implemented on SOC. Finally, the face detector increases to 36fps, and energy power decreases to 27.78mw power per frame. This is the highest accuracy(85.8%) face detector hardware implementation on the WILDER FACE dataset.

Index Terms—edge-computing, SOC, energy efficient, face detection

I. Introduction

Face detector CNN has been widely used due to its necessity and security. In the computer vision field, its accuracy can reach nearly 100% in some larger scale face datasets with delicated structure and training tricks. However, most of CNNs cannot be implemented on a small scale platform for their computation overhead and power consumption. For example, a BNN for face detector is proposed and implemented on a mobile device [1]. The current high accuracy CNN is difficult to quantize into a binary domain without a complex flow (pruning, quantization, and retrain). In this paper, the quantization scheme from float32 to int8 with tiny loss is realized based on Pytorch [2]. It can support all CNN structures. In this work, Retinaface [3], which is the lightest CNN until now with 1M parameters and fewer layers is selected.

For the energy consumption in CNN, some improvements have been proposed. For example, similar features are clustered as the same value to reduce the computation overhead, bringing considerable computation to cluster algorithms [4,5]. These methods will bring an extra circuit design work burden in the calculation of each layer. Thus, a new method for energy efficiency is designed in this paper by changing the computation graph dynamically according to the input images.

In this paper, an edge-computing SOC for face detection is designed and reaches the state-of-the-art face detector hardware system. It has mainly 4 advantages as follows. 1) Its CNN accelerator can support the Pytorch quantization scheme directly, containing all complex structures in all common CNN. This shortens the gap between the CV field and the hardware field. In this work, the int8 quantization scheme on the WIDER FACE TEST dataset of Retinaface achieves 85.8% accuracy. 2) For energy efficiency, an energy-awaring retinaface is proposed. Specifically, an energy-warning structure is added to dynamically regulate the energy consumption according to the input image. The retinaface is decoupled according to different sizes of faces (16x32, 64x128, and 256x512), and an extra classifier is added to indicate the branch to be executed. By skipping some redundant layers, its energy consumption can be reduced to 27.78mw for CNN inference on the Wider Face Dataset [5] on average. 3) For this quantization scheme, a hybrid architecture SOC is proposed to realize the end-to-end solution. A CNN accelerator is designed to perform normal convolution operations, depthwise convolution and branch determination is completed in RISCV core. A ping-pong architecture is used to improve the MAC ARRAY UNIT pipeline utilization. Otherwise, the video flow process is embedded in ISP. 4) In our platform, the energy-awaring retinaface can run as fast as 36fps in QVGA resolution with 27.78mw power per frame.

The rest of this paper is arranged as follows. In section II, the quantization scheme including fusion layers and weights statistics method is demonstrated. In section III, the energy-awaring Retinaface and its performance are explained. Then, the hybrid-structure SOC and some hardware-software collaborations are introduced in section IV. Next, the performance and Comparison of the proposed SOC are presented in section V. Finally, a conclusion is drawn in section VI.

978-1-6654-3867-4/21/$31.00 2021 IEEE

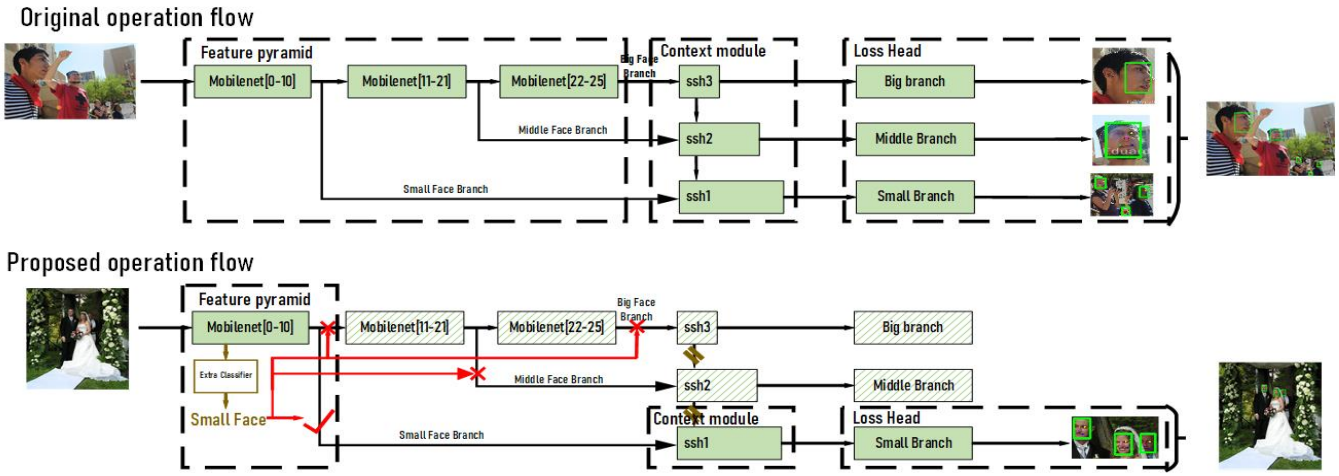

Fig. 1. Proposed energy-awaring retinaface flow.

II. Quantization Scheme and Operator Execution on SOC

As the first step to making the edge-computing to be possible, quantization scheme is crucial to hardware platforms.

A. Post Training Quantization Scheme in Retinaface

In Retinaface, all convolution layers are followed by a batchnorm layer, and some layers have activation layers. In this study, fusion combination includes convolution layer, batchnorm layer, convolution layer, batchnorm layer, and activation layer. For example, in Fig.3, $act()$ denotes the activation function in this model. Then, the retinaface model transforms from 148 layers to 54 layers.

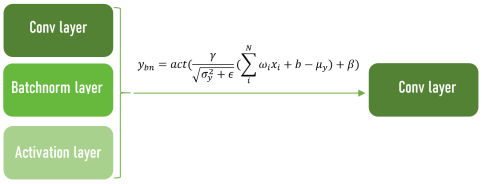

Fig. 2. Fusion scheme in efficient-retinaface network.

B. Quantization parameters determination in Retinaface

Statistics on the floating point range for every layer is conducted based on the WIDER FACE training dataset and all scale and zero point are calibrated by L2 normalization error minimization. A statistic process in conv10-3 in Retinaface and optimal parameters determination is presented in Fig.2. The calibration strategy can be summarized as follows:

•Run floating point inference on Calibration Dataset.

In each layer:

- Create the histogram of the incoming inputs.
- Compute the histogram continuously and the ranges of each bin changes with every new tensor observed.
- Search for the distribution in the histogram for optimal ranges to ensure the minimization of the quantization error related to the floating point model. An approximation for L2 error minimization for selecting min/max values is adopted.
- Calculate the scale and zero point as equation1 and equation2.

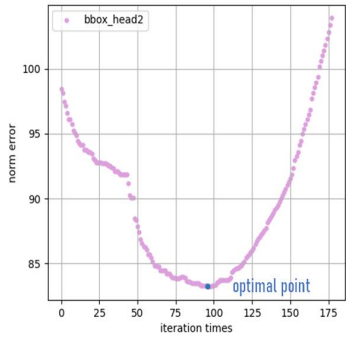

Fig. 3. Iterations versu norm error in layer of bbox_head2 energy-awaring retinaface

In this work, the retinaface of light weight is adopted, with the mobileNet [7] as its backbone. It employs the single-stage model [8] to achieve a lightweight framework. For the original Retinaface, three branches are utilized to detect all sizes of faces for all input images. Nevertheless, these three branches are clearly responsible for different sizes of faces. For example, in Fig.1, the input image includes three sizes of faces, which need three branches to cover all detected archors. Then, these final results are composed of these results from separate branches. As a result, redundant information will be introduced for some input only containing small face, middle face, or big face. Such a method will bring a computation burden for edge-computing platforms. For a limited situation, reducing

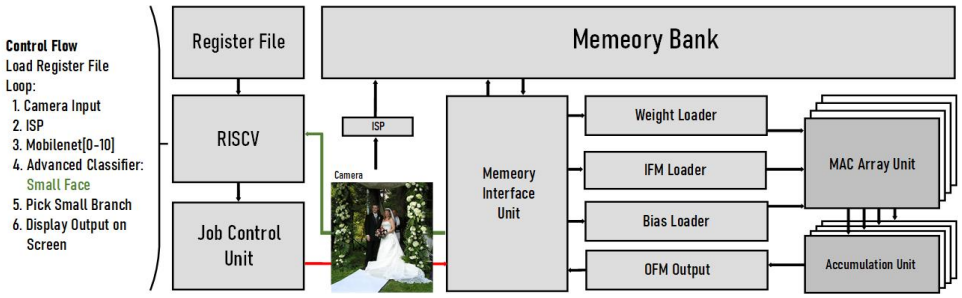

Fig. 4. Proposed SOC architecture.

unnecessary energy consumption will make this solution more appliable in reality.

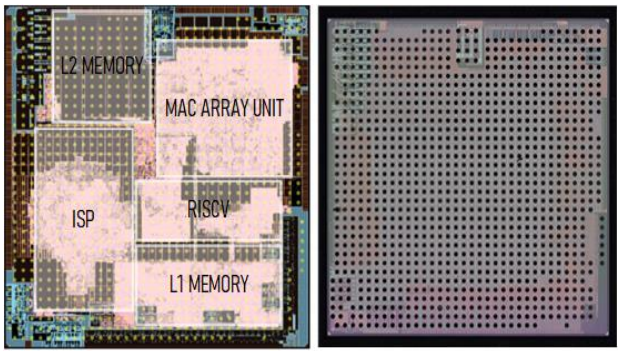

Fig. 5. Annotated layout and die graph in X-ray.

III. Proposed energy-awaring Retinaface Flow

An energy-awaring retinaface is proposed in this paper to solve these problems above. Firstly, in context module the connection between ssh1 and ssh2 and ssh2 and ssh3 is pruned. Consequently, the model is separable completely, and no accuracy reduction on the WIDER FACE test dataset is caused. Secondly, an advance classifier which is a two-layer CNN is added to indicate the size of faces in this input image. The output feature from the first part of mobile backbone is selected as input. That can judge which branch to run in the next. Next, this classifier is trained on the WIDER FACE dataset. Thirdly, according to the result from advanced classifier, these unnecessary branches are abandoned. For example, in Fig.3(b), if 'Small Face' is determined, both Big Face Branch and Middle Face Branch can be skipped, leaving no effect on the final result. By reducing these unavoidable operators, this network can save energy depending on the input image. Thus, this model can be called energy-awaring retinaface. In real situations, such as attendance check and facial identification, they usually only contain few faces, contributing to a further energy efficiency. The model complexity is analyzed, as illustrated in TABLE I. In total, RetinaFace has 0.4165MB parameters and 0.599GB Madd operations. However, these three branches cost 0.278GB, 0.369GB, and 0.225GB Madd operations, respectively. For specific scenarios, this strategy can save almost half of the sum operations.

IV. Proposed SOC structure

The whole architecture of the proposed SOC is illustrated in Fig.4, including: 1) a RISCV top control unit with a 24KB Register memory, 2) three data loader units for int8 weight, int8 IFM, and int32 bias, 3) a channel-wise accumulation core, and 4) a 1M on chip SRAM for storing all quantized parameters. To improve the efficient of feature map and weight loader, a dual-buffer ping-pong design is adopted. Shown as Fig.6, to take advantages of the MAC ARRAY Unit, weights and feature map should be loader advanced in L1 memeory. So the store memory footprint is duplicated to complish the ping-pong structure.

TABLE I
Computation load in Retinaface. The green blanks stand for common module used in all cases; The yellow, red and purple blanks stand for specific module for Small, Middle and Big faces seperately.

Retinaface	params(KB)	Madd(Mega)
mobilenet[0-10]	9.890	90.494
Advanced Classifier	9.887	4.280
mobilenet[11-21]	97.0625	95.9674
mobilenet[22-25]	101.125	26.5485
ssh1	74.1875	187.297
ssh2	78.1875	185.6314
ssh3	50.0625	13.1794
Small Headloss	0.7617	2.988
Middle Headloss	0.7617	1.494
Big Headloss	0.7617	0.1978
Small Branch	94.7266	258.0596
Middle Branch	195.7891	377.8675
Big Branch	268.789	230.6677

As illustrated in Fig.4, the software flow runs on the RISCV core retinaface. It integrates the camera management, ISP processing, the job splitting of CNN, and final NMS (non-maximum suppression) post processing. Data Loader is responsible for generating memory access re-quests and feeding read-back data into execution unit(MAC ARRAY UNIT or Accumulator). Thus,

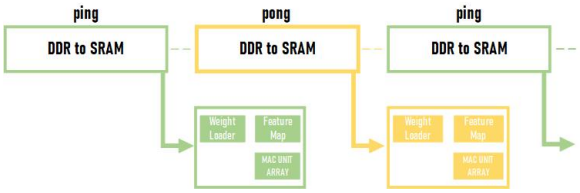

Fig. 6. Ping-pong buffer pipeline.

there are three types of data loader. Specifically, weight Loader loads weights into MAC ARRAY UNIT in the convolution mode. IFM(Input Feature Map) Loader loads IFM into MAC ARRAY UNIT in the convolution mode. Bias Loader loads partial sums into Accumulator. Data Writer(DW) is responsible for writing partial sum or convolution results into memory. The Job Control Unit can split a convolution operation into some sub-operations. The MAC ARRAY UNIT contains 4 processing engines (PE), with each PE having 4 int8 by int8 multipliers. In this SOC normal convolution and depthwise convolution are supported. The RISCV core controls the CNN flow, and the branch determination is also executed in it. For example, at the beginning, the layer 0-10 of retinaface backbone is performed. Then, the advanced classifier give its result back to RISCV to run the corresponding branch. Particularly, RISCV controller changes the pointer address using the classifier result, realizing choosing different branch. In Fig.4, the input image only contains small faces, and the green lines denote writing the 'small face' to RISCV. The following branch completes the left layers to give the final localizations of faces. RISCV core runs the post process including NMS and show results on display screen.

TABLE II
Performance Summary

Tech.	40nm
Algorithm	Retinaface
Area	5.7mm x 5.7mm
NPU Suppply Voltage	1.1V
NPU Frequency	400M
Resulotion	QVGA
Frame rate	36fps
Power	27.78mW per frame

V. SOC Implementation and performance analysis

The die micrograpth is dipicted in Fig.5, which is fabricated in 40nm CMOS technology. All configurations are included in TABLE II. After profiling the energy-awaring retinaface on SOC, this model can reach 36fps. Different from other CNN accelerators with an activation block, this module is reduced from the quantization scheme. Regarding depthwise convolution structure and normal convolution structure, variant tiling scheme is adopted to improve the utilization of the MAC ARRAY UNIT. Compared with the normal hardware implementation, for example [9,10], we keep the complex architectures in original model design like cube concatenation and upsampling. Concerning these functions, the DMA buffer is used to realize these operations, saving the time consuming.

VI. Conclusion

In this work, the retinaface network to int8 domain is quantized and decoupled to three branches. Then, an advanced classifier is trained to determine the running branch dynamically. An edge-computing SOC, which is compatible with our quantization scheme in CNN accelerator, is designed. It also has ISP, RISCV core to perform post process like NMS and control the video flow. Finally, the network can run up to 36fps on our SOC. This is the first one CNN edge-computing SOC that implements a large CNN above 50 layers and achieves a real-time performance in QVGA resolution. Meanwhile, all functions on our platforms, including picture capturing, ISP pre-processing and results displaying, are integrated. This is an end-to-end solution for face detection. This SOC consumes 27.78mw to detect multiple faces in 36fps with an average precision equal to 85.8% on the WIDER FACE dataset with 5.6% accuracy drop compared with the float32 model.

References

[1] Kang S , Join Lee, Kim C , et al. B-Face: 0.2 MW CNN-Based Face Recognition Processor with Face Alignment for Mobile User Identification[C]. 2018 IEEE Symposium on VLSI Circuits. IEEE, 2018.

[2] Paszke A , Gross S , Massa F , et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[J]. 2019.

[3] Deng J , Guo J , Ververas E , et al. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[4] Kim S , Lee J , Kang S , et al. A 15.2 TOPS/W CNN Accelerator with Similar Feature Skipping for Face Recognition in Mobile Devices[C]. 2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019.

[5] Yang S , Luo P , Loy C C , et al. WIDER FACE: A Face Detection Benchmark[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016:5525-5533.

[6] Kim S , Lee J , Kang S , et al. A Power-Efficient CNN Accelerator With Similar Feature Skipping for Face Recognition in Mobile Devices[J]. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2020, PP(99):1-13.

[7] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.zhuzhang

[8] Zhang S , Zhu X , Zhen L , et al. S^3FD: Single Shot Scale-Invariant Face Detector[C]. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[9] Bong K , Choi S , Kim C , et al. Low-Power Convolutional Neural Network Processor for a Face-Recognition System[J]. IEEE Micro, 2017, 37(6):30-38.

[10] Ding R , Su G , Bai G , et al. A FPGA-based Accelerator of Convolutional Neural Network for Face Feature Extraction[C]. 2019 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC). IEEE, 2019.