STUDYING THE MOST EFFECTIVE FPGA NEURAL NETWORKS: ECONOMIC AND MODELING ISSUES

Sergo HARUTYUNYAN, Garnik VOSKANYAN
Andranik GALSTYAN, Davit GABRIELYAN
A&MS Circuit Design Engineer, NPUA, MA student
Chair of Microelectronic Circuits and Systems Interfaculty
Armen SAFARYAN
Ph.D, Doctor of Science in Economics

Keywords: FPGA, Neural Networks, Modeling, Simulation

Introduction. Computational imaging techniques, including object recognition, photo recognition, computational methods, and object tracking, typically use convolutional neural networks (CNNs). Using an FPGA-based diagnosis microprocessor with equal processing speed and efficiency is more complicated the more complex the CNN gets. Sophisticated and low-cost CNN designs such as ShuffleNet and Mobilenet are suitable for FPGA-based systems. On the other hand, researchers are still focused on maintaining as much processing and analytical power as possible [1]. As a solution to this matter, we provide Ad-MobileNet, a system constructed on the conventional MobileNet topology that utilizes fewer computer complexity while achieving classification techniques superior to some well-known FPGA-based platforms. When compared to the standard MobileNet concept, Ad-MobileNet offers something new.

Literature review. The Ad-depth module we develop outperforms the traditional feature maps convolution in this architecture. An embedded system can benefit from this component's improved precision without much raising the underlying hardware's calculation complexity. Research studies later discovered that making hardware design easier by removing several layers with a negligible impact on the prediction performance had been removed [2]. Third, effectiveness appears to be influenced by choice of convolution operation independent from precision. Hardware design uses simple evolutionary algorithms like Tanh, Rectified Linear Unit (ReLU), and sigmoid. Alternative input nodes work better than algorithms to help MobileNet generalize. As a result, we're adding Mish, Tanh Exponential (TanhExp), and ReLU to the list of acceptable non-linear activations.

Methodology. We applied the exponential approach to approximate the nonlinear transfer components TanhExp and Mish in the Ad-MobileNet design. A network is activated when using a learning algorithm if the summation is more significant than zero or if a bias is added. It's like a gate that checks to see whether an inbound number exceeds a predetermined limit [3]. The activation function's most important job is to make a neuron's response less linear. A neural network was little more than a linear regression approach without a convolution operation, and linear regression models cannot address complicated issues like image recognition. Overall, the activation function enhances the learning algorithm, making it more capable of learning and doing more complex tasks [4]. For practical systems, nowadays, most researchers employ an essential nonlinear process that needs minimal computing power.

The figure below shows the overview of the proposed design:

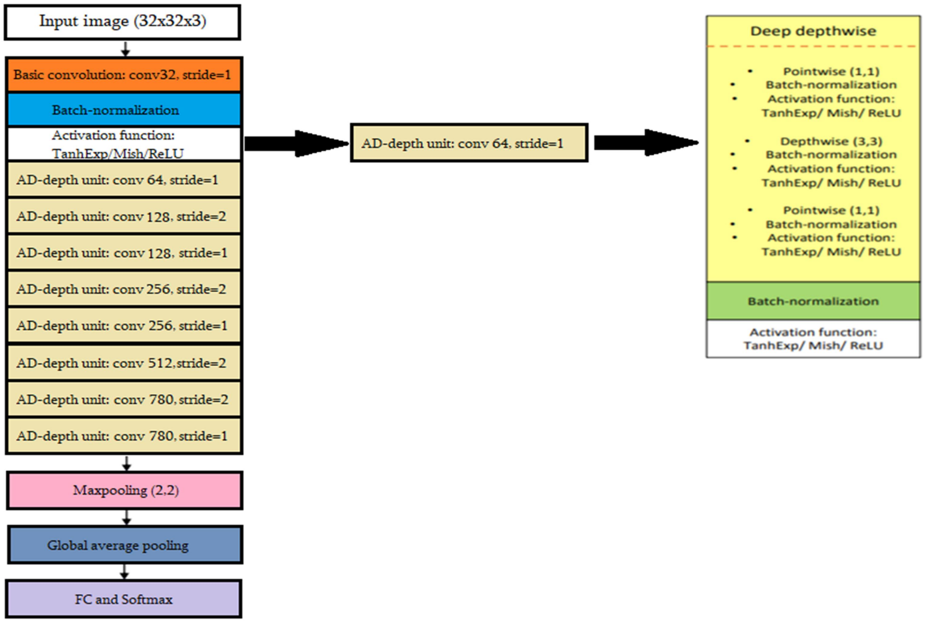

Input image (32x32x3)
Basic convolution: conv32, stride=1
Batch-normalization
Activation function: TanhExp/Mish/ReLU
AD-depth unit: conv 64, stride=1
AD-depth unit: conv 128, stride=2
AD-depth unit: conv 128, stride=1
AD-depth unit: conv 256, stride=2
AD-depth unit: conv 256, stride=1
AD-depth unit: conv 512,stride=2
AD-depth unit: conv 780, stride=2
AD-depth unit: conv 780, stride=1
Maxpooling (2,2)
Global average pooling
FC and Softmax
AD-depth unit: conv 64, stride=1
Deep depthwise
Pointwise (1,1)
Batch-normalization
Activation function: TanhExp/ Mish/ ReLU
Depthwise (3,3)
Batch-normalization
Activation function: TanhExp/ Mish/ ReLU
Pointwise (1,1)
Batch-normalization
Activation function: TanhExp/ Mish/ ReLU
Batch-normalization
Activation function: TanhExp/ Mish/ ReLU

Figure 1. *Ad-MobileNet Schematic Structure*

Advantages of FGA over GPU and CPU: FPGA has three significant benefits over GPU and CPU: adaptability, network connectivity, and energy accuracy. Because FPGA enables developers to modify the underpinning hardware design down to the pixel level, demonstrating adaptability, this ability is what sets FPGAs apart from other types of hardware technologies. Because of this, the slowness of an FPGA is measured in nanoseconds [5]. When it comes to GPU, we're talking about milliseconds here. FPGA typically operates at a slower clock frequency relative to GPU and CPU, but it has a higher or equal calculating capability. As a result, FPGA may be more energy efficient. Xilinx's Virtex Ultimate FPGA platform and Nvidia's Tesla P4 GPU, both of which were released in 2017, are two examples of embedded systems released in 2017.

Beyond the FPGA board, CNN is an efficient algorithm for speeding up computation. CNN uses a considerable number of parallelizable exponentiation processes. A convolutional element's calculations are heavily reliant on the subsequent layer's output.

However, the same CNN model algorithms are primarily autonomous, laying the foundations for massively parallel computing. CNN methodology has currently flourished with the introduction of new arithmetic designs from that other standpoint. This ability means that each person's data is routed in a certain way. FPGA is capable of meeting the designer's particular needs and of achieving a high-frequency band. Binarized data and weights are used in BinaryNet, a predictive algorithm. XOR computations replace the standard CNN's amplification algorithms. FPGA could construct the system since it can express a structure at the bit-level with precision. On the other hand, average accuracy data formats are compatible on broad sense platforms like GPUs and CPUs. The binary image neural framework may use fewer requirements than integrating the network on GPU or CPU.

Scientific novelty. Deep Neural networks (DNNs) have been widely applied to finance and economic forecasting as a powerful modeling technique. Economic funda-mentals are important in driving exchange rates, stock market index price and economic growth. Most neural network inputs for exchange rate prediction are univariate, while those for stock market index prices and economic growth predictions are multivariate in most cases. Prediction performance of neural networks can be improved by being integ-rated with other technologies. Nonlinear combining forecasting by neural networks also provides encouraging results. Finally, this method is compared and contrasted with standard (statistical) approach on real economic data to show the potential of using deep neural network in modelling economic variables. For all these reasons the low-cost neural networks important in economics.

$$Mish(X) = \left\{ \begin{array}{l} x, 2 < x \\ 1.08273 \times x - 0.22053, 1 < x \leq 2 \\ 1.02872 \times x - 0.11989, 0 < x \leq 1 \\ 0.29798 \times x - 0.04853, -1 < x \leq 0 \\ -0.05617 \times x - 0.37574, -2 < x \leq -1 \\ -0.10802 \times x - 0.46768, -3 < x \leq -2 \\ -0.07294 \times x - 0.36108, -4 < x \leq -3 \\ -0.03876 \times x - 0.22528, -5 < x \leq -4 \\ -0.01855 \times x - 0.12508, -6 < x \leq -5 \\ -0.00838 \times x - 0.06455, -7 < x \leq -6 \\ -0.00049 \times x - 0.00632, x \leq -7 \end{array} \right\}$$

$$TanhExp(X) = \left\{ \begin{array}{l} x, 1 < x \\ 1.00522 \times x - 0.01850, 0 < x \leq 1 \\ 0.34275 \times x - 0.06743, -1 < x \leq 0 \\ -0.08953 \times x - 0.45752, -2 < x \leq -1 \\ -0.12067 \times x - 0.50794, -3 < x \leq -2 \\ -0.07575 \times x - 0.37267, -4 < x \leq -3 \\ -0.03929 \times x - 0.22795, -5 < x \leq -4 \\ -0.01864 \times x - 0.12562, -6 < x \leq -5 \\ -0.00840 \times x - 0.06466, -7 < x \leq -6 \\ -0.00049 \times x - 0.00632, x \leq -7 \end{array} \right\}$$

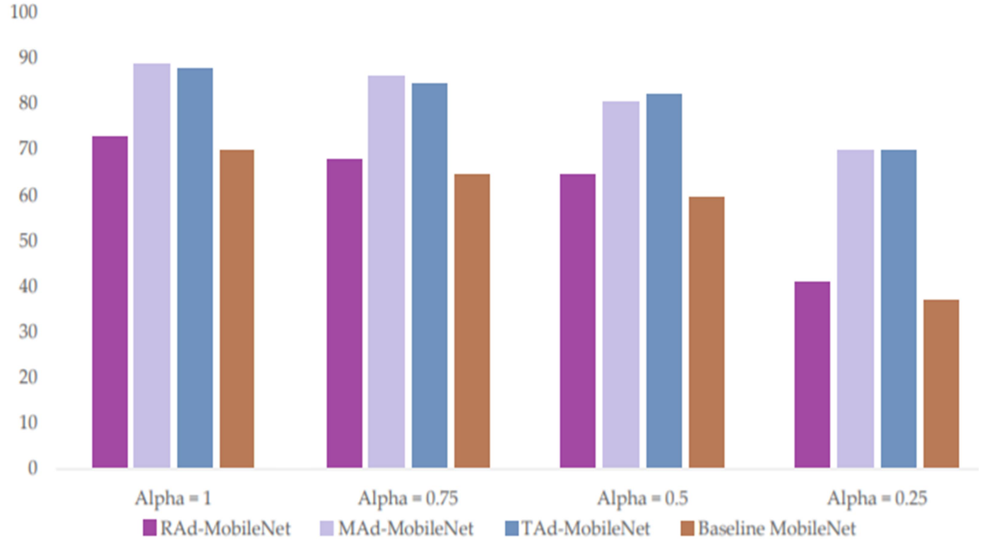


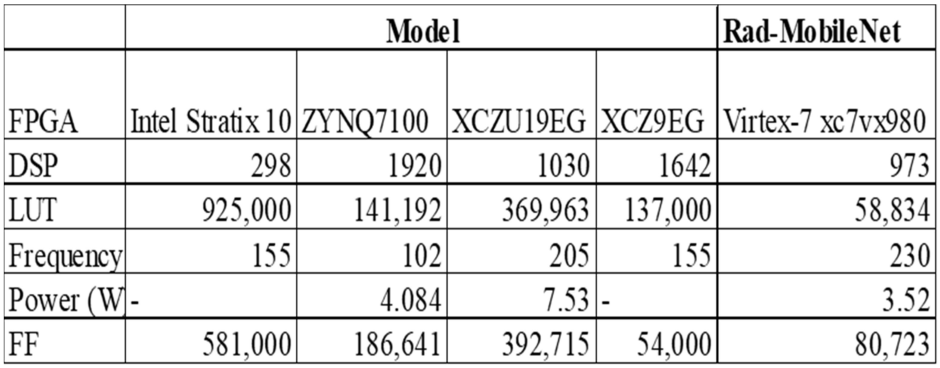
Figure 2. *Accuracy of the Models*

Research design. We use float-point encoding to computerize actual quantities like parameters, variables, and distortions to build the Ad-MobileNet model mostly on FPGA. It is possible to achieve high resolution while using a reasonable amount of physical servers with float-point notation. We Ad then set up the convolution layer in FPGA form with the execution of activation functions. Using Activation Modules ReLU, TanhExp, and Mish, we developed the RAd-MoileNet, TAd-MobileNet, and MAd-MobileNet model on Virtex-7 FPGA xc7vx980t. We employed an image recognition application running on the CIFAR-10 database to show the functionality of models. This resource is divided into ten different categories. VHDL was the hardware modeling language applied in this research. We created the Ad-MobileNet framework to put the changes put into place. We decided to use the PWL methodology to depict nonlinear systems. Thus we experimented with three different convolution layers: TanhExp, Mish, and ReLU. Over the last DSC unit, we reduced the range level from 1000 to 700 by erasing the five new layers. The Ad-depth module has taken on the role of the DSC unit due to its increased thickness. The Ad-recognition MobileNet's performance is improved due to these changes, which use fewer system resources. PWL, TanhEx Mish approximations are developed on an integrated FPGA platform utilizing registers, one multiplier, and one adder for each technique un et al., 2010).

Experimental Results. The Figure below illustrates a comparison of MAd-MobileNet, TAd-MobileNet, RAd-MobileNet, and MobileNet using CIFAR-10 database:

Results Discussion. A look at Figure 3 shows the MobileNet simulation to be the most inaccurate (lowest accuracy). As a result, TAd-MobileNet and MAd-MobileNet have high classification rates than the other two approaches (87.88 percent against 88.76 percent). In terms of hardware requirements, the RAd-MobileNet comes in last. It saves around 41% of the DSP required relative to the MobileNet solution. Additionally, the results show that employing a sophisticated non - linear transformation on an FPGA to build the system uses more physical hardware than utilizing ReLU. In this study, we compare the performance and efficiency of our suggested models' embedded system to other MobileNet-related designs that have been described in the past. They activated the device with ReLU.We attained a frequency of 100 MHz, the lowest value in Table 1 and 130 MHz less than the proposed methodology for RAd-MobileNet.

Table 1 . Comparing MobileNet Models



	Model				**Rad-MobileNet**
FPGA	Intel Stratix 10	ZYNQ7100	XCZU19EG	XCZ9EG	Virtex-7 xc7vx980
DSP	298	1920	1030	1642	973
LUT	925,000	141,192	369,963	137,000	58,834
Frequency	155	102	205	155	230
Power (W	-	4.084	7.53	-	3.52
FF	581,000	186,641	392,715	54,000	80,723

In addition, their system used the most physical hardware (1920 DSP), making it the most powerful in the comparison. They employed 989 times as much DSP than was required for RAd-implementation. MobileNet's Additionally, they consumed 0.8 W more energy than RAd-MobileNet. According to researchers, there are two degrees of heterogeneity in the RAd-obileNet concept, one at the database level and the other at the desirable rate. These ideas attained the second-lowest wavelength on the table, at 155 MHz. In particular, they utilized 1642 DSP, the second-highest analytical hardware capability in the chart. Besides that, they utilized 137 K logic elements (LUTs), which is far more than twice as much as we used in our simulations. Tomato's researchers presented a methodology for automating the creation of effective CNN multipliers called Tomato. As a result of their advice, the MobileNet-V1 architecture used the minuscule DSP system resources of any. They did, however, use the most FFs and LUTs of everyone in the study. Compared to our approach, they employed around 16 times more LUTs. Their model also ran slower, clocking in at 69 MHz slower than ours—the scientists proposed a typical synthesis as an interpretation accelerator on the MobileNet-V2 connection. Their network reached the second-ranked speed of RAd-MobileNet, which attained an actual speed of 205 MHz. The DSP frequency used was higher than the RAd-

MobileNet execution by 83 DSP, however. In addition, those who used six times as many LUTs as we did.

Additionally, they consumed 7.53 W of power, which would be 4.1 W, far more RAd-MobileNet did. This trend suggests that our simulations successfully inaccuracy, energy consumption, and speed while requiring fewer system resources, as shown in Table 1. We choose the model that best fulfills our computing hardware capacity, efficiency, or reliability needs.

Conclusions. We aimed to create a structure with little physical servers but with the highest accuracy rate that didn't result in loss or destruction using an FPGA. To accomplish this, we enhanced the MobileNet-V1 framework. We commenced by removing any elements that were duplicating the resulting shape. Secondly, we experimented with different combinations of the training algorithm of ReLU, TanExp, and Mish to see what happened. All designs in this study employed the convolution layers of ReLU, TanhExp, and Mish. FPGA implementations of TanhExp and Mish non-linear transfer function became greatly facilitated by using PWL approximations. We used float-point integers to digitize the actual data. Thirdly, we amplified the classification rate of the design by replacing the complexity convolution component with a broader one called the Ad-depth module. We compared FPGA implementations on the Virtex-7 platform and other platforms regarding the number of physical servers, power efficiency, information processsing, and reliability. Also, with MAd-MobileNet modeling, we were able to attain a recognition performance of 88.76%, 18.72% more than for the basic MobileNet framework. In addition, the RAd-MobileNet design allowed us to achieve a bandwidth of 225 MHz, which is 120 MHz better than the primary method. It's essential to remember that we obtained the preceding performance with a simple local database CIFAR-10.

References

1. Huang, C. H. An FPGA-Based Hardware/Software Design Using Binarized Neural Networks for Agricultural Applications: A Case Study. IEEE Access. (2021) .
2. Jung, J., Park, J., & Kumar, A. "A verification Framework of Neural Processing Unit for super resolution". In 20th International Workshop on Microprocessor/SoC Test, Security and Verification (MTV), (2019).
3. Zhang, W., Li, X., & Ding, Q. "Deep residual learning-based fault diagnosis method for rotating machinery." ISA transactions (2019).
4. Wu, S., Zhong, S., & Liu, Y. "Deep residual learning for image steganalysis. Multimedia tools and applications". (2018).
5. Lee, J., Kim, C., Kang, S., Shin, D., Kim, S., & Yoo, H. J. "An energy-efficient deep neural network accelerator with fully variable weight bit precision". IEEE Journal of Solid-State Circuits (2018).

Sergo HARUTYUNYAN, Garnik VOSKANYAN, Andranik GALSTYAN, Davit GABRIELYAN, Armen SAFARYAN

Studying The Most Effective FPGA Neural Networks: Economic and Modeling Issues

Key words: FPGA, Neural Networks, Modeling, Simulation .

Field Programmable Gate Arrays (FPGAs) are semiconductor devices that are based around a matrix of configurable logic blocks (CLBs) connected via programmable interconnects. FPGAs can be reprogrammed to desired application or functionality requirements after manufacturing. This feature distinguishes FPGAs from Application Specific Integrated Circuits (ASIC), which are custom manufactured for specific design tasks. The purpose of this paper is to study and compare various FPGA neural networks and determine the most effective among them in terms of energy efficiency. Even though various research studies worked on different FPGA neural model designs, very few researchers have considered comparing these designs to determine energy efficiency. This paper provides Ad-MobileNet as an efficient solution, a system constructed on the conventional MobileNet topology that utilizes fewer computer complexity while achieving classification techniques superior to some well-known FPGA-based platforms. Because of this, we've investigate the Ad-MobileNet model with larger datasets and a range of uses as part of continuing research.