Embedded Neural Networks for Medical Image Classification

Nanda Kumar E L[1], M Sai Charan Teja[2], Manoj K[3], Manoj S[4], Dr. Kiran Bailey[5]

[1][2][3][4] UG Students, DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING, BMS College of Engineering (Affiliated to VTU), Bangalore, India.
[5] Assistant Professor, DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING, BMS College of Engineering (Affiliated to VTU), Bangalore, India.

---***---

Abstract - This paper "Embedded Neural Networks for Medical Image Classification" aims to improvise speed of execution of Neural networks by using specialized hardware architecture for computation in FPGA. With recent advancements in medical technology, the biomedical field has ushered in the era of big data. The applications of deep learning in medical image analysis, electronic health record, genomics and drug development has been found. Deep learning has obvious advantages in making full use of biomedical data and improving medical health level. Thus, there is a need for a hardware framework for inference in embedded medical applications which use deep learning. Deep learning is computationally intensive, real time embedded systems which use deep learning require fast computing hardware to meet its timing requirements. To meet these requirements, the project utilizes and surveys FINN framework which is a compiler framework for generating data flow accelerators for Neural Network models.

Key Words: Image classification, FINN, FPGA, Quantized Neural Networks, Inference engine.

1.INTRODUCTION

The field of medical health has ushered in the era of big data in recent years, thanks to the explosive increase of biomedical data and the rapid development of medical technology and computer technology. Computational medicine emerged as a new discipline in this environment. Computational medicine is an interdisciplinary field that combines medicine, computer science, biology, mathematics, and other disciplines. It is based on massive biomedical data and computer technology. Deep learning applications in medical imaging, electronic health records, genomics, and drug discovery are investigated, with the notion that deep learning has a clear advantage in making full use of biological data and enhancing medical health.

Unlike typical machine learning algorithms, deep learning eliminates the need to manually extract features, which saves time and resources. To extract features, deep learning must process a huge number of biomedical datasets.

Medical image analysis is one of the Deep Learning applications in the world of medicine and diagnosis. Deep learning can be used to help radiologists make diagnoses. Deep learning with convolutional neural networks (CNNs) has lately gained widespread attention due to its superior performance in image recognition.

Dataflow accelerators can be used to accelerate deep learning model inference, such as CNN. Nodes in the hidden layers of CNNs do not necessarily share their output with every node in the following layer. The development of CNNs solutions has accelerated in many applications. This was enabled by the advent of fast GPUs, which could satisfy huge bandwidth and computational complexity caused by the ever-increasing size of CNNs. The deep learning solutions not only limit their applications to heavy computing machines, but there is also growing interest in deploying CNNs on edge devices with limited hardware resources and energy.

2.LITERATURE SURVEY

Radiological imaging diagnosis is crucial in clinical patient care. Deep learning with convolutional neural networks (CNNs) has lately gained widespread attention due to its superior performance in image recognition. CNNs can assist radiologists in achieving diagnostic excellence and improving patient care [1]. According to contemporary estimates, one billion radiologic tests are performed worldwide each year, the most majority of which are evaluated by radiologists [2]. Most professional organizations would agree that all imaging operations should include an experienced radiologist's opinion, which should be provided in the form of a written report [3]. According to [4], computer aided radiology can be utilized to reduce the possibility of missing some radiographic abnormalities in mammography and lung nodule identification on CT. In [5-8] different authors have performed medical image classification for diagnosis using convolutional networks and various other deep learning models. In other clinical image classification tasks, the convolution neural network also achieved the performance of doctors including skin cancer diagnosis, knee osteoarthritis diagnosis, bone age assessment [9 – 11]. In this paper, chest X-ray image classification is undertaken considering its simplicity for analyzing the effectiveness of the Hardware implementation.

Deep learning needs to process such large amounts of biomedical data to obtain efficient information. Thus, training and inference of NNs are computationally intensive. There are many hardware-software based computational architectures for training and inference of NNs. To balance flexibility and efficiency, a CPU and accelerator combo is a popular choice.

Manufacturers, including those of GPU accelerators and DSP accelerators, are now researching the architecture of various deep learning accelerators [12].

The main goal of current research has been to decrease data transfer while retaining accuracy, throughput, and cost because it accounts for the majority of energy use. Choosing designs with advantageous memory hierarchies, such as a spatial array, and creating dataflows that promote data reuse at the low-cost levels of the memory hierarchy are necessary to achieve this. Reduced bit width precision, higher sparsity, and compression are employed to reduce the amount of required data movement through collaborative algorithm and hardware design [13]. Field programmable gate arrays (FPGA) offer an intriguing alternative to the general-purpose general-purpose processors (GPGPU) that have been used as the present answer. FPGAs are now more compatible with high-level software practices often used in the deep learning community thanks to current developments in design tools, making them more available to individuals who create and use models. Due to the flexibility of FPGA architectures, researchers may be able to investigate model-level optimizations in ways that are not achievable with inflexible architectures like GPUs. Additionally, FPGAs frequently offer excellent performance per watt of power consumption, which is crucial for application scientists interested in embedded applications with constrained resources or large-scale server-based deployment.

The Xilinx team proposed the FINN Framework in [14] to operate neural networks on FPGA. It was suggested for creating FPGA-accelerated BNNs that are quick and adaptable. Finn designed fully connected, non-padded convolutional and pooling layers, with per-layer computation resources being matched to user-provided throughput needs, using a novel set of optimizations that enable efficient mapping of BNNs to hardware. Binarized neural networks (BNNs) with one-bit weights and activations are explored in [15]. The deep learning community is becoming more interested in binarized neural networks (BNNs) as a result of their much lower computational and memory costs. They work especially well with reconfigurable logic devices since these devices have a lot of fine-grained computing power and can either produce smaller, lower power implementations or, conversely, higher classification rates. The quick investigation of Binary Neural Networks on Reconfigurable Logic using a C++ library is covered in [16].

3.DATASET

The dataset is divided into the following 3 folders: train, test, and val. Each image category (Pneumonia/Normal) has its own subfolder within the dataset. There are 2 categories (Pneumonia/Normal) and 5,863 X-Ray images in JPEG format.

From retrospective cohorts of pediatric children aged one to five at the Guangzhou Women and Children's Medical Centre in Guangzhou, chest X-ray images (anterior-posterior) were chosen. All chest X-ray imaging was done as part of the regular clinical treatment provided to patients.

All chest radiographs were initially checked for quality control before being removed from the study of the chest x-ray pictures. Before the diagnosis for the photos could be used to train the AI system, they were graded by two experienced doctors. A third expert reviewed the evaluation set.

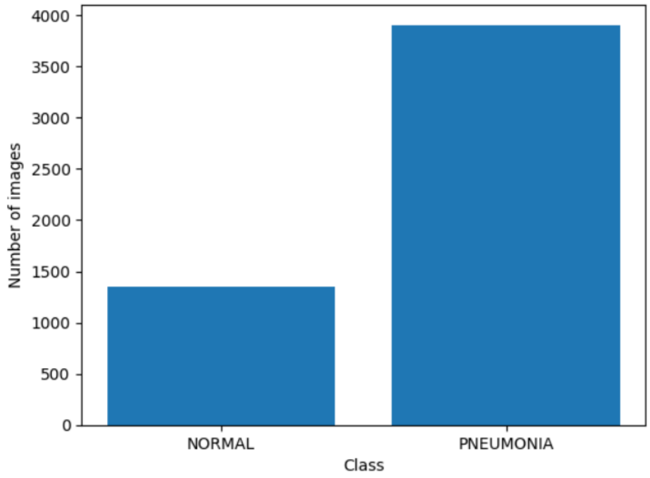

Fig. 1. Class distribution in the train dataset.

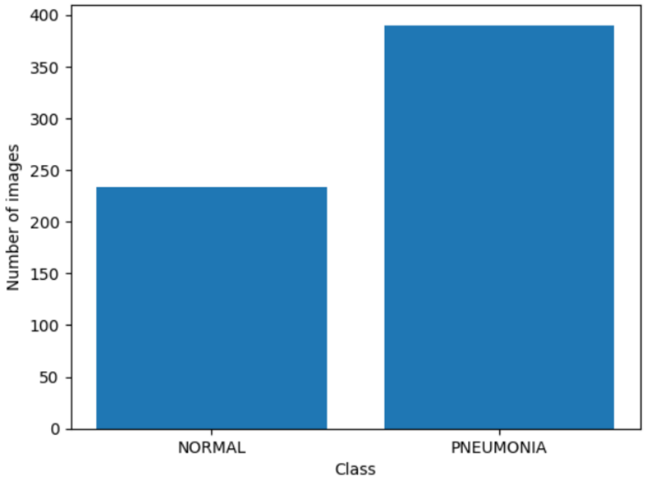

Fig. 2. Class distribution in the test dataset.

4.MODEL

The model was designed with fully connected layers and ReLu activation function. The weights, activations and biases are 8 bit quantized. Quantization is performed with the brevitas library. It is a PyTorch library for quantization of neural networks with support for quantization aware training and post quantization training [14]. The model is optimized for implementation in FPGAs for providing High throughput and low latency. High throughput is achieved through quantization of fully connected layers, which also helps in single fold implementation of all the layers in the FPGA since reduced bits mean easier computation and less area will be required for hardware implementation in FPGA. Streamlining architecture reduces the latency of the inference.

A. Quantization:

A crucial stage in the implementation of Neural Networks on FPGA and ASIC platforms is the choice of data type. Arithmetic operations on floating-point data have a high computational complexity, necessitating the use of specialized hardware (such as floating-point units). The machine learning community initially preferred neural networks that used floating point computations, although learnt parameters can have a lot of duplicate information. The fixed-point format offers an alternative to the floating-point format. The Q-format specification Q m. f, where m specifies the number of integer bits and f denotes the number of fractional bits, can be used to specify a fixed-point number format. Quantization is a prospective essential strategy for using redundancy in neural networks. It has been shown that switching from floating-point to low-precision integer arithmetic has only a small influence on the accuracy of the network, particularly if the network is retrained. Numerous software libraries have been designed and optimized to support the quantization of neural networks down to 8-bit integers, including gemmlowp and the ARM Compute Library. The extreme example of Binary Neural Networks (BNNs) with binary representations for synapse weights, input and output activations was used in this study as a further decrease of precision because it has the lowest hardware cost.

B. Dataflow Accelerator:

Instead of scheduling operations on top of a fixed architecture, a heterogeneous streaming architecture creates a tailored architecture for a certain topology. Each layer has its own specialized computation engine, and these engines talk to one another through on-chip data streams. As soon as the preceding engine begins to provide output, the subsequent engines begin to compute. Additionally, all neural network parameters are stored in on-chip memory because of the small model size of BNNs. By overlapping processing and communication, this minimizes the start interval, eliminates the majority of accesses to off-chip memory, and reduces latency (the time it takes to finish categorizing one image). By customizing compute arrays for each layer's needs separately, "one-size-fits-all" inefficiencies can be avoided, allowing for more reconfigurable computing benefits

.

C. Vivado IP Generation:

Vivado IP can be generated using FINN compiler framework. Xilinx Research Labs created the experimental framework FINN to investigate the deep neural network (DNN) interface on FPGAs. It focuses on creating data flow type topologies unique to each network and specifically targets quantized neural networks (QNN). The primary benefit of adopting the FINN framework is its suitability for real-time and low-power applications due to its high performance and low latency for neural network inference on FPGAs. Additionally, the framework gives designers freedom to create unique neural network designs that are hardware-optimized for FPGA. The Brevitas tools for training QNNs, the FINN compiler, and the FINN-hlslib Vivado HLS library of FPGA components for QNNs are the main elements of the FINN framework. The FINN framework's flow is depicted in Figure 3.

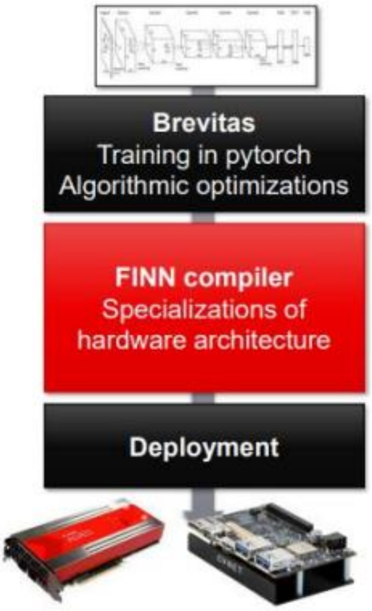

Fig. 3. FINN Flow

5.RESULT

Model is trained for 3 epochs and the training loss at the end of the third epoch is 0.1776.

The test accuracy of the designed model is 84.13%., with weight, activation and bias quantization. The confusion matrix for the designed model is shown in Fig. 4. This result was obtained after reducing the bit size to 8.

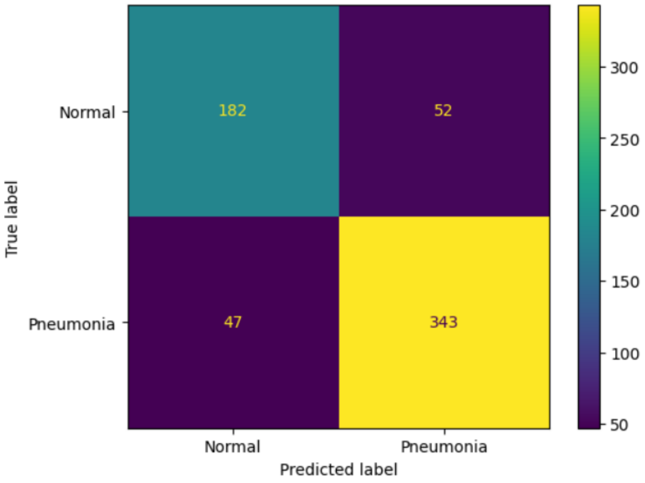

Fig. 4.Confusion matrix

6.CONCLUSION AND FUTURE WORK

Quantization upto 8 bits doesn't significantly affect the accuracy of image classification models with fully connected layers. Further reduction in bit size affects the accuracy. Quantization exploits the redundancy in neural networks. For critical image classification tasks such as medical image classification the reduction in bitsize affects the accuracy very slightly, which is a compromise for faster inference rate and reduced latency.

REFERENCES

[1] Koichiro Yasaka et al., "Deep learning and artificial intelligence in radiology: Current applications and future directions",2018.

[2] Bruno MA, Walker EA, Abujudeh HH (2015) Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radio graphics 35:1668–1676

[3] Royal College of Radiologists (2006) Standards for the reporting and interpretation of imaging investigations. RCR, London

[4] Adrian P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?", November 2016.

[5] Cao, Y., Liu, C., Liu, B., Brunette, M. J., Zhang, N., Sun, T., et al. (2016). *Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. In 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE).* New York, NY: IEEE, 274–281. doi: 10.1109/CHASE.2016.18

[6] Cicero, M., Bilbily, A., Colak, E., Dowdell, T., Gray, B., Perampaladas, K., et al. (2017). Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investigat. Radiol.* 52, 281–287. doi: 10.1097/rliYuan,

[7] D., Zhu, X., Wei, M., and Ma, J. (2019). *Collaborative deep learning for medical image analysis with differential privacy. In 2019 IEEE Global Communications Conference (GLOBECOM).* New York, NY: IEEE, 1–6.

[8] Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., and Pu, J. (2019). SDFN: segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Comp. Med. Imag. Graphics* 75, 66–73. doi: 10.1016/j.compmedimag.2019.05.005

[9] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J. M., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. doi: 10.1038/nature21056

[10] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 29, 1836–1842.

[11] Rakhlin, A., Shvets, A., Iglovikov, V., and Kalinin, A. A. (2018). "Deep convolutional neural networks for breast cancer histology image analysis," in International Conference on Image Analysis and Recognition, (Montreal: ICIAR), 737–744. doi: 10.1007/978-3-319-93000-8_83

[12] V. Sze, Y. H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: challenges and opportunities," in Proceedings of the IEEE conference on Custom Integrated Circuits Conference, April 2017.

[13] Vivienne et al., "Hardware for Machine Learning Challenges and Opportunities", IEEE Custom Integrated Circuits Conference, April 2017

[14] Alessandro, Pappalardo, "Xilinx/brevitas, zenodo, 2023, doi: 10.5281/zenodo.3333552