An Energy-Efficient and Reconfigurable CNN Accelerator Applied To Lung Cancer Detection

Yi Hsin Liao
Electrical Engineering
National Tsinghua University
Hsinchu, Taiwan
asd58400664@gmail.com

Shu You Lin
Electrical Engineering
National Tsinghua University
Hsinchu, Taiwan
terry900531@gmail.com

Yu-Chiao Chen
Electrical Engineering
National Tsinghua University
Hsinchu, Taiwan
richard5512278@gmail.com

Hsin-Han Chen
Electrical Engineering
National Tsinghua University
Hsinchu, Taiwan
ian900623@gmail.com

Ding Xiao Wu
Electrical Engineering and Computer Science
National Tsinghua University
Hsinchu, Taiwan
dingxiao1230@gmail.com

Hong Wen Luo
Electrical Engineering and Computer Science
National Tsinghua University
Hsinchu, Taiwan
michellewenlooo@gmail.com

Kea-Tiong Tang
Electrical Engineering
National Tsinghua University
Hsinchu, Taiwan
kttang@mx.nthu.edu.tw

2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 979-8-3503-3267-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICAS57966.2023.10168583

***Abstract*—We propose a system to fast and easily detect lung cancer by breathing into the device, which is not invasive. Some particular substances only exist in lung cancer patients' breathing. Based on this, we use the CNN model to extract the feature in the gas exhaled by the testee. Then, the neural network will give out the prediction of lung cancer. To accelerate the computation of CNN, we design a hardware accelerator and implement it with FPGA (Field Programmable Gate Array). By comparing the performance, like power consumption and energy efficiency of different architectures, we could find the most appropriate architecture for us. Ultimately, we could reduce memory access by about 20% and reduce 12% of the energy consumption, achieving low power at edge devices. The performance of the CNN model is with a training accuracy 88.41%, a testing accuracy 85.29%, a false negative rate 5.8%, and a false positive rate 41.17%**

Keywords—lung cancer detection, CNN accelerator

I. INTRODUCTION

In recent years, lung cancer has become one of Taiwan's top 10 causes of death. Patients diagnosed with terminal cancer receive treatments, such as chemotherapy, targeted therapy, and radiation therapy, which only control cancer. However, cancer in its early stage could be effectively cured, which reduces the fatality rate. Under normal circumstances, lung cancer could be detected by a CT scan (Computed Tomography scan), which takes much time and a vast space to accommodate the medical equipment. CT scan is not convenient enough for people who overlook early cancer, leading to cancer usually being diagnosed in the end-stage. To increase the rate of detecting lung cancer earlier, we aim to make a quick sieve, which is feasible for people.

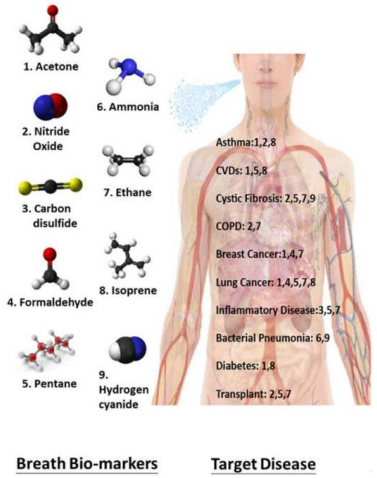

Fig. 1. Breath Bio-markers and target disease.

According to the research, some chemical materials in breath, such as Acetone, Formaldehyde, Pentane, Ethan, and Isoprene, only exist in those who get lung cancer. As shown in Fig 1., some lung diseases are on the right side. The number means that the bio-markers correspond to the disease. Based on this concept, we can tell whether the person gets lung cancer by identifying these specific substances from breath. In other words, we need to extract these bio-markers from breath.

Therefore, we propose a CNN system based on GasNet, which can detect lung cancer quickly and inexpensively. The system could function as a "quick sieve for lung cancer. " It could not only effectively release healthcare pressure but also enhance the quality and efficiency of healthcare in Taiwan.

Since we want to make a quick sieve, we design a hardware accelerator with FPGA to make the whole system operate more efficiently, fast, and portable. The organization in the following section is SectionII System Flow, SectionIII FPGA Architecture, SectionIV RESULT and ADVANTAGES, SectionV Application, and SectionVI Demonstration Setup.

II. SYSTEMFLOW

The system flow is shown in Fig2.

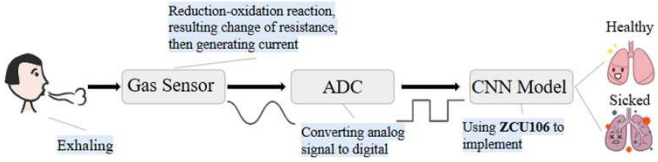

Fig. 2. system flow

Authorized licensed use limited to: INSTITUTE OF ENGINEERING & MANAGEMENT TRUST. Downloaded on February 18,2026 at 08:56:02 UTC from IEEE Xplore. Restrictions apply.

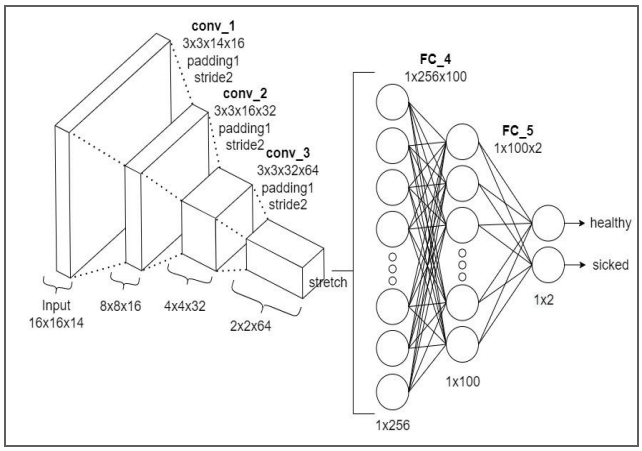

Fig. 3. CNN model with layer size

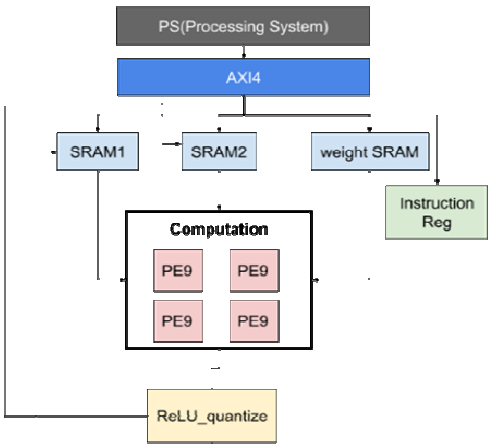

Fig. 4. The block diagram of the FPGA architecture

A. *Gas Sensor Array and ADC*

First, we use a sensor array to convert the exhaled gas into digital data. When the gas comes in, the materials inside the sensor will go through redox reactions, which change its resistivity and generate the current. Then, the sensor array records the magnitude of voltage according to resistivity and converts these values into 14x16x16 digital data, each containing 4 bits. So far, we have completed data preprocessing.

B. *CNN model*

As shown in Fig. 3, we design our CNN model with three convolution layers and two fully connected layers based on the GasNet. The layer size is labeled on Fig. 3. The CNN model has been trained under 35 healthy samples and 102 LCA samples which the corporate hospital provides. To avoid overfitting the dataset, we design our CNN model in a relatively small size since the dataset is not large enough. As we get more and more samples, we will modify the CNN model to reach higher accuracy.

The training accuracy is 100%, and the testing accuracy is 83.82% before quantization. However, we want to implement the CNN model on hardware devices, so we quantize our parameters into 4-bit fix points using dorefa. After that, we get a training accuracy of 88.41% and a testing accuracy 85.29%.

TABLE I. TRAINING ACCURACY

Confusion matrix (training)		**Prediction**			
		Before quantization		***After quantization***	
		Healthy	***LCA***	***Healthy***	***LCA***
Ground truth	***Healthy***	18	0	12	6
	LCA	0	51	2	49
Training Accuracy		100%		88.41%	

TABLE II. TESTING ACCURACY

Confusion matrix (testing)		**Prediction**			
		Before quantization		***Before quantization***	
		Healthy	***LCA***	***Healthy***	***LCA***
Ground truth	***Healthy***	7	10	10	7
	LCA	1	50	3	48
Training Accuracy		83.82%		85.29%	

To accelerate training convolutional neural networks in low bitwidth parameters, we choose “DoReFa-Net”[5] to train our model. In DoReFa quantization, weights, activations, and gradients are first normalized to the limit range in [-1, 1]. Then, they are quantized to a low bitwidth using a function that applies an affine mapping between these low bitwidth numbers and fixed-point integers. Finally, parameter gradients are stochastically quantized to low bitwidth numbers before propagating to convolutional layers. Overall, it saves run-time memory and forward and backward computation complexity, then can be accelerated significantly, and also has a novel gradient quantization method to train the network effectively.

It is essential to recognize sick people, especially quick sieves. From the confusion matrix shown in tableI and tableII, the false negative rate at testing after quantization is 3 / (3 + 48) = 5.8%, and the false positive rate after quantization is 7 / (10 + 7) = 41.17%. The false negative rate is much less than the false positive.

Now, as we get 14x16x16 binary data from the sensor, the data can be viewed as a 14x16x16 image. Then, we can apply our CNN model implemented on the FPGA to tell if the patients get lung cancer quickly.

III. FPGA ARCHITECTURE

The block diagram of the designed FPGA architecture is shown in Fig. 4.

A. *Ping-Pong method*

We adopt the Ping-Pong method for storing data. SRAM will take over less area in this way. Otherwise, we will have numerous SRAMs according to the number of layers in the model. Then, how does Ping-Pong method work? In Fig. 3, assuming SRAM1 is the input RAM storing the input data, SRAM2 is the output RAM storing the output data at the first layer. When finishing the computation for the first layer, the hardware will exchange the role of SRAM1 and SRAM2. In other words, SRAM1 will be the output RAM, while SRAM2 will be the input RAM. SRAM in the rest layer of the model will follow this kind of rule.

Authorized licensed use limited to: INSTITUTE OF ENGINEERING & MANAGEMENT TRUST. Downloaded on February 18,2026 at 08:56:02 UTC from IEEE Xplore. Restrictions apply.

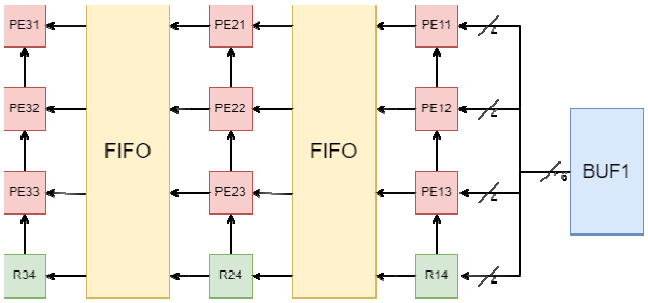

Fig. 5. Computation core

B. *Instruction Register*

The functionality of instruction registers in Fig. 4 is to provide the layer information to the hardware, such as input size, output size, and layer type (Convolution or Fully Connected). With this instruction register, adding more layers to the model is easy. What we need to do is add more instructions to the instruction register.

C. *Computation Core*

We design our computation core as Fig. 5. We apply weight stationary and use FIFO to increase the pixel-reuse ability. Since the current input pixel is the pixel from their neighbor, we use a shift register to store those pixels. As the convolution goes to the next row, most of its input pixel is the pixel used in that column, so we use a FIFO to store these pixels. With the design of these registers and the FIFO, we don't need to fetch the same data from RAM. It is fantastic to achieve less memory access that will reduce power consumption with pixel reuse in the architecture. Moreover, zero-padding is improved by the arrangement of dataflow, and we don't need to store zero-padding pixels in RAM, which reduces the storage size of RAM.

D. *PS(Processing System) and AXI4*

The processing system on FPGA allows us to control the hardware resources by Python. In this way, we could input the gas data to the hardware on the computer with the communication protocol AXI4.

IV. RESULT AND ADVANTAGES

In our accelerator, there are some advantages.

*1) Reduction on storage:*Attributing to designate of architecture, it is needless to store zero padding in RAM, so that we could reduce 20 % storage of RAM(586710 um^2 → 459913 um^2)

*2) Robust to size variation:*Since we have the instruction register, it is robust to input size variation.

*3) Acceleration:*With the accelerator, we can implement it in real-time. Compared with CPU, which costs 50ms, we only use 0.6ms to calculate.

*4) Model accuracy:*In model accuracy, we discuss the accuracy before and after quantization. The training accuracy before quantization can reach 100%, and the testing accuracy can reach 83.82%; after quantization, the training accuracy can reach 88.41%, and the testing accuracy can reach 85.29%.

V. APPLICATION

There are a lot of large instruments for lung cancer inspection, which is not common to see outside the exceptional hospitals. Besides, these instruments should primarily be operated by professional medical personnel. It is almost impossible for the general public to use the devices themselves. Furthermore, the majority of people need help to afford the costly expense.

Compared with the abovementioned instruments, it has a smaller volume and is more convenient. By only exhaling to the gas sensor array, which is NOT invasive, we quickly finish the lung cancer inspection. In the era of covid-19 epidemic, this work effectively releases the loading of healthcare. Moreover, people could also frequently conduct inspections. This way, we attain the goal of "Finding cancer earlier, being cured earlier. "

VI. DEMOSTRATION SETUP

We will bring a laptop, a gas sensor, an airbag for collecting exhaled gas, and the FPGA of type ZCU106. Our gas sensor takes roughly half an hour to set up the baseline.

The laptop functions as the controller of our accelerator and the gas sensor. We have a unique app for controlling the operation of the gas sensor. After that, the tester uses the processing system on FPGA with the Jupyter Notebook to input the gas data recorded by the gas sensor to our accelerator. Then, we could get the output prediction for lung cancer.

REFERENCES

[1] haofang Li, Syuan-Hao Sie, Jye-Luen Lee, Yi-Ren Chen, Ting-I Chou, Ping-Chun Wu, Yu-Ting Chuang, Yu-Te Lin, I-Cherng Chen, Chih-Cheng Lu, Ying-Zong Juang, Shih-Wen Chiu, Chih-Cheng Hsieh, Meng-Fan Chang and Kea-Tiong Tang, "A Miniature Electronic Nose for Breath Analysis", 2021 IEEE International Electron Devices Meeting (IEDM))

[2] Ting-I Chou, Shih-Wen Chiu, Kwuang-Han Chang, Yi-Ju Chen, Chen-Ting Tang, Chung-Hung Shih, Chih-Cheng Hsieh, Meng-Fan Chang, Chia-Hsiang Yang, Herming Chiueh, and Kea-Tiong Tang, "Design of a 0.5V 1.68mW Nose-on-a-Chip for Rapid Screen of Chronic Obstructive Pulmonary Disease", the IEEE 2016 Biomedical Circuits and System Conference (BioCAS 2016), Shanghai, China, 2016

[3] Shih-Wen Chiu, Jen-Huo Wang, Kwuang-Han Chang, Ting-Hau Chang, Chia-Min Wang, Chia-Lin Chang, Chen-Ting Tang, Chien-Fu Chen, Chung-Hung Shih, Han-Wen Kuo, Li-Chun Wang, Hsin Chen, Member, IEEE, Chih-Cheng Hsieh, Meng-Fan Chang, Yi-Wen Liu, Tsan-Jieh Chen, Chia-Hsiang Yang, Herming Chiueh, Juyo-Min Shyu, and Kea-Tiong Tang, "A Fully Integrated Nose-on-a-Chip for Rapid Diagnosis of Ventilator-Associated Pneumonia", IEEE Transaction on Biomedical Circuits and Systems, vol. 8(6), pp. 765-778, 2014

[4] K.-T. Tang, S.-W. Chiu, C.-H. Shih, C.-L. Chang, C.-M. Yang, D.-J. Yao, J.-H. Wang, C.-M. Huang, H. Chen, K.-H. Chang, C.-C. Hsieh, T.-H. Chang, M.-F. Chang, C.-M. Wang, Y.-W. Liu, T.-J. Chen, C.-H. Yang, H. Chiueh, J.-M. Shyu, "A 0.5V 1.27mW Nose-on-a-Chip for Rapid Diagnosis of Ventilator-associated Pneumonia", 2014 International Solid-State Circuits Conference (ISSCC), San Francisco, United State

[5] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, Yuheng ZouMegvi i Inc. "DOREFA-NET: TRAINING LOW BITWIDTH CONVOLUTIONAL NEURAL NETWORKS WITH LOW BITWIDTH GRADIENTS"

Authorized licensed use limited to: INSTITUTE OF ENGINEERING & MANAGEMENT TRUST. Downloaded on February 18,2026 at 08:56:02 UTC from IEEE Xplore. Restrictions apply.