International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 1, February 2022, pp. 997~1002
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i1.pp997-1002

A 15-Gbps BiCMOS XNOR gate for fast recognition of COVID-19 in binarized neural networks

Rosana W. Marar[1], Hazem W. Marar[2]
[1]Computer Graphics and Animation Department, Princess Sumaya University for Technology, Amman, Jordan
[2]Computer/Electrical Engineering Department, Princess Sumaya University for Technology, Amman, Jordan

Article Info

Article history:

Received Apr 30, 2020
Revised Jul 18, 2021
Accepted Aug 7, 2021

Keywords:

BiCMOS
Corona-virus
COVID-19
Deep neural networks
XNOR

ABSTRACT

The COVID-19 pandemic is spreading around the world causing more than 177 million cases and over 3.8 million deaths according to the European Centre for Disease Prevention and Control. The virus has devastating effects on economies, health, and well-being of worldwide population. Due to the high increase in daily cases, the available number of COVID-19 test kits in under-developed countries is scarce. Hence, it is vital to implement an effective screening method of patients using chest radiography since the equipment already exists. With the presence of automatic detection systems, any abnormalities in chest radiography that characterizes COVID-19 can be identified. Several artificial-intelligence algorithms have been proposed to detect the virus. However, neural networks training is considered to be time-consuming. Since computations in training neural networks are spent on floating-point multiplications, high computational power is required. Multipliers consume the most space and power among all arithmetic operators in deep neural networks. This paper proposes a 15 Gbps high-speed bipolar-complementary-metal-oxide-semiconductor (BiCMOS) exclusive-nor (XNOR) gate to replace multipliers in binarized neural networks. The proposed gate can be implemented on BiCMOS-based field-programmable gate arrays (FPGAs). This will significantly improve the response time in identifying chest abnormalities in CT scans and X-rays.

Corresponding Author:

Hazem W. Marar
Computer/Electrical Engineering Department, Princess Sumaya University for Technology
Amman, Jordan
Email: h.marar@psut.edu.jo

1. INTRODUCTION

The widespread of corona-virus pandemic has become a serious public health problem worldwide [1], [2]. Devastating effects on economies, health, and well-being of the global population is actively occurring across the world. The virus infection results a severe acute respiratory syndrome, multi-organ failures, or sometimes death [1]-[3]. COVID-19 infected more than 177 million cases and over 1.8 million deaths around the globe. Figures 1 and 2 illustrate the fast exponential rate at which the number of cases and total deaths are growing.

A vital step in the fight against COVID-19 is effective screening of infected patients. This implies the immediate treatment and isolation of those who are infected to hamper the spread of the virus [4]-[6]. The main screening method recommended for detecting COVID-19 cases is reverse transcriptase-polymerase chain reaction (RT-PCR) testing [7]. However, due to the long time consumption, labor-intensive, and complicated manual process, this method usage is limited in under-developed countries. An alternative, more efficient screening method is radiography examination. In such method, chest radiography imaging like chest

X-ray or computed tomography (CT) imaging are used and analyzed. Looking for distinguishable visual indicators associated with COVID-19, radiography imaging recognition could be used as a primary tool for COVID-19 screening in epidemic areas [8]-[10]. Several recent studies focused on implementing computer aided diagnostic systems that can aid radiologists to rapidly and accurately examine and interpret radiography images to detect COVID-19 cases. This is due to the scarcity of expert radiologists to interpret the radiography images during this pandemic. Hence, automated radiography examination can be of great help in conducting faster diagnostics [11], [12]. Furthermore, due to the availability of chest radiology imaging systems in under-developed healthcare systems, utilizing an automated recognition system can be vital and of higher sensitivity and precision [13], [14].

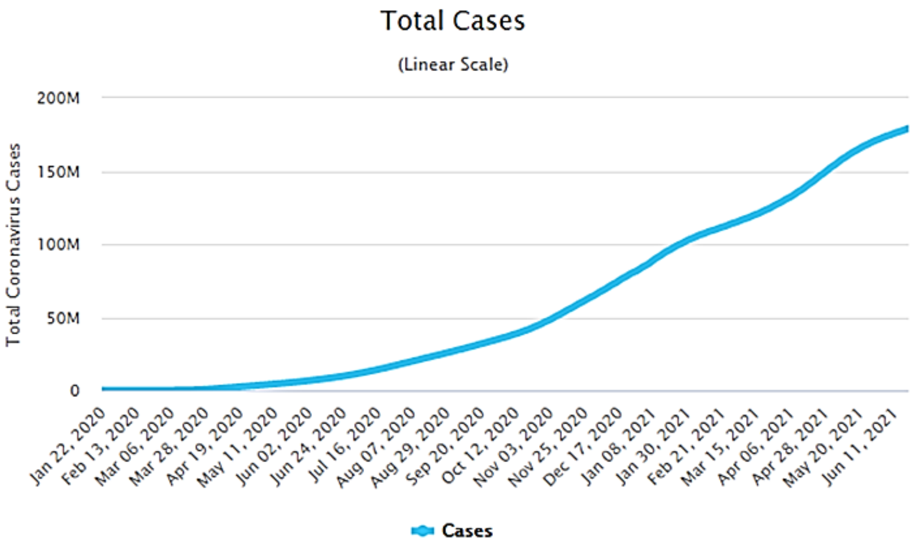

Figure 1. Total number of COVID-19 cases [3]

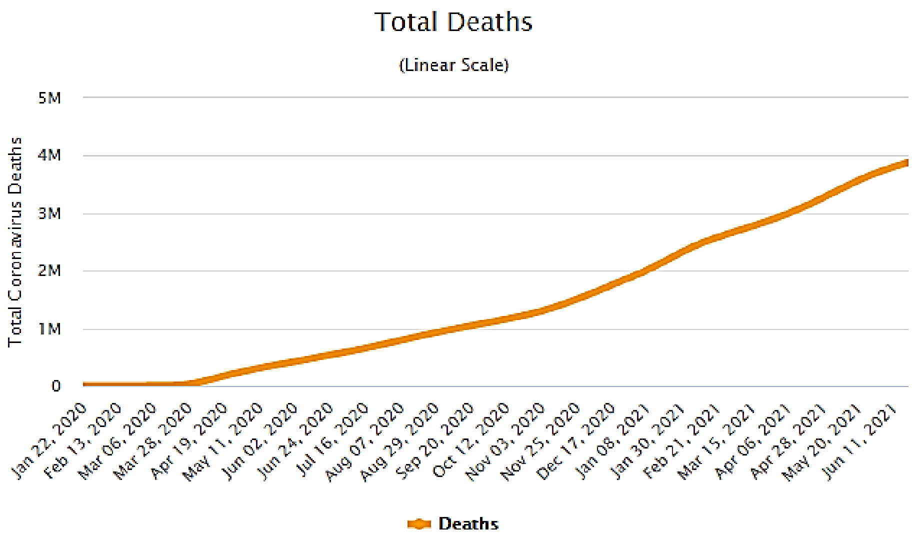

Figure 2. Total number of COVID-19 deaths [3]

A deep learning-based system for automatic segmentation of lungs and infection sites using chest CT is developed in [15]. Changes in the follow-up CT scans of COVID-19 patients are tracked and analyzed. The system is able to measure the severity of cases based on a deep learning system. An early screening model to differentiate COVID-19 pneumonia and Influenza-A viral pneumonia from healthy cases utilizing pulmonary CT images and deep learning techniques [16]. A deep learning method that can extract the graphical features of COVID-19 is proposed in [17]. Based on the COVID-19 radiographic changes from CT images, a system to provide clinical diagnosis prior to pathogenic testing is presented. Consequently, saving critical time for the disease diagnosis. It is evident that all research is focused on improving the speed of the virus recognition. Deep neural networks (DNN) and convolutional neural networks (CNN) are often used for this purpose. Training sets of X-ray images or CT scan images are fed into the neural network. However, with the fast spread of the pandemic along with the enormous number of training sets and diagnostics, neural networks will consume heavy time processing data. X-ray is the most popular diagnostic imaging tool used for COVID-19 recognition. Its broad availability and accessibility make it the primal choice of testing. X-rays are a 2-dimensional form of radiation, when passed through a patient's body, bone and other dense objects block the radiation and appear white on the film of the X-ray. The less dense tissues are harder to recognize and appear gray. While being the most popular diagnostic tool, X-rays accuracy might not be high due to the limited image quality. On the contrary, CT scans generates 3-dimensional, high-quality, detailed images of the body. It is considered a more powerful and sophisticated tool that scans a 360-degree image of internal organs [18], [19]. However, the output image of a CT is much larger than an X-ray. This causes neural networks training to consume heavy time that is vital for patients' treatment.

For the majority of deep learning algorithms, training is considered a notorious time consumer. This is due to fact that most of the computation in training binarized neural networks is typically spent on matrix multiplications. Multipliers are considered to consume the most space and power among all arithmetic operators of the digital implementation of deep neural networks [20], [21]. A solution using a novel BiCMOS XNOR gate to immensely speed up the process of neural networks training is proposed. Figure 3 illustrates how XNOR gates with a use of a simple look-up table can replace multipliers in binarized neural networks [22]-[24].

$$[-1 \quad +1 \quad +1] \times \begin{bmatrix} -1 & -1 & +1 \\ +1 & -1 & +1 \\ +1 & -1 & +1 \end{bmatrix} = [-1\times-1+1\times1+1\times1 \quad -1\times-1+1\times-1+1\times-1 \quad -1\times1+1\times1+1\times1]$$

$$= [3 \quad -1 \quad +1]$$

(a)

$$[0 \quad 1 \quad 1] \otimes \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} Bcount(XNOR(011,011)) \\ Bcount(XNOR(011,000)) \\ Bcount(XNOR(011,111)) \end{bmatrix}^T = [3 \quad -1 \quad +1]$$

(b)

Input	Computation	Output
000	-1-1-1=-3	101
001	-1-1+1=-1	111
010	-1+1-1=-1	111
011	-1+1+1=+1	001
100	+1-1-1=-1	111
101	+1-1+1=+1	001
110	+1+1-1=+1	001
111	+1+1+1=3	011

(c)

Figure 3. Using XNOR gates to replace multipliersa: (a) example of binarized matrix multiplication, (b) example of binarized matrix multiplication using XNOR and Bcount function. (-1 is represented using 0), (c) Bcount function (output is in 2's complement form)

2. RESEARCH METHOD

Since computations in training neural networks are mainly spent on floating-point multiplications, multipliers are considered to be the bottleneck of performance in deep neural networks. A model in which multipliers are replaced with XNOR operations is proposed [25]. XNOR operations are considered free in Yao’s garbled circuits (GC) protocol [26]. Therefore, performing oblivious inference on binary neural networks (BNNs) using GC results in the removal of costly multiplications. Hence, converting multiplication operations into faster XNOR operations will dramatically cut down the processing time of BNNs and remarkably improve response time that is extremely crucial in COVID-19 recognition. Figure 4 illustrates the proposed novel BiCMOS XNOR gate.

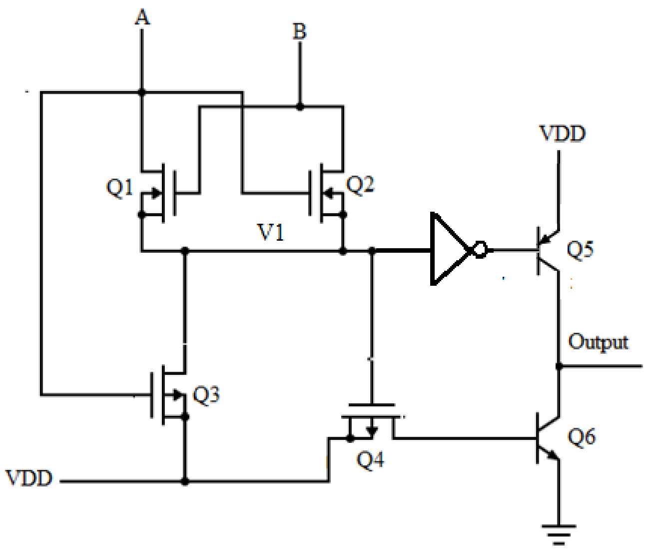

Figure 4. Proposed BiCMOS XNOR gate

Consisting of only 6 MOSFETs and 2 BJTs, this XNOR gate has minimal size compared against available BiCMOS logic gates. This allows it to be an ideal option to be integrated within DNNs. Multiple BiCMOS-based field-programmable gate array (FPGA) modules have been designed [27]-[29]. Hence, a neural network can be modeled on FPGA modules using the proposed XNOR gate. Transistors Q1 and Q2 are native transistor. Native transistors are a special variety of the metal oxide semi-conductor field-effect transistors (MOSFET). They operate in an intermediate state between enhancement and depletion modes. Replacing the layer of insulating oxide under the Gate terminal with a thin oxide film formed over silicon during processing of other layers causes the threshold voltage to drop. Hence, a native MOSFET is considered as a transistor with nearly zero threshold voltage [26]. If both inputs A and B are low, the large native negatively metal oxide semiconductor (NMOS) transistor Q1 and the native NMOS transistor Q2 are off. The positive-channel metal oxide semiconductor (PMOS) transistor Q3 is ON passing high voltage (VDD) to the node V1. This, in turns, turns off the PMOS device Q4 and turns on the PNP transistor Q5. The high output voltage is expressed in (1):

$$V_{OH_{WithoutShunt}} = VDD - V_{CE\_SAT} \tag{1}$$

where V_{CE_SAT} is the saturation collector-emitter voltage. $V_{CE\_SAT} < 200mV$.

In case of having A as high input and B as low input, both Q1 and Q3 transistors are off. Q2 will discharge the voltage on node V1. This allows the PMOS device Q4 to activate the NPN transistor Q6, and hence discharging the output to logic low as illustrated in (2):

$$V_{OL} = V_{CE\_SAT} \tag{2}$$

where V_{CE_SAT} is the saturation collector-emitter voltage. $V_{CE\_SAT} < 200mV$.

In case of having A as low input and B as high input, the large Q1 transistor causes the node V1 to be low, allowing the PMOS device Q4 to activate the NPN transistor Q6. This generates a low output voltage described in (3). If both inputs A and B are high, both Q1 and Q2 pass high voltage to node V1 switching the output voltage to VDD. Usually, complementary metal oxide semiconductor (CMOS) technology affords benefits of having less power dissipation and low noise margin with high packing density. While bipolar junction transistor (BJT) technology affords high switching with good noise performance. The BiCMOS technology uses MOSFETs to provide high input impedance logic gates and bipolar transistors to provide high current gain. This results a very fast gate that is ideal to use in CT images recognition in neural networks.

3. SIMULATION RESULTS

28 nm CMOS design kit was used to build and simulate the novel BiCMOS XNOR gate. A clocked binary sequence is fed to the input of the logic gate. Figure 5 illustrates the output waveform of the proposed design. As seen from Figure 5, the proposed BiCMOS XNOR gate operates at 15-Gbps. This gate will dramatically reduce the computational requirements of multiplication in BNNs.

Figure 5. Simulation output waveform of proposed BiCMOS XNOR gate

4. CONCLUSION

In collaboration with the global efforts against the COVID-19 pandemic, neural networks play a significant role in identifying chest abnormalities in CT scans and X-rays. However, due to the large data size of CT scans, performance of neural networks can be compromised. This paper proposes a 15 Gbps high-speed BiCMOS XNOR gate to replace multipliers in BNNs. The proposed gate can be implemented on BiCMOS-based FPGAs to improve the performance. Training a neural network consumes heavy time, this solution will significantly improve the training time and response time of neural networks. These aids identifying chest abnormalities in CT scans by eliminating the power-hungry multiplication operations.

REFERENCES

[1] K. Rossa *et al.*, “Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020,” *Infectious Disease Modelling*, vol. 5, pp. 256-263, 2020, doi: 10.1016/j.idm.2020.02.002.

[2] L.Yan *et al.*, “Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan,” *medRxiv preprint*, pp. 1-18, 2020, doi: 10.1101/2020.02.27.20028027.

[3] “COVID Live Update: 199,782,673 Cases and 4,252,833 Deaths from the Coronavirus - Worldometer. (n.d.).” Worldometers. https://www.worldometers.info/coronavirus/ (accessed Jul. 1, 2021).

[4] L. Wangand and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *Scientific Reports*, vol. 10, 2020, Art. no. 19549, doi: 10.1038/s41598-020-76550-z.

[5] E. Mahase, “Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate,” *The BMJ*, vol. 368, 2020, doi: 10.1136/bmj.m641.

[6] R. Weissleder, H. Lee, J. Ko, and M. J. Pittet, "COVID-19 diagnostics in context," *Science Translational Medicine*, vol. 12, no. 546, 2020, doi: 10.1126/scitranslmed.abc1931.

[7] W. Wang *et al.*, "Detection of SARS-CoV-2 in different types of clinical specimens,” *JAMA*, vol. 323, no. 18, pp. 1843-1844, 2020, doi: 10.1001/jama.2020.3786.

[8] M. Y. Ng *et al.*, "Imaging profile of the COVID-19 infection: Radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, 2020, doi: 10.1148/ryct.2020200034.

[9] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497-506, 2020, doi: 10.1016/S0140-6736(20)30183-5.

[10] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. 32-40, 2020, doi: 10.1148/radiol.2020200642.

[11] M. J. Binnicker, "Challenges and controversies to testing for COVID-19," *Journal of Clinical Microbiology*, vol. 58, no. 11, 2020, doi: 10.1128/JCM.01695-20.

[12] C. P. West, V. M. Montori, and P. Sampathkumar, "COVID-19 testing: the threat of false-negative results," *Mayo Clinic Proceedings*, vol. 95, no. 6, pp. 1127-1129, 2020.

[13] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, vol. 24, pp. 1207-1220, 2021, doi: 10.1007/s10044-021-00984-y.

[14] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology*, vol. 296, no. 2, pp. 115-117, 2020, doi: 10.1148/radiol.2020200432.
[15] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," *Computer Vision and Pattern Recognition*, 2020.
[16] X. Xu *et al.*, "Deep learning system to screen coronavirus disease 2019 Pneumonia," *Medical Physics*, vol. 6, no. 10, pp. 1122-1129, 2020, doi: 10.1016/j.eng.2020.04.010.
[17] S. W. Bo *et al.*, "A deep learning algorithm using CT images to screen for corona virus disease (COVID-19)," *European Radiology*, vol. 31, no. 8, pp. 6096-6104, 2021, doi: 10.1101/2020.02.14.20023028.
[18] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *Medical Physics*, vol. 44, no. 10, pp. 360-375, 107, doi: 10.1002/mp.12344
[19] W. Xia, Q. Jin, C. Ni, Y. Wang, and X. Gao, "Thorax x-ray and CT interventional dataset for non-rigid 2D/3D image registration evaluation," *Medical Physics*, vol. 45, no. 11, pp. 5343-5351, 2018, doi: 10.1002/mp.13174.
[20] Z. Lin, Zhouhan, M. Courbariaux, R. Memisevic, and Y. Bengio, "Neural networks with few multiplications," 2015, *arXiv:1510.03009*.
[21] M. Courbariaux, Y. Bengio, and J. P. David, "Training deep neural networks with low precision multiplications," 2014, *arXiv:1412.7024*.
[22] Y. Umuroglu *et al.*, "FINN: a framework for fast, scalable binarized neural network inference," in *International Symposium on Field-Programmable Gate Arrays*, 2017, pp. 65-74, doi: 10.1145/3020078.3021744.
[23] E. Nurvitadhi, D. Sheffield, J. Sim, A. Mishra, G. Venkatesh and D. Marr, "Accelerating binarized neural networks: comparison of FPGA, CPU, GPU, and ASIC," *2016 International Conference on Field-Programmable Technology (FPT)*, 2016, pp. 77-84, doi: 10.1109/FPT.2016.7929192.
[24] S. Mittal, "A survey of FPGA-based accelerators for convolutional neural networks," *Neural Computing and Applications*, vol. 31, pp. 1109-1139, doi: 10.1007/s00521-018-3761-1.
[25] M. S. Riazi, M. Samragh, H. Chen, K. Laine, K. Lauter, and F. Koushanfar, "XONN: XNOR-based oblivious deep neural network inference," in *Proceedings of the 28th USENIX Conference on Security Symposium*, 2019, pp. 1501-1518.
[26] V. K. T. Schneider, "Improved garbled circuit: free XOR gates and applications," in *35th International Colloquium on Automata, Languages, and Programming*, 2008, pp. 486-498, doi: 10.1007/978-3-540-70583-3_40.
[27] C. You *et al.*, "A 5-10 GHz SiGe BiCMOS FPGA with new configurable logic block," *Microprocessors and Microsystems*, vol. 29, no. 2-3, pp. 121-131, 2005, doi: 10.1016/j.micpro.2004.06.008.
[28] K. Zhou *et al.*, "Multi-GHz SiGe BiCMOS FPGAs with new architecture and novel power management techniques," *Journal of Circuits, Systems, and Computers*, vol. 14, no. 2, pp. 179-193, 2005, doi: 10.1142/S0218126605002234.
[29] Y. Tsividis and C. Mc Andrew, "Operation and modeling of the MOS transistor," 3rd Ed., Oxford Univ. Press, 2011.

BIOGRAPHIES OF AUTHORS

Rosana W. Marar received her Master's Degree in Computer Science in 2016. Currently, she is a lecturer at the Computer Graphics and Animation Department at King Hussein School of Computing Sciences at Princess Sumaya University for Technology. Her research interests include neural networks, fuzzy logic, and virtual and augmented reality. She can be contacted at email: r.marar@psut.edu.jo.

Hazem W. Marar received his Master's Degree in Embedded Systems Design/Computer Engineering in 2012. Currently, he is a lecturer at Computer Engineering/Electrical Engineering Department at King Abdullah II School of Engineering at Princess Sumaya University for Technology. His research interests include neural networks, on-chip communication, and high-speed integrated circuits. He can be contacted at email: h.marar@psut.edu.jo.