multiplication, and then combining them. After designing the proposed variable precision multiplier with Verilog HDL, the operation verification was completed for all input data using ModelSim.

After it was synthesized for Altera Cyclone III EP3C16F484C6 using Quartus II 13.1.0 Web Edition, the area, and speed were compared with general 8-bit and 16-bit boot multiplication. The number of registers and combinational functions were four times and five times more than the 8-bit multiplier, respectively, and two times and three times more than the 16-bit multiplier, respectively. The combinational function is more than the value proportional to the number of bits because the proposed multiplier adds a logic circuit in the process of combining them after partial multiplication. The added combinational function makes the clock speed slower than other multipliers.

However, because it is based on parallel processing, four 8-bit multiplications can be processed within 1.68 times the processing time of one 8-bit multiplication, and 16-bit multiplication can be performed at 75% of the processing time of the 16-bit multiplier. Therefore, the proposed multiplier is expected to increase speed and energy efficiency by selecting bit precision according to the layer in the CNN model that requires different precision for each layer.

REFERENCES

[1] T. S. Alemayehu, and We. -D. Cho, "Distributed Edge Computing for DNA-Based Intelligent Services and Applications: A Review," *Journal of Information Processing Systems*, vol. 9. no. 12, pp. 291-306, 2020, doi: 10.3745/KTCCS.2020.9.12.291.

[2] J. H. Hong, K. C. Lee, and S. Y. Lee, "Trends in Edge Computing Technology," *Electronics and Telecommunications Trends*, vol. 35, no. 6, pp. 78-87, Dec. 2020, doi: 10.22648/ETRI.2020.J.350608.

[3] Y. B. Zikria, M. K. Afzal, S. W. Kim, A. Marin, and M. Guizani, "Deep learning for intelligent IoT: Opportunities, challenges and solutions," *Computer Communications*, vol. 164, pp. 50-53, Dec. 2020, doi: 10.1016/j.comcom.2020.08.017.

[4] H. Li, K. Ota and M. Dong, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," *IEEE Network*, vol. 32, no. 1, pp. 96-101, Jan.-Feb. 2018, doi: 10.1109/MNET.2018.1700202.

[5] T. Han, K. Muhammad, T. Hussain, J. Lloret and S. W. Baik, "An Efficient Deep Learning Framework for Intelligent Energy Management in IoT Networks," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3170-3179, 1 March1, 2021, doi: 10.1109/JIOT.2020.3013306.

[6] J. Kim, J. Jeon, M. Kee and G. H. Park, "The Method Using Reduced Classification Models for Distributed Processing of CNN Models in Multiple Edge Devices," *Journal of KIISE*, vol. 47, no. 08, pp. 787-792, Aug. 2020, doi: 10.5626/jok.2020.47.8.787.

[7] K. Cao, Y. Liu, G. Meng and Q. Sun, "An Overview on Edge Computing Research," *IEEE Access*, vol. 8, pp. 85714-85728, 2020, doi: 10.1109/ACCESS.2020.2991734.

[8] J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655-1674, Aug. 2019, doi: 10.1109/JPROC.2019.2921977.

[9] T. Sledevič, and A. Serackis, "mNet2FPGA: A Design Flow for Mapping a Fixed-Point CNN to Zynq SoC FPGA," *Electronics*, vol. 9, no. 11: 1823, 2020, doi: 10.3390/electronics9111823.

[10] M. Merenda, C. Porcaro, D. Iero, "Edge Machine Learning for AI-Enabled IoT Devices: A Review," *Sensors*, vol. 20, no. 9, 2533, 2020, doi: 10.3390/s2009253.

[11] Y. Byun, M. Ha, J. Kim, S. Lee and Y. Lee, "Low-Complexity Dynamic Channel Scaling of Noise-Resilient CNN for Intelligent Edge Devices," in *2019 DATE*, Florence, Italy, 2019, pp. 114-119, doi: 10.23919/DATE.2019.8715280.

[12] M. Shao, J. Dai, J. Kuang, and D. Meng, "A dynamic CNN pruning method based on matrix similarity," *SIViP*, vol. 15, pp. 381–389, March 2021, doi: 10.1007/s11760-020-01760-x

[13] S. K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K. R. Müller, and W. Samek, "Pruning by explaining: A novel criterion for deep neural network pruning," *Pattern Recognition*, vol. 115, July 2021, 107899, doi: 10.1016/j.patcog.2021.107899.

[14] Y. Liang, L. Lu, Y. Jin, J. Xie, R. Huang, J. Zhang, and W. Lin, "An Efficient Hardware Design for Accelerating Sparse CNNs With NAS-Based Models," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 3, pp. 597-613, March 2022, doi: 10.1109/TCAD.2021.3066563.

[15] C. Y. Lo, F. C. M. Lau and C. Sham, "Fixed-Point Implementation of Convolutional Neural Networks for Image Classification," in *2018 International Conference on Advanced Technologies for Communications*, Ho Chi Minh City, Vietnam, 2018, pp. 105-109.

[16] T. Sledevič, A. Serackis, "mNet2FPGA: A Design Flow for Mapping a Fixed-Point CNN to Zynq SoC FPGA," *Electronics*, vol. 9, no. 11, 1823, 2020, doi: 10.3390/electronics9111823.

[17] Z. Nie, Z. Li, L. Wang, S. Guo, Y. Deng, R. Deng and Q. Dou, "Laius: an energy-efficient FPGA CNN accelerator with the support of a fixed-point training framework," *International Journal of Computational Science and Engineering*, vol. 21, no. 3, pp. 418-428, 2020, doi: 10.1504/IJCSE.2020.106064.

[18] D. Lin, S. Talathi, and S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, 2016, pp. 2849–2858.

[19] Z. Bao, G. Fu, W. Zhang, K. Zhan and J. Guo, "LSFQ: A Low-Bit Full Integer Quantization for High-Performance FPGA-Based CNN Acceleration," *IEEE Micro*, vol. 42, no. 2, pp. 8-15, 1 March-April 2022, doi: 10.1109/MM.2021.3134968.

[20] M. Sailesh, K. Selvakumar, P. Narayanan, "A novel framework for deployment of CNN models using post-training quantization on microcontroller," *Microprocessors and Microsystems*, vol. 94, 104634, 2022, doi: 10.1016/j.micpro.2022.104634.

[21] I. Hubara, M. Courbariaux, D. Soudry, R. EI-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6869–6898, Jan. 2017. [Online]. Available: https://dl.acm.org/citation.cfm?id=3122009.3242044.

[22] L. Cavigelli and L. Benini, "Origami: A 803-GOp/s/W convolutional network accelerator," *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, vo. 27, no. 11, pp. 2461–2475, Nov. 2017, doi: 10.1109/TCSVT.2016.2592330.

[23] Y. Park, Y. Kang, S. Kim, E. Kwon, and S. Kang, "GRLC: Grid-based Run-length Compression for Energy-efficient CNN Accelerator", in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, Boston, MA, USA, Aug. 2020, pp. 91-96, doi: 10.1145/3370748.3406576.

[24] D. Kim, E. Park, and S. Yoo, "Energy-efficient neural network accelerator based on outlier-aware low-precision computation," In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, Los Angeles, CA, USA, July 2018, pp. 688–698, doi: 10.1109/ISCA.2018.00063.

[25] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU:An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019. doi: 10.1109/JSSC.2018.2865489.

[26] J. Wang, S. Fang, X. Wang, J. Ma, T. Wang and Y. Shan, "High-Performance Mixed-Low-Precision CNN Inference Accelerator on FPGA," *IEEE Micro*, vol. 41, no. 4, pp. 31-38, 1 July-Aug, 2021, doi: 10.1109/MM.2021.3081735.

[27] S. -N. Tang, "Area-Efficient Parallel Multiplication Units for CNN Accelerators With Output Channel Parallelization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 3, pp. 406-410, March 2023, doi: 10.1109/TVLSI.2023.3235776.

[28] W. Liu, J. Lin, and Z. Wang, "A Precision-Scalable Energy-Efficient Convolutional Neural Network Accelerator," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 10, pp.3484-3497, Oct. 2020, doi: 10.1109/TCSI.2020.2993051.

[29] V. Camus, L. Mei, C. Enz, and M. Verhelst, "Review and Benchmarking of Precision-Scalable Multiply-Accumulate Unit Architectures for Embedded Neural-Network Processing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 4, pp.697-711, Dec. 2019, doi: 10.1109/JETCAS.2019.2950386.

[30] A. D. Booth, "A Signed Binary Multiplication Technique," Quarterly *Journal of Mechanics and Applied Mathematics*, vol. 4, pp. 236-240, 1951.