UNIVERSITAS GADJAH MADA
AN EMPIRICAL STUDY OF THE IMPACT OF A SINGLE BIT-FLIP ERROR ON LOW PRECISION QUANTIZED CONVOLUTIONAL NEURAL NETWORKS
GABRIEL KAUNANG, Muhammad Alfian Amrizal, B.Eng., M.I.S., Ph.D.
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

REFERENCES

Banner, R., Nahshan, Y., & Soudry, D. (2019). Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems*, *32*.

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems*, *2015-January*, 1135–1143.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*. https://doi.org/10.1109/CVPR.2016.90

He, Y., Balaprakash, P., & Li, Y. (2020). Fidelity: Efficient resilience analysis framework for deep learning accelerators. *Proceedings of the Annual International Symposium on Microarchitecture, MICRO*, *2020-October*. https://doi.org/10.1109/MICRO50266.2020.00033

Hong, T., Li, Y., Park, S. B., Mui, D., Lin, D., Kaleq, Z. A., Hakim, N., Naeimi, H., Gardner, D. S., & Mitra, S. (2010). QED: Quick error detection tests for effective post-silicon validation. *Proceedings - International Test Conference*. https://doi.org/10.1109/TEST.2010.5699215

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets. *ArXiv Preprint ArXiv:1704.04861*.

Idrissi, I., Azizi, M., & Moussaoui, O. (2022). A Lightweight Optimized Deep Learning-based Host-Intrusion Detection System Deployed on the Edge for IoT. *International Journal of Computing and Digital Systems*, *11*(1). https://doi.org/10.12785/ijcds/110117

UNIVERSITAS GADJAH MADA
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2018.00286

Khasoggi, B., Ermatita, & Samsuryadi. (2019). Efficient mobilenet architecture as image recognition on mobile and embedded devices. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(1). https://doi.org/10.11591/ijeecs.v16.i1.pp389-394

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11). https://doi.org/10.1109/5.726791

Leveugle, R., Calvez, A., Maistri, P., & Vanhauwaert, P. (2009). Statistical fault injection: Quantified error and confidence. *Proceedings -Design, Automation and Test in Europe, DATE*. https://doi.org/10.1109/date.2009.5090716

Li, G., Hari, S. K. S., Sullivan, M., Tsai, T., Pattabiraman, K., Emer, J., & Keckler, S. W. (2017). Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, *2017-November*. https://doi.org/10.1145/3126908.3126964

Litvinenko, A., Kucherov, D., & Glybovets, M. (2022). Decomposition Method for Calculating the Weights of a Binary Neural Network. *Cybernetics and Systems Analysis*, *58*(6). https://doi.org/10.1007/s10559-023-00522-0

Mirkhani, S., Mitra, S., Cher, C. Y., & Abraham, J. (2015). Efficient soft error vulnerability estimation of complex designs. *Proceedings -Design, Automation and Test in Europe, DATE*, *2015-April*. https://doi.org/10.7873/date.2015.0367

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van, M., Qualcomm, B., Research, A. I., & Blankevoort, T. (2021). *A White Paper on Neural Network Quantization*.

Nguyen, H. T., Yagil, Y., Seifert, N., & Reitsma, M. (2005). Chip-level soft error estimation method. In *IEEE Transactions on Device and Materials Reliability* (Vol. 5, Issue 3). https://doi.org/10.1109/TDMR.2005.858334

Qian, C., Zhang, M., Nie, Y., Lu, S., & Cao, H. (2023). A Survey of Bit-Flip Attacks on Deep Neural Network and Corresponding Defense Methods. In *Electronics (Switzerland)* (Vol. 12, Issue 4). https://doi.org/10.3390/electronics12040853

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Solovyev, R. A., Kalinin, A. A., Kustov, A. G., Telpukhov, D. V., & Ruhlov, V. S. (2018). FPGA implementation of convolutional neural networks with fixed-point calculations. *ArXiv*.

Sun, X., Wang, N., Chen, C. Y., Ni, J. M., Agrawal, A., Cui, X., Venkataramani, S., El Maghraoui, K., Srinivasan, V., & Gopalakrishnan, K. (2020). Ultra-low precision 4-bit training of deep neural networks. *Advances in Neural Information Processing Systems*, *2020-December*.

UNIVERSITAS GADJAH MADA
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

Véstias, M. P. (2019). A survey of convolutional neural networks on edge with reconfigurable computing. In *Algorithms* (Vol. 12, Issue 8). https://doi.org/10.3390/a12080154

Yadav, Prof. O., Sharma, A., Philip, D., & Alexander, B. (2021). A Self Driving Car using Machine Learning and IOT. *International Journal Of Trendy Research In Engineering And Technology*, *05*(04). https://doi.org/10.54473/ijtret.2021.5401

Yin, P., Lyu, J., Zhang, S., Osher, S., Qi, Y., & Xin, J. (2019). Understanding straight-through estimator in training activation quantized neural nets. *7th International Conference on Learning Representations, ICLR 2019*.

Zaniolo, L., & Marques, O. (2020). On the use of variable stride in convolutional neural networks. *Multimedia Tools and Applications*, *79*(19–20). https://doi.org/10.1007/s11042-019-08385-4