DAFTAR PUSTAKA

Agarwal, V. (2015). Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis. *International Journal of Computer Applications*, *131*(4), 30–36. https://doi.org/10.5120/ijca2015907309

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., Devito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., … Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS*, *2*, 929–947. https://doi.org/10.1145/3620665.3640366

Baghbanbashi, M., Raji, M., & Ghavami, B. (2023). Quantizing YOLOv7: A Comprehensive Study. *2023 28th International Computer Conference, Computer Society of Iran, CSICC 2023*. https://doi.org/10.1109/CSICC58665.2023.10105310

Balakrishnan, B., Chelliah, R., Venkatesan, M., & Sah, C. (2022). Comparative Study On Various Architectures Of Yolo Models Used In Object Recognition. *3rd IEEE 2022 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2022*, 685–690. https://doi.org/10.1109/ICCCIS56430.2022.10037635

Bondarenko, Y., Nagel, M., & Blankevoort, T. (2021). *Understanding and Overcoming the Challenges of Efficient Transformer Quantization*. http://arxiv.org/abs/2109.12948

Chen, L., & Lou, P. (2022). Clipping-Based Post Training 8-Bit Quantization of Convolution Neural Networks for Object Detection. *Applied Sciences (Switzerland)*, *12*(23). https://doi.org/10.3390/app122312405

Endrawati, D. N., Ibad, S. I., Syafalni, I., Sutisna, N., Mulyawan, R., & Adiono, T. (2021). YOLOv3-Tiny's Weight Size Reduction using Pruning and Quantization. *Proceeding of*

15th International Conference on Telecommunication Systems, Services, and Applications (TSSA) : 18-19 November 2021, Bali, Indonesia.

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2021). *A Survey of Quantization Methods for Efficient Neural Network Inference*. http://arxiv.org/abs/2103.13630

Gupta, V., Mishra, V. K., Singhal, P., & Kumar, A. (2022). An Overview of Supervised Machine Learning Algorithm. *Proceedings of the 2022 11th International Conference on System Modeling and Advancement in Research Trends, SMART 2022*, 87–92. https://doi.org/10.1109/SMART55829.2022.10047618

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

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*, 270–281. https://doi.org/10.1109/MICRO50266.2020.00033

Horowitz, M. (2014). 1.1 Computing's Energy Problem (and what we can do about it). *2014 IEEE International Solid-State Circuits Conference*, 10–14.

Huang, X., Shen, Z., & Cheng, K.-T. (2023). *Variation-aware Vision Transformer Quantization*. http://arxiv.org/abs/2307.00331

Jocher, G. (2020, May 29). *YOLOv5 by Ultralytics*. Https://Github.Com/Ultralytics/Yolov5.

Jocher, G., Chaurasia, A., & Qiu, J. (2023, January 10). *Ultralytics YOLO*. Https://Ultralytics.Com.

Kusumawati, R. (2008). KECERDASAN BUATAN MANUSIA (ARTIFICIAL INTELLIGENCE): TEKNOLOGI IMPIAN MASA DEPAN. In *Ulul Albab* (Vol. 9, Issue 2).

Li, Y., Xu, S., Zhang, B., Cao, X., Gao, P., & Guo, G. (2022). *Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer*. https://github.com/YanjingLi0202/Q-ViT.

Li, Z., Wang, P., Wang, Z., & Cheng, J. (2021). Fixed-point Quantization for Vision Transformer. *Proceeding - 2021 China Automation Congress, CAC 2021*, 7282–7287. https://doi.org/10.1109/CAC53003.2021.9728246

Liqun, C., & Lei, H. (2023). Clipping-based Neural Network Post Training Quantization for Object Detection. *2023 IEEE International Conference on Control, Electronics and Computer Technology, ICCECT 2023*, 1192–1196. https://doi.org/10.1109/ICCECT57938.2023.10141287

Liu, Z., Wang, Y., Han, K., Ma, S., & Gao, W. (2021). *Post-Training Quantization for Vision Transformer*. http://arxiv.org/abs/2106.14156

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., van Baalen, M., & Blankevoort, T. (2021). *A White Paper on Neural Network Quantization*. http://arxiv.org/abs/2106.08295

Shi, L., Huang, H., Song, B., Tan, M., Zhao, W., Xia, T., & Ren, P. (2023). TAQ: Top-K Attention-Aware Quantization for Vision Transformers. *2023 IEEE International Conference on Image Processing (ICIP)*, 1750–1754. https://doi.org/10.1109/icip49359.2023.10222721

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research*. http://arxiv.org/abs/2106.10270

Vara Prasad, P. (2024). A Comprehensive Review on Neural Networks. *International Research Journal of Engineering and Technology*. www.irjet.net

Yuan, Z., Xue, C., Chen, Y., Wu, Q., & Sun, G. (2021). *PTQ4ViT: Post-Training Quantization Framework for Vision Transformers with Twin Uniform Quantization*. http://arxiv.org/abs/2111.12293