Paper Review Assignment 2
Quantization

Chia-Chi Tsai (蔡家齊)
cctsai@gs.ncku.edu.tw
AI System Lab
Department of Electrical Engineering
National Cheng Kung University

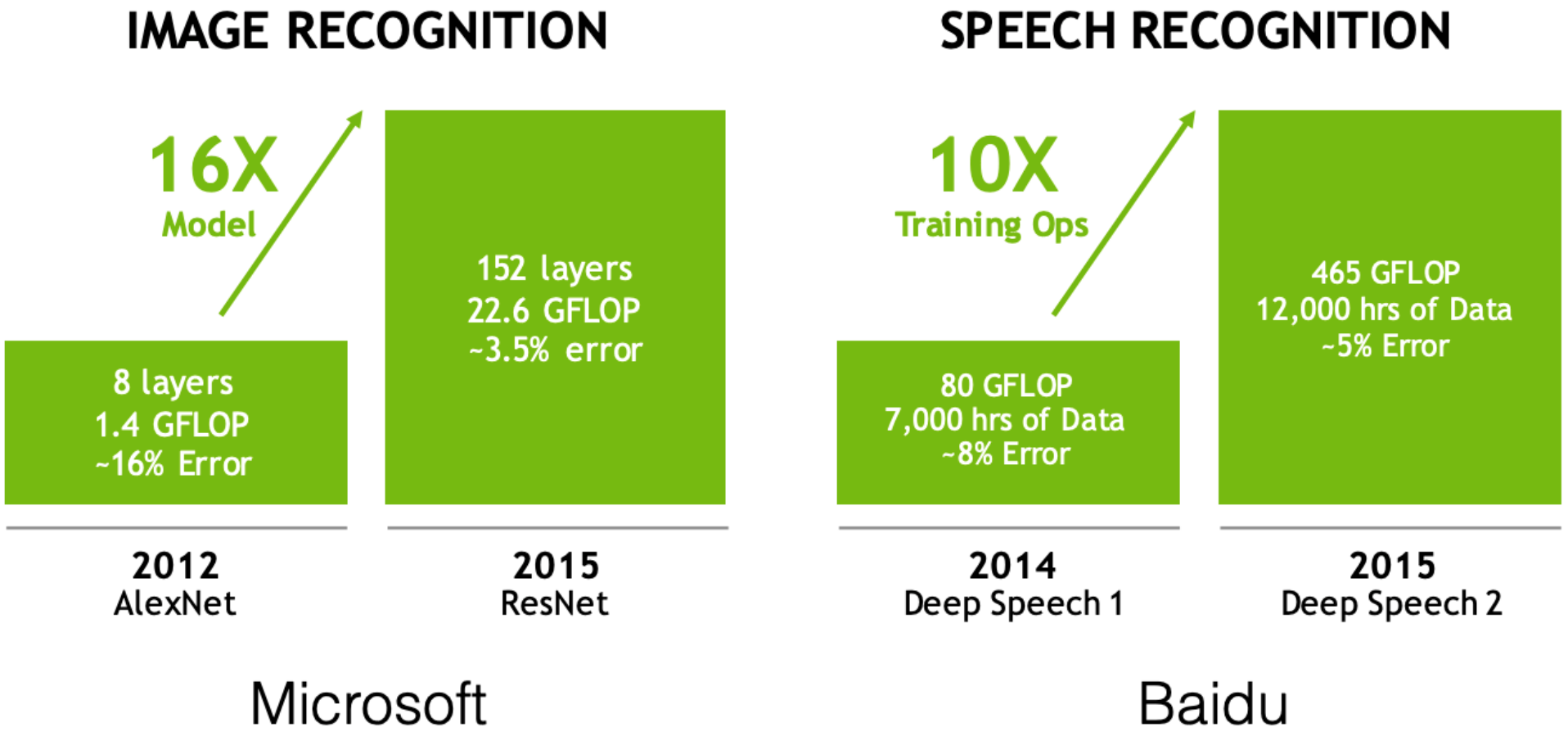
Outline

- Short Introduction to Quantization
- Paper Review Assignment 2 - Quantization

Outline

- Short Introduction to Quantization
- Paper Review Assignment 2 - Quantization

Models Are Getting Larger



Approaches

- Reduce size of operands for storage/compute
  - Floating point → Fixed point
  - Bit-width reduction
  - Non-linear quantization
- Reduce number of operations for storage/compute
  - Exploit Activation Statistics (Compression)
  - Network Pruning
  - Compact Network Architectures

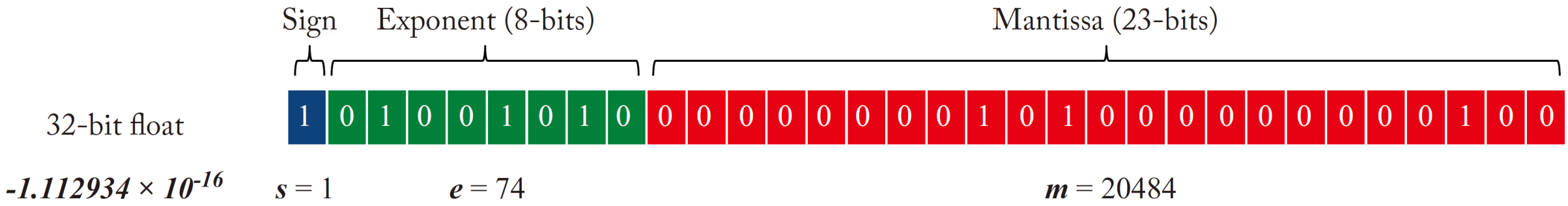
Taxonomy

- Precision refers to the number of levels
  - Number of bits $= \log_2(number\ of\ levels)$
- **Quantization:** mapping data to a smaller set of **levels**
  - Linear
    - e.g., fixed-point
  - Non-linear
    - Computed (e.g., floating point, log-domain)
    - Table lookup (e.g., learned)

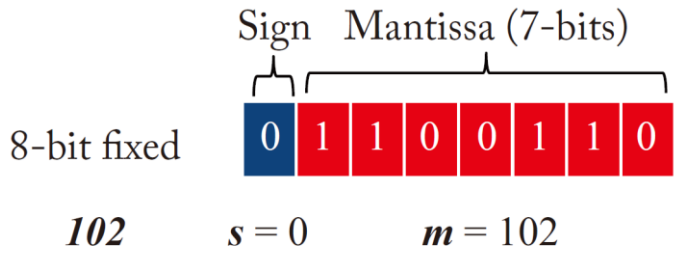
Objective: Reduce size to improve speed and/or reduce energy while preserving accuracy

Floating-Point vs. Fixed-Point

32-bit floating point



8-bit fixed point



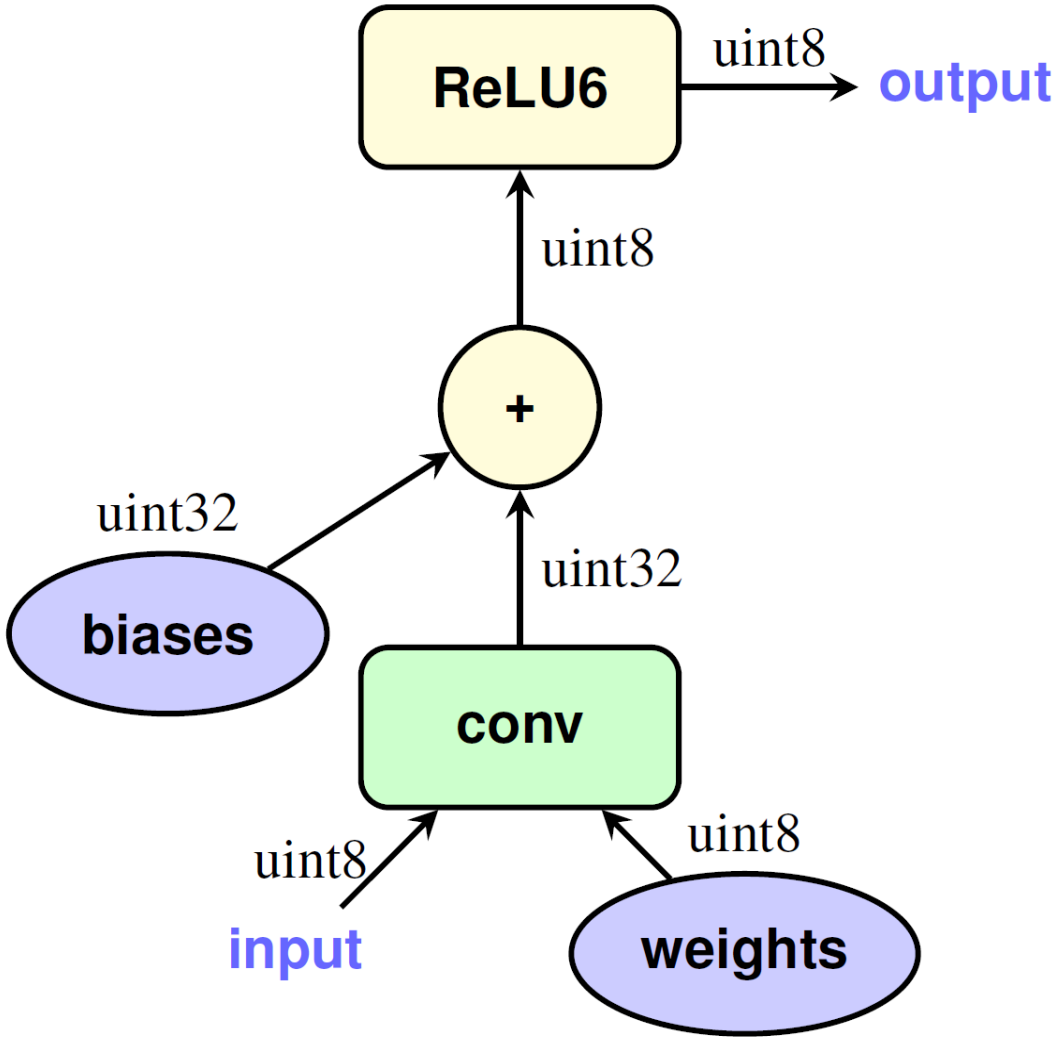
INT8 Inference

- NT8 has significantly lower precision and dynamic range compared to FP32

	Dynamic Range	Min Positive Value
FP32	$-3.4 \times 10^{38} \sim +3.4 \times 10^{38}$	1.4×10^{-45}
FP16	-65504 ~ +65504	5.96×10^{-8}
INT8	-128 ~ +127	1

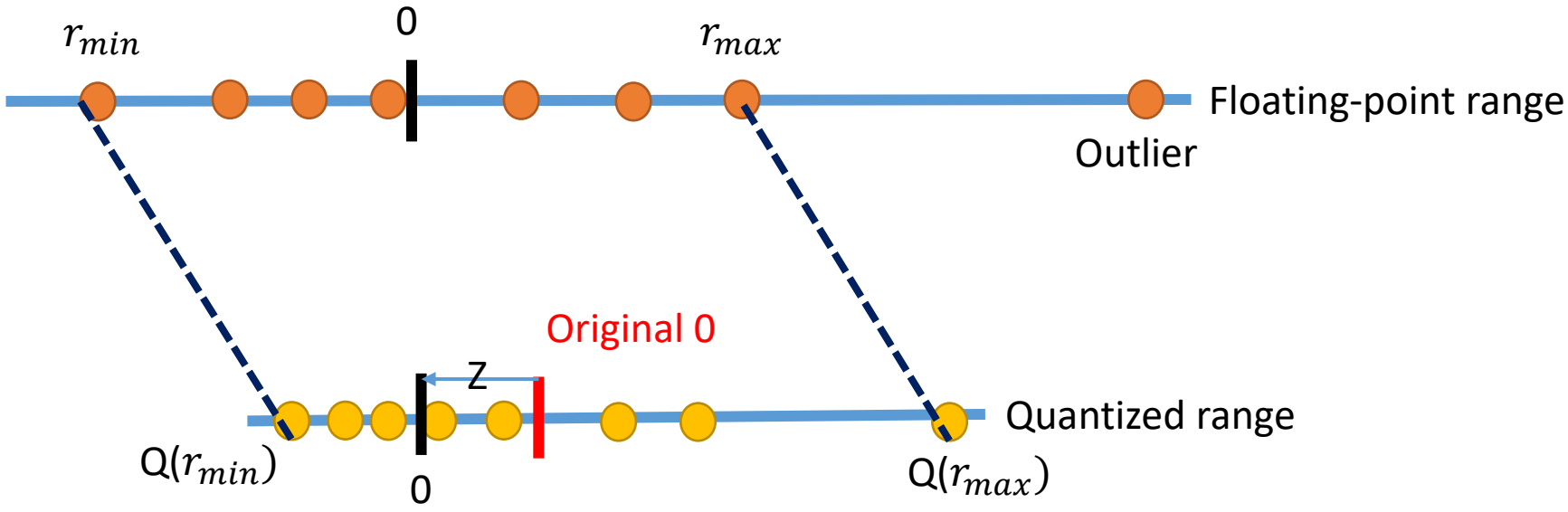
- Requires more than a simple type conversion from FP32 to INT8.

INT8 Inference Scheme

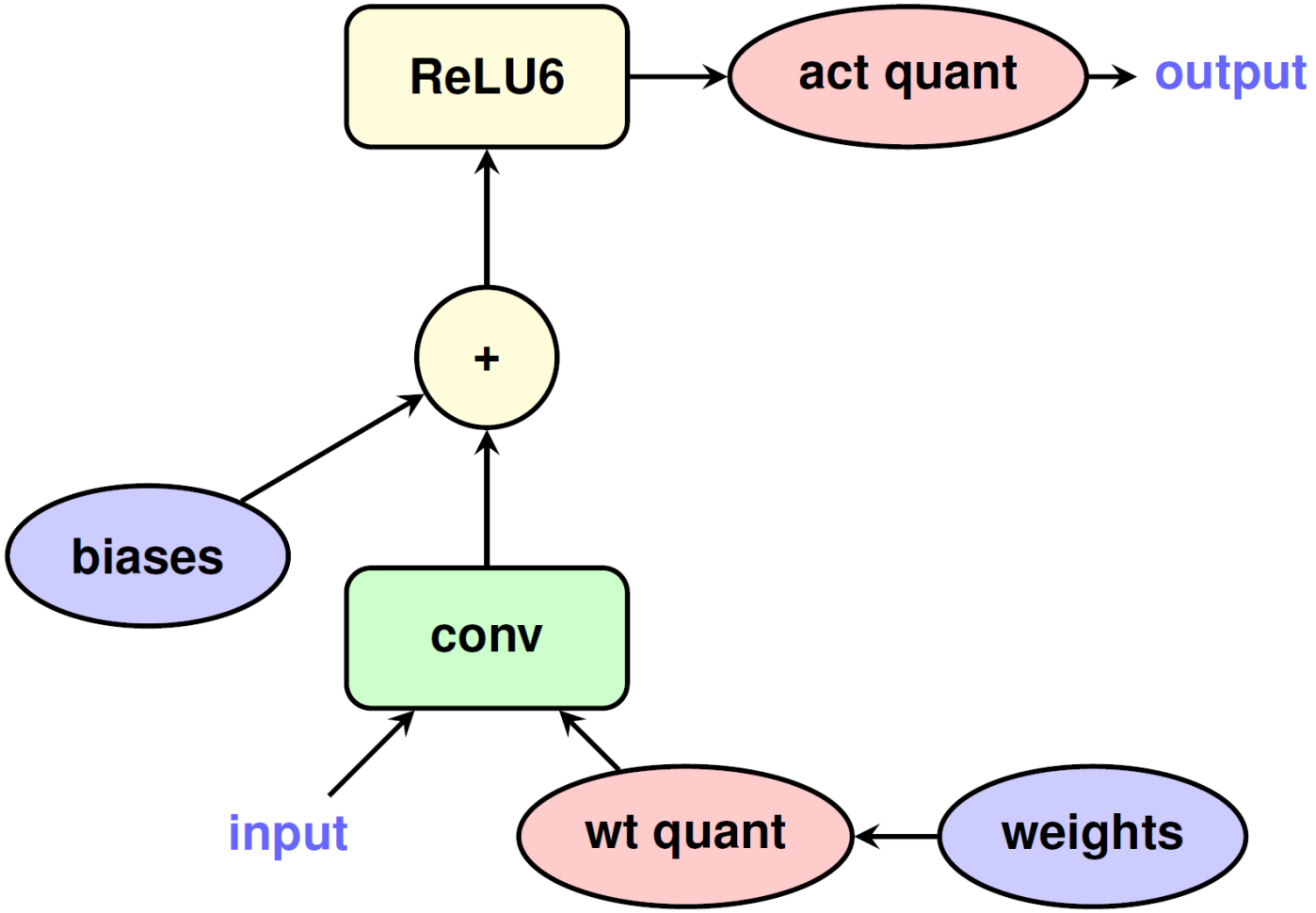


INT8 Quantization Scheme

- $r = S(q - Z)$
  - r: real number value
  - q: quantized number value
  - Z: zero-point shifting
  - S: scale factor



INT8 Quantization-Aware Training Scheme



More about Quantization

Will discuss in later courses…

Outline

- Short Introduction to Quantization
- Paper Review Assignment 2 - Quantization

Paper Readings and Review

- Paper related to AI Models Quantization
  - To learn why and what is Quantization in DNNs
  - To understand state-of-the-art Quantization method
- **Due**
  - **4/8 23:59**
- Requirement
  - Choose **at least one or more** papers
    - From recommended paper list
    - **Or any other paper as long as it related to the topics**
  - Summarize and write paper review in word/latex format
    - **LaTeX format is highly recommended**
  - Hand in **compiled pdf files** on moodle

Paper Readings and Review

- Reading reviews are free of format
- But the following review questions guide you through the paper reading process.
  - What are the **motivations** for this work?
  - What is the **proposed solution**?
  - What is the work's **evaluation** of the proposed solution?
  - What is your **analysis** of the identified problem, idea, and evaluation?
  - What are **future directions** for this research?
  - What **questions** are you left with?

Recommended Paper List

- Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference
  - Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2704-2713).
- Quantization Networks
  - Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., ... & Hua, X. S. (2019). Quantization networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7308-7316).
- ADMM
  - Leng, C., Dou, Z., Li, H., Zhu, S., & Jin, R. (2018, April). Extremely low bit neural network: Squeeze the last bit out with admm. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Ternary Weight Networks
  - Li, F., Zhang, B., & Liu, B. (2016). Ternary weight networks. *arXiv preprint arXiv:1605.04711*.
- Trained ternary quantization
  - Zhu, C., Han, S., Mao, H., & Dally, W. J. (2016). Trained ternary quantization. *arXiv preprint arXiv:1612.01064*.

Recommended Paper List

- Discovering low-precision networks close to full-precision networks for efficient embedded inference
  - McKinstry, J. L., Esser, S. K., Appuswamy, R., Bablani, D., Arthur, J. V., Yildiz, I. B., & Modha, D. S. (2019, December). Discovering low-precision networks close to full-precision networks for efficient inference. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)* (pp. 6-9). IEEE.
- Two-step quantization for low-bit neural networks
  - Wang, P., Hu, Q., Zhang, Y., Zhang, C., Liu, Y., & Cheng, J. (2018). Two-step quantization for low-bit neural networks. In *Proceedings of the IEEE Conference on computer vision and pattern recognition* (pp. 4376-4384).
- Training deep neural networks with low precision multiplications
  - Courbariaux, M., Bengio, Y., & David, J. P. (2014). Training deep neural networks with low precision multiplications. arXiv preprint arXiv:1412.7024.
- Estimating or propagating gradients through stochastic neurons for conditional computation
  - Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432.

Recommended Paper List

- Binarized Neural Networks
  - Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks. Advances in neural information processing systems, 29.
- LogNet
  - Lee, E. H., Miyashita, D., Chai, E., Murmann, B., & Wong, S. S. (2017, March). Lognet: Energy-efficient neural networks using logarithmic computation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5900-5904). IEEE.
- XNOR-Net
  - Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016, October). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision* (pp. 525-542). Springer, Cham.
- Binary Connect
  - Courbariaux, M., Bengio, Y., & David, J. P. (2015). Binaryconnect: Training deep neural networks with binary weights during propagations. Advances in neural information processing systems, 28.
- Fully quantized network for object detection
  - Li, R., Wang, Y., Liang, F., Qin, H., Yan, J., & Fan, R. (2019). Fully quantized network for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2810-2819).