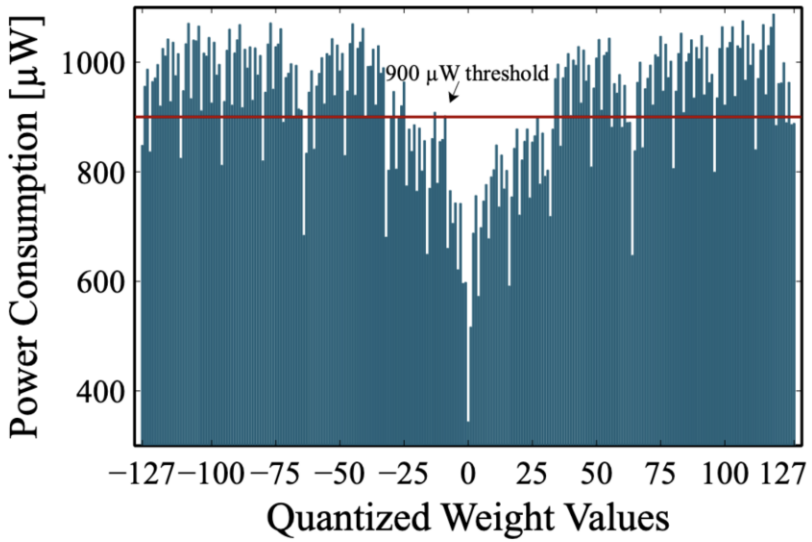
WEIGHT SELECTION FOR CNN INFERENCE ENERGY REDUCTION

MOTIVATION

- Neural network **inference is power intensive**
- Edge devices are power constrained
- Not always possible to deploy new hardware to save energy
- Reduction of energy via software

- Idea: **Identify and fine-tune model to prefer low-power weights during MAC**



APPROACH (CHOOSE ONE)

Reduce Switching Activity

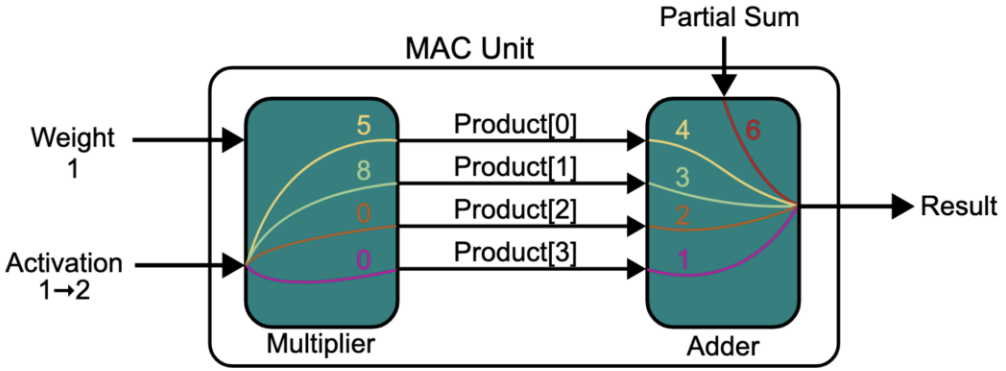
- Different weights exhibit different average switching
- Find low-power weights
  - Switching activity depends on MAC input transitions
  - Characterize and sample from input transition distribution
  - Simulate switching activity to compute average power
- Prune model to use low-power weights
- Fine-tune to regain accuracy
- Calculate possible power saving

Reduce Operating Voltage

- Propagation delay of MAC unit differs by weight value
  - Excluding worst-case weights can decrease propagation delay
  - Allows voltage reduction without timing errors
- Strategy: Analyze worst-case delay of weights and select subset
- Prune model to use “fast” weights
- Fine-tune to regain accuracy
- Calculate possible voltage reduction and power saving

TASKS

- **Develop** either of the two energy reduction techniques
- **Evaluate** them on some CNNs
- Provide clean **code** and write a short **report**



Reference paper:

Petri, R.; Zhang, G. L.; Chen, Y.; Schlichtmann, U.; Li, B. PowerPruning: Selecting Weights and Activations for Power-Efficient Neural Network Acceleration.

https://doi.org/10.48550/arXiv.2303.13997.