HAILO
Empowering Intelligence

Same, Same But Different

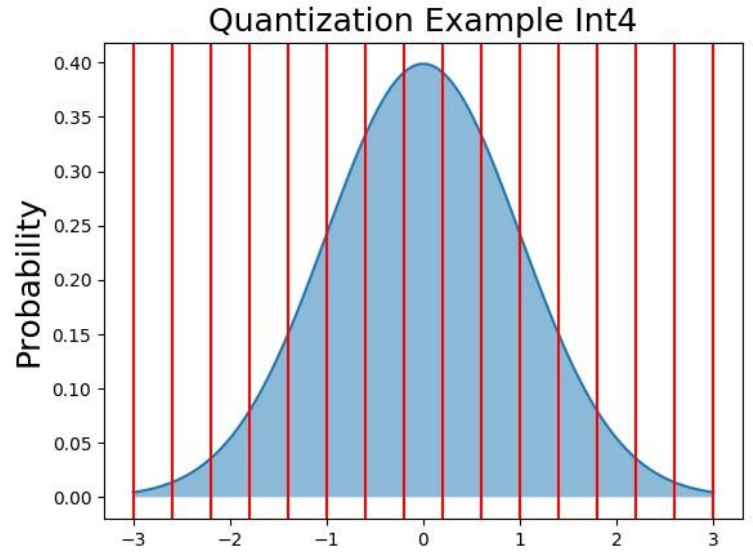
Recovering Neural Network Quantization Error Through Weight Factorization

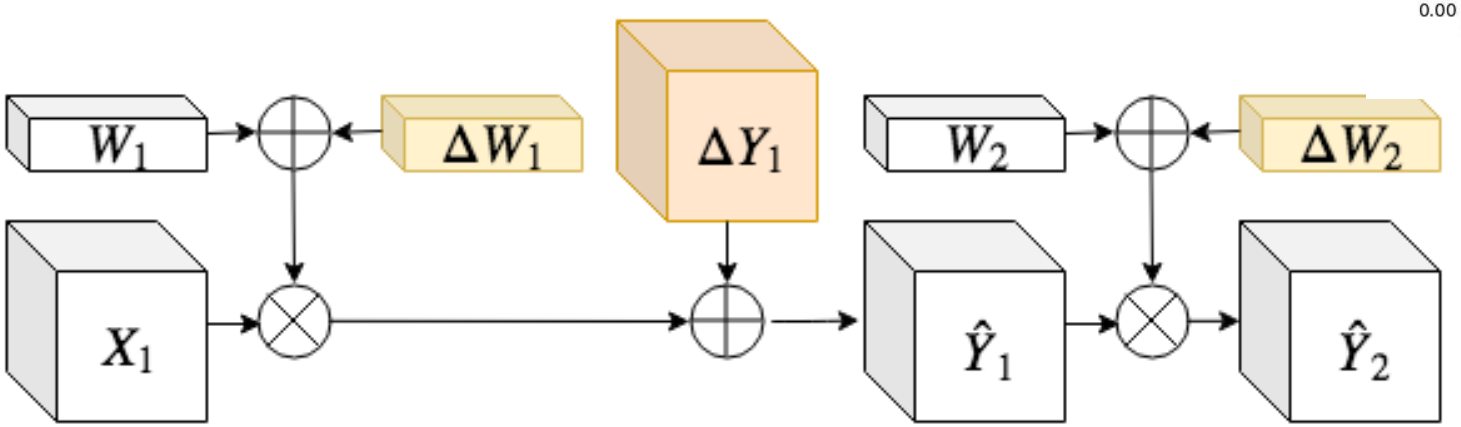
Eldad Meller

ICML 2019

Neural Network Quantization

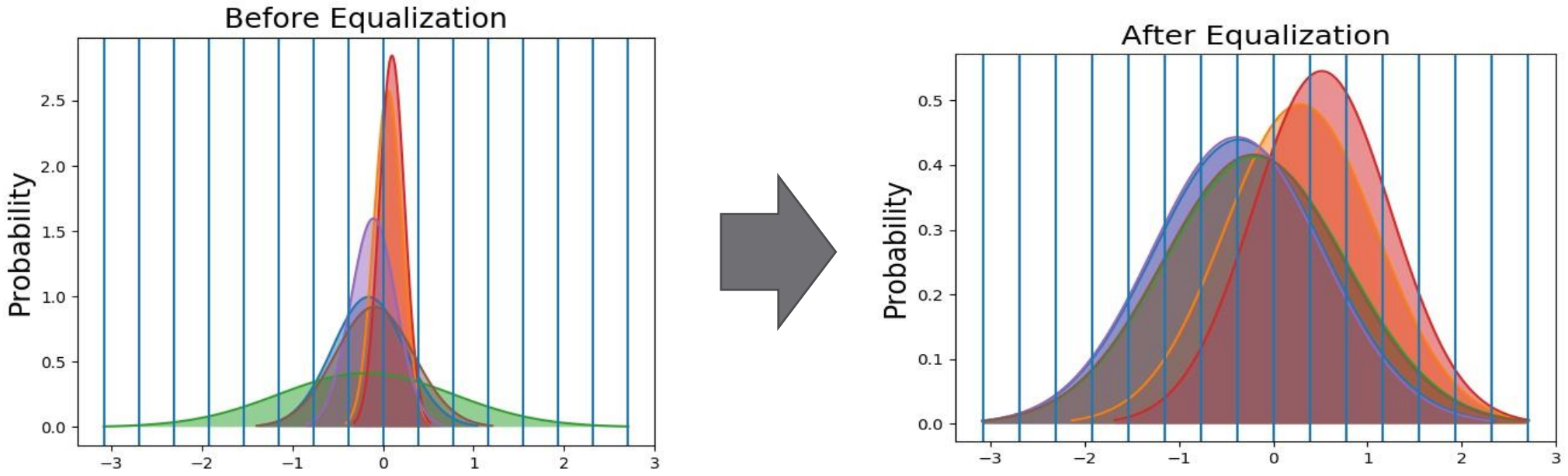
- **Quantization** of Neural Networks is needed for efficient inference
- Quantization **adds noise to the network** and **degrades its performance**





Quantization Dynamic Range

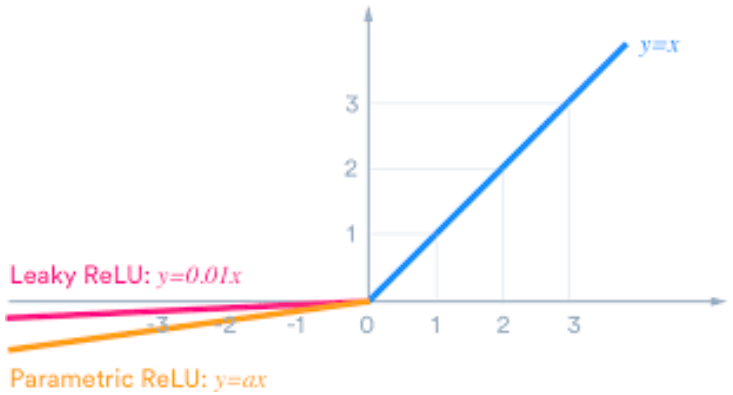
- The most common quantization setting is **layer-wise** quantization where all the channels in a layer are quantized using the same dynamic range
- **Equalizing** the dynamic range of all the **channels** in a layer by amplifying channels with small dynamic range **will reduce overall quantization noise**



A simple trick to amplify channels

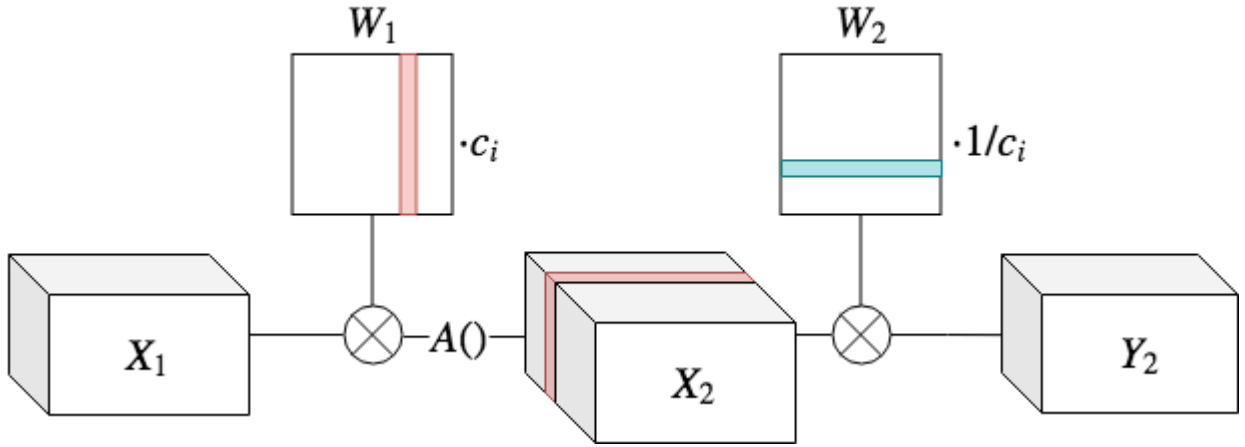
- For any homogeneous activation functions

$$A(\alpha \cdot x) = \alpha \cdot A(x) \quad \forall \alpha > 0$$

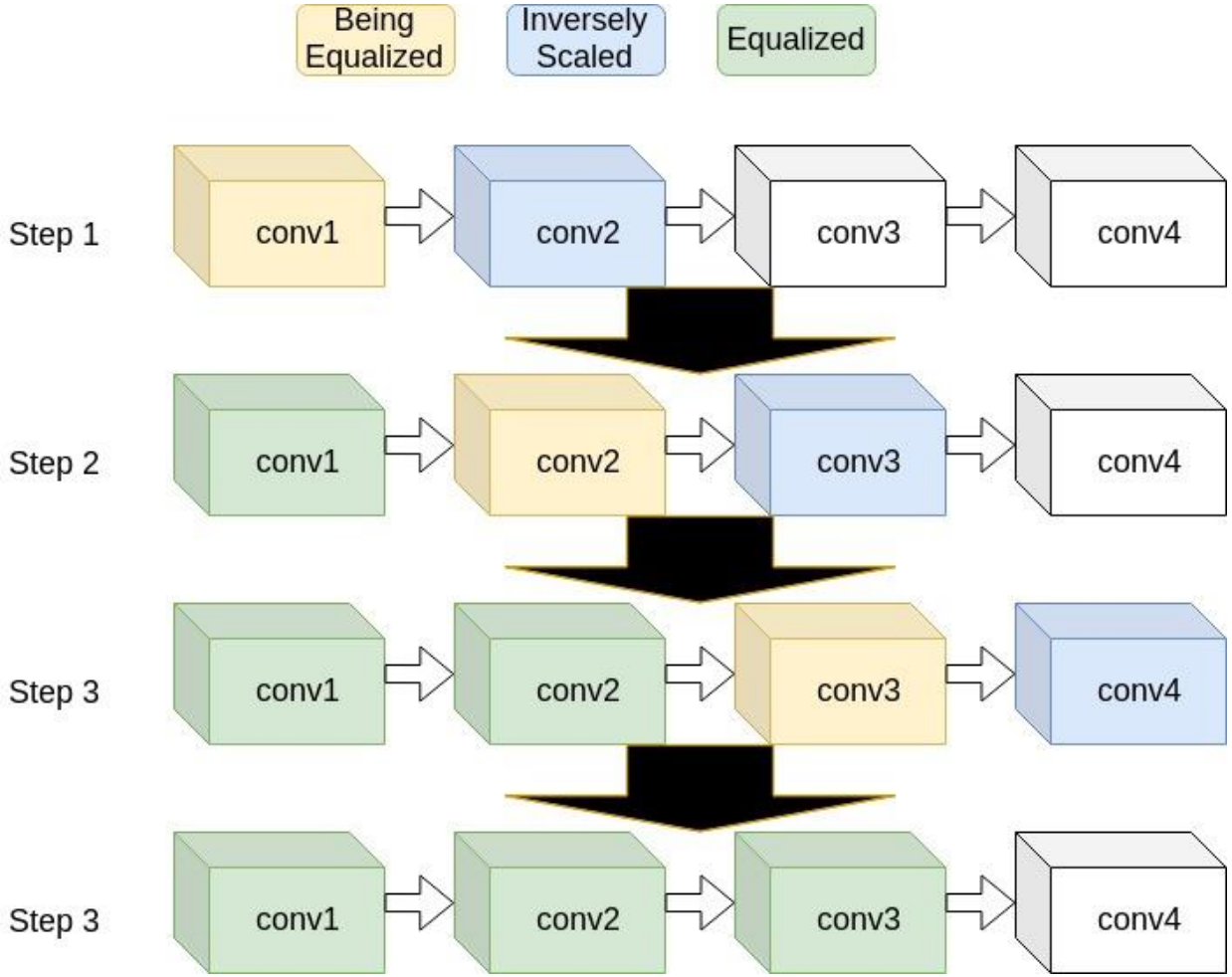


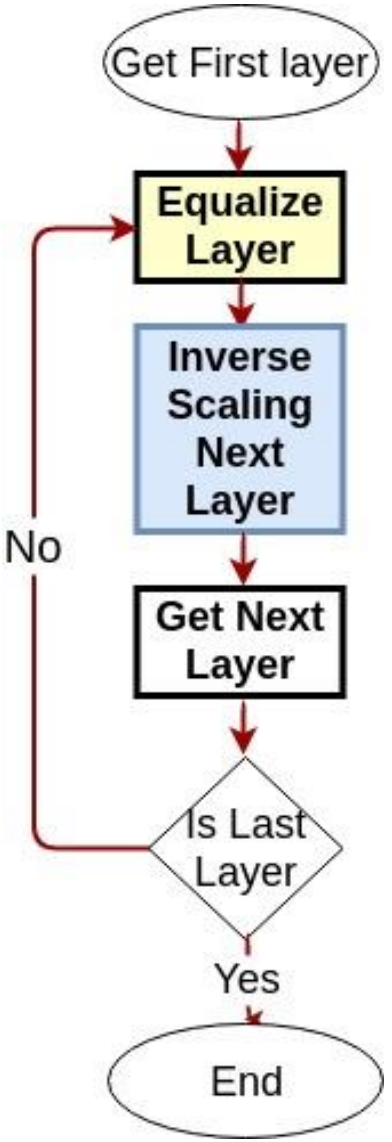
- Any channel in the network can be scaled by any positive scalar if the weights in the consecutive layer are properly inversely scaled

- **The network's output remains unchanged**

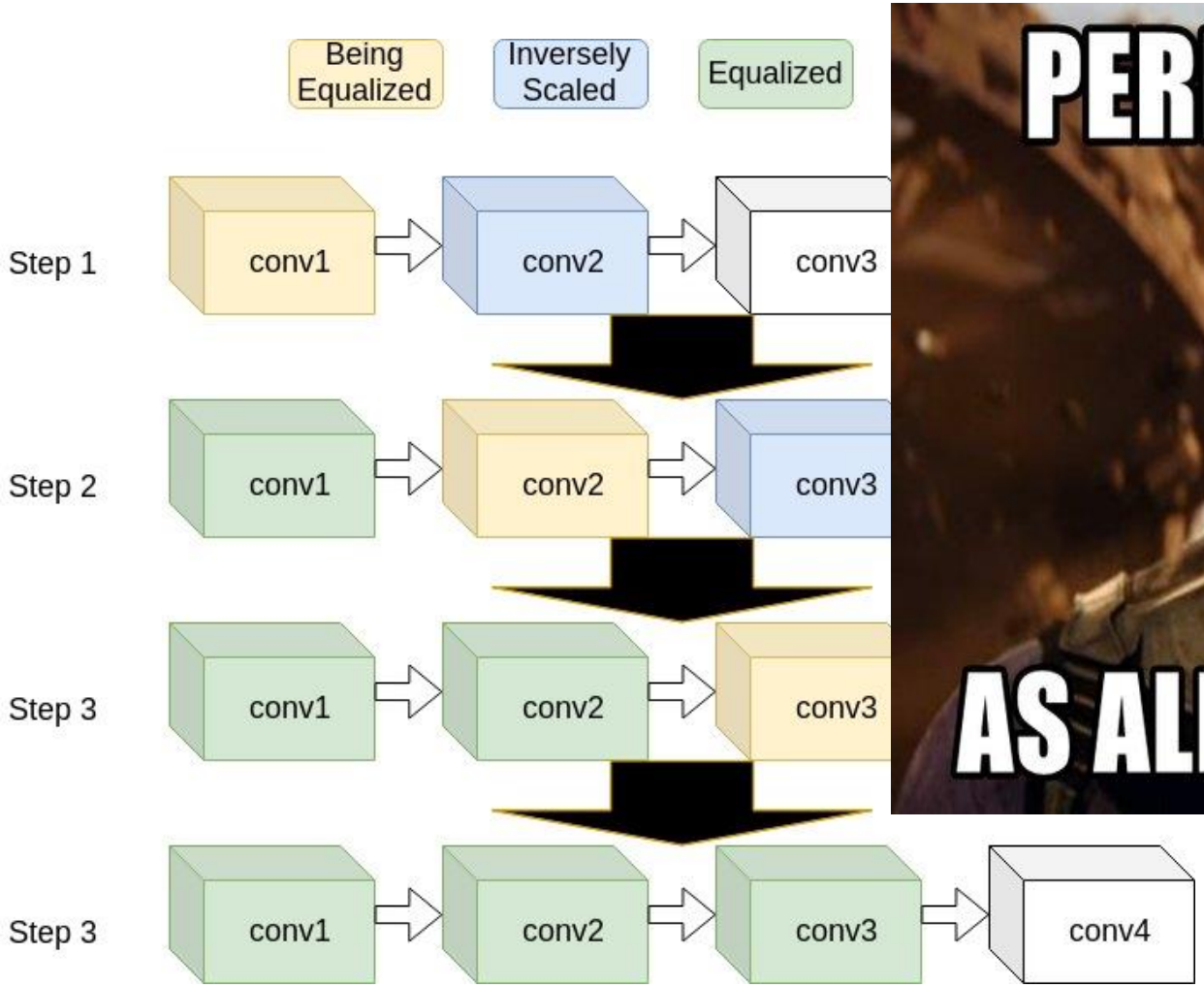


Network Equalization

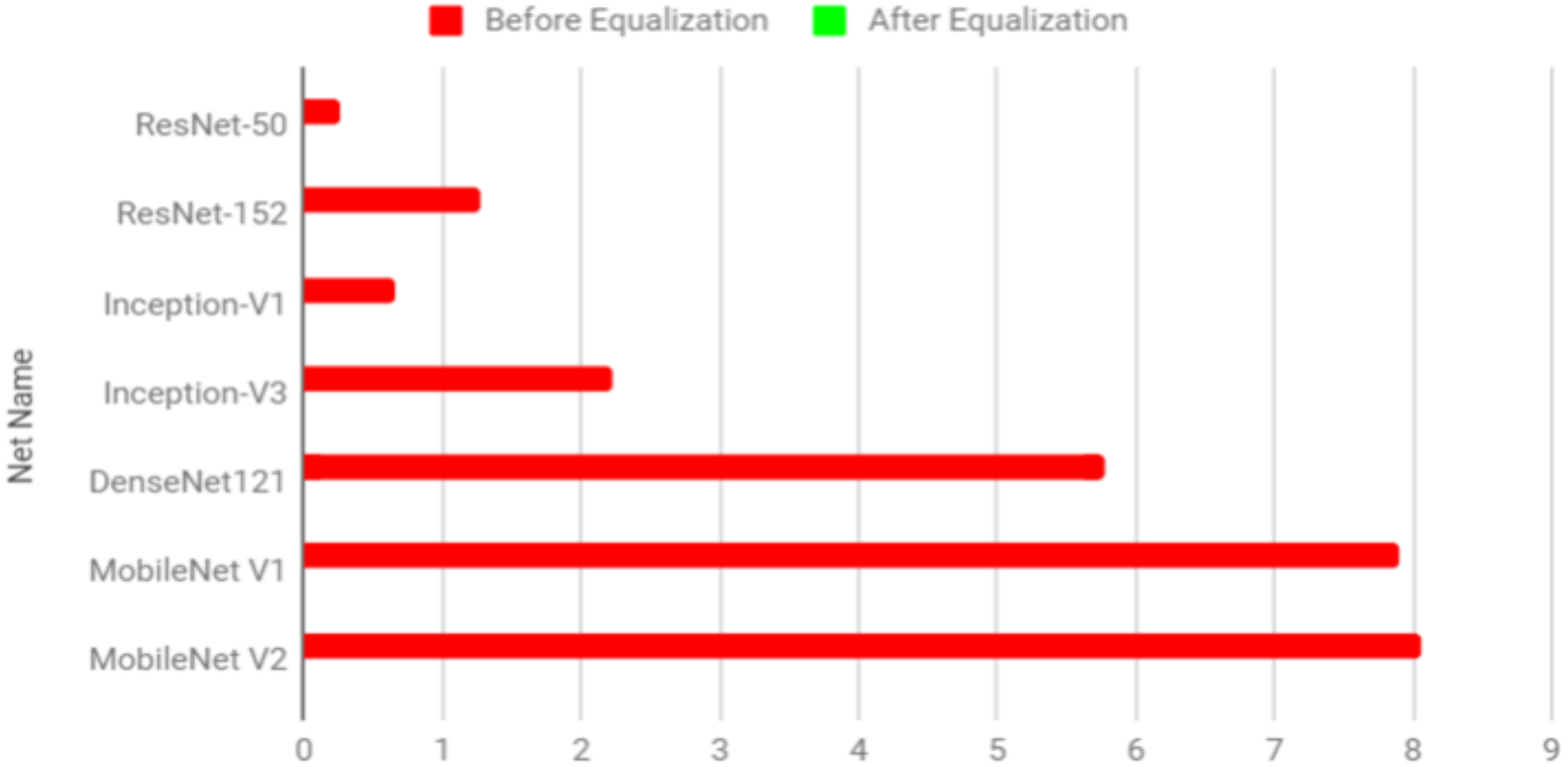




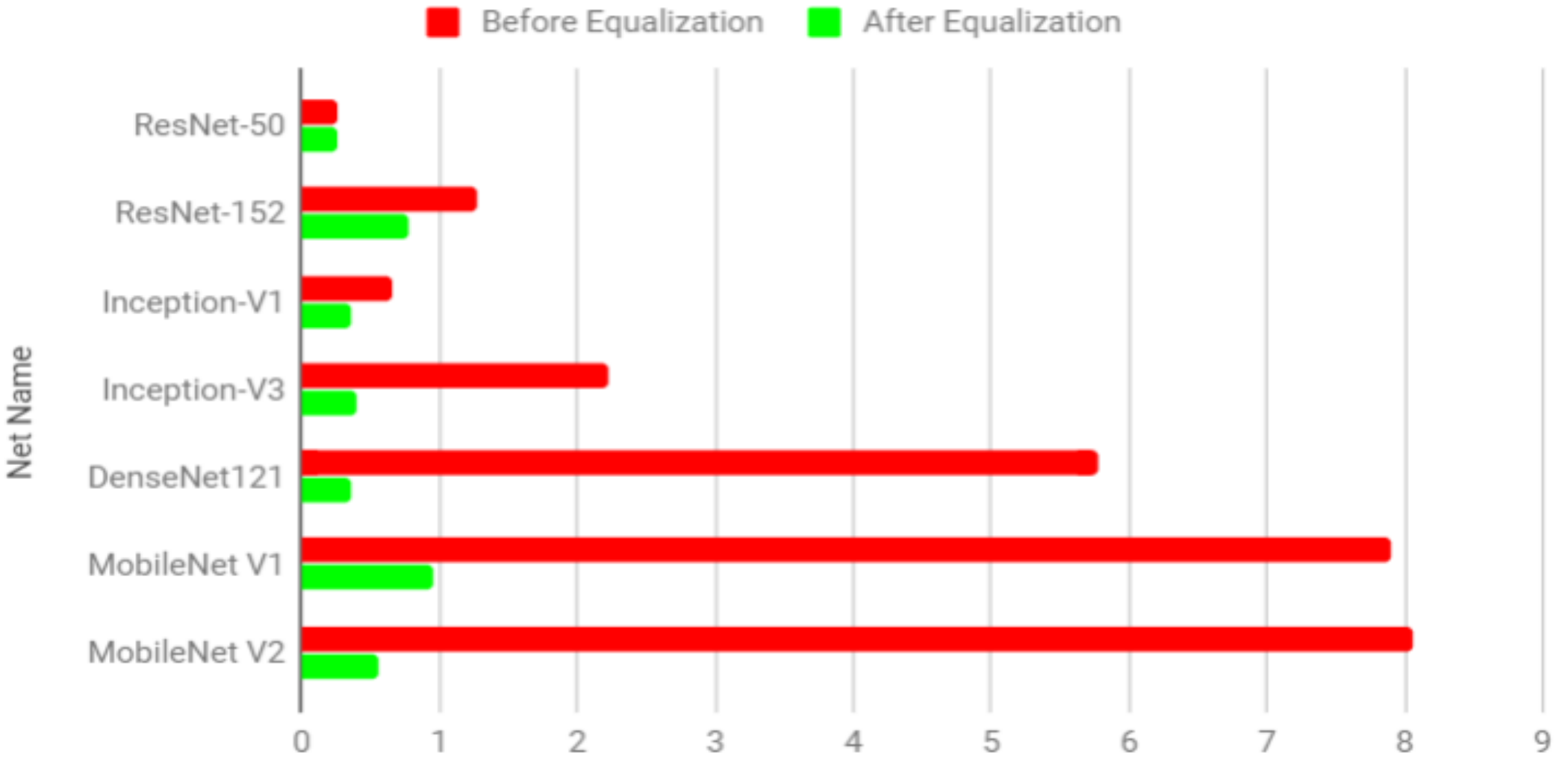
Network Equalization



Quantization Degradation on Imagenet[%]



Quantization Degradation on Imagenet[%]



Summary

- Equalization is an easy to use **post-training quantization** method to recover quantization noise in neural networks
- Can be applied to any network
- A novel approach to quantization by searching for the best **equivalent representation**
- The method can be combined with other quantization methods - e.g. quantization-aware training and smart clipping