Background

Motivation
- DNNs have the disadvantage of high computational and memory costs
- Especially a problem for embedded and mobile
- Can be improved with fixed-point low bit-width computations

Problem statement
- SGD cannot optimize quantized variable (gradients are zero)
- Quantizing weights of pre-trained model yields poor results

Previous approaches
- Using low-precision weights for forward pass and shadow full-precision weights for backward pass
- Applying soft thresholding during training and hard thresholding at inference
- Usually using uniform quantization, not optimal for the empirically normally distributed weights:

Results

- Measuring complexity in num. of bit operations (BOPs)
- We quantize all layers (unlike competitors), thus our ResNet50 is more efficient that their ResNet34 (and the same for our ResNet34 vs. their ResNet18)
- SOTA ImageNet accuracy for a fixed budget of BOPS (and SOTA BOPs for a required accuracy)
- Our method favorable both when more sensitive to complexity or accuracy (in the complexity vs. accuracy trade-off)

UNIQ: Uniform Noise Injection for Quantization of neural networks

Chaim Baskin

Joint work with Eli Schwartz, Evgenii Zheltonozhskii, Natan Liss, Raja Giryes, Alex Bronstein and Avi Mendelson

Visualization Theory and Applications Lab