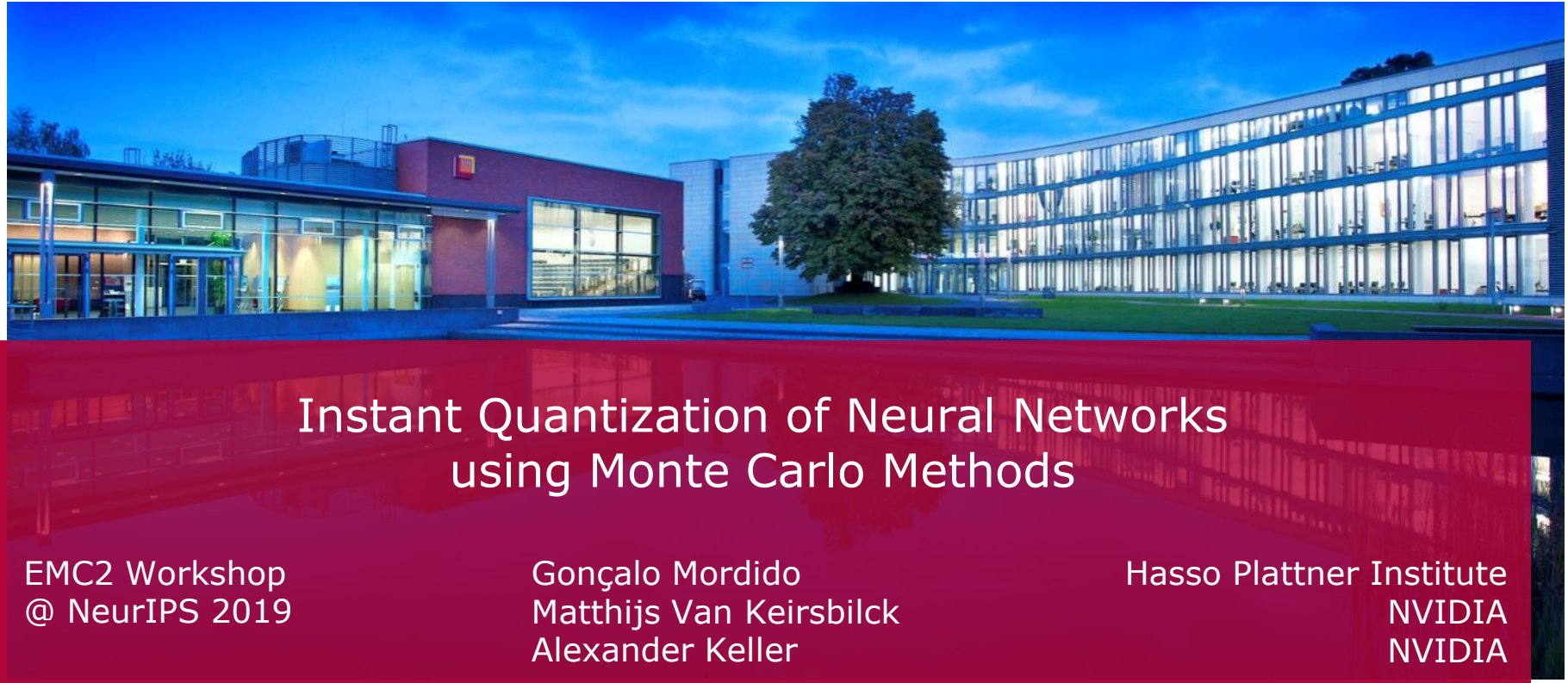


Instant Quantization of Neural Networks using Monte Carlo Methods

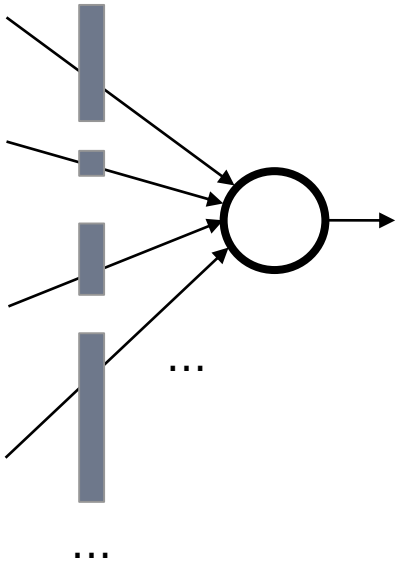
EMC2 Workshop
@ NeurIPS 2019

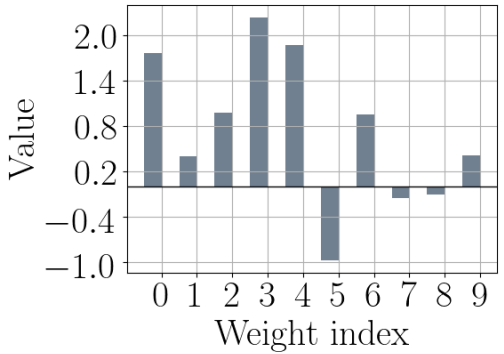
Gonçalo Mordido — Hasso Plattner Institute
Matthijs Van Keirsbilck — NVIDIA
Alexander Keller — NVIDIA

Motivation and idea

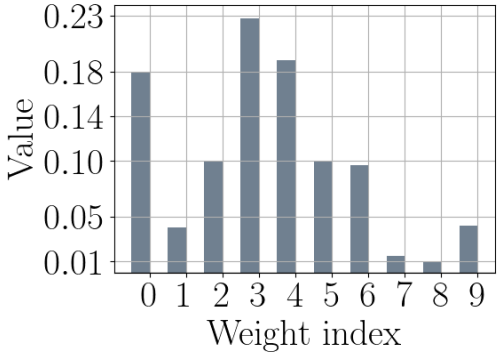
- neural network quantization/sparsity
  - lower cost: compute, memory, power, bandwidth, ...
- quantization usually requires retraining
- idea: use importance sampling
  - fast and efficient due to stratified sampling
  - sparsity and bit-width adjustable by the number of samples
  - **no additional training**

Monte Carlo Quantization (MCQ)



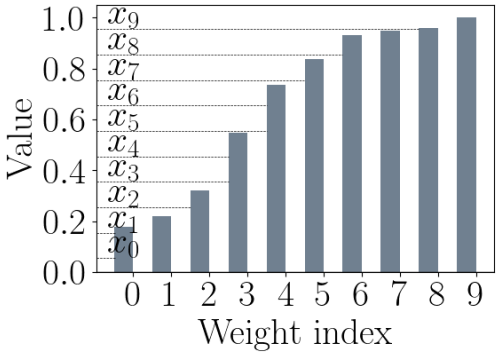


full precision values

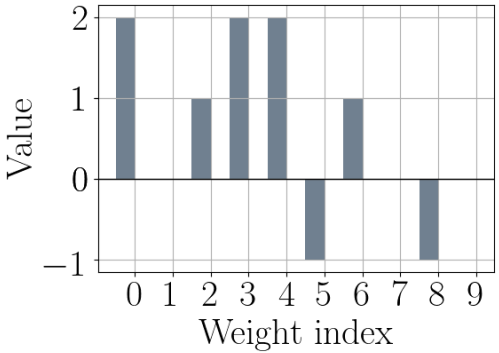


PDF

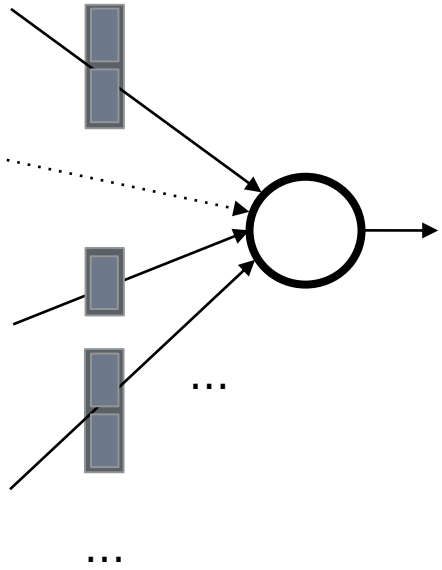
Monte Carlo Quantization (MCQ)



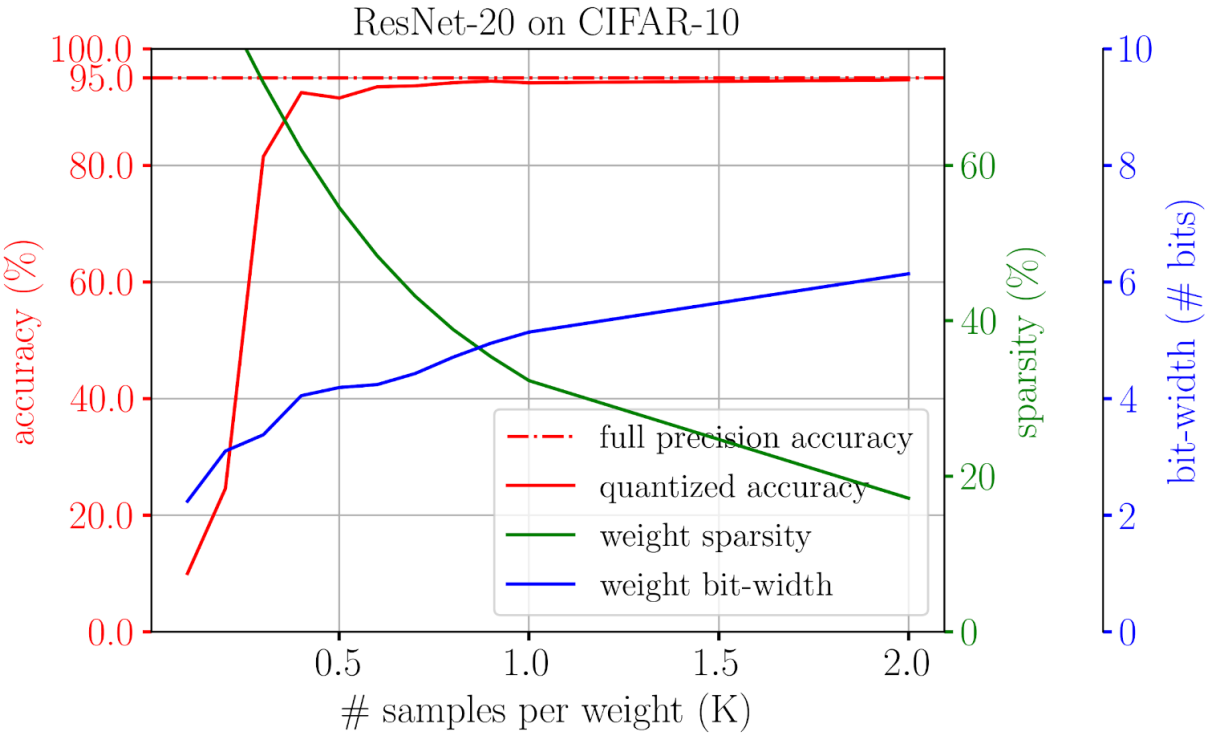
CDF



integer values



Results


ResNet-20 on CIFAR-10
accuracy (%)
sparsity (%)
bit-width (# bits)
samples per weight (K)
full precision accuracy
quantized accuracy
weight sparsity
weight bit-width
100.0
95.0
80.0
60.0
40.0
20.0
0.0
60
40
20
0
10
8
6
4
2
0
0.5
1.0
1.5
2.0

Monte Carlo Neural Networks

- simple method to quantize/sparsify models
  - low accuracy loss
  - no retraining
- general applicability
  - weights and/or activations
  - related to random walks
- future work
  - quantized gradients
  - integer neural networks