Integer Quantization for Deep Learning Inference
: Principles and Emprical Evaluation

인공지능융합전공 222AIG16 박지은

Q1. What is the problem being solved?

- The mathematical aspects of quantization parameters
  - Evaluate their choices on a wide range of neural network models for different application domains (vision, speech, and language)
- Quantization tachniques that are amenable to acceleration by processors with high-throughput integer math pipelines
- Present a workflow for 8-bit quantization that able to maintain accuarcy within 1% of the floating-point baseline on all networks studied, including models that are more difficult to quantize like MobileNets and BERT-large

Q2. What is unique about the suggested solution?

- Integer quantization for neural network inference, where trained networks are modified to use integer weights and activations so that integer math pipelines can be used for many operations
- Math-intensive tensor operations executed on 8-bit integer types showed better performance than in fp32 version and maintain model accuracy within 1% of each baseline floating-point network.
  - image processing, language modeling, language translation, speech recognition (With convolutional networks, recurrent networks, attention-based networks)
- Quantization Aware Training (QAT) to be sufficient for int8 quantization on the models they evaluated, and as such they chose not to include these methods in theire evaluation of int8 quantization.
- Quantization Fundamentals
  - Range mappling (Affine quantization, Scale quantization), Tensor quantization Granularity, Computational cost of affine quantization, Calibration
- Post Training Quantization
  - Weight quantization, Activation quantization
- Techniques to Recover Accuracy
  - Partial quantization, Quantization-Aware training, Leargning quantization parameters

Q3. How is the ieda evaluated?

- They empirically evaluated various choices for int8 quantization of a variety or models, leading to a quantization workflow proposal
- All models they studied can be quantized to int8 with accuacy that either matches or is within 1% of the floating-point model accuarcy including challenging models for quantization like MobileNets and BERT