Big Models Quantization

Alexander Alekseev, Mikhail Seliugin, Alexander Stepikin
Mentors: Daniil Merkulov, Nazarii Tupitsa

Contents

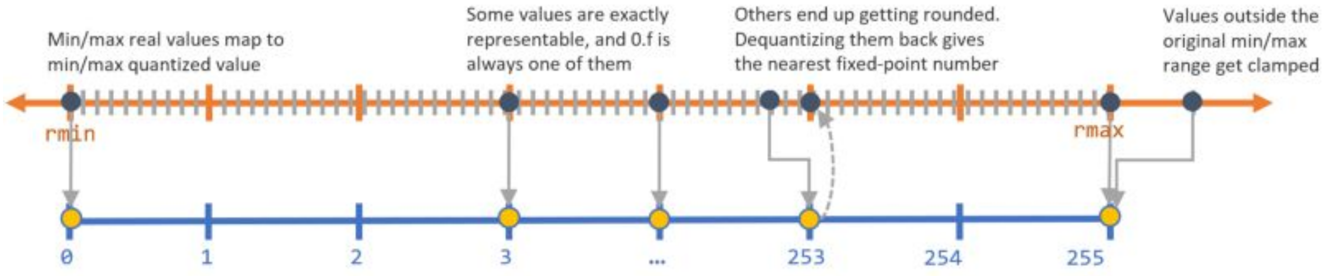
- ➔ Intro
- ➔ Training Pipelines
- ➔ F8Net
- ➔ StatQuant
- ➔ Automatic Mixed Precision

Quantization Introduction

Idea: Use an integer weight and activation representation, reducing the precision.

Motivation:
1. Reduce memory footprint
2. Reduce inference latency
3. Reduce training time*

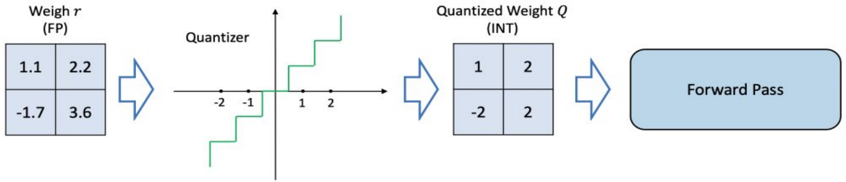
INT8 compared to FP32:
- 4x size reduction
- 2-4x memory reduction
- 2-4x inference acceleration



Quantization Training Pipelines

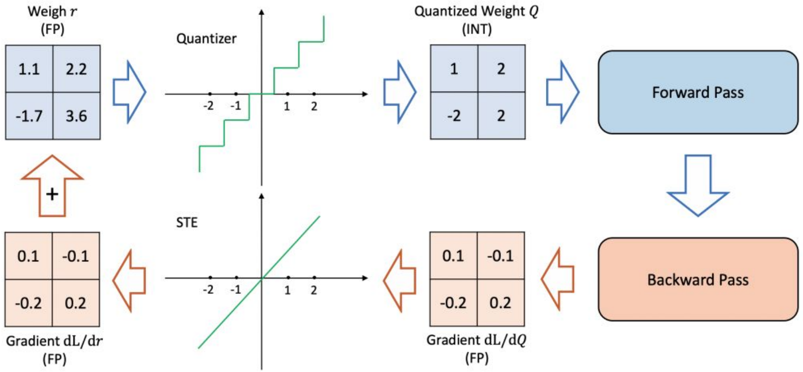
Post-training Quantization:

➔ collecting statistics
➔ no re-training



Quantization-Aware Training:

➔ fine-tuning quantized model
➔ using STE* to approximate gradients



*Fully Quantized Training:

➔ fine-tuning quantized model
➔ gradients are quantized

F8Net Framework

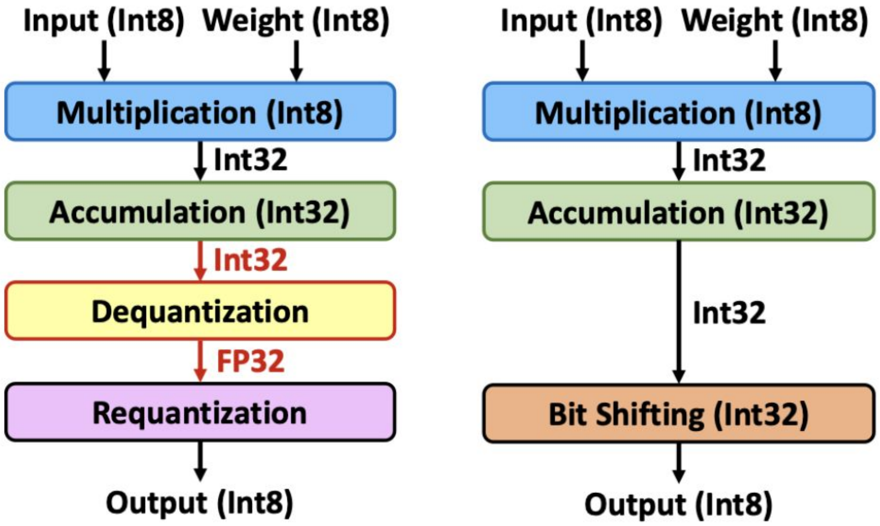
Idea: Quantize inputs and weights, perform all operations with 8-bit integers

Pros:

➔ hardware-friendly forward pass
➔ adjustable tensor representation in fixed-point format

Cons:

➔ full-precision gradients



Considered quantization settings: Simulated quant. (left) and Fixed-point quant. (right)

StatQuant Framework

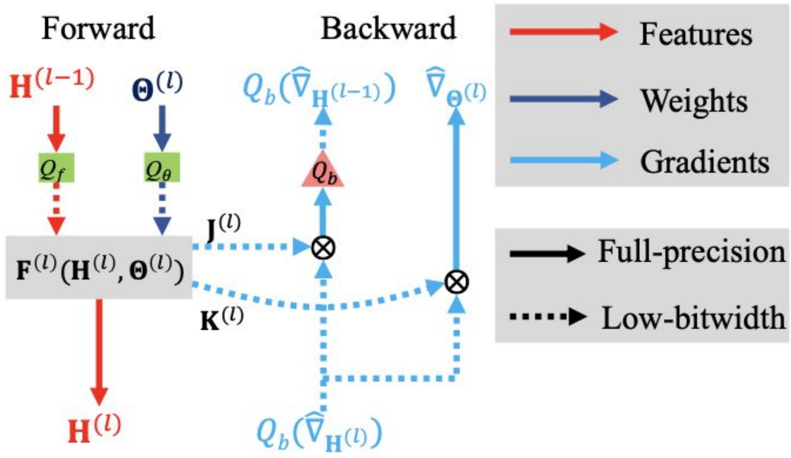
Idea: Consider FQT gradient as an unbiased estimate of QAT gradient and reduce its variance via Per-sample and Block Householder quantizers

Pros:

- ➔ speed-up on specialized hardware
- ➔ almost no performance degradation
- ➔ solid theoretical foundations

Cons:

- ➔ no significant memory reduction



StatQuant FQT pipeline

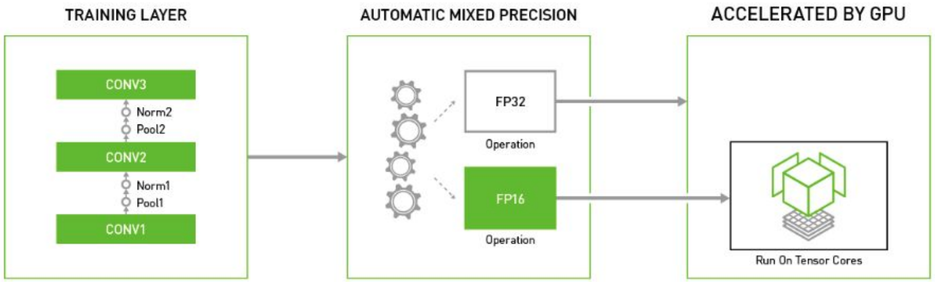
Automatic Mixed Precision

Idea: use 16-bit precision to speed up training
Problem: small values vanish (underflowing)

Tricks:
- ➔ Some operations in fp16, some in fp32
- ➔ Apply updates to fp32 weights
- ➔ Loss scaling

AMP results on ResNet-18 on ImageNet (B - Baseline)

Algorithm	Test Accuracy	GPU Memory	Total Training Time
B - 1080 Ti	94.13	10737MB	64.9m
B - 2080 Ti	94.17	10855MB	54.3m
AMP - 1080 Ti	94.07	6615MB	64.7m
AMP - 2080 Ti	94.23	7799MB	37.3m



Our Results: Fine-tuning GPT2 on Max Korzh texts 5 epochs

	16-bit AMP	32-bit
Training time	01:01	01:32
GPU Mem	6691 MB	8325 MB
Loss	2.7372	2.7370

Quantization Team

Alexander Stepikin

Mikhail Seliugin

Alexander Alekseev

Special Greetings: Alexander Bredikhin Yulia Koniushenko Dmitrii Zudin

Thank you for your attention!