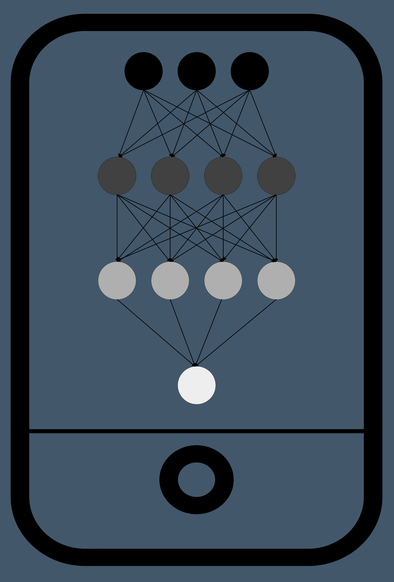

AUTOMATED QUANTIZATION of Neutral Networks

Miroslav Šafář (xsafar23@stud.fit.vut.cz)

Supervisor: Ing. Vojtěch Mrázek Ph.D. (mrazek@fit.vut.cz)

Proposed solution

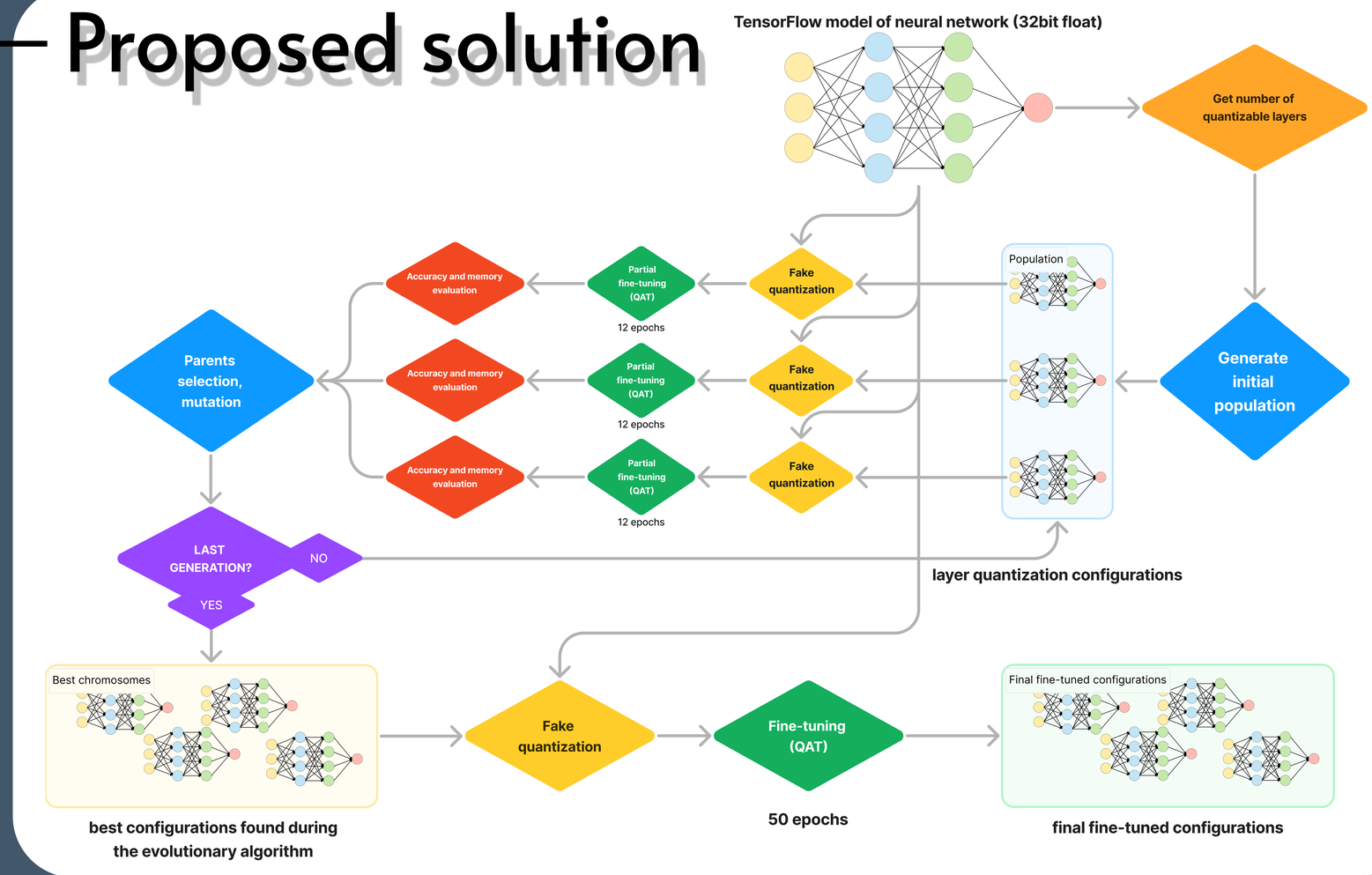

Figure 1: Scheme of proposed system design

Summary

We propose a system for automated mixed-precision quantization of neural networks.

We utilize the evolutionary algorithm NSGA-II and quantization-aware training for fine-tuning quantized models. We conducted experiments with a tiny-imagenet dataset and MobileNetv1 0.25.

We achieved accuracies comparable to those of floating-point models while making the model about 30 % smaller for per-tensor asymmetric quantization and 40% smaller for per-channel symmetric quantization in comparison to 8-bit models. This beats uniform quantization by 10 %, respectively by 6 %.

TensorFlow mixed-precission quantization

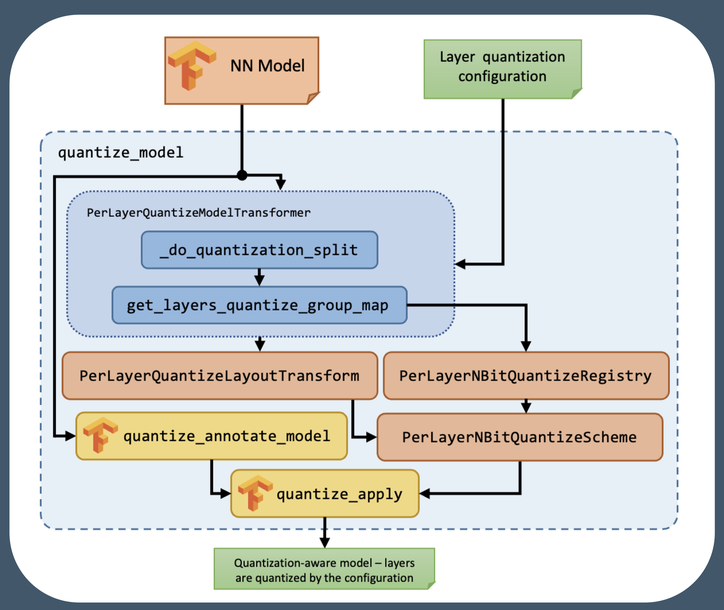

Figure 2: Scheme of quantize_model function that supports mixed-precision quantization

Results

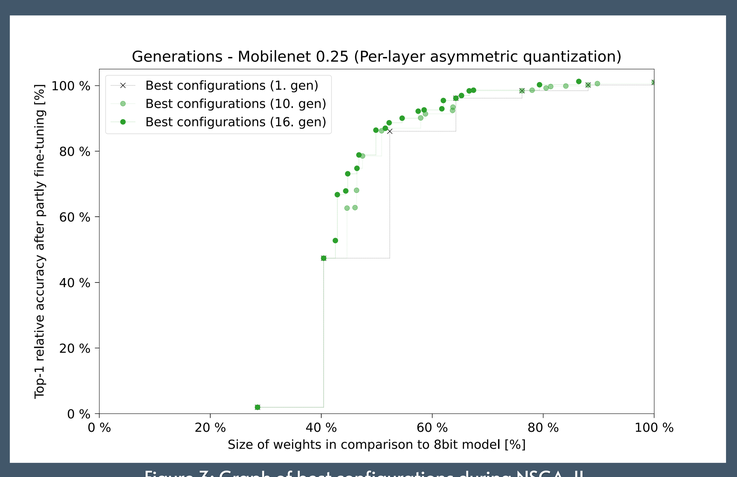

Figure 3: Graph of best configurations during NSGA-II for per-tensor asymmetric quantization of weights

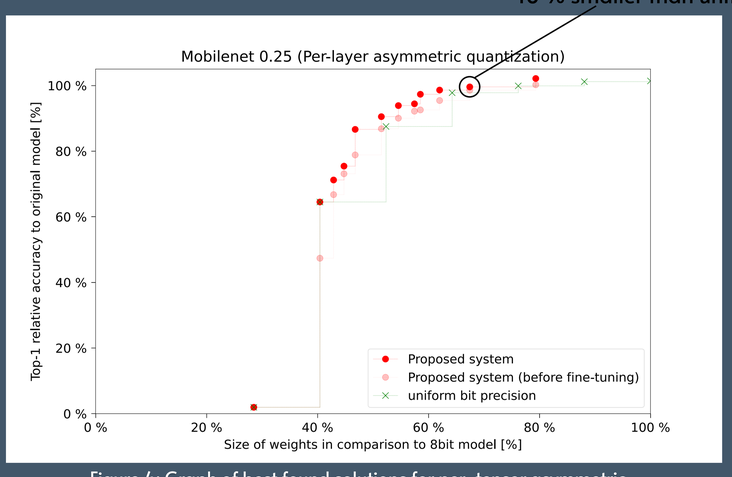

Figure 4: Graph of best found solutions for per-tensor asymmetric quantization of weights

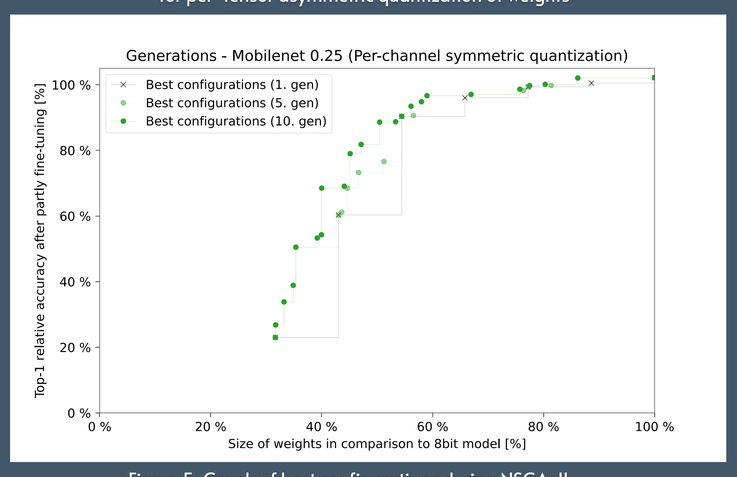

Figure 5: Graph of best configurations during NSGA-II for per-channel symmetric quantization of weights

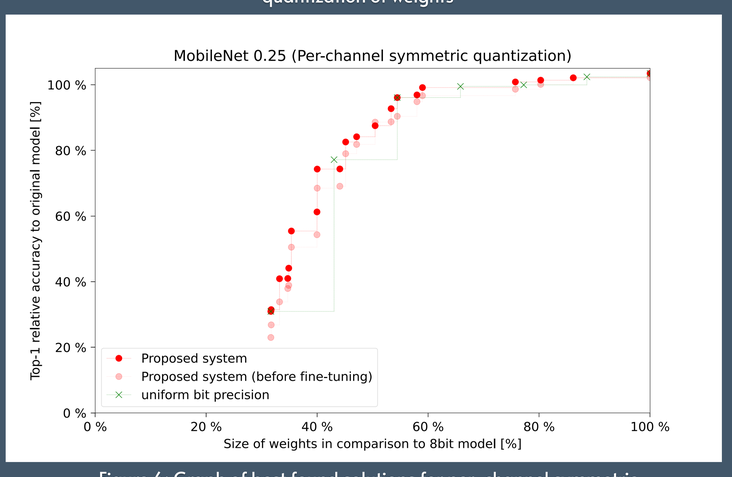

Figure 6: Graph of best found solutions for per-channel symmetric quantization of weights

Contributions

- TensorFlow extension for per-tensor quantization-aware training and mixed-precission model quantization
- framework for searching layer quantization level configuration
- up to 10% smaller model than uniform solution without any accuracy loss

BRNO UNIVERSITY OF TECHNOLOGY | FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023