Raghuraman Krishnamoor
xxx

Title: Quantization for efficient inference in edge devices

Problem to be solved: For TinyML to become a reality, complexity (latency and power) of inference needs to be reduced.

Technical Approach: We present techniques for quantizing deep networks and showcase potential savings in model size and latency for convolutional neural networks.

Results:
We show the benefits of the following techniques for quantizing CNNs for inference:

1. Per-channel quantization of weights and per-layer quantization of activations to 8-bits of precision post-training produces classification accuracies within 2% of floating point networks for a wide variety of CNN architectures
2. Model sizes can be reduced by a factor of 4 by quantizing weights to 8- bits, even when 8-bit arithmetic is not supported. This can be achieved with simple, post training quantization of weights
3. We benchmark latencies of quantized networks on CPUs and DSPs and observe a speedup of 2x-3x for quantized implementations compared to floating point on CPUs. Speedups of up to 10x are observed on specialized processors with fixed point SIMD capabilities, like the Qualcomm QDSPs with HVX.
4. Quantization-aware training can provide further improvements, reducing the gap to floating point to 1% at 8-bit precision. Quantization-aware training also allows for reducing the precision of weights to four bits with accuracy losses ranging from 2% to 10%, with higher accuracy drop for smaller networks.
5. We review best practices for quantization-aware training to obtain high accuracy with quantized weights and activations.
6. We recommend that per-channel quantization of weights and per-layer quantization of activations be the preferred quantization scheme for hardware acceleration and kernel optimization. We also propose that future processors and hardware accelerators for optimized inference supports 4,8 and 16 bit precisions for weights and activations.

Note that this poster is a summary of the work at: https://arxiv.org/abs/1806.08342

Significance: Our hope is that the techniques presented here are used widely to build lightweight models for inference.