Here’s why quantization matters for AI

Shrinking machine learning models without losing accuracy

MAR 12, 2019

Qualcomm products mentioned within this post are offered by Qualcomm Technologies, Inc. and/or its subsidiaries.

Whether you realize it or not, AI is already being widely deployed in commercial applications. Many people are experiencing the benefits of AI through the better photos they’re taking, the increased security on their phones, and the more personalized responses coming from their voice assistants. And that’s only the beginning. Along with improved user experiences, the computation demands of AI training and inference are growing. Inference efficiency, in particular, has become a burning issue for machine learning.

Reducing computation demand and increasing power efficiency

One way to reduce the AI computation demands and increase power efficiency is through quantization. Quantization is an umbrella term that covers a lot of different techniques to convert input values from a large set to output values in a smaller set. You can think of quantization through the following analogy. Someone asks you what time it is. You’d look at your watch and say “10:21” but that’s not 100% accurate. Hours, minutes, and seconds are a convention that we use in order to quantize, or approximate, the continuous variable that is time. We simplify time into discrete numbers.

Another example is capturing a digital image by representing each pixel by a certain number of bits, thereby reducing the continuous color spectrum of real life to discrete colors. For example, a black and white image could be represented with one bit per pixel, while a typical image with color has twenty-four bits per pixel (see GIF below). Quantization, in essence, lessens the number of bits needed to represent information.

GIF 1: The number of bits to represent each pixel in an image, whether it be 24, 8, or 1, is a good example of quantizing data.

Getting back to AI, artificial neural networks consist of activation nodes, the connections between the nodes, and a weight parameter associated with each connection. It is these weight parameters and activation node computations that can be quantized. For perspective, running a neural network on hardware can easily result in many millions of multiplication and addition operations. Lower-bit mathematical operations with quantized parameters combined with quantizing intermediate calculations of a neural network results in large computational gains and higher performance.

Besides the performance benefit, quantized neural networks also increase power efficiency for two reasons: reduced memory access costs and increased compute efficiency. Using the lower-bit quantized data requires less data movement, both onchip and off-chip, which reduces memory bandwidth and saves significant energy. Lower-precision mathematical operations, such as an 8-bit integer multiply versus a 32-bit floating point multiply, consume less energy and increase compute efficiency, thus reducing power consumption. In addition, reducing the number of bits for representing the neural network’s parameters results in less memory storage. (see GIF 2).

GIF 2: Quantization for AI provides critical benefits, such as higher performance and lower power.

Quantizing without sacrificing accuracy

What’s the catch of using low-bit networks? Typically, the accuracy of the quantized AI model tends to drop. As a leader in power-efficient on-device AI processing, Qualcomm Technologies has research dedicated to improving quantization techniques and solve this accuracy challenge. We are particularly interested in quantizing 32-bit floating point weight parameters to 8-bit integers in neural networks without sacrificing accuracy. Outside of our ongoing research in Bayesian deep learning for model compression and quantization, our two accepted papers at ICLR 2019 focus on the execution of low-bit AI models.

The “Relaxed Quantization for Discretized Neural Networks” paper showcases a new method that better prepares the neural network for quantization during the training phase. This allows the neural network to adapt to the quantized computations that will happen during the deployment of the model. The method produces quantized models that perform better and retain more accuracy than alternative state-of-the-art approaches.

The ”Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets” paper contributes to the theoretical understanding of the straight-through estimator (STE), which is widely used in quantization-aware model training. The paper proves that with a properly chosen STE, a quantized network model converges to a critical point of the training loss function, while a poor choice of STE leads to an unstable training process. The theory was verified with experiments.

Model compression (including Bayesian learning, quantization, and decomposition) is just one example of the research directions that Qualcomm AI Research is currently focusing on. Other topics include: equivariance of convolutional neural networks, audio to speech compression, machine learning for autonomous vehicles, computational photography, and model training optimized for low power devices. Our goal is to make fundamental AI research breakthroughs so that we as well as our customers can scale the technology across industries and ultimately enrich our daily lives. Find out more about Qualcomm AI Research and see our list of published papers here.

Discover our machine learning vacancies >

Sign up for our newsletter to receive the latest information about mobile computing >

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Engage with us on
TwitterandFacebook

Artificial Intelligence

Research

Opinions expressed in the content posted here are the personal opinions of the original authors, and do not necessarily reflect those of Qualcomm Incorporated or its subsidiaries (”Qualcomm”). Qualcomm products mentioned within this post are offered by Qualcomm Technologies, Inc. and/or its subsidiaries. The content is provided for informational purposes only and is not meant to be an endorsement or representation by Qualcomm or any other party. This site may also provide links or references to non-Qualcomm sites and resources. Qualcomm makes no representations, warranties, or other commitments whatsoever about any non-Qualcomm sites or third-party resources that may be referenced, accessible from, or linked to this site.

Jilei Hou
Vice President, Engineering, Qualcomm Technologies

More articles from this author ›

About this author ⊕

Related News

OnQ Blog DEC 19, 2019

Three key technologies for next-gen smart speakers

OnQ Blog DEC 9, 2019

NeurIPS 2019: Experience the latest breakthrough in AI

OnQ Blog NOV 25, 2019

One year of Qualcomm AI Research at a glance [video]

C

r
ri
V

○ ○

Qualcomm

Language

About Qualcomm Careers Offices Contact Us Support Subscription Center

Terms of Use Privacy Cookies

References to ”Qualcomm” may mean Qualcomm Incorporated, or subsidiaries or business units within the Qualcomm corporate structure, as applicable.

Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its products and services businesses. Qualcomm products referenced on this page are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

Materials that are as of a specific date, including but not limited to press releases, presentations, blog posts and webcasts, may have been superseded by subsequent events or disclosures.

Nothing in these materials is an offer to sell any of the components or devices referenced herein.