How algorithmic advances make power-efficient AI possible

Pioneering Bayesian deep learning for model compression and quantization

AUG 13, 2018

Qualcomm Technologies is inventing technology at scale to realize the promise of AI on trillions of connected devices, which will enrich people's lives and transform industries. For AI to run on these connected devices, the power

transform industries. For AI to run on these connected devices, the power efficiency of the processing must continue to advance. At Qualcomm AI Research, power efficiency is one of our core research areas. In this blog post, I'm providing an example of our latest research in power-efficient AI algorithms and how it goes hand-in-hand with the design of power-efficient hardware.

Intelligence per joule will become the AI benchmark

AI is being powered by the explosive growth of Deep Neural Networks (DNNs). However, we are seeing small, yet significant, improvements in accuracy for exponential increases in energy consumption. At our current trajectory, a neural network in 2025 is projected to have approximately 100 trillion weight parameters (Figure 1), which is similar to the number of synapses in the human brain. The brain does give us hope and inspiration that we can do much better since it is 100x more power efficient than current digital hardware.

Deep neural networks are energy

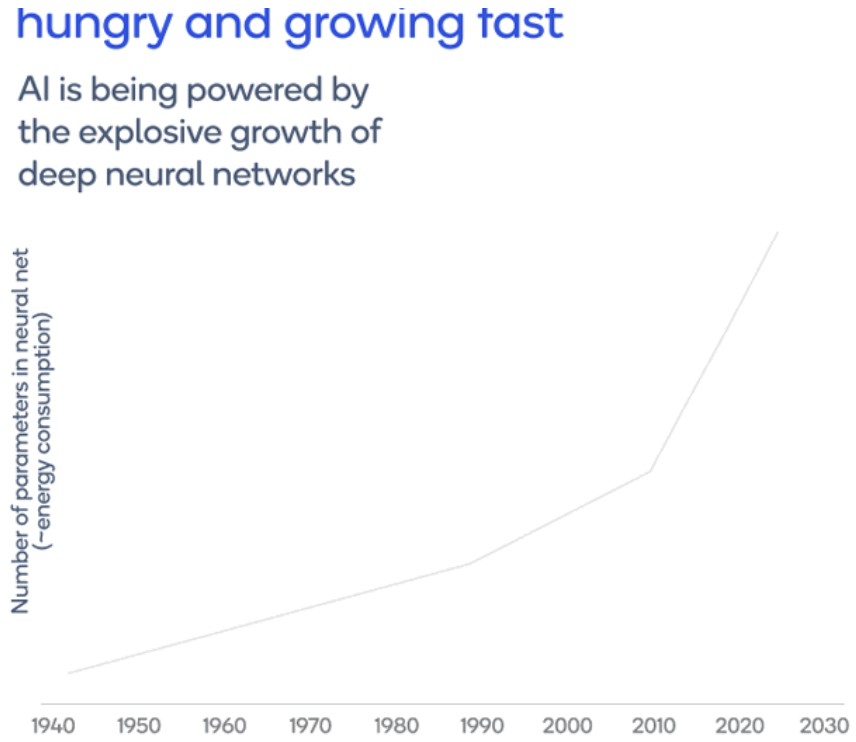

Figure 1: The exponential growth in deep neural networks is driving increased energy consumption. (Source: Welling)

This is important to note because I believe that the benchmark for AI processing will soon change and that AI algorithms will be measured by the amount of intelligence they provide per joule. There are two key reasons for this:

- First, broad economic viability requires energy-efficient AI since the value created by AI must exceed the cost to run the service. To put this in perspective, the economic feasibility of applying AI per transaction may require a cost as low as a micro dollar (1/10,000th of a cent). An example of this is using AI for personalized advertisements and recommendations.
- Second, on-device AI processing in sleek, ultra-light mobile form factors requires power efficiency. Processing always-on compute-intensive workloads in power- and thermal-constrained form factors that require all-day battery life is part of making AI broadly adopted by consumers. The same power efficiency attributes are needed for other classes of devices, such as autonomous cars, drones, and robots.

| AI algorithms will be

> AI algorithms will be measured by the amount of intelligence they provide per joule.

Diving deeper into deep neural networks

DNNs, which are primarily composed of Convolutional Neural Networks (CNNs), are powering the current AI revolution. I like to think about the properties of CNNs in terms of the good, the bad, and the ugly. It is only by understanding the bad and the ugly at a fundamental level that we can improve CNNs.

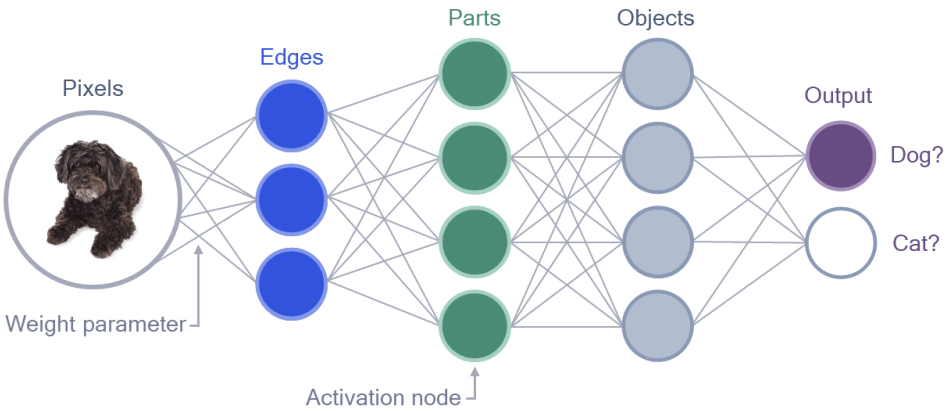

Figure 2: A simple deep neural network for image classification.

In terms of the good, CNNs extract learnable features with state-of-the art results, encode location invariance (e.g., can classify the dog in figure above regardless of its location in an image), share parameters for data efficiency, and execute quickly on modern hardware with parallel processing. On the bad and ugly side, the biggest problem is that CNNs use too much memory, compute, and energy. They also do not encode additional symmetries, such as rotation invariance (imagine the dog in the Figure 2 above now rotated upside down and it no longer works), do not reliably quantify the confidence in a prediction, and are easy to fool by changing the input only slightly, such as adversarial

are easy to fool by changing the input only slightly, such as adversarial examples. We're researching techniques to address these challenges, and the one with very promising results is Bayesian deep learning.

Noise can be a good thing for AI

Bayesian deep learning is stochastic, meaning that we add noise or random values to the weights of the neural network, which also propagates noise to the activation nodes. This noise is a good thing and is inspired by how the brain works. One key benefit of Bayesian deep learning is compression and quantization, which reduces complexity of the neural network model. Quantization reduces the bit-width of parameters (such as using four bits instead of eight bits), and compression prunes the number of activations in the model, which ultimately improves power efficiency. You'll have to take my word for it or tune in to my webinar where I'll explain this in more detail.

Bayesian deep learning is not just theory — in our research, we've applied it to real use cases. For example, we measured the size and accuracy of ResNet-18, a neural network for image classification that has already been optimized for size, on a large set of labeled images as a baseline data point. We then compared the accuracy and compression ratio of state-of-the-art pruning methods versus Bayesian pruning. Bayesian pruning had the best results and provided 3x the compression ratio as baseline while maintaining close to the same accuracy.

A holistic approach to AI power efficiency

What does the future look like for AI hardware? The key to making efficient hardware comes from a deep understanding of real AI workloads at a system level — in other words, how real applications run on real devices. We do this by focusing on the intersection of hardware, algorithms, and software.

For AI hardware acceleration research, we investigate the appropriate compute architecture and memory hierarchy for the given task as well as remove bottlenecks that would decrease utilization and prevent performance from reaching theoretical peaks. Our focus on algorithmic advances, such as Bayesian deep learning, help us to optimize the hardware. Our software tools

Bayesian deep learning, help us to optimize the hardware. Our software tools, such as our Snapdragon Neural Processing SDK, help unlock the built-in optimizations of the hardware designed to provide high performance per watt.

From seeing how applications are written, understanding popular neural networks, and examining the bottlenecks in the system, we repeatedly apply these learnings to our hardware design. It is this system expertise that allows us to continue to advance in each area — hardware, algorithms, and software — to provide efficient and holistic solutions. It is this relentless passion about power-efficient AI that drives our vision of making ubiquitous on-device AI a reality.

Register for my webinar to learn more >

Download the webinar presentation >

Stay up to date on mobile computing news — sign up for our newsletter >

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. Qualcomm Snapdragon is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

Engage with us on
TwitterandFacebook

Artificial Intelligence | Machine Learning | Heterogeneous Computing

Dr. Max Welling

Vice President, Technology

More articles from this author >

About this author

Related News

OnQ Blog FEB 20, 2020

Teachir cars to see with AI [video]

OnQ Blog FEB 4, 2020

5G+AI: The ingredie fueling tomorrc tech innovat

OnQ Blog DEC 19, 2019

Three key technol for next-gen smart speake

OnQ Blog DEC 9, 2019

NeurIPS 2019: Experie the latest breakth in AI

OnQ Blog NOV 25, 2019

On yec Qu AI Researc at a glance [video]

Qualcomm

Language