April 2019

@qualcomm_tech

Qualcomm

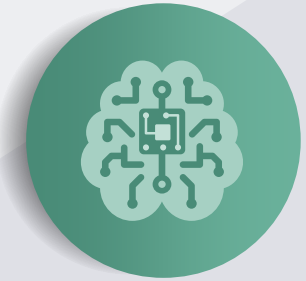
Leading research across the AI spectrum

Qualcomm Technologies, Inc.

Advancing research to make AI ubiquitous

Perception

Object detection, speech recognition, contextual fusion



Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

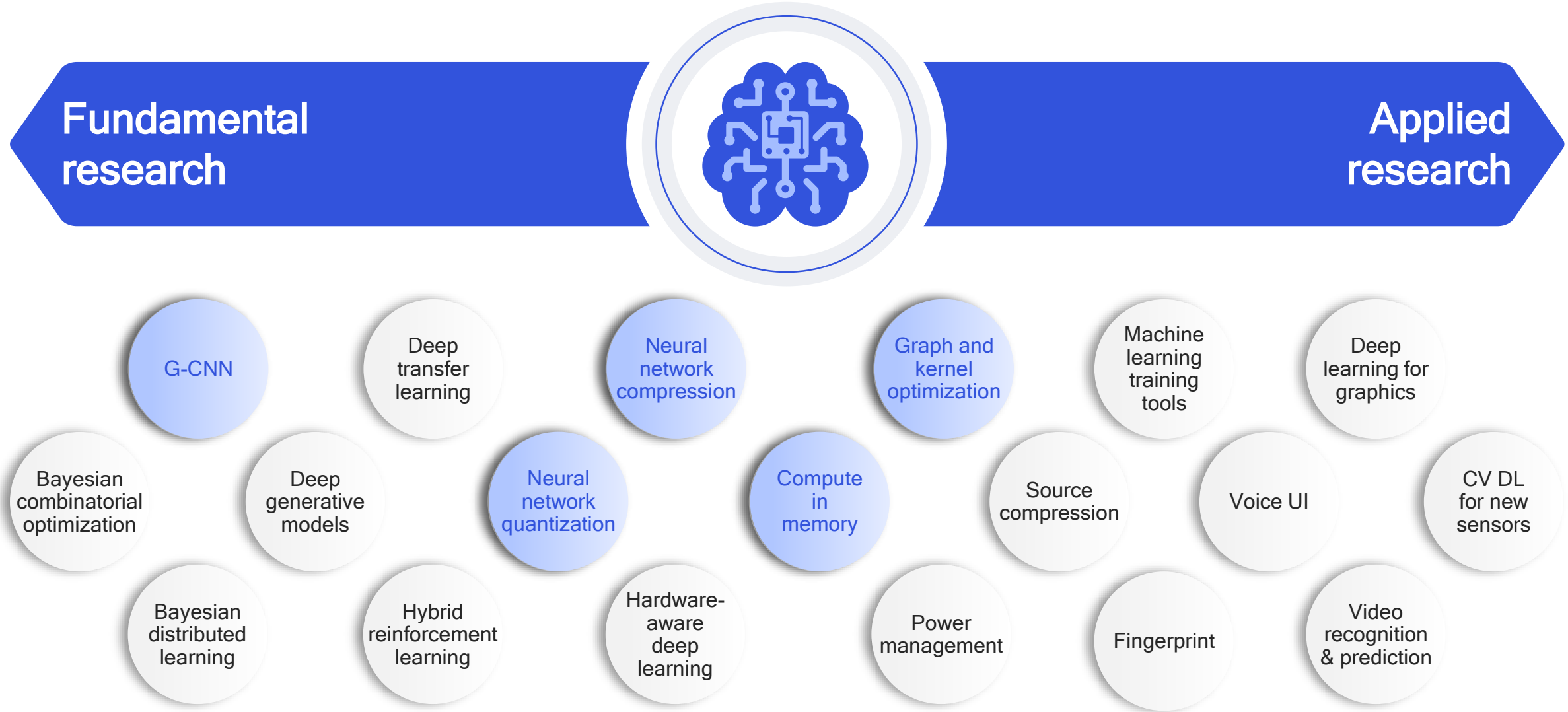
Power efficiency

Personalization

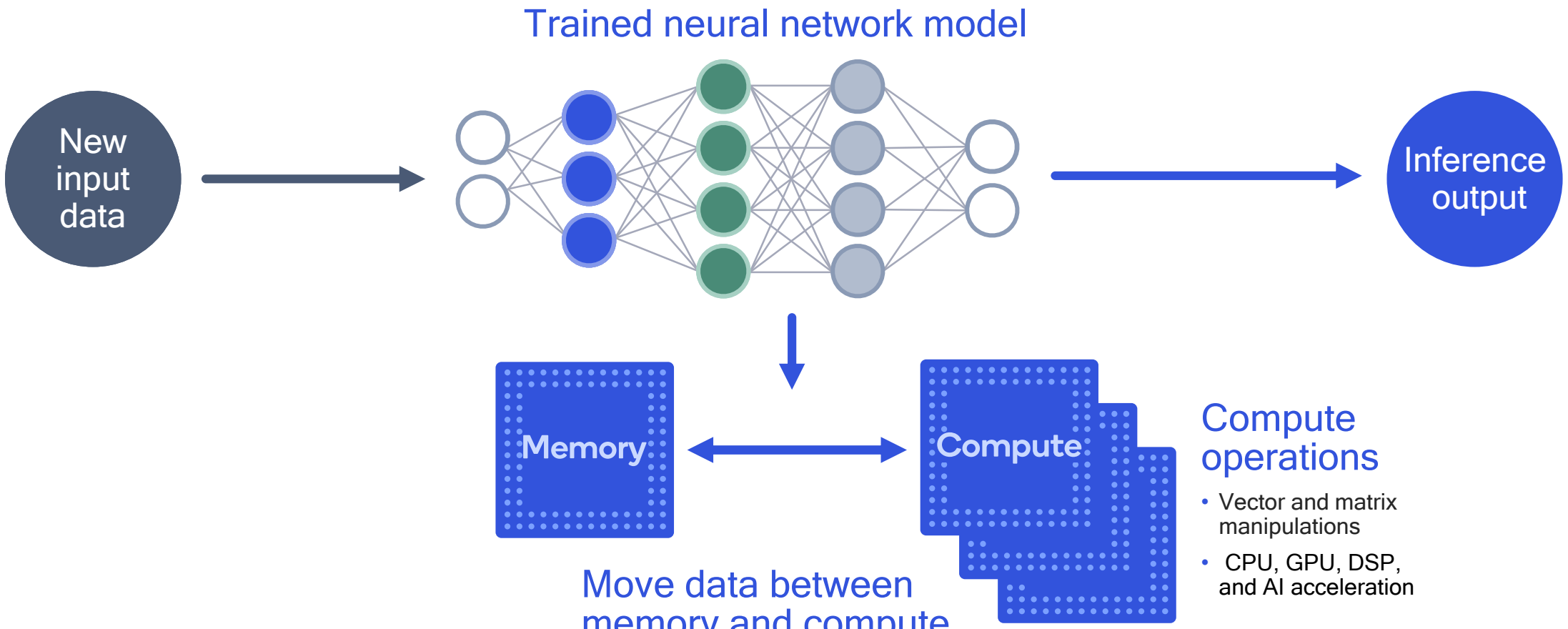
Efficient learning

We are creating platform innovations to scale AI across the industry

Leading research and development across the entire spectrum of AI



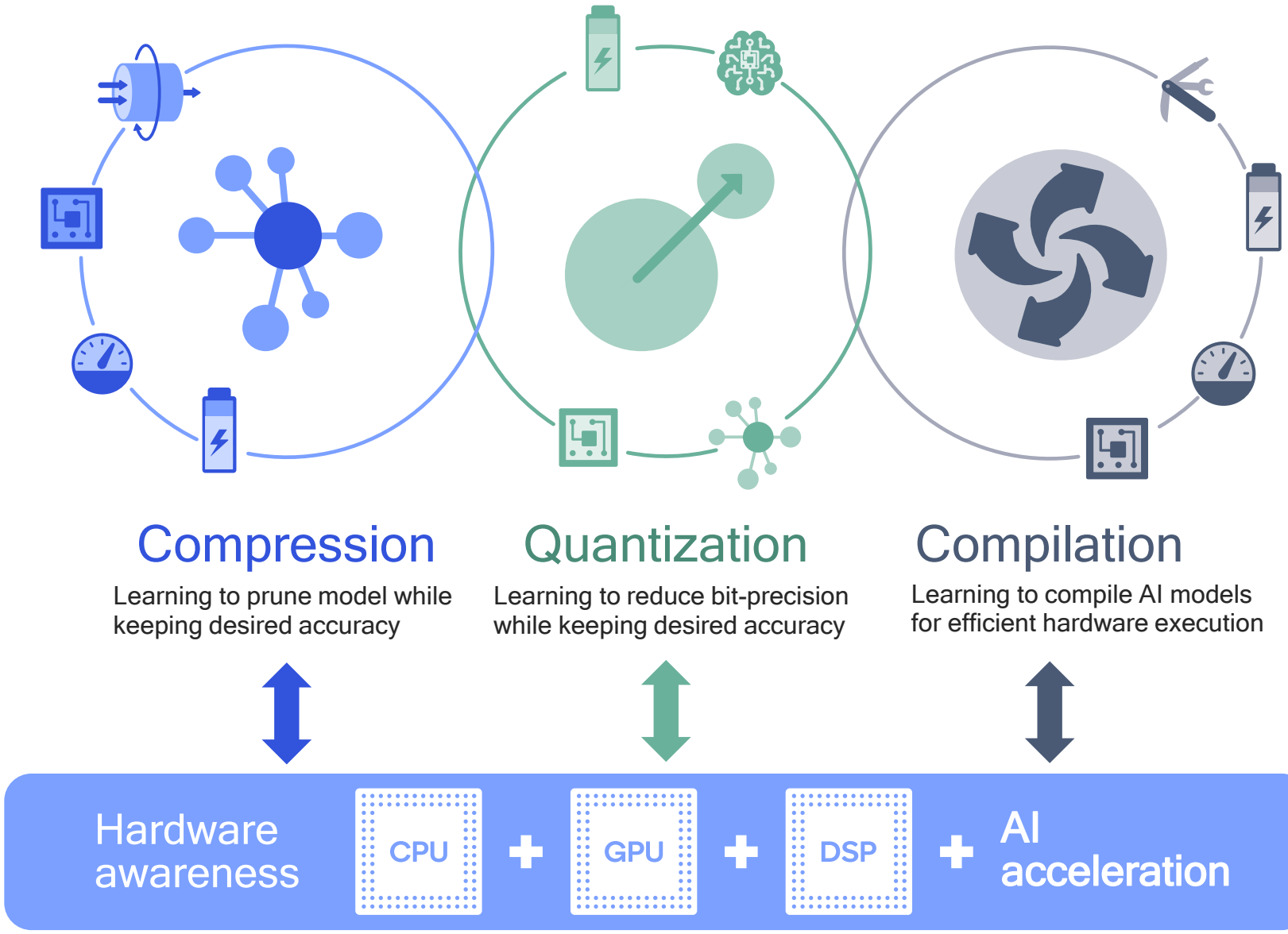
Advancing AI research to increase power efficiency



AI model optimization research for power efficiency

Applying AI to optimize AI models through automated techniques

Reduced time-to-market and engineering cost



Compression of AI model architectures

Automated removal of insignificant/redundant elements while maintaining accuracy

Tensor decomposition

Decomposing a single layer into two or more efficient layers

Spatial SVD

Channel pruning

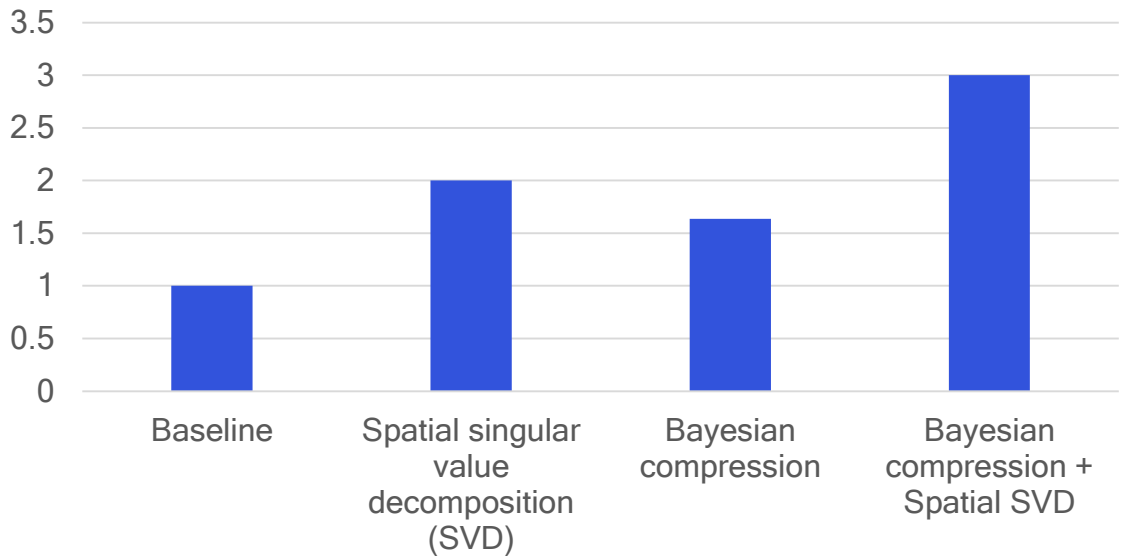
Removing channels from the network

L2 filter magnitude and Bayesian techniques

Hardware aware compression

3x

Compression with less than 1% loss in accuracy*

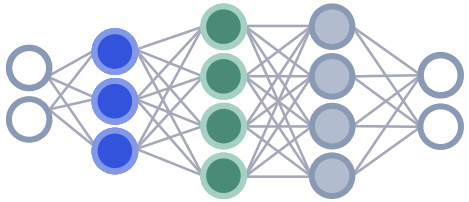


*: Comparison between baseline and compression with both Bayesian compression and spatial SVD. Example uses ResNet18 as baseline.

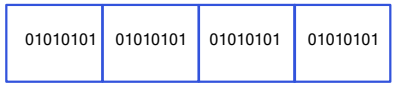
Quantization for power efficiency

Automated reduction in precision of weights and activations while maintaining accuracy

Models typically trained at high precision

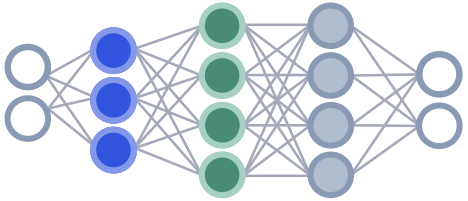


32-bit

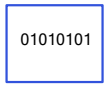


Floating point 3452.3194

Inference at lower precision



8-bit



Integer 3452

>4x

increase in perf. per watt from savings in memory and compute*

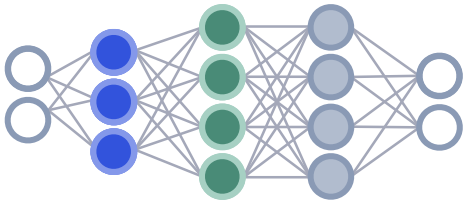
Promising results show that 8-bit AI models can become ubiquitous

Virtually same accuracy for FP32 and quantized INT8

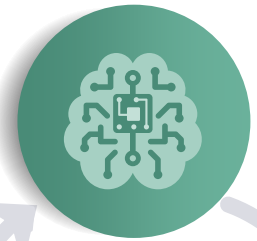
*: Compared to a FP32 model that is not quantized

Compiler research for efficient hardware usage

Reinforcement learning for automated HW compilation–as there are billions of potential configurations



Quantized and compressed AI model



AI optimization agent
Update model based on feedback

Schedule compilation[2]

Execute compiled code
On target hardware CPU, GPU, DSP, AI accelerators

Feedback: execution time

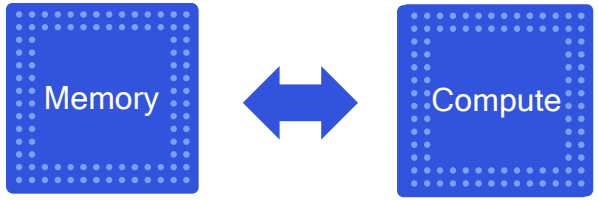
4x
Speedup improvement over TensorFlow Lite[1]

1) On average improvement of tested AI models
2) Schedule kernels and graphs, tile size, reorder, unroll, parallelize, vectorize,…

AI hardware acceleration research

Example: compute-in-memory AI research

- Analog compute
- New memory design
- Need low bit-width AI models



Traditional computer architecture

- Compute and memory are separate and data has to be shuffled back and forth
- Good for general purpose operations

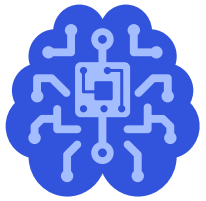
Compute-in-memory

- Computations, like add and multiply, are done in memory
- Good for simple math operations and when memory becomes bottleneck

10-100x
Power efficiency improvement for 1-bit ops*

A paradigm shift from traditional computer architecture can bring orders of magnitude increase in power efficiency

* Compared to traditional Von Neumann architectures today



Deep generative model research for unsupervised learning

Given unlabeled training data, generate new samples from the same distribution

Generative models

Variational auto encoder (VAE)*

Generative adversarial network (GAN)

Auto-regressive

Invertible

Powerful capabilities

Feature extraction: Learn a low-dimension feature representation from unlabeled data

Sampling: Compression, restoration, generation, or prediction of audio, speech, image, or video

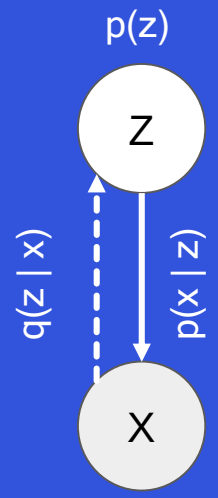
Broad applications

- Speech/video compression
- Text to speech
- Graphics rendering
- Computational photography
- Voice UI

* VAE first introduced by D. Kingma and M. Welling in 2013

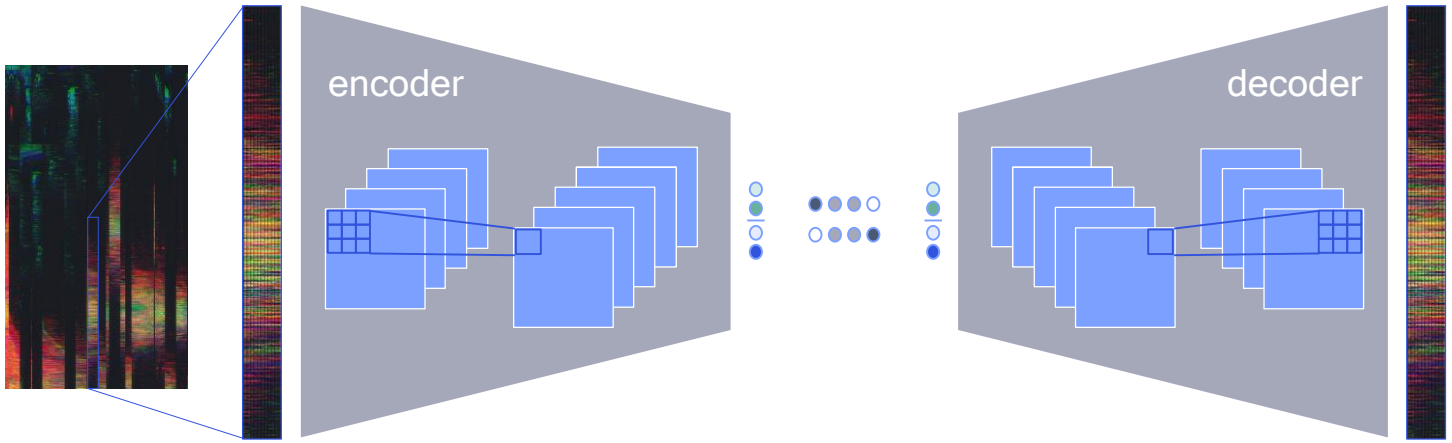
Achieving state-of-the-art data compression

Applying VAE[1] for end-to-end data compression at a lower bit rate



1 VAE first introduced by D. Kingma and M. Welling in 2013

End-to-end speech compression example



Input speech

Compressed speech

Output speech

State-of-the-art traditional speech coding algorithm

2.6x

Bit-rate compression at same speech quality*

Same speech quality via our AI speech compression

*: Comparison between traditional speech coding algorithm and AI speech compression

Can we apply foundational mathematics of physics, like quantum field theory, to deep learning?

G-CNN

Video

Pioneering deep learning research in G-CNNs

Generalized CNNs

- Generalized input, such as rotated objects, applicable to drones, robots, cars, XR,…
- Generalized geometry, such as curved image objects for fisheye lenses, 3D gaming,..

Broader societal benefits:

State-of-the-art accuracy on climate pattern segmentation

G-CNNs are generalized CNNs: Gauge equivariant, Group, and Steerable CNN

Qualcomm

We are advancing AI research to make AI power efficient

We are conducting leading research and development across the entire spectrum of AI

We are creating AI platform innovations that are fundamental to scaling AI across the industry

Qualcomm

Thank you!

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog