Leading research across the AI spectrum

Qualcomm Technologies, Inc.
Advancing research to make AI ubiquitous

We are creating platform innovations to scale AI across the industry
Leading research and development across the entire spectrum of AI

**Fundamental research**
- G-CNN
- Deep transfer learning
- Deep generative models
- Bayesian combinatorial optimization
- Bayesian distributed learning
- Neural network quantization
- Hardware-aware deep learning
- Hybrid reinforcement learning

**Applied research**
- Neural network compression
- Graph and kernel optimization
- Compute in memory
- Compute in memory
- Machine learning training tools
- Deep learning for graphics
- CV DL for new sensors
- Voice UI
- Fingerprint
- Video recognition & prediction
- Power management
- Source compression
- Video recognition & prediction
- Fingerprint
- Machine learning training tools
Advancing AI research to increase power efficiency

- New input data
- Trained neural network model
- Inference output

- Compute operations:
  - Vector and matrix manipulations
  - CPU, GPU, DSP, and AI acceleration

- Move data between memory and compute:
  - Move pieces of input data and AI model from memory to compute
  - Send partial results back to memory
AI model optimization research for power efficiency

Applying AI to optimize AI models through automated techniques

Reduced time-to-market and engineering cost

Compression
Learning to prune model while keeping desired accuracy

Quantization
Learning to reduce bit-precision while keeping desired accuracy

Compilation
Learning to compile AI models for efficient hardware execution

Hardware awareness
CPU + GPU + DSP + AI acceleration
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

01010101 01010101 01010101 01010101

Floating point 3452.3194

Inference at lower precision

8-bit

01010101

Integer 3452

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

>Virtually same accuracy for FP32 and quantized INT8

>4x increase in perf. per watt from savings in memory and compute*

*: 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

AI optimization agent
Update model based on feedback

Schedule compilation
Quantized and compressed AI model

Feedback:
execution time

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

4x Speedup improvement over TensorFlow Lite

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

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

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

* 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

Variational auto encoder (VAE)*

Generative adversarial network (GAN)

Auto-regressive

Invertible

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

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

Powerful capabilities

Speech/video compression

Text to speech

Graphics rendering

Computational photography

Voice UI

Broad applications

* 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

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,…

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

Broader societal benefits:
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
Thank you!

Follow us on:  

For more information, visit us at:
www.qualcomm.com & www.qualcomm.com/blog

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

©2019 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark of Qualcomm Incorporated, registered in the United States and other countries. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other 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 product and services businesses, including its semiconductor business, QCT.