Dec.'22

MIT Guest Lecture

AI Model Efficiency Toolkit (AIMET)

Chirag Patel
Principal Eng./Mgr.
Qualcomm AI Research

The challenge of AI workloads

- Very compute intensive
- Large, complicated neural network models
- Complex concurrencies
- Real-time
- Always-on

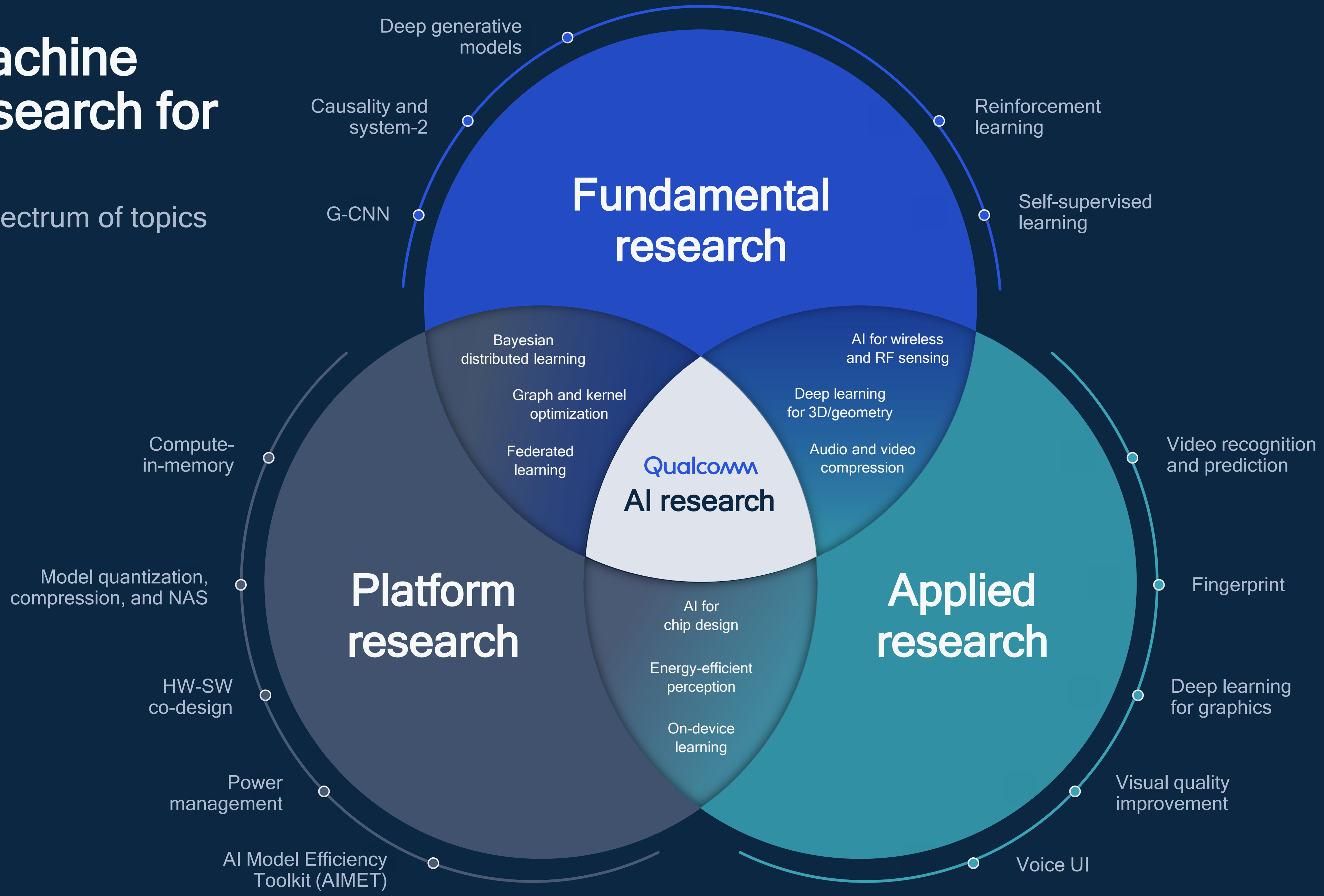
Power and thermal efficiency are essential for on-device AI

Constrained mobile environment

- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations

Leading machine learning research for edge AI

across the entire spectrum of topics


Deep generative models
Causality and system-2
G-CNN
Reinforcement learning
Self-supervised learning
Fundamental research
Bayesian distributed learning
Graph and kernel optimization
Federated learning
AI for wireless and RF sensing
Deep learning for 3D/geometry
Audio and video compression
Qualcomm
AI research
Compute-in-memory
Model quantization, compression, and NAS
HW-SW co-design
Power management
AI Model Efficiency Toolkit (AIMET)
Platform research
AI for chip design
Energy-efficient perception
On-device learning
Applied research
Video recognition and prediction
Fingerprint
Deep learning for graphics
Visual quality improvement
Voice UI

AIMET is a product of Qualcomm Innovation Center, Inc.

AI Firsts

Brought to you by Qualcomm AI Research

Model quantization

Invented the best techniques for fast deployment of 8-bit quantization

Best power-efficiency toolkit in the industry

On-device learning

Invented continuous learning techniques for SOTA on-device voice-UI

First demonstration of 30% improvement to keyword spotting

Federated learning

Invented methods for combining differential privacy and compression

First end-to-end research software framework deployable on mobile

Video semantic segmentation

Top the Cityscape leaderboard with loss function innovation for boundary-awareness

First real-time SS at FHD on mobile

Group equivariant CNN

Pioneer for rotational equivariance; best paper at ICLR’18

First G-CNN segmentation for health on mobile

AI for wireless

Invented neural augmentation to enhance physical layer algorithms

First weakly supervised method for real-world passive RF sensing

Video super resolution

Full stack optimization for visual quality improvement at 4K resolution

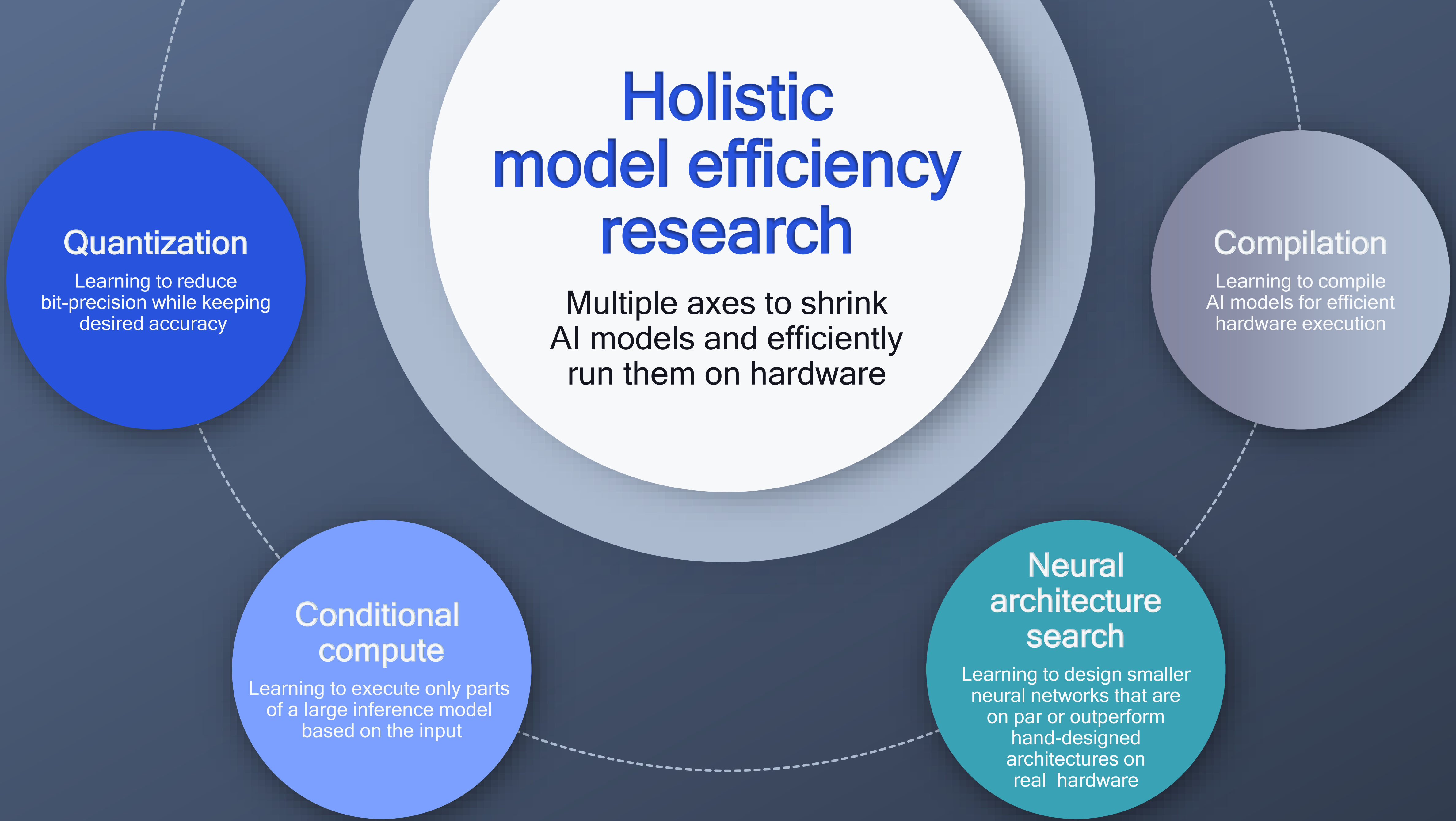
First 4K SR at 100+ FPS on mobile

Neural video compression

Invented instance-adaptive for SOTA performance & new deployment scenarios

First real-time HD decoding on mobile

SOTA: State-of-the-art; Cityscapes Benchmark: https://www.cityscapes-dataset.com/



Holistic model efficiency research

Multiple axes to shrink AI models and efficiently run them on hardware

Quantization

Learning to reduce bit-precision while keeping desired accuracy

Compilation

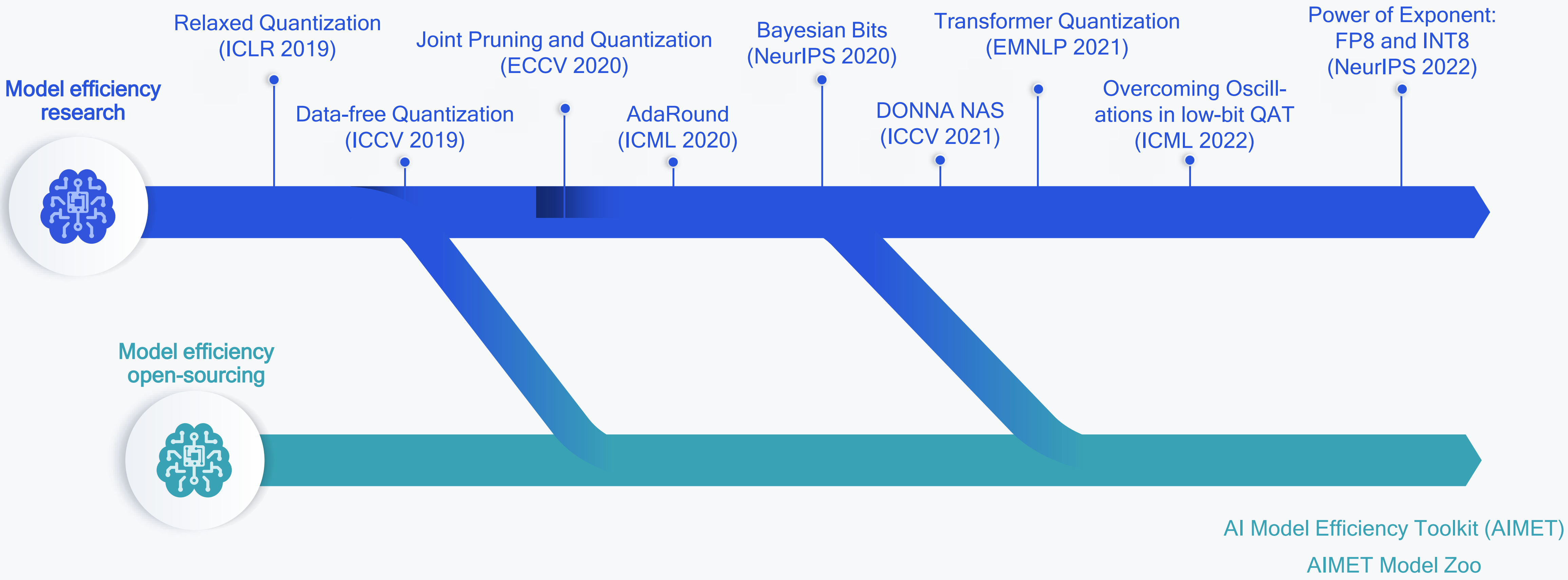
Learning to compile AI models for efficient hardware execution

Conditional compute

Learning to execute only parts of a large inference model based on the input

Neural architecture search

Learning to design smaller neural networks that are on par or outperform hand-designed architectures on real hardware


Model efficiency research
Relaxed Quantization (ICLR 2019)
Data-free Quantization (ICCV 2019)
Joint Pruning and Quantization (ECCV 2020)
AdaRound (ICML 2020)
Bayesian Bits (NeurIPS 2020)
DONNA NAS (ICCV 2021)
Transformer Quantization (EMNLP 2021)
Overcoming Oscill-ations in low-bit QAT (ICML 2022)
Power of Exponent: FP8 and INT8 (NeurIPS 2022)
Model efficiency open-sourcing
AI Model Efficiency Toolkit (AIMET)
AIMET Model Zoo

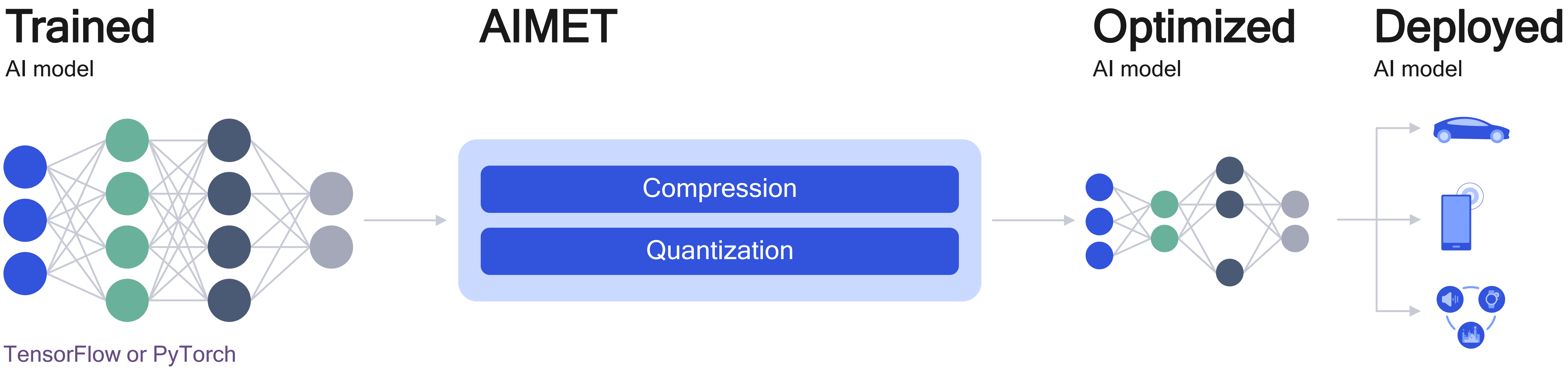
Driving the industry toward integer inference and power-efficient AI

Leading model efficiency research and fast commercialization

AIMET and AIMET Model Zoo are products of Qualcomm Innovation Center, Inc.

AIMET makes AI models small

State-of-the-art quantization and compression techniques from Qualcomm AI Research



Github: https://github.com/quic/aimet

Features:

- State-of-the-art network compression tools
- State-of-the-art quantization tools
- Support for both TensorFlow and PyTorch
- Benchmarks and tests for many models
- Developed by professional software developers

AIMET

Providing advanced model efficiency features and benefits

Benefits

- Lower power
- Lower memory bandwidth
- Maintains model accuracy
- Lower storage
- Higher performance
- Simple ease of use

Features

Quantization

State-of-the-art INT8 and INT4 performance

Quantization simulation

Post-training quantization (PTQ) methods:

- Data-Free Quantization
- Adaptive Rounding (AdaRound),
- Automatic Mixed Precision (AMP)
- AutoQuant

Quantization-aware training (QAT)

Compression

Efficient tensor decomposition and removal of redundant channels in convolution layers

Spatial singular value decomposition (SVD)

Channel pruning

Visualization

Analysis tools for drawing insights for quantization and compression

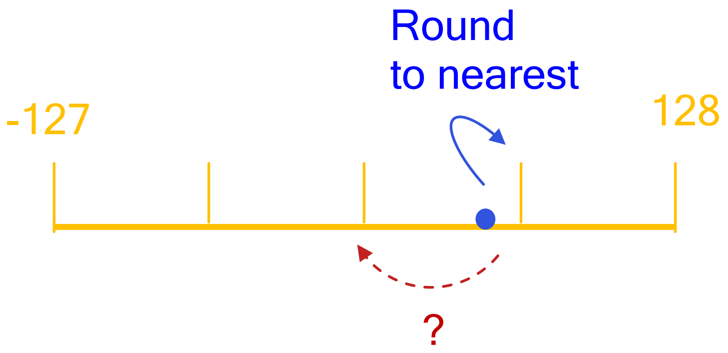
Weight ranges

Per-layer compression sensitivity

AdaRound: Adaptive Rounding for Better Quantization

ICML'20 paper

Rounding-to-the-nearest is not optimal



Object Detection

Configuration	mAP
Floating point	82.20
Nearest Rounding - 8-bit weights, 8-bit activations	49.85
AdaRound - 8-bit weights, 8-bit activations	81.21

mAP: Mean Average Precision

Semantic Seg. (Deeplabv3)

Configuration	mIOU
Floating point	72.94
Nearest Rounding - 4-bit weights, 8-bit activations	6.09
AdaRound - 4-bit weights, 8-bit activations	70.86

mIOU: Mean Intersection Over Union

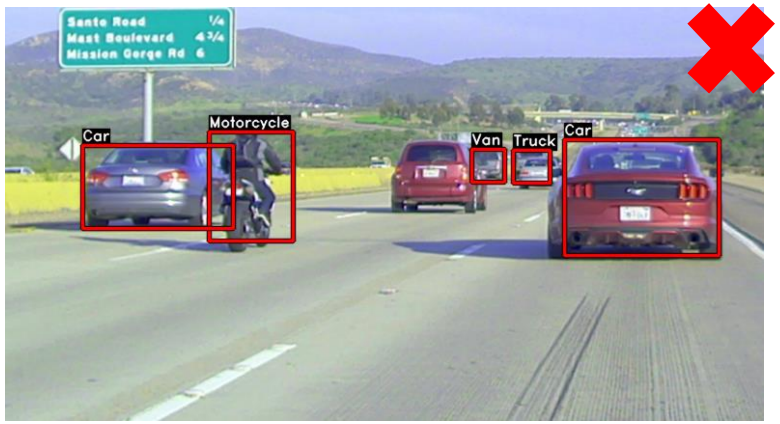
AdaRound optimizes the network weights without model fine-tuning

$$\underset{\mathbf{V}}{\arg\min} \quad \left\|\mathbf{Wx} - \widetilde{\mathbf{W}}\mathbf{x}\right\|_F^2 + \lambda f_{reg}(\mathbf{V})$$

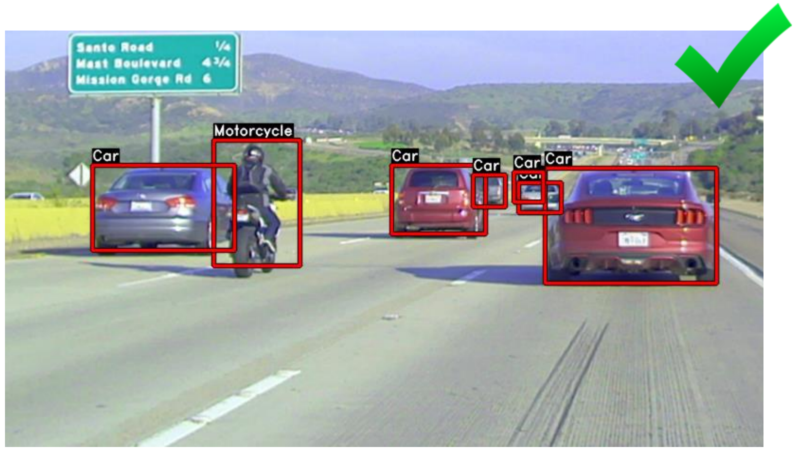
AdaRound Results

- Poor baseline INT8 quantization performance
- AdaRound performance within 1% of FP32

INT8, Baseline

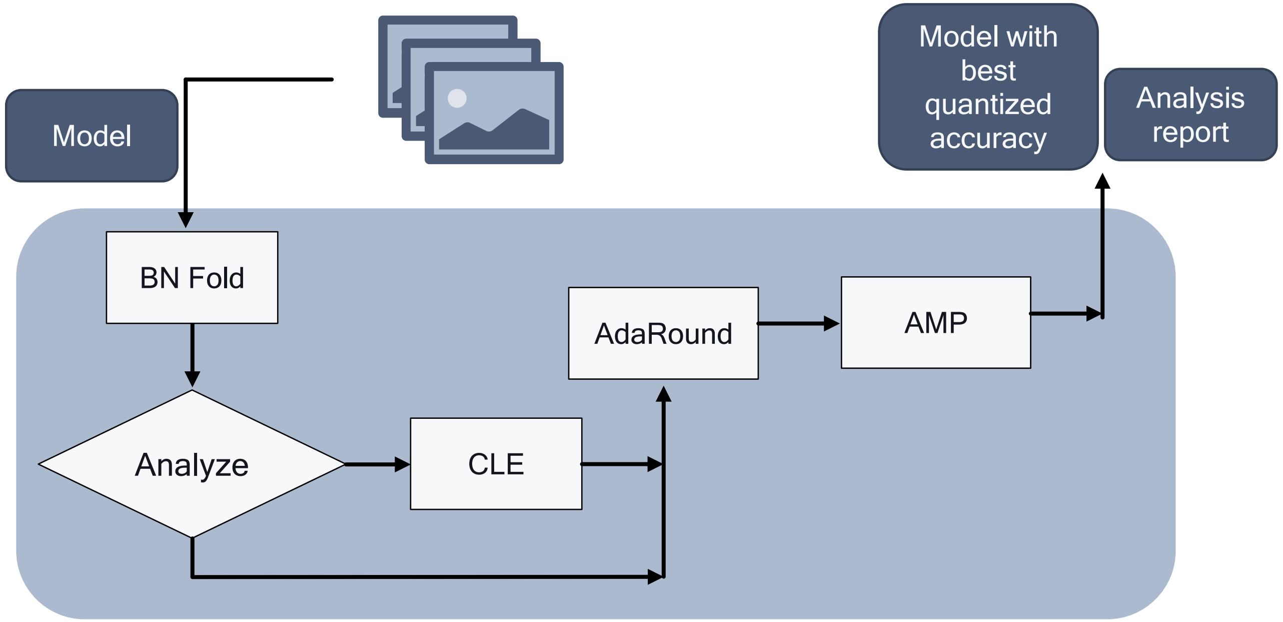


INT8, AdaRound



AutoQuant simplifies post-training quantization

- Analyzes the model
- Applies the best sequence of already existing post-training quantization (PTQ) features
- Returns the best accuracy model with analysis report
- A simple, blackbox, push-button solution

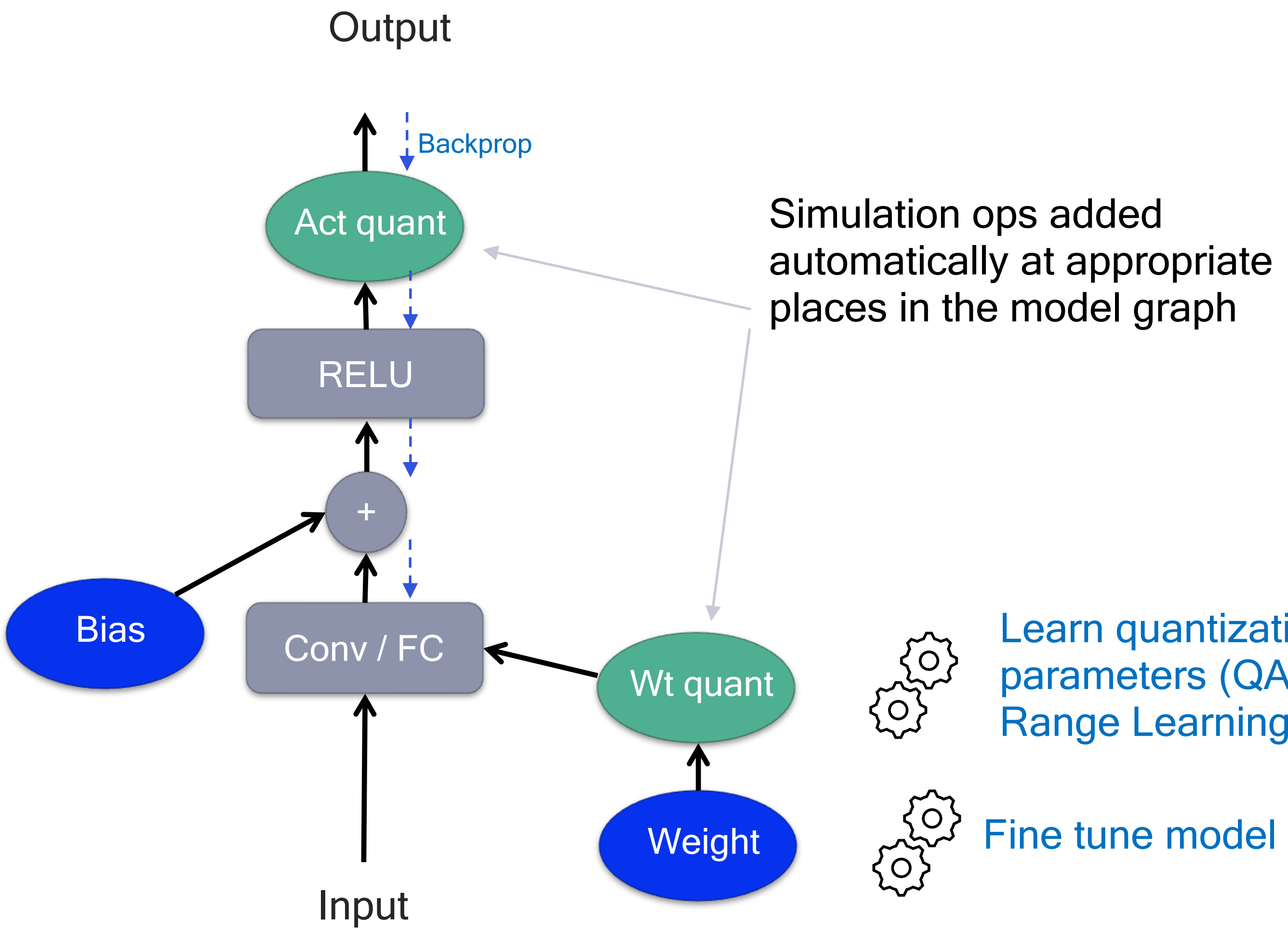


AIMET Quantization Aware Training

Simulate quantization noise in the forward pass and fine-tune for improved robustness

Provides accurate prediction of on-target performance by HW/run-time awareness

INT8 performance typically within 0.5-1% of FP32 performance



Simulation ops added automatically at appropriate places in the model graph

Learn quantization parameters (QAT with Range Learning)

Fine tune model weights

AIMET Model Zoo includes popular quantized AI models

Accuracy is maintained for INT8 and INT4 models – less than 1% loss*

<1% Loss in accuracy*

TensorFlow

ResNet-50 (v1)	Top-1 accuracy*	FP32	75.21%	INT8	74.96%
MobileNet-v2-1.4	Top-1 accuracy*	FP32	75%	INT8	74.21%
EfficientNet Lite	Top-1 accuracy*	FP32	74.93%	INT8	74.99%
ResNet-50 (v1)	mAP*	FP32	0.2469	INT8	0.2456
RetinaNet	mAP*	FP32	0.35	INT8	0.349
Pose estimation	mAP*	FP32	0.383	INT8	0.379
SRGAN	PSNR*	FP32	25.45	INT8	24.78

PyTorch

MobileNetV2	Top-1 accuracy*	FP32	71.67%	INT8	71.14%
EfficientNet-lite0	Top-1 accuracy*	FP32	75.42%	INT8	74.44%
DeepLabV3+	mIoU*	FP32	72.62%	INT8	72.22%
MobileNetV2-SSD-Lite	mAP*	FP32	68.7%	INT8	68.6%
Pose estimation	mAP*	FP32	0.364	INT8	0.359
SRGAN	PSNR	FP32	25.51	INT8	25.5
DeepSpeech2	WER*	FP32	9.92%	INT8	10.22%
ABPN	PSNR	FP32	32.75	INT8	32.69

*: Comparison between FP32 model and INT8 model quantized with AIMET. For further details, check out: https://github.com/quic/aimet-model-zoo/

Transformer Quantization

Model	FP32	INT8
BERT-base-uncased	82.73 (GLUE)	82.53
DistilBERT-base-uncased	80.35 (GLUE)	79.81
mobileBERT	81.24 (GLUE)	81.27
VIT (vision transformer)	81.30	81.50

AIMET quantizes transformers with high accuracy, comparable to FP32

Comparison between FP32 model and model quantized with AIMET

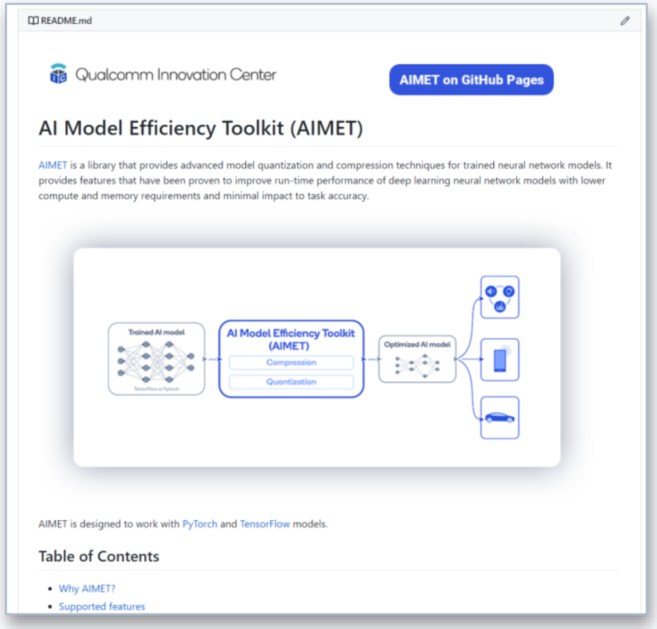
AIMET enables accurate INT4 (4-bit weights, 8-bit activations) for wide range of use cases

Task	Model	FP32	INT W4A8
Classification	ResNet50	76.10%	75.4%
	ResNet18	69.75%	68.96%
	EfficientNet-Lite	75.31%	74.33%
	Regnext	78.30%	77.20%
Segmentation	Deeplabv3 (RN-50)	76.07%	75.91%
Super-resolution	ABPN	31.97 dB	31.67 (dB)
Pose detection	PoseNet (HRNet-32)	0.765	0.763

With better PTQ and QAT techniques, more models will achieve better power efficiency

AIMET

State-of-the-art quantization and compression techniques



README.md

Qualcomm Innovation Center

AIMET on GitHub Pages

AI Model Efficiency Toolkit (AIMET)

AIMET is a library that provides advanced model quantization and compression techniques for trained neural network models. It provides features that have been proven to improve run-time performance of deep learning neural network models with lower compute and memory requirements and minimal impact to task accuracy.

AIMET is designed to work with PyTorch and TensorFlow models.

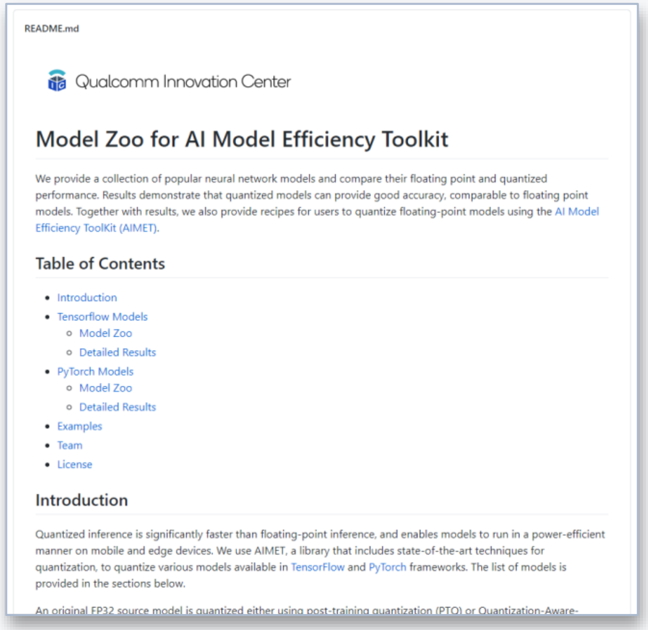
Table of Contents

- Why AIMET?
- Supported features

github.com/quic/aimet

AIMET Model Zoo

Accurate pre-trained 8-bit quantized models



README.md

Qualcomm Innovation Center

Model Zoo for AI Model Efficiency Toolkit

We provide a collection of popular neural network models and compare their floating point and quantized performance. Results demonstrate that quantized models can provide good accuracy, comparable to floating point models. Together with results, we also provide recipes for users to quantize floating-point models using the AI Model Efficiency ToolKit (AIMET).

Table of Contents

- Introduction
- Tensorflow Models
  - Model Zoo
  - Detailed Results
- PyTorch Models
  - Model Zoo
  - Detailed Results
- Examples
- Team
- License

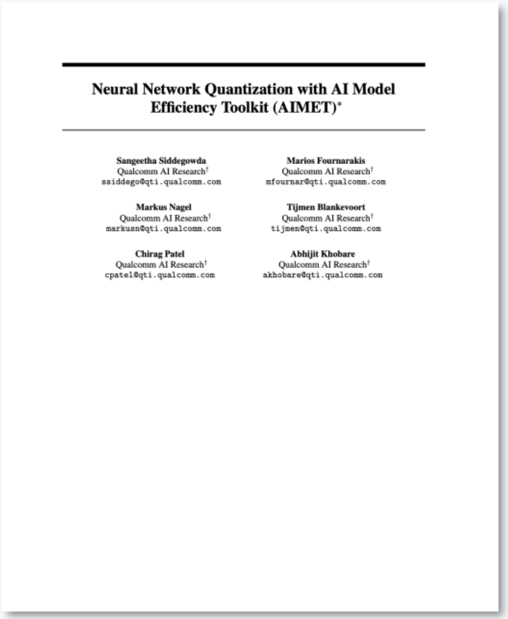
Introduction

Quantized inference is significantly faster than floating-point inference, and enables models to run in a power-efficient manner on mobile and edge devices. We use AIMET, a library that includes state-of-the-art techniques for quantization, to quantize various models available in TensorFlow and PyTorch frameworks. The list of models is provided in the sections below.

An original FP32 source model is quantized either using post-training quantization (PTQ) or Quantization-Aware-

github.com/quic/aimet-model-zoo

Quantization whitepaper



Neural Network Quantization with AI Model Efficiency Toolkit (AIMET)*

Sangeetha Siddegowda
Qualcomm AI Research†
ssiddego@qti.qualcomm.com

Marios Fournarakis
Qualcomm AI Research†
mfournar@qti.qualcomm.com

Markus Nagel
Qualcomm AI Research†
markusn@qti.qualcomm.com

Tijmen Blankevoort
Qualcomm AI Research†
tijmen@qti.qualcomm.com

Chirag Patel
Qualcomm AI Research†
cpatel@qti.qualcomm.com

Abhijit Khobare
Qualcomm AI Research†
akhobare@qti.qualcomm.com

arxiv.org/abs/2201.08442

Explore our open-source projects and tools

Qualcomm AI Stack

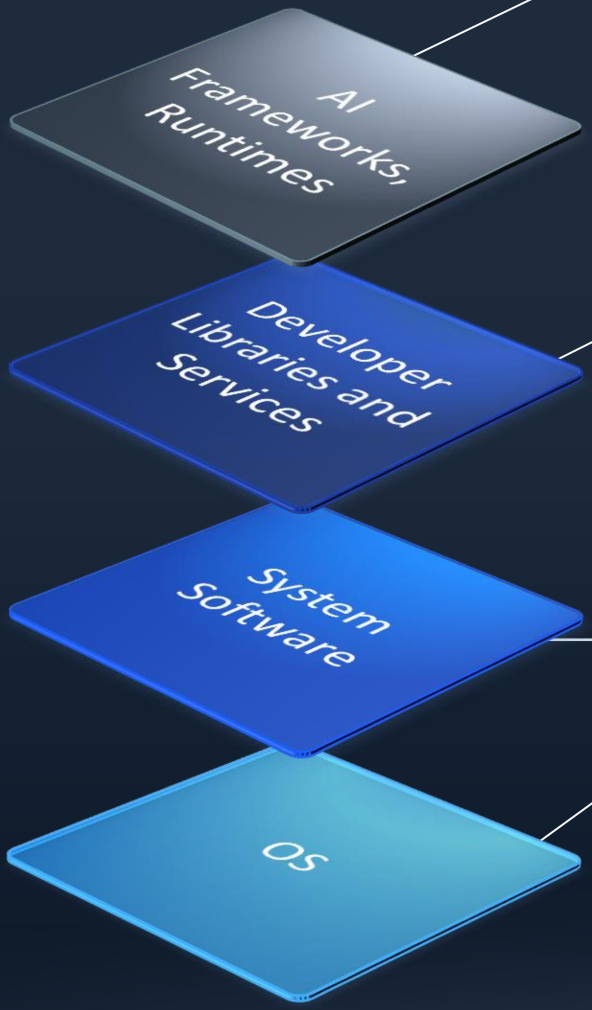
Tools:

Qualcomm AI Model Studio

- AIMET
- AIMET Model Zoo
- NAS
- Model analyzers

Infrastructure:

- Prometheus
- kubernetes
- docker



AI Frameworks

TensorFlow PyTorch ONNX

AI Runtimes

Qualcomm® Neural Processing SDK ONNX RUNTIME TF Lite Micro Direct ML TF Lite

Qualcomm® AI Engine Direct (QNN)

Math Libraries Compilers Virtual platforms

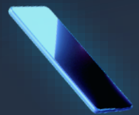
Profilers & Debuggers Programming Languages Core Libraries

System Interface SoC, accelerator drivers Emulation Support

android Zephyr ubuntu® CentOS QNX

Platforms

Smartphones



XR

ACPC

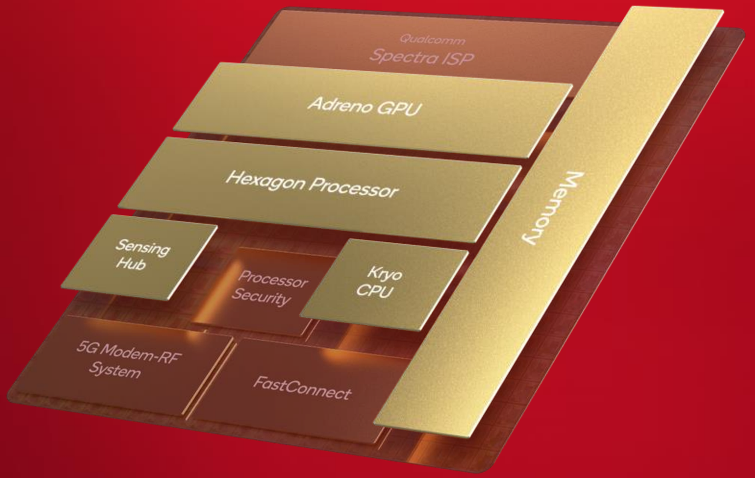
IoT

Robotics

Auto

Cloud

Snapdragon® 8 Gen 2 Mobile Platform Qualcomm® AI Engine



Micro-tile Inferencing

More Hardware Acceleration

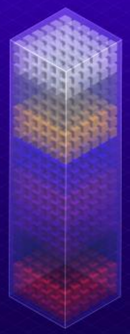
2X Tensor Accelerator Performance

up to **4.35X** Performance improvement*

60% More power efficient

Multi-language Translation

World's 1ST **INT4** support



HEXAGON DIRECT LINK

ArcSoft®

AI Cinematic Mode

Qualcomm AI Stack

Feature updates

Performance Improvements

INT4 Support

UNREAL ENGINE

AI bot plug-in

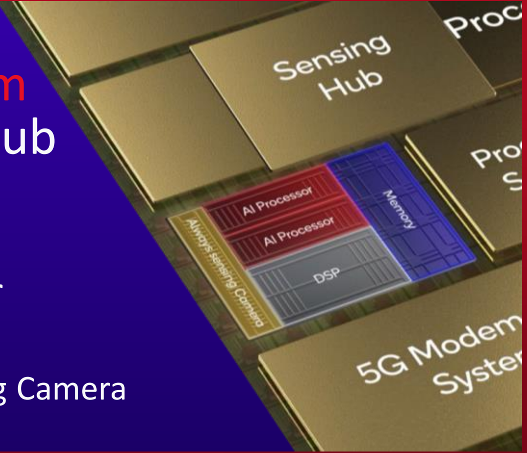
AI Frameworks & Runtimes
Developer Libraries and Services
System Software
OS

Qualcomm Innovators Developer Kit

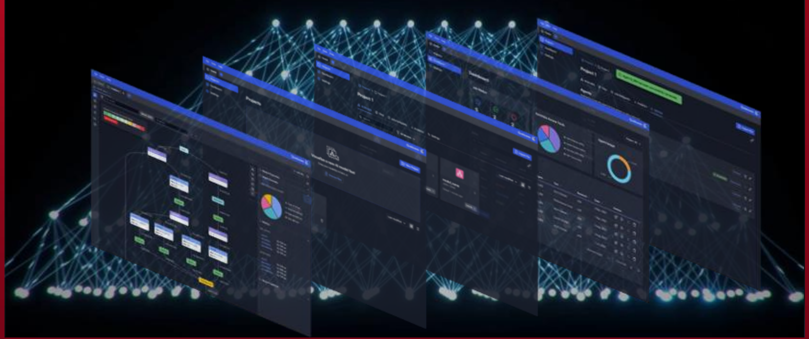
Qualcomm Sensing Hub

Dual AI Processor

Always-sensing Camera



Qualcomm AI Studio



Thank you

Qualcomm

Follow us on:

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

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

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 our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.