2023 embedded VISION SUMMIT

Accelerating Newer ML Models Using Qualcomm® AI Stack

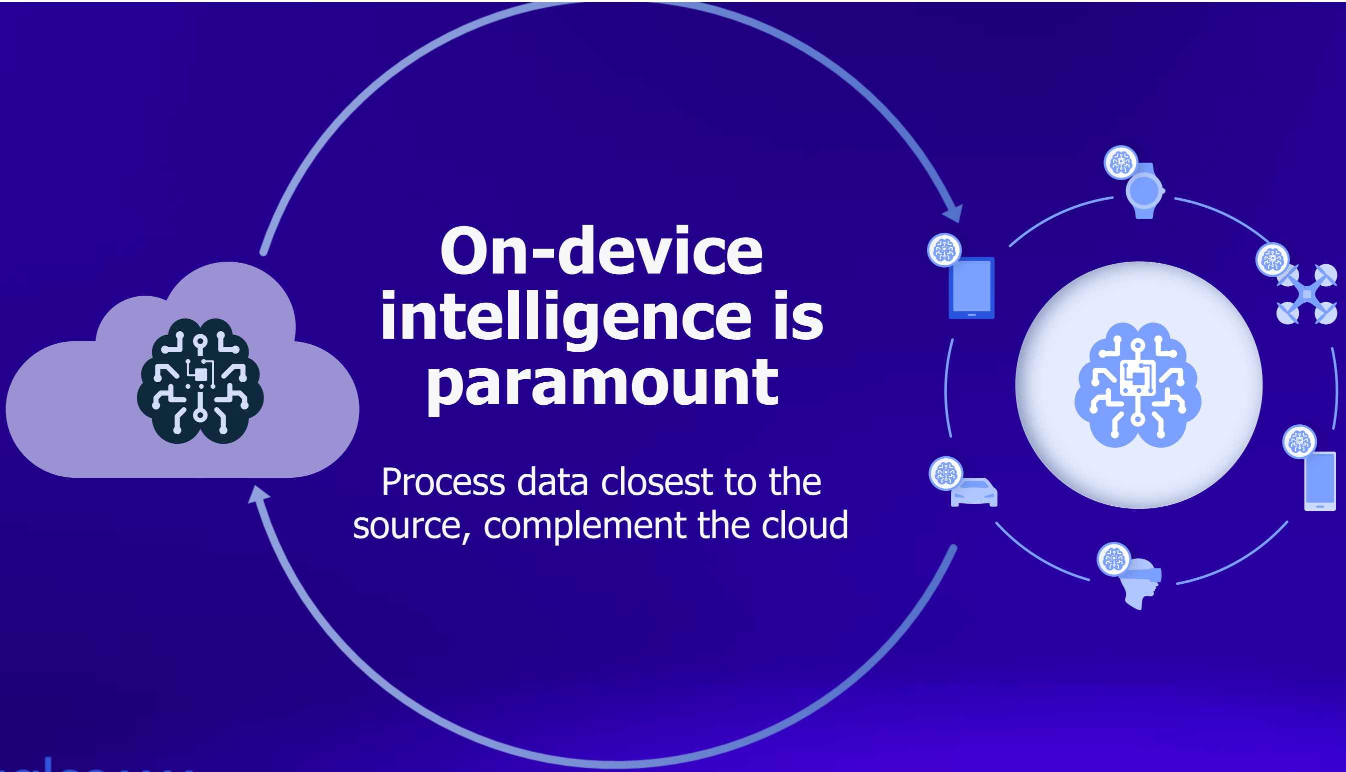
Dr. Vinesh Sukumar

Sr Director – AI/ML Product
Qualcomm Technologies, Inc.

Qualcomm

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

Center of Gravity Moving to the Edge…



On-device intelligence is paramount

Process data closest to the source, complement the cloud

Historically

- Privacy
- Reliability
- Low latency
- Efficient use of network bandwidth

Increased Demand

- Personalization
- Security
- Autonomy
- Efficiency

AI Applications : Across Various Segments

Mobile

AI Assisted Imaging

- AI 3A
- Scene-based Camera Selection

Image Understanding

- Face Detection / Tracking / Features
- Object Detection / Tracking
- Body Detection / Tracking / Pose
- Human Segmentation

Beautify / Augment / Gaming

- Scene-based Image Enhancement

Image Processing

- AI based NR or Image SR
- Scene-based Camera Selection

Audio

- Real time language
- Natural language processing (NLP)

Modem

- Sensor Fusion (Cont. awareness)
- Modem RF E2E (Tuners..)

CSS

Robotics

- Autonomous navigation
- Obstacle Avoidance

Retail

- Visitor/Face/Gesture Recognition
- Object/People Detection and Counting
- Barcode decoding

Transportation

- License plate recognition
- Face and facial landmark detection
- Drowsiness detection

Smart Devices

- Object/People detection
- Speaker detection

Smart Buildings

- People Tracking
- Access Control

Manufacturing/Logistics

- Predictive maintenance
- Energy management with Asset demand

Compute

Productivity

- Background based noise cancellation on Audio (inbound and outbound)
- Segmentation/Blur/Super Resolution on Video

Privacy & Security

- Automatic screen unlock and login
- Privacy alert
- Guard mode

Content Creation & Gaming

- Gaming with gesture control
- Gaming with voice commands
- Intelligent highlight videos
- Game play improvement

Performance & Efficiency

- Power and Screen optimization

Cloud

Data Centers

- Natural language processing
- Computer vision
- Recommendation system

Edge Compute

- Theft detection
- Face/body/license plate detection / recognition
- Image classification and segmentation

XR

Metaverse

- Person and Object Detection
- Recommendation Engine & Chatbots
- Multilingual translation (speech-to-speech)
- Neural Super Resolution
- Content Summarization

Auto

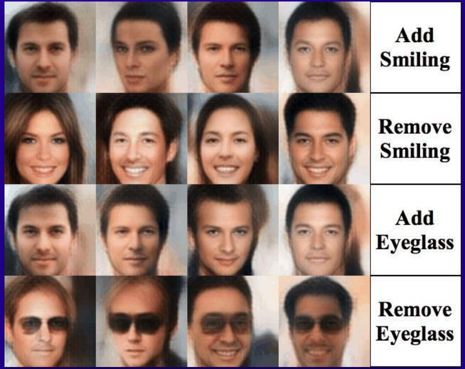
IVI

- Occupancy monitoring system (OMS)
- Driver monitoring system (DMS)
- Surround perception
- Audio Command & Control

ADAS (Up to L4)

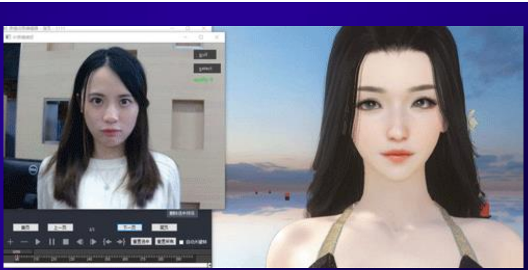
- Highway driving assist
  - Front collision warning
  - lane departure,
  - Traffic jam assist
  - Auto lane change
  - Auto lane merge
  - Traffic light recognition
  - Construction zones
  - Urban autonomous driving
- Parking assist
  - Person detection,
  - Perception
  - Valet parking
- Driver monitoring

Emerging AI Models – For the Various Markets

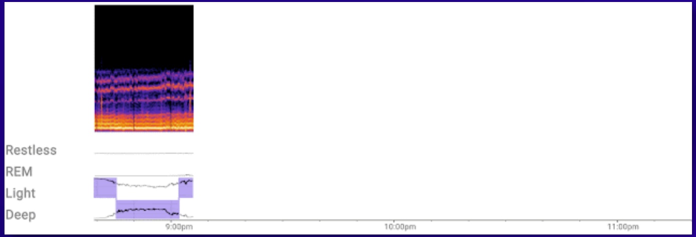


Generative networks
(Image to Image Transformation)

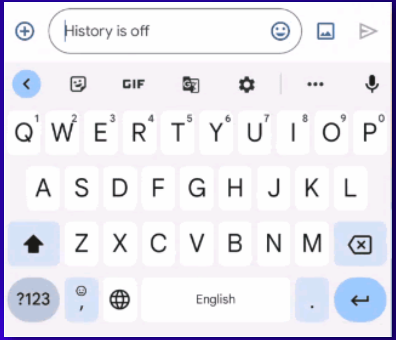
Emerging Deep Learning Models



Canvas networks
(Virtual Transformation for Avatars)



Time series networks
(Behavior to Text Transformation)



Transformer networks (NLP/NLU)
(Sequence to Sequence transformation)

Vision: Accelerate Solution Deployment

Performance

Accelerate "out of box" operator functionality and performance

Scalability

Ability to have programming consistency from Cloud to Edge

Tools

Accelerate AI solution deployment with investment in tools

Innovation

Innovation to drive product leadership (Pre-emption, DFS, Multi chaining)

STEP: 1 -> Model Optimization Using NAS

Integrated into Qualcomm® Software Stack

How →

Search Space

Space of allowable architectures (Structure, operations, connectivity)

Search Algorithm

Sampling populations of good architecture candidates

Evaluation Strategy

Estimate performance of sampled architecture

NAS Results: Observations from ML Models

Category	Model	Task	Dataset	Results
CNNs	EfficientNet-B0	Image Classification	ImageNet	**+1.0%** accuracy **33%** latency reduction
	ResNet-18		ImageNet	**+2.2%** accuracy **31%** latency reduction
	RetinaNet	2D Object Detection	Pascal	**+1.5** mAP accuracy **11%** latency reduction
	EfficientDet-D0		COCO	**+0.8** mAP accuracy **30%** latency reduction
RNNs	CRNN	Keyword Spotting	Google Speech Commands v2	**+1.0%** accuracy similar model size
Transformers	MobileBERT	Question & Answering	SQuAD v1.1	On-par accuracy **12%** latency reduction

STEP: 2 -> New Techniques to Quantize Models

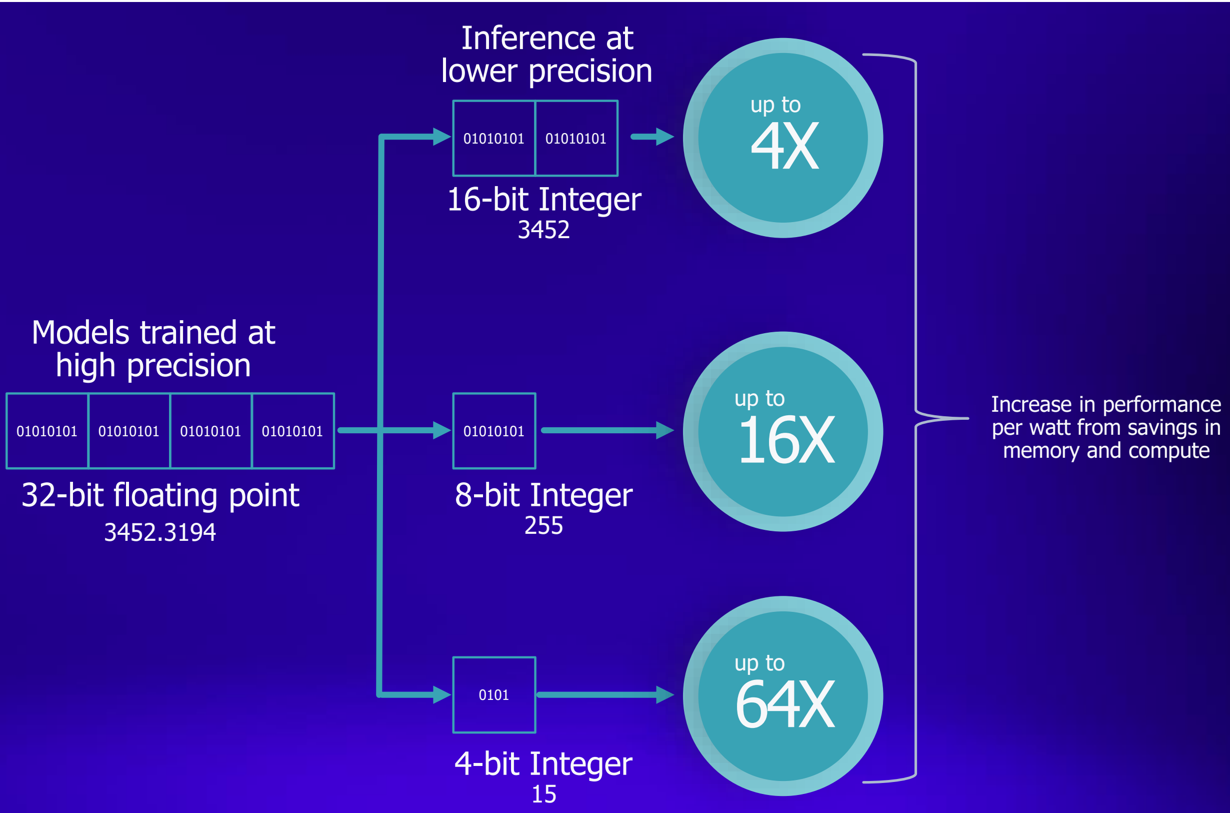
Integrated into Qualcomm Software Stack

Automated reduction in precision of weights and activations while maintaining accuracy

Promising results show that low-precision integer inference can become widespread

Virtually the same accuracy between a FP32 and quantized AI model through:

- Automated, data free, post-training methods
- Automated training-based mixed-precision method



© 2023 Qualcomm Technologies Inc.

1: FP32 model compared to quantized model

Pushing the Limits – For Quantization & Pruning

embedded VISION SUMMIT

Highest Focus of Attention

Data-free quantization

How can we make quantization as simple as possible?

Created an automated method that addresses bias and imbalance in weight ranges:

- ☑ No training
- ☑ Data free

SOTA 8-bit results

Making 8-bit weight quantization ubiquitous

<1% Accuracy drop for MobileNet V2 against FP32 model

AdaRound

Is rounding to the nearest value the best approach for quantization?

Created an automated method for finding the best rounding choice:

- ☑ No training
- ☑ Minimal unlabeled data

SOTA 4-bit weight results

Making 4-bit weight quantization ubiquitous

<2.5% Accuracy drop for MobileNet V2 against FP32 model

Bayesian bits

Can we quantize layers to different bit widths based on precision sensitivity?

Created a novel method to learn mixed-precision quantization:

- ☑ Training required
- ☑ Training data required
- ☑ Jointly learns bit-width precision and pruning

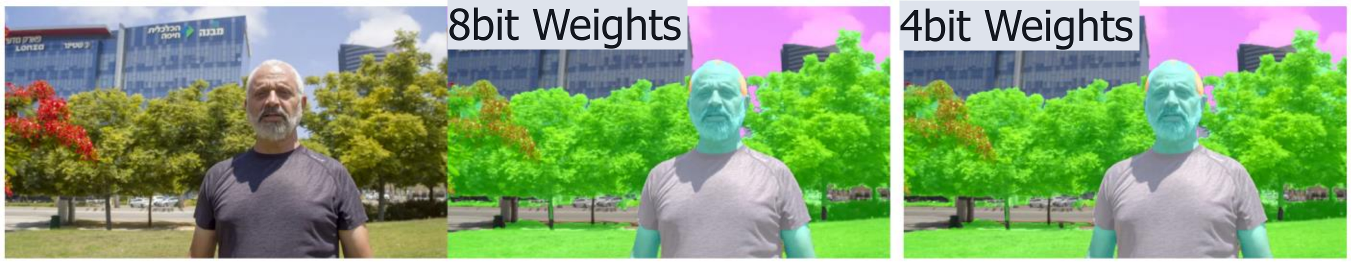
SOTA mixed-precision results

Automating mixed-precision quantization and enabling the tradeoff between accuracy and kernel bit-width

<1% Accuracy drop for MobileNet V2 against FP32 model for mixed precision model with **computational complexity equivalent to a 4-bit weight model**

Qualcomm

Moving towards W4A8 – Newer ML Models



Segmentation Models: Seeing >20% power + >40% in memory footprint saving

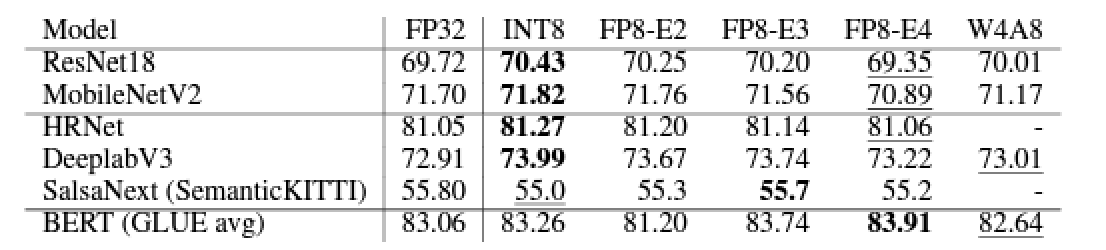
Model	FP32	INT4 Accuracy	Comments
ResNet50	76.1%	75.4%	Using Post-training Quantization (PTQ)
ResNet18	69.8%	69%	
EfficientNet-Lite	75.3%	74.3%	
Regnext	78.3%	77.2%	
Mobilenet-v2	71.7%	71.3%	Using Quantization Aware Training (QAT)

With better PTQ and QAT techniques, increasingly more models will be able to use W4A8, resulting in better energy efficiency → **This is going to be major push for AI solution deployment on the edge**

Need for FP8
– Is this Needed for ML Model Inference?

- Strong participation from many silicon vendors on driving FP8 engagements
  - Various E/M (exponent/mantissa) ratios to support dynamic range for data representation
- FP8 is an appealing potential speed-up for the costly and time-intensive training procedures in deep learning
- **Need for Inference (observations) :**
  - The hardware implementation of the FP8 format is somewhere between 50% to 180% less efficient than INT8 in terms of chip area and energy usage
  - Can we convert FP8 to INT8 with good accuracy?

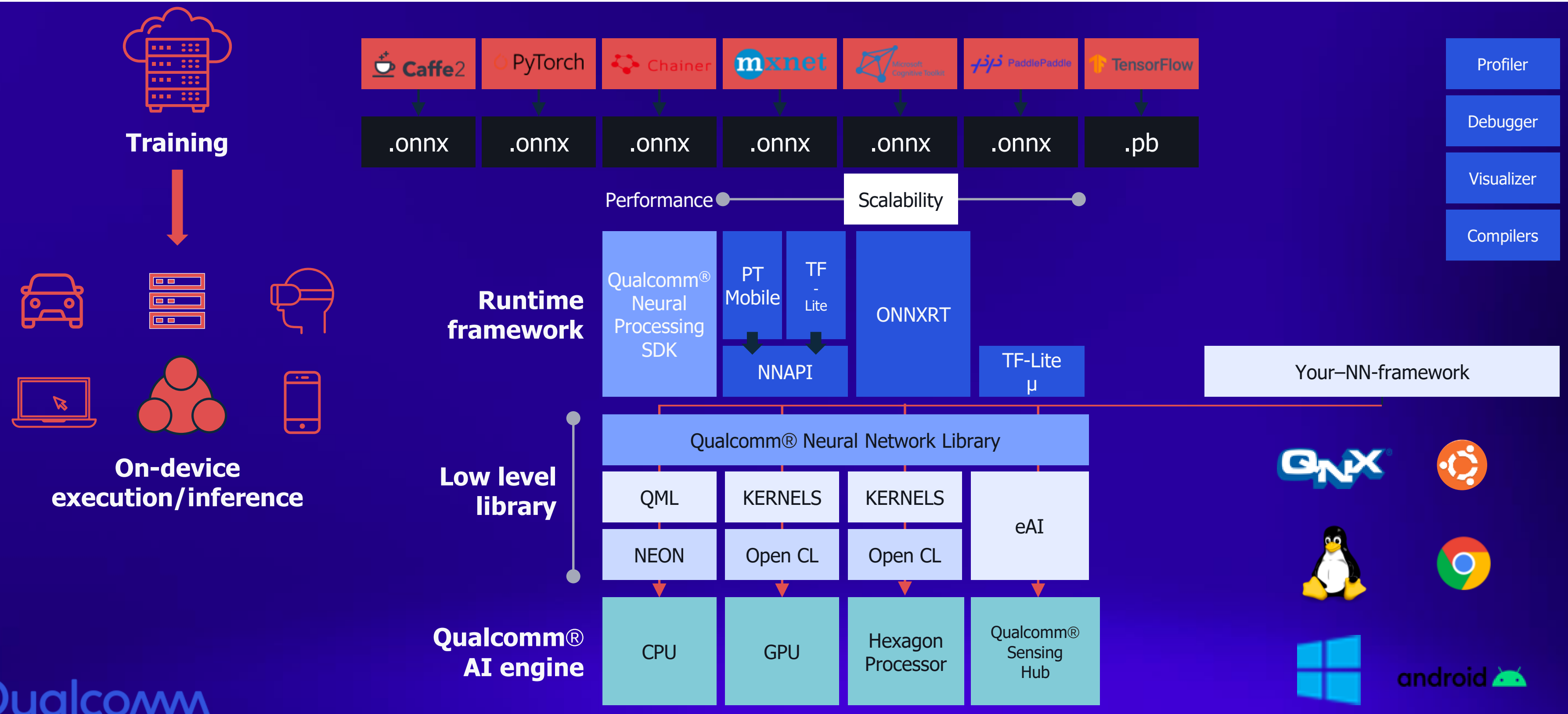
Published in the Qualcomm Technologies "FP8" White Paper



Model	FP32	INT8	FP8-E2	FP8-E3	FP8-E4	W4A8
ResNet18	69.72	**70.43**	70.25	70.20	69.35	70.01
MobileNetV2	71.70	**71.82**	71.76	71.56	70.89	71.17
HRNet	81.05	**81.27**	81.20	81.14	81.06	-
DeeplabV3	72.91	**73.99**	73.67	73.74	73.22	73.01
SalsaNext (SemanticKITTI)	55.80	55.0	55.3	**55.7**	55.2	-
BERT (GLUE avg)	83.06	83.26	81.20	83.74	**83.91**	82.64

STEP: 3 → Performance and Scalability Support - Application Deployment

Integrated into Qualcomm Software Stack



Training

On-device execution/inference

Caffe2 | PyTorch | Chainer | mxnet | Microsoft Cognitive Toolkit | PaddlePaddle | TensorFlow

.onnx | .onnx | .onnx | .onnx | .onnx | .onnx | .pb

Performance — Scalability

Profiler

Debugger

Visualizer

Compilers

Runtime framework

Qualcomm® Neural Processing SDK

PT Mobile

TF - Lite

NNAPI

ONNXRT

TF-Lite μ

Your–NN-framework

Low level library

Qualcomm® Neural Network Library

QML | KERNELS | KERNELS | eAI

NEON | Open CL | Open CL

Qualcomm® AI engine

CPU | GPU | Hexagon Processor | Qualcomm® Sensing Hub

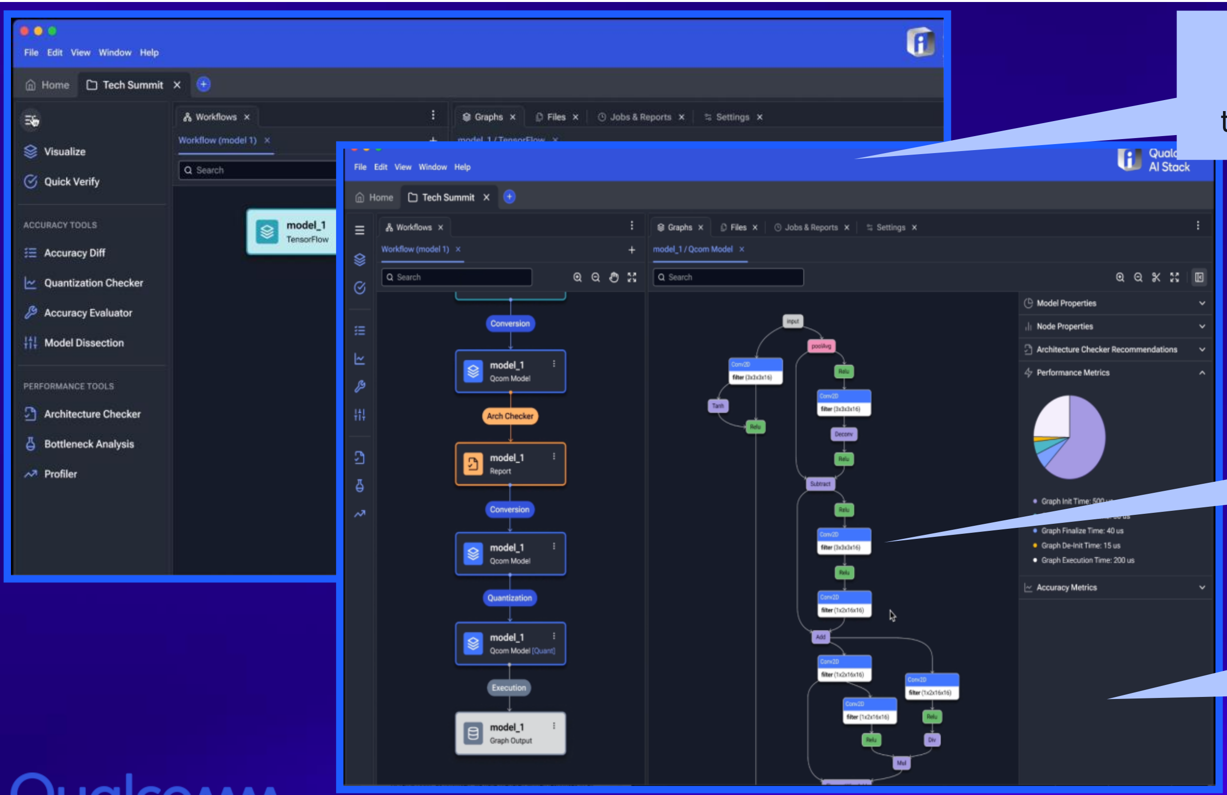
QNX

android

Qualcomm

Qualcomm Model Studio: Accelerating ML Model Deployment

Integrated into Qualcomm Software Stack



- Graph Init Time: 500 us
- Graph Finalize Time: 40 us
- Graph De-Init Time: 15 us
- Graph Execution Time: 200 us

Workflow panel

Shows steps in a workflow including tools, artifacts and their relationships

Graph panel

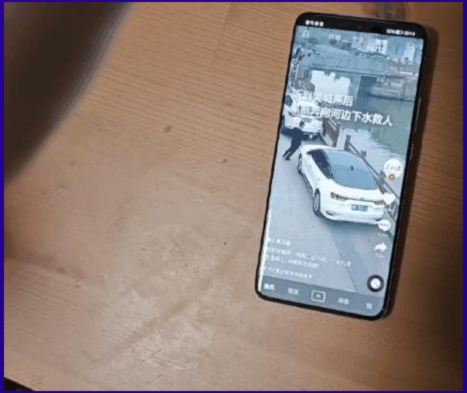
Model visualization, node information (precision, etc.)

Metrics panel

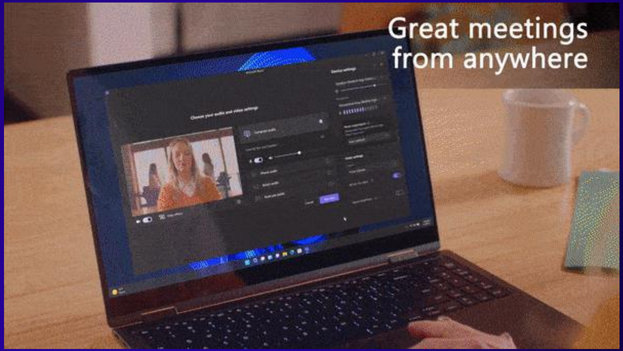
Detailed information on selected model, nodes including performance info from execution

Qualcomm AI Studio
Qualcomm
AI Stack

Recently Deployed Applications – Using Qualcomm AI Stack



Industry’s first low power gesture control + context awareness to service recommendation – Launched on Honor



Windows 11 features for video + audio AI – Launched on ThinkPad X13S

Conclusions

- AI applications expanding beyond modalities of computer vision to linguistics, communication, commerce and language understanding

- With evolution of AI applications, this continues to stress on support for new DL architectures & models

- Qualcomm AI Stack expands to enable support for any developer and drive innovation in performance, latency, QoS among others. Focus on
  - Advanced quantization mechanics
  - Support for newer data types
  - Neural architecture support
  - Flexible run time for performance & portability

Resources

Qualcomm® Mobile AI

Mobile AI | On-Device AI | Qualcomm

Qualcomm Technologies & Google NAS

Qualcomm Technologies and Google Cloud Announce Collaboration on Neural Architecture Search for the Connected Intelligent Edge | Qualcomm

Dr. Vinesh Sukumar
Senior Director, Product Management – AI/ML
vinesuku@qti.qualcomm.com

2023 Embedded Vision Summit

4:15 pm: Develop Next-Gen Camera Apps Using Snapdragon Computer Vision Technologies
- Judd Heape, VP of Product Management for Camera, Computer Vision and Video Technology, Qualcomm Technologies

Qualcomm Wireless Academy

Fundamentals of AI

Available for free until October 2023

THANK YOU

Qualcomm