AI Firsts from Qualcomm AI Research

The group has pioneered new technologies from on-device learning to wireless AI and has ambitious plans for the next set of “firsts”.

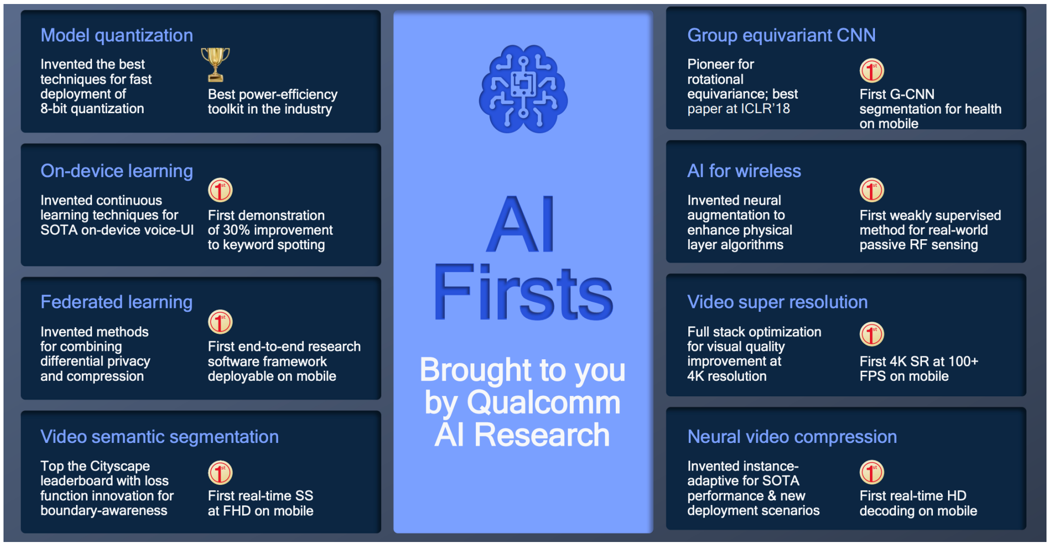
Whenever people take photos or speak to a digital assistant using a mobile phone, they often don’t realize that they just took advantage of Artificial Intelligence (AI). If they think of AI at all, it is typically in the context of Autonomous Vehicles or perhaps Facebook’s (Meta’s) massive data centers. While AI is becoming ubiquitous and distributed across edge devices and cloud servers, many challenges remain to realize the connected intelligent edge vision CEO Cristiano Amon has for AI to enable automated perception, reasoning, and action on edge devices.

For AI to enable the levels of automation and personalization Mr. Amon and his team envisions, AI hardware and software must become much smaller, faster, more efficient, lower power, and able to learn continuously at the edge in the real world. This provides the perfect complement to remote processing in the cloud, whose reach has been further advanced through Qualcomm’s 5G technology. These are not just engineering challenges; progress and breakthroughs must come from fundamental scientific, applied, and platform research. This is where Qualcomm AI Research comes in, built with Qualcomm's core culture of innovation and evolved from over a decade of investment in R&D focused on machine learning and AI.

AI Firsts from Qualcomm AI Research

Qualcomm AI Research recently held a webinar where Jilei Hou, VP of engineering and the head of AI R&D at Qualcomm AI Research, outlined “firsts” in eight research areas for which they are justifiably proud. Their novel AI research and full-stack AI optimizations have pushed the AI industry forward and enabled first-ever proof-of-concept demonstrations on commercial mobile devices.

Their AI firsts include 8-bit model quantization leading to the best power-efficiency toolkit in the industry (the AI Model Efficiency Toolkit, or AIMET), on-device learning demonstrating a 30% improvement in keyword detection, federated learning with an end-to-end software framework, video semantic segmentation in real-time, group equivariant CNNs, AI for wireless with the first weakly supervised method for passive RF sensing, video super-resolution at 4K at 100 FPS on mobile, and neural video compression with the first real-time HD decoding on mobile. You can check out the rest of Qualcomm AI Research’s firsts by downloading this presentation.



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

AI Firsts

Brought to you by Qualcomm AI Research

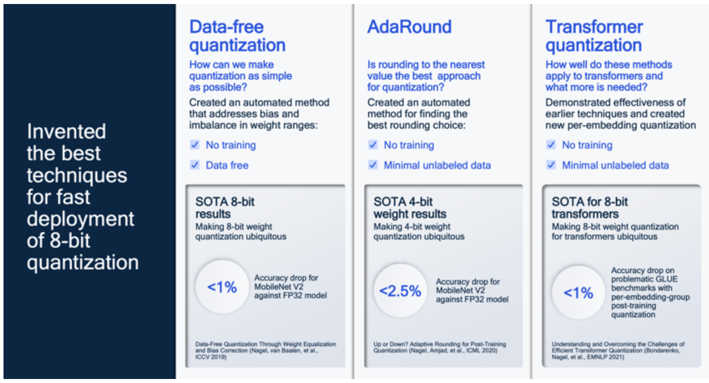
Group equivariant CNN
Pioneer for rotational equivariance; best paper at ICLR’18
First G-CNN segmentation for health on mobile

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

Artificial Intelligence at Qualcomm, like most everything else in the company, begins with mobile. And that heritage sets a high bar, since mobile devices are battery-powered, lightweight consumer devices with limited processing, memory, and I/O. The ongoing research underway at Qualcomm AI Research is centered around three primary domains: fundamental research, applied research, and platform research. The latter covers topics such as power efficiency and on-device learning, which are essential in mobile but have also benefited Qualcomm in other vertical markets like IOT, XR, and automotive. But efficiency doesn't stop with hardware; Qualcomm AI Research is also advancing model design, compression, quantization, algorithms, and software tools to attack the problem through full-stack research.



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

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

Data-Free Quantization Through Weight Equalization and Bias Correction (Nagel, van Baalen, et al., ICCV 2019)

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

Up or Down? Adaptive Rounding for Post-Training Quantization (Nagel, Amjad, et al., ICML 2020)

Transformer quantization

How well do these methods apply to transformers and what more is needed?

Demonstrated effectiveness of earlier techniques and created new per-embedding quantization

- ✓ No training
- ✓ Minimal unlabeled data

SOTA for 8-bit transformers

Making 8-bit weight quantization for transformers ubiquitous

<1% Accuracy drop on problematic GLUE benchmarks with per-embedding-group post-training quantization

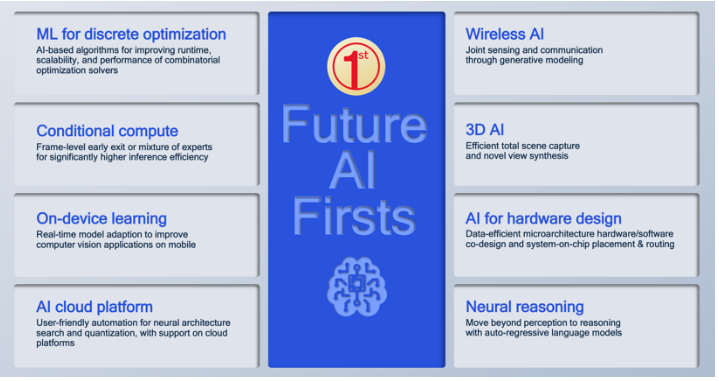
Understanding and Overcoming the Challenges of Efficient Transformer Quantization (Bondarenko, Nagel, et al., EMNLP 2021)

Efficiency comes in part from using less data in the model and quantization. If one can run a model with 4-bit integer math, it will be up to 64 times more efficient than using 32-bit floating point. But all that efficiency cannot come at the cost of reduced accuracy. To that point, Qualcomm AI Research has demonstrated a state-of-the-art 8-bit transformer model with less than 1% accuracy degradation.

Today, some AI processing has to be offloaded from the mobile device to a cloud service, but the trend is toward more and more processing at the edge devices. Faster response times, improved privacy, better personalization, and improved understanding within the context of the request are all benefits of on-device learning.

LOOKING TO THE FUTURE

The researchers at Qualcomm AI Research are pursuing additional new "firsts" that could result in significant value and differentiation for customers and partners. Targeting improvements in AI efficiency on mobile devices and the rest of the connected intelligent edge can pave the way for improved personalization and automation of many tasks, enabling new use cases and improving the overall user experience.



Future AI Firsts

ML for discrete optimization

AI-based algorithms for improving runtime, scalability, and performance of combinatorial optimization solvers

Conditional compute

Frame-level early exit or mixture of experts for significantly higher inference efficiency

On-device learning

Real-time model adaption to improve computer vision applications on mobile

AI cloud platform

User-friendly automation for neural architecture search and quantization, with support on cloud platforms

Wireless AI

Joint sensing and communication through generative modeling

3D AI

Efficient total scene capture and novel view synthesis

AI for hardware design

Data-efficient microarchitecture hardware/software co-design and system-on-chip placement & routing

Neural reasoning

Move beyond perception to reasoning with auto-regressive language models

CONCLUSIONS

Increased AI processing on edge devices is both inevitable and extremely valuable to improve user experience and device functionality. The challenges remain daunting, given the limited amount of processing and memory capacity on mobile platforms. Qualcomm AI Research is rising to the challenge, and these 8 "AI Firsts" demonstrate that the organization is already making significant progress, with more innovations in sight.

Important Information About This Paper

Author: Karl Freund, Founder, and Principal Analyst at Cambrian-AI Research

Inquiries:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be mentioned in the context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

This document, including any supporting materials, is owned by Cambrian-AI Research. This publication may not be reproduced, distributed, or shared in any form without Cambrian-AI Research's prior written permission.

DISCLOSURES

This document was developed with Qualcomm Technologies, Inc. (QTI) funding and support. Although the paper may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The views expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.