Advanced network quantization and compression through the AI Model Efficiency Toolkit (AIMET)

Abhijit Khobare  
Director, Engineering  
Qualcomm Technologies, Inc.

Chirag Patel  
Principal Eng./Mgr.  
Qualcomm Technologies, Inc.
Deep neural networks are energy hungry and growing fast

AI is being powered by the explosive growth of deep neural networks

2025 | Will we have reached the capacity of the human brain?
Energy efficiency of a brain is 100x better than current hardware
Power and thermal efficiency are essential for on-device AI

The challenge of AI workloads

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

Constrained mobile environment

- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations
Breadth of our research across model efficiency

Quantization
Learning to reduce bit-precision while keeping desired accuracy

Compilation
Learning to compile AI models for efficient hardware execution

Compression
Learning to prune model while keeping desired accuracy

Neural architecture search
Learning to design neural networks that are on par or outperform hand-designed architectures

Multiple vectors of attack to improve model efficiency
Leading quantization research and fast commercialization

Driving the industry towards integer inference and power-efficient AI
AIMET makes AI models small
Open-sourced GitHub project that includes state-of-the-art quantization and compression techniques from Qualcomm AI Research

If interested, please join the AIMET GitHub project: 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 storage
- Lower memory bandwidth
- Higher performance
- Maintains model accuracy
- Simple ease of use

Quantization
- State-of-the-art INT8 and INT4 performance
- Post-training quantization methods, including Data-Free Quantization and Adaptive Rounding (AdaRound) [coming soon]
- Quantization-aware training
- Quantization simulation

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
Easy to use features and APIs

- **Invoke directly from your pipeline**
  - Supports TensorFlow & PyTorch
  - Direct Algorithm API frameworks

- **User-Friendly APIs**
  ```python
  compress_model(model,
  eval_callback=object_eval,
  compress_scheme=Scheme.spatial_svd, ...)
  
equalize_model(model, ...)
  ```
AIMET

Available as an open-source project

github.com/quic/aimet

Includes documentation, code examples, video tutorials
AIMET Model Zoo includes popular quantized AI models
Accuracy is maintained for INT8 models

**TensorFlow**
- Loss in top-1 accuracy*: <0.3%
- Loss in top-1 accuracy*: <1%
- Loss in top-1 accuracy*: <0.1%
- Loss in mAP*: <0.01
- Loss in mAP*: <0.01
- Loss in mAP*: <0.7
- Loss in PSNR*: <0.7

ResNet-50 (v1)  MobileNet-v2.1.4  EfficientNet Lite  SSD MobileNet-v2
RetinaNet  Pose Estimation  SRGAN

**PyTorch**
- Loss in top-1 accuracy*: <0.6%
- Loss in top-1 accuracy*: <1%
- Loss in mIoU*: <0.4%
- Loss in mAP*: <0.1%
- Loss in mAP*: <0.05
- Loss in PSNR*: <0.01
- Loss in WER*: <0.3%

MobileNetV2  EfficientNet-lite0  DeepLabV3+
Pose Estimation  SRGAN  DeepSpeech2

* Comparison between FP32 model and INT8 model quantized with AIMET. For further details, check out: https://github.com/oppi/aimet-model-zoo
Visual difference in model accuracy is telling

For DeepLabv3+ semantic segmentation, AIMET quantization results in accurate quantization, while the baseline quantization method is inaccurate.

FP32

INT8 (AIMET quantization)

INT8 (Baseline quantization)

Baseline quantization: Post-training quantization using min-max based quantization grid
AIMET optimized quantization: Model fine-tuned using Quantization Aware Training in AIMET
Quantization

Enable AI ecosystem for efficient, fixed-point edge inference
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

Leading research to efficiently quantize AI models
What is neural network quantization?

For any given trained neural network:
- Store weights in n bits
- Compute calculations in n bits

Quantization analogy
Similar to representing the pixels of an image with less bits
Quantization in practice

\[ \text{scale} = \frac{\text{max} - \text{min}}{2^{\text{bitwidth}} - 1} \]

\[ \text{offset} = -\frac{\text{min}}{\text{scale}} \]
Model quantization problem

Classify quantization techniques into levels indicating increased effort

Bring most models into “Level 2” category using algorithm innovations, minimize “Level 3”

- **Level 1**
  - Simple 8-bit
    - Simply quantize floating point model parameters to 8-bit with no other changes.

- **Level 2**
  - Post-training
    - Model can undergo some in-place changes such as rescaling of weights or setting of better quantization ranges.

- **Level 3**
  - Fine-tuning
    - Model can be quantized but needs significant fine-tuning with datasets afterwards.

- **Level 4**
  - Architecture changes
    - Architecture needs to be modified to be suitable for quantization, and then trained from scratch.
Making quantization practical for the masses

Our research focuses on quantization techniques that maintain accuracy while reducing engineering effort.

Ideal properties for simple quantization

**No data**
Data might not be available

**No back propagation**
Retraining is time intensive and sensitive to hyper parameters

**No architecture changes**
Requires training from scratch with quantization in mind
Data-Free Quantization

Optimize models without needing labelled data or fine-tuning

Cross-Layer Equalization
Rescale weights in neighboring layers

Bias Correction
Adjust layer biases to account for quantization bias

Model
PyTorch or TF Floating-point model

AIMET
Unlabeled data samples

Model
PyTorch or TF Floating-point model

Quantized Runtime
Cross-layer equalization scales weights in neighboring layers for better quantization

We can scale two layers with a (P)Relu together to optimize it for quantization.
Highlight: Data Free Quantization technique

- Cross-layer equalization
- Bias absorption
- Weight quantization
- Bias correction
- Activation range estimation

Equalize weight ranges
Avoid high activation ranges
Measure and correct shift in layer outputs
Estimate activation ranges for quantization

<table>
<thead>
<tr>
<th>Model</th>
<th>FP32 model</th>
<th>INT8 model with DFQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet-v2 (top-1 accuracy)</td>
<td>71.72%</td>
<td>71.08%</td>
</tr>
<tr>
<td>ResNet-50 (top-1 accuracy)</td>
<td>76.05%</td>
<td>75.45%</td>
</tr>
<tr>
<td>DeepLabv3 (mIoU)</td>
<td>72.65</td>
<td>71.91</td>
</tr>
</tbody>
</table>

Enables INT8 inference with very minimal loss in accuracy
AdaRound: Adaptive rounding for better quantization

ICML’20 paper

Semantic Seg. (DeepLabv3)

<table>
<thead>
<tr>
<th>Configuration</th>
<th>mIOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floating point</td>
<td>72.94</td>
</tr>
<tr>
<td>Nearest Rounding - 4-bit weights, 8-bit activations</td>
<td>6.09</td>
</tr>
<tr>
<td>AdaRound - 4-bit weights, 8-bit activations</td>
<td>70.86</td>
</tr>
</tbody>
</table>

mIOU: Mean Intersection Over Union

AdaRound optimizes the network weights in minutes without model fine-tuning

\[
\arg \min_V \|Wx - \bar{W}x\|^2_F + \lambda f_{reg}(V)
\]
AdaRound results
Coming soon to the AIMET GitHub repo

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

### Object Detection

<table>
<thead>
<tr>
<th>Configuration</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floating point</td>
<td>82.20</td>
</tr>
<tr>
<td>Nearest Rounding - 8-bit weights, 8-bit activations</td>
<td>49.85</td>
</tr>
<tr>
<td>AdaRound - 8-bit weights, 8-bit activations</td>
<td>81.21</td>
</tr>
</tbody>
</table>

mAP: Mean Average Precision
AIMET quantization simulation

- Simulates quantization noise in the forward pass
- Helps determine quantized accuracy
- Used for Quantization-aware Training to improve quantized accuracy
How accurate is the quantization simulation?

Very accurate – the rounding errors are tiny

<table>
<thead>
<tr>
<th>Model</th>
<th>Top1 simulated</th>
<th>Top1 on-device</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet 50</td>
<td>75.76%</td>
<td>75.67%</td>
</tr>
<tr>
<td>MobileNetV2</td>
<td>70.12%</td>
<td>70.01%</td>
</tr>
</tbody>
</table>

Hardly any difference between quantization simulation and real hardware
Simple APIs

```python
model = models.resnet18(pretrained=True)
equalize_model(model, input_shape=(1, 3, 224, 224))

sim = QuantizationSimModel(model)
sim.compute_encodings(forward_pass_callback=evaluate_model,
                       forward_pass_callback_args=500)

evaluate_model(sim.model)
```

cross-layer equalization & high-bias absorption
User pipeline will provide quantized accuracy
Compression

Automate the tedious process of compression

- Research proves compression is possible - there is lot of redundancy in the network!
- Requires automation with innovative algorithms!!
Compression Features

Spatial SVD

- Weight tensor (mxn) decomposed into two tensors (mxk) and (kxn)

Channel Pruning

- Used for 2D conv layers
- Prune input channels to a layer

Automatic selection of compression-ratio per layer
Techniques can be combined to get more compression gains
Compression flow

Selects how much to compress each layer
- Greedy Comp-Ratio Selection

Make the model smaller (MACs or Memory)
- Spatial SVD (Level 3)
- Channel Pruning (Level 3)
Greedy compression-ratio selection

Chooses optimal compression per-layer

<table>
<thead>
<tr>
<th>Layer</th>
<th>Layer 2</th>
<th>Layer 3</th>
<th>Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.01</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>0.2</td>
<td>0.06</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>0.3</td>
<td>0.30</td>
<td>0.49</td>
<td>0.17</td>
</tr>
<tr>
<td>0.4</td>
<td>0.51</td>
<td>0.55</td>
<td>0.22</td>
</tr>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>0.64</td>
<td>0.35</td>
</tr>
<tr>
<td>0.6</td>
<td>0.88</td>
<td>0.71</td>
<td>0.41</td>
</tr>
<tr>
<td>0.7</td>
<td>0.92</td>
<td>0.83</td>
<td>0.55</td>
</tr>
<tr>
<td>0.8</td>
<td>0.94</td>
<td>0.97</td>
<td>0.59</td>
</tr>
<tr>
<td>0.9</td>
<td>0.95</td>
<td>0.99</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Compression-ratio vs. model performance table
Built-in visualizations

Cross-channel variance in amplitude
Features

- State-of-the-art quantization tools
- State-of-the-art network compression tools
- Support for both TensorFlow and PyTorch
- User-friendly APIs to automate optimization workflow
- Growing feature set to address different AI use cases

Feature details: Visit the Qualcomm Innovation Center channel on YouTube: [https://www.youtube.com/channel/UCrTuUKoSwqPCU_95STeReg](https://www.youtube.com/channel/UCrTuUKoSwqPCU_95STeReg)
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.

©2018-2021 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated. 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 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.
We thank the authors for their presentations and everyone who participated in the tinyML Summit 2021.

Along with a special thank you to the sponsors who made this event possible!
Executive Sponsors
Arm: The Software and Hardware Foundation for tinyML

1. Connect to high-level frameworks
2. Supported by end-to-end tooling
3. Connect to Runtime

- Profiling and debugging tooling such as Arm Keil MDK
- Optimized models for embedded
  - Runtime (e.g. TensorFlow Lite Micro)
- Optimized low-level NN libraries (i.e. CMSIS-NN)
- RTOS such as Mbed OS
- Arm Cortex-M CPUs and microNPUs

Stay Connected
- @ArmSoftwareDevelopers
- @ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
Advancing AI research to make efficient AI ubiquitous

Power efficiency
Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning
Robust learning through minimal data, unsupervised learning, on-device learning

Perception
Object detection, speech recognition, contextual fusion

Reasoning
Scene understanding, language understanding, behavior prediction

Action
Reinforcement learning for decision making

A platform to scale AI across the industry

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Samsung brings AI in the hands of everyone, with >300M Galaxy phones per year. Fingerprint ID, speech recognition, voice assistant, machine translation, face recognition, AI camera; the application list goes on and on.

In the heart of AI applications is the NPU, the neural processor that efficiently calculates AI workloads. Samsung NPU is a home grown IP that was employed since 2018 inside Samsung Exynos SoC.

Samsung NPU is brought by global R&D ecosystem that encompasses US, Korea, Russia, India, and China. In US, we are the fore-runner to guide the future directions of Samsung NPU, by identifying major AI workloads that Samsung’s NPU needs to accelerate in 3-5 years. For this, we collaborate with world-renowned academia research groups in AI and NPU.
Platinum Sponsors
**Eta Compute** creates energy-efficient AI endpoint solutions that enable sensing devices to make autonomous decisions in energy-constrained environments in smart infrastructure and buildings, consumer, medical, retail, and a diverse range of IoT applications.

[www.etacompute.com](http://www.etacompute.com)
Lattice Semiconductor (NASDAQ: LSCC) is the low power programmable leader. We solve customer problems across the network, from the Edge to the Cloud, in the growing communications, computing, industrial, automotive and consumer markets. Our technology, relationships, and commitment to support lets our customers unleash their innovation to create a smart, secure and connected world. www.Latticesemi.com.
Gold Sponsors
AKIDA™ Neuromorphic Technology: Inspired by the Spiking Nature of the Human Brain

- Supports ultra-low power applications (microwatts to milliwatts)
- Edge capabilities: on-chip training, learning, and inference
- Designed for AI Edge applications: vision, audio, olfactory, and smart transducer applications
- Licensed as IP to be designed into SoC or as silicon
- Sensor inputs are analyzed at the point of acquisition rather than through transmission via the cloud to the data center. Enables real time response for power-efficient systems
- Software Development Platform
BabbleLabs AI speech wizardry in Cisco Webex

AI meets speech - deep experience in speech science, AI/ML, embedded systems

Massive compute

- 300 TFLOPS per engineer

Novel deep neural networks

- Silicon-optimized software

Massive data corpus

- 40K hours of speech
- 15K hours of music
- 10K hour of noise
- 100K room models

Speech enhancement

- Conferencing
- Call centers

Speech recognition

- Digital Assistants
- Calling
DSP Group, Inc. develops wireless communications and voice processing chipsets, algorithms, and software solutions for converged communications and smart-enabled devices. Core competencies include, but are not limited to, voice processing. Its technology supports the development and integration of voice user interfaces (VUIs) for applications ranging from smartphones to the smart home. Its Ultra-Low Energy (ULE, per the ULE Alliance) wireless solutions enable low-power, long-range, secure communication applications for the IoT and are distinguished by their native support of two-way voice communication. On-going development efforts include the application of machine learning (ML) and artificial intelligence (AI) hardware and algorithms to address the need for accurate AI solutions at the edge for applications such as sound detection, proximity detection, and acoustic beacons.
TinyML for all developers

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time
Open source SDK

Embedded and edge compute deployment options

Test

Test impulse with real-time device data flows

www.edgeimpulse.com
The Eye in IoT
Edge AI Visual Sensors

CMOS Imaging Sensor
- Ultra Low power CMOS imager
- AI + IR capable

Computer Vision Algorithms
- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers

IoT System on Chip
- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators
GrAI Matter Labs has created an AI Processor for use in edge devices like drones, robots, surveillance cameras, and more that require real-time intelligent response at low power. Inspired by the biological brain, its computing architecture utilizes sparsity to enable a design which scales from tiny to large-scale machine learning applications.

www.graimatterlabs.ai
Enabling the next generation of **Sensor and Hearable products** to process rich data with energy efficiency

- Visible Image
- Sound
- IR Image
- Radar
- Bio-sensor
- Gyro/Accel

Wearables / Hearables

Battery-powered consumer electronics

IoT Sensors
Himax Technologies, Inc. provides semiconductor solutions specialized in computer vision. Himax’s WE-I Plus, an AI accelerator-embedded ASIC platform for ultra-low power applications, is designed to deploy CNN-based machine learning (ML) models on battery-powered AIoT devices. These end-point AI platforms can be always watching, always sensing, and always listening with on-device event recognition.

Imagimob AI SaaS

• End-to-end development of tinyML applications
• Guides and empowers users through the process
• Support for high accuracy applications requiring low power and small memory
• Imagimob AI have been used in 25+ tinyML customer projects
• Gesture control

imagimob.com
LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com
Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.
Qeexo AutoML
Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

For more information, visit: www.qeexo.com
Add Advanced Sensing to your Product with Edge AI / TinyML

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars “see with sound”

Reality AI Tools® software
- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

sensiml.com
Silicon Labs (NASDAQ: SLAB) provides silicon, software and solutions for a smarter, more connected world. Our technologies are shaping the future of the Internet of Things, Internet infrastructure, industrial automation, consumer and automotive markets. Our engineering team creates products focused on performance, energy savings, connectivity, and simplicity. silabs.com
**Syntiant Corp.** is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant’s chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a **CES® 2021 Best of Innovation Awards Honoree**, **shipped over 10M units worldwide**, and **unveiled the NDP120** part of the NDP10x family of inference engines for low-power applications.

[www.syntiant.com](http://www.syntiant.com)  
[@Syntiantcorp](https://twitter.com/Syntiantcorp)
TensorFlow is an end-to-end open source platform for machine learning. Our ecosystem of tools, libraries, and community resources help users push the state-of-the-art in building and deploying ML powered applications.
JOIN OUR SESSIONS DURING THE TINYML SUMMIT

Performing inference on BNNs with xcore.ai
Tuesday, March 23 at 12pm (PST)

TinyML: The power/cost conundrum
Thursday, March 25 at 12pm (PST)

VISIT XMOS.AI TO FIND OUT MORE
Silver Sponsors

EDGE
CORTIX

HOTG

SynSense
Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® Summit 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at the tinyML Summit. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org