Autonomous Driving AI Workloads:
Technology Trends and Optimization Strategies

Ahmed K. Sadek
Senior Director of Engineering
Qualcomm Technologies, Inc.
The Need for Intelligent, Personalized Experiences Powered by AI is Ever-growing

Smartphone  
Smart homes  
Video conferencing  
Autonomous vehicles

Smart factories  
Extended reality  
Smart cities  
Video monitoring

Qualcomm
What Makes an Autonomous Vehicle (AV)?

Configurable set of camera, radar and lidar sensors.
### SAE Levels of Autonomy

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0</td>
<td>Driver only</td>
<td>Driver operates vehicle</td>
</tr>
<tr>
<td>L1</td>
<td>Assisted Driving</td>
<td>Vehicle steers or controls speed</td>
</tr>
<tr>
<td>L2</td>
<td>Partial Autonomy</td>
<td>Vehicle drives itself but not 100% safely</td>
</tr>
<tr>
<td>L3</td>
<td>Conditional Autonomy</td>
<td>Vehicle drives itself but may give up control</td>
</tr>
<tr>
<td>L4</td>
<td>Significant Autonomy</td>
<td>Vehicle drives itself in specific cases (e.g.: urban streets) * Driver not required at all times</td>
</tr>
<tr>
<td>L5</td>
<td>Complete Autonomy</td>
<td>Vehicle drives itself in all situations</td>
</tr>
</tbody>
</table>

*SAE is Society of Automotive Engineers*
Autonomous Driving Stack
Solving the complex autonomous driving problem flow

**High-precision Localization**
- GNSS, IMU, CAN
- VEPP
- Map fusion
  - Ego Pose, Ego Velocity

**Robust Perception Stack**
- Cameras
- Radars
- Lidars
- Multicamera Perception
- Radar Perception
- Lidar Perception
- Sensor Fusion
  - Lane, Traffic Signs
  - Camera Objects
  - Radar Objects
  - Position, Vel, PointCloud
  - PointCloud, Lane detections
  - Object position, velocity, acceleration

**Human-like Driving Planner Assertiveness vs Safety**
- Behavior Prediction and Planning
- Motion Planning
- Lateral and Longitudinal Controls
- C-V2X
- Predicted trajectories of objects, desired plan for Ego vehicle

**Car actuation**
- Steering
- Brake
- Throttle
What Drives ADAS/Autonomy Workloads Complexity

1 KPI is Key Performance Indicator
Peak TOPs is not Enough

Autonomy KPIs

- **Latency**
  - BatchSize=1, reacting to a sudden event / high speed delta

- **Throughput / TOPs\(^1\)**
  - Architecture / Compiler Efficiency

- **Quantization Accuracy**
  - mAP\(^2\) (average) is not enough. Need to handle low SNR scenario

- **TOPs/Watt**
  - Cost / Thermal mitigation

- **Safety and Security**
  - ASIL B/D Protecting Network Weights

1 TOPs is Tera Operations Per Second  2 mAP is mean Average Precision
Robust Quantization Techniques

• mAP is a typical KPI used in studying accuracy before and after quantization

• **Low SNR scenarios (tail scenarios) are critical for autonomy**

• Low SNR in context of DL could mean the FP32 performance (e.g., decision boundaries) are barely meeting performance, and quantization noise results in failure for these scenarios, e.g.:
  • Objects represented by few pixels such as far away objects or small objects
  • Rare objects (e.g., non-conventional trailer trucks, animal on the road, ....)
  • Bad weather/lighting conditions

• **Significant improvement from quantization-aware training**

Example for Low SNR Scenario

- Far object
- Uncommon object
Driving the Industry Towards Integer Inference and Power-efficient AI
Leading model efficiency research and fast commercialization

Qualcomm® Neural Processing SDK and Qualcomm® AI Model Efficiency Toolkit (AIMET) Pro

Model efficiency research

Model efficiency commercialization

Model efficiency open-sourcing

Qualcomm Nexus Process (ICCV 2019)
Relaxed Quantization (ICLR 2019)
Joint Pruning and Quantization (ECCV 2020)
AdaRound (ICML 2020)
Bayesian Bits (NeurIPS 2020)
DONNA-NAS (ICCV 2021)
Transformer Quantization (EMNLP 2021)

Data-free Quantization (ICCV 2019)

AdaRound (ICML 2020)

Bayesian Bits (NeurIPS 2020)

DONNA-NAS (ICCV 2021)
Transformer Quantization (EMNLP 2021)

AI Model Efficiency Toolkit (AIMET) OS
Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc.
AIMET is a product of Qualcomm Innovation Center, Inc.
AdaRound Results

Post-training technique that makes INT8 quantization more accurate and INT4 quantization possible

<table>
<thead>
<tr>
<th>Bit width</th>
<th>Mean AP (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32</td>
<td>Baseline</td>
</tr>
<tr>
<td>INT8 baseline quantization</td>
<td>&gt;10%</td>
</tr>
<tr>
<td>INT8 AdaRound quantization</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Reduction in accuracy between FP32 and INT8 AdaRound quantization

INT8, baseline quantization

INT8, AdaRound quantization
Multi-task architecture with common shared backbone

Architecture trends:

- Higher resolution → Pushing requirements on compute and memory
- High dilation factors
- Transformer heads
- Low level fusion across multiple cameras
Low level fusion and sparse point cloud signals

• **3D Sparse Convolution** can reduce the computation and speed up the inference

• Efficient and novel approaches in both hardware and software architecture to handle high sparsity from both data movement and compute

• Camera/radar low level fusion between high input resolution and high sparse multi-modal signals

TY Lim, et.al, Radar and Camera Early Fusion for Vehicle Detection in Advanced Driver Assistance Systems,
Scheduling / Graph Optimizations

- Graph scheduler optimization to minimize spillage to DDR
- Reuse intermediate activations for next layer(s) processing on chip
- Boost inferences/second
- Reduce MBytes/inference and preserve DDR BW for other applications
Behavior Planning: Model-Based RL

Highly dynamic dataflows
Neural Architecture Search (NAS)

• Existing techniques require lots of computational time and has large search space
  • e.g., RL and Evolutionary Algorithms
• The need for efficient techniques has emerged
• Weight-sharing
• Differentiable Architecture Search
• These methods often suffer from an instability issue, and in many cases require careful training methods or search space design
Improving NAS Efficiency

Trust Region Aware Sample-Efficient Architecture Search for Distillation

We have improved NAS in

• Search phase
  • Trust Region Bayesian Optimization for sample-efficient search
  • Knowledge Distillation-guided score to perform more efficient and teacher-aware search

• Query phase
  • Orthogonality regularization
Snapdragon Ride SDK

Middleware
- Production Ready Platform Support Package
- Safe Operating System with Hypervisors
- High Performance Multi-SoC Compute

Auto-Imaging System
- Rich suite of camera support with multi-high-resolution cameras

Neural Processing Toolkit
- AIMET, NAS, Compiler, Quantizer, Simulator and profiling tools for optimizing AI perception, planning

Embedded Vision and AD Libraries
- Rich set of math and vision library functions optimized for DSP, CPU, GPU and Vision Accelerators

Middleware and Multi-SoC Infrastructure
- Production Ready safety Platform, Automotive Multi-SoC middleware for seamless low latency high speed data movement

Tools
- Profiler Tools to analyze processing blocks utilization, latency, memory bandwidth, power and thermal management

Development Platform
- Reference hardware design with up to 16 cameras, radars, lidars, location for L2 to L3 system design
- Multi-SoC architecture with Safety MCUs and Storage, with production ready thermal design

Partners:
- QNX
- RedHat
- Infineon
- Vector

Snapdragon Ride is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.
Conclusion

• Autonomous driving AI workloads are increasing in complexity requiring SW-HW co-design for increased efficiency
  • Both for HW accelerator architecture and data flow optimizations
• New quantization methods presented that optimize for both average and low SNR regimes
• Further improvements required to increase NAS efficiency and speed innovation cycle
• Among topics not covered that will impact future network architecture design: causality and ability to reason
For More Information

Qualcomm AI
http://www.qualcomm.com/ai

Qualcomm ADAS
Qualcomm ADAS

Qualcomm Technology
Qualcomm YouTube

Qualcomm @ 2022 Embedded Vision Summit:

“A Practical Guide to Getting the DNN Accuracy You need and the Performance You Deserve” – Felix Baum – Wed, May 18, 2:40 PM

“Tools for Creating Next-Gen Computer Vision Apps on Snapdragon” – Judd Heape - Wed, May 18, 10:50 AM

“The Future of AI is Here Today: Deep Dive into Qualcomm’s On-Device AI Offerings” – Vinesh Sukumar - Wed, May 18, 12:00 PM

“Seamless Deployment of Multimedia and Machine Learning Applications at the Edge” – Megha Daga - Tuesday, May 17, 2:40 PM
Thank You