Bringing Edge AI to Life

Ali Osman Örs – NXP Semiconductors
David Steele – Arcturus
July 13, 2021
Presenters

Ali Osman Örs
Director, AI ML Strategy and Technologies
NXP Semiconductors

David Steele
Director of Innovation
Arcturus
Agenda

• Introduction
• Tools and Enablement
• Building and Scaling Applications
• Optimization Techniques
• Demo
• 2x $250 Amazon Gift Card Draw
Introduction to Edge AI Applications

Boundary Crossing and Intrusion

Package Analytics

Characterization

Behaviour Analysis
Development Challenges

- Development paradigm
  - Experimentation and adaptation
  - Heavily reliant on optimization
  - Complex interdependencies
  - Requires broad expertise

- Starting points
  - Do not move to production seamlessly
  - Dependencies / edge runtime limitations
  - Lab vs field data
  - Missing components / specialized requirements

- Considerations
  - Move an application to the edge
  - Improve accuracy and performance
  - Develop a scalable/flexible architecture
Edge Enablement and Tools

Ali Osman Örs
NXP Semiconductors
NXP Broad-based Machine Learning Solutions and Support

**eIQ™ ML SW Development Environment**
- eIQ Toolkit with eIQ Portal GUI to:
  - Import/create, convert, optimize, validate and deploy ML models
  - Dataset curation tools to create new, augment, label/annotate datasets
- eIQ inference with: TensorFlow Lite, TensorFlow Lite Micro, Arm NN, ONNX Runtime, Glow and DeepViewRT
- Support for i.MX 8 family, i.MX RT family
- Integrated into NXP development environments (MCUXpresso, Yocto/Linux)

**Third Party SW and HW**
- Google Coral Dev Board
- i.MX 8M Mini Development Kit for Amazon® Alexa Voice Service
- Au-Zone Value-add packages for NXP eIQ Toolkit
- Arcturus video applications
- SensiML tools for sensor analysis

**eIQ™ Auto AI Enablement**
- Deep Learning toolkit for S32V processors
- Auto Quality: A-SPICE qualified inference engine
- Optimization: Prunes, quantizes, compresses the Neural Network
- Automated neural net layer deployment to optimum available compute resource

**Turnkey Solutions**
- Alexa Voice Services (AVS) solution
  - i.MX RT106A (kit – SLN-ALEXA-IOT)
  - i.MX RT106L (kit – SLN-LOCAL-IOT)
- Face & emotion recognition solution with Anti-Spoofing
  - i.MX RT106F (kit – SLN-VIZN-IOT)

---

DIY

Automotive Grade

.... And more

Fully Tested
Scalable/reusable software, tools and solutions

Arm Cortex-A and Cortex-M cores
APPLICATIONS PROCESSORS
S32x | i.MX | Layerscape
Voice / Audio | Graphics | Video | ML | Networking | 5G Access

CROSSOVER
i.MX RT
i.MX ULP

Arm Cortex-M cores
MCUs
LPC | Kinetics | S32K
Motors, Control, Sensors, Power, Wireless Connectivity

With Industrial-grade reliability, security, longevity

Future i.MX 9 MPUs
Example Applications:
- Smart Cities
- Smart Homes
- Smart Buildings
- Smart Factories
Cortex-A cores & Cortex-M cores
First use of Arm Ethos-U65 NPU

Example Applications:
- Machine vision
- Industrial computer
- Security and vision

i.MX 8M Plus MPU
Quad/Dual Cortex-A53 and M7
2.3 TOPS NPU, GPU, DSP
Camera, display, video, audio
CAN-FD, GbE, TSN, PCIe
Machine learning use cases and accelerators

<table>
<thead>
<tr>
<th></th>
<th>Vision</th>
<th>Audio</th>
<th>Time Series Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face and still image recognition, person detection (images)</td>
<td>Multi-face recognition, object detection (video)</td>
<td>Live video face and object recognition</td>
<td>Anomaly detection (environmental sensors)</td>
</tr>
<tr>
<td>Multi-object surveillance (people, cars, animals)</td>
<td>40,000 words vocabulary, multiple speaker recognition</td>
<td>Complex real-time motion analysis</td>
<td>Pose estimation</td>
</tr>
<tr>
<td>Wake word, 10 Word speech, speaker recognition</td>
<td>Automatic speech recognition (basic command phrases)</td>
<td>Gesture recognition</td>
<td>Gesture recognition</td>
</tr>
<tr>
<td>Speech accents interpretation</td>
<td></td>
<td></td>
<td>Complex real-time motion analysis</td>
</tr>
</tbody>
</table>

Cortex-M  4x Cortex-A53  GPU  Neural Processing Unit (NPU)
eIQ Machine Learning SW Development Environment

NXP’s **eIQ ML Software** provides a collection of development tools, utilities and libraries for building ML applications using NXP MCUs and applications processors (MPUs).

eIQ ML software can be leveraged as part of a user’s existing flow or can be used for the complete flow depending on the ML application targeted.
eIQ™ ML SW Development Environment

- Embedded Developers
- Data Scientists
- ML Experts

eIQ Toolkit

Bring Your Own Data Workflow
- Data Curation
- Model Selection
- Model Training, Optimization, Quantization
- Model Validation

Bring Your Own Model Workflow

- eIQ Portal
- eIQ Toolkit

Applications running on EdgeVerse processors
- ANOMALY DETECTION
- VOICE RECOGNITION
- FACE RECOGNITION
- MULTI-OBJECT CLASSIFICATION
- COMPLEX DIAGNOSIS

- eIQ inference

- eIQ inference with...
- TFLite-Micro
- Glow
- DeepViewRT™
- TFLite
- Arm NN
- ONNX Runtime
- DeepViewRT™

- Value-Add Solutions and Services
- eIQ Marketplace

- eIQ Toolkit
- eIQ Portal

- eIQ inference with...
- Arm® Cortex®-M, DSP
- Arm Cortex-A, GPU, DSP, NPU
eIQ Toolkit and eIQ Portal

- Eases the ML development experience with the eIQ Portal as well as with command line host tools like; Glow tools, TensorFlow, PyTorch and other third-party tools
- Enables graph-level profiling capability with runtime insights to help optimize neural network architectures for execution on target NXP processors
- Includes ML application examples
- **eIQ Portal** intuitive graphical user interface (GUI) that simplifies ML development:
  - Creates, optimizes, debugs, converts, and exports ML models
  - Imports datasets and models, rapidly trains and deploys neural network models and ML workloads
  - Output seamlessly feeds into DeepViewRT, TensorFlow Lite, TensorFlow Lite Micro, Glow, Arm NN, and ONNX Runtime inference engines
  - Can import models from TensorFlow and PyTorch ML frameworks
  - Includes object detection and image classification models for computer vision applications
- **eIQ Marketplace** offers value-add solutions, professional support and design services from trusted eco-system partners and NXP
- Delivered with a single click from the eIQ Toolkit at [www.nxp.com/eIQ](http://www.nxp.com/eIQ)
eIQ inference

- eIQ inference with TensorFlow Lite, ARM NN, Glow and ONNX Runtime
  - Optimized support for open community-based engines on NXP target devices

- eIQ inference with DeepViewRT runtime
  - Platform-optimized, proprietary runtime inference engine that scales across a wide range of NXP devices and neural network compute engines
  - Stable and longer-term maintained solution to complement the open community-based inference solutions.
  - Supports EdgeVerse™ processors, including the i.MX RT crossover MCUs (Arm® Cortex®-M cores), i.MX applications processors (Cortex-A and Cortex-M cores, dedicated Neural Processing Units (NPU) and GPUs)
  - Enables compact code size for resource-constrained devices with ease of analysis and fine-tuning of model performance using the eIQ Toolkit
  - Delivered via NXP standard Yocto BSP release for Linux® OS-based development, and MCUXpresso SDK release for embedded RTOS-enabled MCU development
  - eIQ inference with DeepViewRT runtime is provided free-of-charge to NXP customers as part of our ML enablement

Applications running on EdgeVerse processors

- ANOMALY DETECTION
- VOICE RECOGNITION
- FACE RECOGNITION
- MULTI-OBJECT CLASSIFICATION
- COMPLEX DIAGNOSIS

- eIQ inference with... 
  - TFLite-Micro
  - Glow
  - DeepViewRT™
  - TFLite
  - Arm NN
  - ONNX Runtime
  - DeepViewRT™

- Platforms:
  - Arm® Cortex®-M, DSP
  - Arm Cortex-A, GPU, DSP, NPU
NXP eIQ™ ML Software Development Environment

Inference Engines and Libraries for Neural Network Model Deployment

DeepViewRT
TensorFlowLite for Microcontrollers
GLOW

Arm® Cortex®-M
- i.MX RT1064
- i.MX RT1060
- i.MX RT1170

DSP
- i.MX RT600

Microcontroller Compute Engines

DSP
- i.MX 8M Plus
- i.MX 8M
- i.MX 8M Nano

GPU
- i.MX 8M Plus

ML Accelerator
- i.MX 8M Plus

Applications Processor Compute Engines

Arm® Cortex®-A
- i.MX 8M Plus
- i.MX 8M
- i.MX 8M Nano
- i.MX 8M Nano UL
- i.MX 8M Mini

* Additional support for devices not listed can be available
eIQ Toolkit availability

Registered users can download eIQ Toolkit from: http://www.nxp.com/eiq
eIQ Portal
Data Curation

- Capture and annotate images for model training and validation
- Import datasets from public or user-defined formats
- Distribute data for training and testing
**Dataset Augmentation**

- Dataset augmentation adjusts image parameters to improve model training.
- Reduce over-fitting and increase robustness to dynamic real-world environments.
- Visualize how augmentation parameters affect images.
- Set augmentation parameters relative to the application.
• Select the appropriate class of model:
  • Classification, Detection
MODEL TOOLS

Base Models

- **mobilenet_v1**
  - Task: classification
  - Problem Type: image

- **mobilenet_v2**
  - Task: classification
  - Problem Type: image

- **mobilenet_v3**
  - Task: classification
  - Problem Type: image

- **fpn_ssd_mobilenet_v2**
  - Task: detection
  - Problem Type: boxes

- **ssd_mobilenet_v3**
  - Task: detection
  - Problem Type: boxes
MODEL TOOLs

- Choose a model version that matches performance and accuracy requirements
MODEL Tools

- Model selected must be correct fit for target resources
  - Application Processors or MCUs
  - Models further optimized for compute units* (MCU, CPU, GPU, NPU)

* Planned future functionality
Model Training and Optimizing for BYOD and BYOM

- Model training should support users of all levels of experience
- Use default settings or fine-tune with hyperparameter selection

- Static model analysis for debug and bottleneck detection
- Automatic graph-level optimizations should improve performance and memory utilization without precision loss (e.g., pruning, fusing, layer folding)
- Configurable optimizations control accuracy tradeoffs (e.g., quantization, layer replacement, weight rounding)
- Quantization converts 32-bit floating point models to 8-bit integer format
Validation to Prove and Improve Model Behavior

- Use validation to uncover areas that can be improved in your data set
- Analyze and compare model accuracy running different optimizations
Validation target

- By default you will validate the model on your local PC
- If you have a Remote Device connected you can validate on target and get profiling information
BYOM: eIQ Model tool
## Model converter support

<table>
<thead>
<tr>
<th>Source/Destination</th>
<th>Float</th>
<th>Quantized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeepviewRT RTM</td>
<td>TensorFlow Lite</td>
</tr>
<tr>
<td>TensorFlow 1.x pb</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Saved Model (Folder/tar)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Keras (.h5)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DeepviewRT RTM</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>TensorFlow Lite (tflite)</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>ONNX</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TensorFlow Lite Quantized</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>ONNX Quantized</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
eIQ Model Converter
Application Examples

Boundary Crossing and Intrusion

Package Analytics

Characterization

Behaviour Analysis
Motion Tracking

- Object Detection Classification Model
- Object Detection
- Bounding Box Localization
- Motion Prediction
- New Identity Created
- Existing Identity Matched
- Compare / Refine

Relies on continuous detections - object cannot leave frame/FoV.
Motion and Appearance Tracking

Visual appearance reidentifies objects irrespective of time or space - but - requires generating embeddings for each object detected.
Building and Scaling Pipelines

Harshad Mahadik
Edge AI and Vision Team Lead – Arcturus

Jonathan Rynne
Data Scientist – Arcturus
Tracking Pipeline (by resource)

- Input Stream
- Frame Data
- Detection Inference (SSD_MobileNet)
- Object Class Localization x/y
- Offload to NPU
- Arm CPU
- NPU
- Embeddings Generation (MobileNet)
- Motion Track (Kalman Filter)
- ID Assignment (Hungarian)
- Appear. Track (N. Neighbor)
- IP/Port:
- UI / UX
- Services / Notifications / Integrations
- IP/Port:
## Optimization Techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
<th>Benefit</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concurrent Processing</td>
<td>Allows multiple models to run at the same time</td>
<td>Decreases latency</td>
<td>Resource intensive</td>
</tr>
<tr>
<td>Batch Processing</td>
<td>Allow multiple images to processed at the same time</td>
<td>Increases throughput</td>
<td>Increases latency</td>
</tr>
<tr>
<td>Model Depth Reduction</td>
<td>Reduces the complexity of the network</td>
<td>Decreases inference time</td>
<td>Decreases accuracy</td>
</tr>
<tr>
<td>Tiling</td>
<td>Breaks down input image into multiple image tiles</td>
<td>Improves accuracy</td>
<td>Increases latency</td>
</tr>
<tr>
<td>Network Quantization</td>
<td>Reduces precision of network e.g. from Float32 to INT8</td>
<td>Improve performance</td>
<td>Decreases accuracy</td>
</tr>
</tbody>
</table>
Concurrent Model Processing

Pros
- Reduced latency
- Increased throughput
  - Increased FPS
- Increased analytic results

Cons
- Increased complexity
  - memory management (IPC)
- Increased memory consumption
- CPU/GPU/NPU intensive

Performance Example
- MobileNet_v1_embeddings network
  - 30ms inference
  - 4 People Detected within Frame
  - Total Inference Time without Concurrent Processing
    - 30ms * 4 = 120ms
  - Total Inference Time with Concurrent Processing
    - 30ms + 30ms (overhead) = 60ms
Batch Processing/Batch Inference

Throughput* vs. Latency for Increased Batch Size (PreProcess)

Pros
- Increased throughput
- Decreased complexity

Cons
- Increased memory consumption
- Model-backend support
- Increased latency

* SSD_MobileNet_v2_300x300_quantized / i.MX 8M Plus
Tiling

Prosg

• Improve detection accuracy
• Improve Tracking
• Can be combined with batch
Tiling

Cons

- Increase false positives
- Add pre/post processing overhead
Model Size Depth Reduction

Definition in source code:

*depth_multiplier*: Float multiplier for the depth (number of channels) for all convolution ops. The value must be greater than zero. Typical usage will be to set this value in (0,1) to reduce the number of parameters or computation cost of the model.

<table>
<thead>
<tr>
<th>Network</th>
<th>Depth Multiplier</th>
<th>Speed* (Embed Time)</th>
<th>Accuracy (mAP on Market 1501)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobilenet v1</td>
<td>1.0 (100%)</td>
<td>51ms</td>
<td>63.49%</td>
</tr>
<tr>
<td>Mobilenet v1</td>
<td>0.75 (75%)</td>
<td>33ms</td>
<td>62.12%</td>
</tr>
<tr>
<td>Mobilenet v1</td>
<td>0.5 (25%)</td>
<td>19ms</td>
<td>60.21%</td>
</tr>
</tbody>
</table>

* Calculated by running model on the CPU (4x Arm Cortex-A53)
Let’s do even more!

David Steele
Arcturus
Characterization

Detection and Colour Modeling
Two networks process the same frame in order to improve precision and fully characterize object.
## Atlas Edge AI Hardware Platform

### Atlas Platform

<table>
<thead>
<tr>
<th>NXP 8M Plus</th>
<th>Imaging and Video</th>
<th>Core and Memory</th>
<th>Connectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core and Memory</td>
<td>2x MIPI CSI</td>
<td>4x Cortex-A53 up to 1.8 GHz Cortex-M7 up to 800 MHz</td>
<td>2x Gigabit Ethernet (UART, I2S, I2C, SPI)</td>
</tr>
<tr>
<td>4x Cortex-A53 up to 1.8 GHz Cortex-M7 up to 800 MHz</td>
<td>2x ISPs up to 12MP, 3D GPU, 2D GPU</td>
<td>Video encode/decode up to 1080p60, h.265/4, HDMI, MIPI-DSI, LVDS</td>
<td>2x USB 3.0, PCIe Gen 3, SDIO, CAN-FD</td>
</tr>
<tr>
<td>Memory</td>
<td>INTERCONNECT BACKPLANE</td>
<td>4GB (64-bit) LP-DDR4</td>
<td>3.3VDC I/O</td>
</tr>
<tr>
<td>6GB LP-DDR4 (32-bit), 512GB eMMC, 32MB SPI NOR</td>
<td>POWER MGMNT</td>
<td>up to 8GB (128-bit) 16GB eMMC</td>
<td>RESET/WATCHDOG</td>
</tr>
<tr>
<td>AI Acceleration</td>
<td>2x M.2 CARD SLOT (M-Key)</td>
<td>Memory</td>
<td>AI Acceleration</td>
</tr>
<tr>
<td>NPU up to 2.3 TOPS</td>
<td>4x MIPI CSI up to 6</td>
<td>4x Cortex-A57 @ 1.4 GHz up to 6x Carmel @ 1.1 GHz</td>
<td>GPU (472 GFLOPS) or GPU (21 TOPs)</td>
</tr>
</tbody>
</table>

### External Terminations

- +12VDC
- RESET
- TAC
- READY LED
- ALARM LED
- LINK / ACT LED
- RS485 ETH
- USB3 (Up to 3)
- HDMI (Up to 2)
- I/O (Hx4)
- MICRO SD CARD
Characterization Pipeline (by resource)

i.MX 8M Plus
High Frame Rate Detection

Data passed over network

Input Stream Decoded

Segmentation Inference (FCN-ResNet18)

Detection Inference (SSD_MobileNet)

Object Class Localization x/y

Pixelwise Probability

Colour Characterization (KNN)

Motion Track (Kalman Filter)

ID Assignment (Hungarian)

Characterization Assignment

NVIDIA
Complex Segmentation

IP/Port:
Localhost:

Arm CPU
NPU
GPU

UI / UX
Services / Notifications / Integrations

© 2020 Arm Limited (or its affiliates)
Characterization Demo
NXP ML/AI training series

1. MACHINE LEARNING CONCEPTS
   - Concepts and Introduction

2. eIQ™ SOFTWARE DEVELOPMENT ENVIRONMENT
   - eIQ Overview
   - Transfer Learning Intro & Lab
   - Handwritten Digit Recognition Example

3. eIQ TOOLKIT
   - eIQ Toolkit: How to BYOD
   - eIQ Toolkit: How to BYOM
   - eIQ Toolkit: Command Line interface

4. MACHINE LEARNING ON MCUS
   - Machine Learning with i.MX RT
   - Get started with eIQ on i.MX RT
   - Create your own Model using Glow
   - eIQ inference with TensorFlow Lite for MCUs – Overview & Lab
   - eIQ inference with Glow NN Compiler – Overview & Lab
   - eIQ inference with DeepViewRT

5. MACHINE LEARNING ON MPUS
   - eIQ inference with ONNX
   - eIQ inference with Arm NN
   - eIQ inference with TensorFlow Lite (for MPUs)

6. PARTNER ML SOLUTIONS
   - Solutions and topics contributed by NXP eIQ partners

- 20+ training modules
- Available at www.nxp.com/mltraining
References and helpful links

- eIQ™ ML Software Development Environment (https://www.nxp.com/eiq)
- eIQ ML/AI Training Series (https://www.nxp.com/mltraining)
- Embedded Linux for i.MX Applications Processors (https://www.nxp.com/design/software/embedded-software/i-mx-software/embedded-linux-for-i-mx-applications-processors:IMX LINUX)
- Brinq™ Edge AI and Vision Analytics (https://www.arcturusnetworks.com/brinq/)
- Brinq Edge AI for Public Safety (White paper)
- Arm AI Tech Talk – The Smart City In Motion – Intelligent Transportation Systems (webinar recording)
- Arm Dev Summit – Using Arm NN to Develop Edge AI in the Smart City (webinar recording)
Closing Remarks and Prize Draw

Thank you for participating in our Arm AI Tech Talk.

Complete your information to be eligible to win one of 2x $250 Amazon Gift cards courtesy of Arcturus and NXP.

The entry deadline is July 13 2021 at 2pm ET. Prize draw will occur on July 13th and winners will be notified by email. One entry per person, per email.

Terms and conditions - by entering you are agreeing to share your information with Arcturus Networks Inc. and NXP Semiconductors for the purpose of contacting you about this and future promotions. Refer to each company's privacy policy and terms of use for additional detail.

https://forms.gle/ciBtRgC3c76BCNVD7
Thank you!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board
## AI Virtual Tech Talks Series

<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 13&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Bringing Edge AI to Life - from PoC to Production</td>
<td>Arcturus &amp; NXP</td>
</tr>
<tr>
<td>July 20&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Easy TinyML with Arduino: taking advantage of machine learning right where things are happening</td>
<td>Arduino</td>
</tr>
</tbody>
</table>

Visit: developer.arm.com/techtalks
Thank you!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board
Where Hardware & Software Join Forces

Call for Papers

Learn More at devsummit.arm.com
Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
tודה