Collaborative Edge-based Machine Intelligence: Promise and Challenges

Archan Misra

Acknowledge the creative contributions of:
PhD: Amit Sharma, Dulanga Weerakoon
Post-Docs: Tran Huy Vu, Kasthuri Jayarajah, Manoj Gulati, Meera Radhakrishnan
Post-doc & Engineers: Vengat Subramaniam, Dhanuja Wanniarachchige
Collaborators: Vigneshwaran Subbaraju, Tarek Abdelzaher, Rajesh Balan
My Research History

**Mobile Sensing & Analytics**
- Indoor Location
- Group Detection
- Queuing Detection

**Wearable Sensing & Systems**
- Eating (*Annapurna*)
- In-Store Shopping (*IRIS, I4S*)
- VR+ mobile (*Empath-D*)

**Wearable + IoT Systems**
- Batteryless Wearables
- Wireless/RFID Sensing
- Fine-grained Gestural Tracking

**Key Research Thrusts**
- Fusion of multi-modal sensing (*inertial*)
- Adaptive sampling & triggered sensing
- Multiple live deployments (*campus, malls, museums*) + licensing

- Optimize (*Energy, Accuracy, Latency*) tradeoffs
- Multi-modal sensor fusion (*inertial, image*)

- Make Batteryless (or Ultra-Low Power) Sensing possible
- Method: Utilize new sensing modalities (*video, wireless*) & collaborative ML at edge
W8-Scope: Exercise Monitoring using IoT Sensors

Goals:
- Quantified insights on weight stack-based exercises → provide personalized digital coaching

Techniques:
- Simple weight stack sensor *(accelerometer+ magnetometer)* to track & understand exercises

Results:
- Longitudinal Data Collection at 2 gyms ➔ 95+% accuracy & adaptation to medium-term evolutionary behavior

Magnetic Sensor on Wt. Stack ➔ *(Weight, Type, User)*
ERICA: Earable-based Real-Time Feedback for Free-weights Exercises

Goals:

- **Associate** User’s Earable with Dumbbell-mounted IoT sensors
- Perform exercise recognition & real-time mistake detection
- Provide “live” corrective feedback

Feedback after every ~4 repetitions results in lower mistakes during set
Some Lessons Learnt

Pure Wearable/Mobile Sensing or Infrastructure Sensing isn’t Enough
- Need to fuse inputs from personal and ambient sensors

Computation vs. Communication Tradeoffs are Changing
- Comms getting cheaper; computation more complex

<table>
<thead>
<tr>
<th>Proximity</th>
<th>NFC</th>
<th>ZigBee</th>
<th>BT</th>
<th>WiFi</th>
<th>LoRa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>1 mm</td>
<td>10 cm</td>
<td>10–100 m</td>
<td>10–100 m</td>
<td>30–50 m</td>
</tr>
<tr>
<td>Data rate</td>
<td>8–32 Gbps</td>
<td>0.021–0.48 Mbps</td>
<td>0.02–0.2 Mbps</td>
<td>0.8–2.1 Mbps</td>
<td>300 Mbps (11g)</td>
</tr>
<tr>
<td>Energy-efficiency</td>
<td>4 pJ/b</td>
<td>1–50 nJ/b</td>
<td>5 nJ/b</td>
<td>15 nJ/b</td>
<td>5 nJ/b</td>
</tr>
</tbody>
</table>

Source: doi: 10.1109/MIC.2018.011581520

Resource Bottlenecks & Trends

1. Where’s the Resource Bottleneck?

2. The Rise of the “Edge”

By 2022, 7 out of 10 bytes of data will never see a data center

This Talk: Summary of Collaborative Machine Intelligence (CMI)

**Collaboration is the Key to Realizing this Vision.** Among:
- Wearable devices & Edge infrastructure
- Multiple IoT devices & Edge infrastructure

**DS:** Distributed & Triggered Sensing
Tightly coordinate Cheaper Expensive Sensor Triggering

**CMI:** Collaborative ML-based Edge Intelligence
Distribute Inferencing Pipelines across multiple pervasive devices & across modalities → (Accuracy, Energy, Latency)
RF/Wireless: A Swiss-Army Knife

Sensing
- Use Radio signal reflections to capture gestures
  - WiSee: Doppler Shifts ➔ Movement Frequency
- Human Motion Artefacts
  - WiBreathe: Breathing Rate
  - Doppler Shift
- Object Composition
  - RFID Phase Shift ➔ Shape & Liquid Detector

Energy Harvesting
- Multiple emerging modalities: light > vibration > temperature > RF
- Factors: size/form factor, on-body position, intrusiveness.
**Vision**

- Utilize battery-free sensors on wearables & IoT devices to provide fine-grained tracking
- **Key breakthrough**: Charge devices wirelessly via WiFi “power packet” transmissions

**Applications**

- Activity Tracking of Workers & Moving Equipment
- Product Monitoring in Warehouses
- Elderly Monitoring in smart homes

**Challenges**

- Low energy density using omnidirectional WiFi antenna (< 1µW at 1.5m)
- WiFi AP coordination to charge multiple devices
The Wearable + AP System

Power Management  Storage  Micro-controller

The Harvester:
- Matching Circuit
- Rectifier

Motion Trigger
Super Capacitor

Accelerometer  RF Comm.

The Wearable

The Beamforming AP

- Ping detection (nRF24L01+)
- Rx Buffer for AoA
- Tx Phase Sync for Beamforming

1st WARP

2nd WARP

Raw Rx Buffer for DoA

Tx Phy & Mac

Rx Phy & Mac

Sync Cable

Pkt det

Tx syn

Rx syn
The WiWear++: Low-power downlink (LPD)

- Base version: Ping triggered by significant motion; No MAC
- New: Use Wake-up Receivers to support low-power downlink (AP to device)
  - Proactive ping request (update orientation)
  - Content-free uplink trx

WiWear++ Prototype

Wakeup Receiver

µController (LPUART)

START

STOP

LPUART compatible (1 START + 2 STOP bits)

Raw OOK signal (From AP)

µController only wakes up to read LPUART data register
The Cloud RAN & The Future of Multi-AP Operation

- Harvesting Power levels drop with multiple wearable devices

- Lots of distributed transmitters (915/964 MHz channels) surrounding the target.
  - Adjust phase → distributed beamforming
  - 24 Trx (1.7W) in 20X20 m² → 0.6-0.7mW power harvested

- Future: What about multiple APs, that coordinate their transmissions?
  - Complex balance between sensing, communication and energy transfer capacity
Takeaways & Reflections

New opportunities:
• Edge-Coordinated Activation of sensing on wearable devices.
• Combination of passive RF sensing + battery-less wearable/IoT devices
• **Edge ML needed to perform real-time multi-modal inferencing**
Collaborative MI: The Solution for Dependable Machine Intelligence

Key Idea: Overcome limitations in resource & fidelity by performing machine intelligence jointly

- **Real-time decision making**
  - Complex ML pipelines being executed on individual IoT devices or with edge-assistance

- **Key Resource & Performance Bottlenecks**
  - Latency of DNN execution
    - 550 msec+ for person recognition/frame on a Movidius co-processor (1W)
  - Low Accuracy
    - Individual sensors subject to environmental artefacts
  - Energy Overhead
    - Need to support battery-less operations
EA1: Collaborative IoT & The Edge: Ongoing Work

- Collaborative Sensing
  - Spatial and/or temporal overlap among sensors
  - Sensor Multiplicity
  - Adjust Inferencing Pipeline *on-the-fly*

- Dependable Systems
  - Resilience to Adversarial Attacks

- Collaborative Sensing
  - Spatial and/or temporal overlap among sensors

- Dependable Systems
  - Resilience to Adversarial Attacks

A’s view is partly occluded

Learning from B can improve A’s accuracy

Malicious C can purposefully perturb shared inferences
Collaborative IoT & The Edge: Ongoing Work

Design Goals

1. Requires **NO re-training** of the DNN models
2. **Backward compatibility** to non-collaborative mode when no collaborators are available
3. **Minimal latency and bandwidth overhead** for infusing collaborative input

**Closing the accuracy gap with collaboration**

- Accuracy on Learned Task vs. Latency
- Collaborative models vs. Low complexity/deep models
- High complexity/very deep models
Approach #1: Run-Time Collaborative Inferencing

(a) CNMS: Collaboration at Decision Stage
- Concat Prior Masks from N Cameras
- Prior Masks from Neighboring Cameras
- Reputation Score Update
- Trusted Inferences

(b) CSSD: Collaboration at Input Stage
- Concat Prior Masks from N Cameras
- Prior Masks from Neighboring Cameras

PETS Dataset (8 cameras)
- Person detection using SSD300; Homographic View mapping

High accuracy improvement with minimal latency

<table>
<thead>
<tr>
<th></th>
<th>Inference Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD Baseline</td>
<td>80ms</td>
<td>71%</td>
</tr>
<tr>
<td>Collaborative SSD</td>
<td>85ms</td>
<td>82.2%</td>
</tr>
<tr>
<td>CNMS</td>
<td>100ms</td>
<td>75.5%</td>
</tr>
</tbody>
</table>
Approach #2.1: Adapting the ML Pipelines “On the Fly” for Improved Accuracy

**Exploration #1**
Improving Accuracy through Sensor Multiplicity

Cam 1

Cam 2

Same person object, perceptively clearer in the collaborator view

Convolutional Layers

Class Confidence boosting

Improved Detections

Detections

<table>
<thead>
<tr>
<th>View 8+5 (33% overlap)</th>
<th>View 8+6 (62%)</th>
<th>View 8+7 (38%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase in TP%</td>
<td>Increase in FP%</td>
<td>Gain in TP for nominal increase in FP</td>
</tr>
<tr>
<td>Gains</td>
<td>Gains</td>
<td>Gains</td>
</tr>
</tbody>
</table>
Approach #2.2: Using Hidden Layer State for Collaborative Classification

**Exploration #2**
Early Estimation and Hybrid Classification

- **Feature Extraction/Selection**
- **Shallow Classification**
- **Class Probabilities**
- **Convolutional Layers**
- **Detections**
- Activated “people” pixels

### Accuracy of Discriminant FMaps in Detecting People Pixels

- PETS Dataset, Camera 7 as Reference View (over 795 frames).
- Precision = Pixels activated by discriminant fmaps that overlap with Ground truth Bboxes

Average precision of over 93% in detecting target-specific pixels as early as Layer 1

### Accuracy of Shallow Person Classification

- AUC > 95%
- as early as Layer 1

**H** – Histogram of Feature values only
**L** – Location of Anchor Boxes
**S** – Scale of anchor boxes
**A** – Aspect ratio of anchor boxes

### Precision vs Recall vs AUC

- **Precision**
- **Recall**
- **AUC**
Coupled Collaboration+ Sensing: CollabCam

Reduce Sensor Sampling Energy by Reducing Camera Image Resolution

Overlapping FoVs

Object Matching

Estimated Overlap

Resolution Reduction

[Diagram showing overlapping FoVs, object matching, and estimated overlap between camera 1 and camera 2.]

Estimated overlap between camera 1 and camera 2.
CollabCam: Mixed Resolution & Accuracy

Overall Networked Vision Sensing Architecture
Opportunity #3: Sensing Energy Savings

Camera Prototype

- CMUCam-5 (Pixy 2 Platform)
- Camera: Aptina MT9M114 CMOS
- NXP LPC4330 dual-core ARM processor @ 204MHz
- 264kb SRAM | 2 Mb Flash Memory
- Firmware modification → mixed resolution capture

Observation from Experiments:

- Reduced Resolution lowers sensing energy
- Energy proportionality requires additional adaptive clocking of sensor

Sensing Energy Savings

Net Energy = Total Energy – Baseline Energy

ML & Network Status Coupling:

- DNN can adapt to differing resolution and data rates from individual sources
- Data rate selection can depend on network congestion+ device state
Takeaways & Reflections

New opportunities:
• Edge-Coordinated Activation of sensing on wearable devices.
• Combination of passive RF sensing + battery-less wearable/IoT devices
• Edge ML needed to perform real-time multi-modal inferencing

• ML Coordination between a set of distributed edge (IoT) & wearable devices
• Run-time Collaboration: Improve Accuracy, Energy & Latency
• Collaborative ML (Training) requires
  • new DNN architectures
  • network-aware DNN adaptation
Edge Computing at Present

- Offload computation to a nearby, powerful-computational entity
  - **Edge provides isolation and resource augmentation**

- **Advantages**
  - Low-latency, real time ML pipelines
  - Data privacy
  - Energy-efficiency

- **Isolated Interaction between individual device & “cloudlet”**
My Vision: Cognitive Edge for IoT

Edge enables CMI

Edge as a
- Matchmaker/broker for IoT (mobile) devices
- ML-as-a-Service
- Monitor for Resiliency

(1) “Scene” summary + feature statistics
(2) Establish Collaborative ML
(3) Training & Model Compression
(4) Priors & Collaborative Inferencing
Challenges for The Cognitive Edge

• **Find Useful Spatiotemporal Correlations among Devices**
  - Minimizing Communication Overhead
  - Handling Disparate Sensing Modalities
  - Handle Redundancy in Dense IoT Deployments

• **Enable trusted interactions among Devices**
  - Find Correlations from non-sensitive Metadata/Features
  - Identify and isolate malicious/non-conformant devices

• **Handle Dynamic Workloads**
  - Mobile devices that temporarily reside in specific areas
  - Changes in spatiotemporal human/event patterns
Takeaways & Reflections

New opportunities:
• Edge-Coordinated Activation of sensing on wearable devices.
• Combination of passive RF sensing+ battery-less wearable/IoT devices
• **Edge ML needed to perform real-time multi-modal inferencing**

• ML Coordination between a set of distributed edge (IoT) & wearable devices
• Run-time Collaboration: Improve Accuracy, Energy & Latency
• **Collaborative ML (Training) requires**
  • new DNN architectures
  • network-aware DNN adaptation

• Edge as a Dynamic Matchmaker & Orchestrator between “dumb” IoT Devices
Conclusion

- **Need for greater interaction between wearable devices & edge computing/network entities**
  - Key to 100-fold decrease in power consumption on pervasive platforms

- **Need for inferencing orchestration among edge devices**
  - Significant opportunities for scaling up ML-based applications
  - Need for standardized models for distributing computational state
  - Need for stackable ML models for accommodating sensing diversity

- **Need for Edge Platforms to be enablers of such multi-device orchestration**
  - Need to rethink the role of edge computing
  - Adaptive computational resources to support DNN vs. network tradeoffs

(E) archanm@smu.edu.sg
(U) https://sites.google.com/view/archan-misra