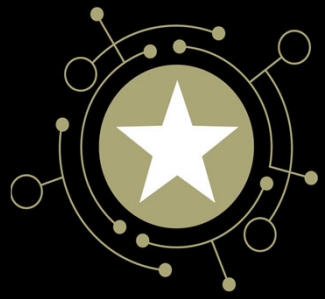
Internet of Battlefield Things

COLLABORATIVE RESEARCH ALLIANCE



Intelligence At The Point Of Need for Tactical Edge Operations

Brian Panneton, Research Area Lead
DEVCOM, Army Research Laboratory

Prashant Shenoy, Research Area Lead
University of Massachusetts Amherst

March 18. 2022

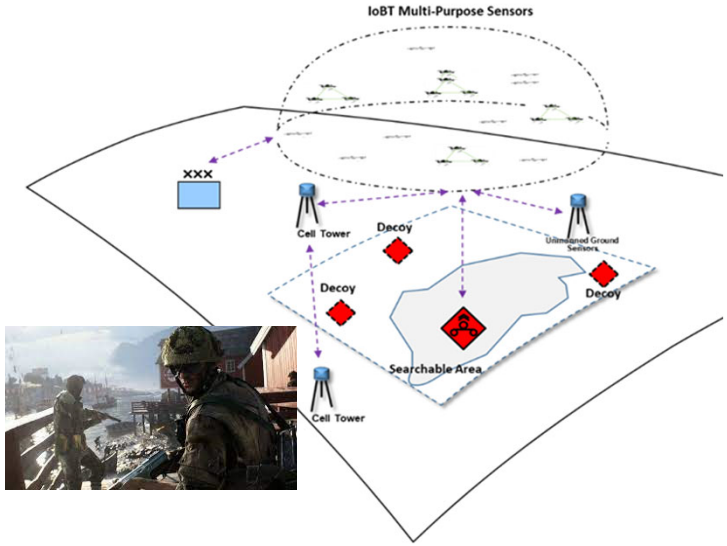
ILLINOIS · USC University of Southern California · UCLA · UMass Amherst · Carnegie Mellon University · Georgetown University · SRI International · Stanford University · The University of Texas at Austin · DEVCOM Army Research Laboratory

OVERVIEW

Intelligence at the Point of Need

GOAL

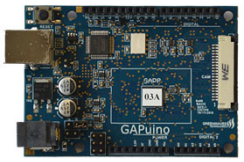
- Provide intelligence to Commanders, Staff and Soldiers through **edge processing and communication** in **resource constrained** environments
- Provide intelligence by **fusing multiple data modalities, unconventional sensing, and distribution of processing** across heterogeneous edge devices while quantifying **uncertainty**



NEW SCIENCE

Fundamental advances in multi-modal sensing, distributed inference, edge computing, and uncertainty quantification

- New techniques to exploit multiple, unconventional sensing modalities to infer side-channel information
- New distributed processing techniques to enable heterogeneity and flexible tradeoffs at the point of need: multi-modal, robust, and adaptive
- Extending uncertainty quantification to edge systems for improved decision making



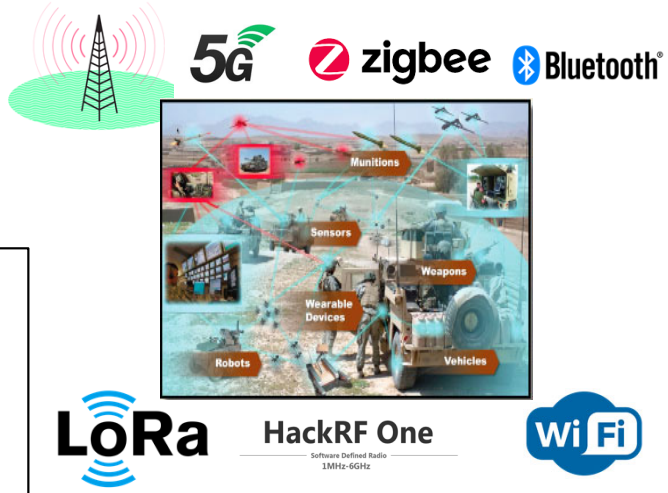
RESEARCH CONTRIBUTIONS

IoBT REIGN

DEVCOM ARMY RESEARCH LABORATORY

Intelligence at the Point of Need

Multi-modal sensing and distributed prediction at the tactical edge

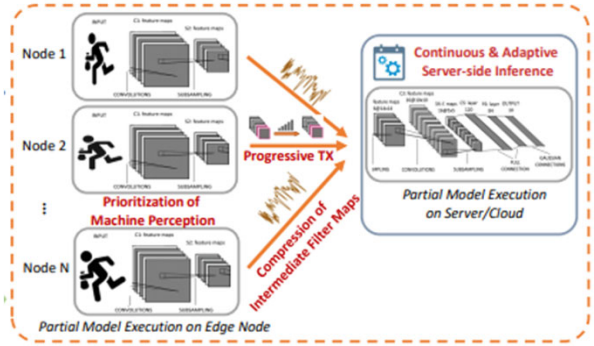


Opportunistic and Unconventional Sensing

- Exploit non-visual commodity sensors to extract information beyond their intended sensing capability
- Non-line-of-sight sensing through walls, fog, and in the dark
- Multi-modal sensing using RF (LTE, LoRa, WiFi), LiDAR, audio, motion, thermal

Adaptive, Distributed Prediction

- Distributed neural network inference across heterogeneous hardware resources in contested environments
- Enable flexible trade-offs between accuracy/latency of prediction services and resources available through run-time adaptation



Theory of Uncertainty Quantification

- URSAbench: Analysis of the relation between uncertainty, latency, and resource use
- Reduce uncertainty by separating complex scenes into regions based on the tolerable uncertainty of the object states
- Algorithms and architectures for uncertainty/cost tradeoffs

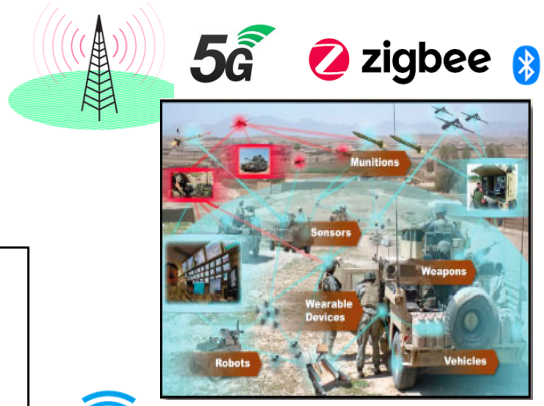
RESEARCH CONTRIBUTIONS

IoBT REIGN

DEVCOM ARMY RESEARCH LABORATORY

Intelligence at the Point of Need

Multi-modal sensing and distributed prediction at the tactical edge

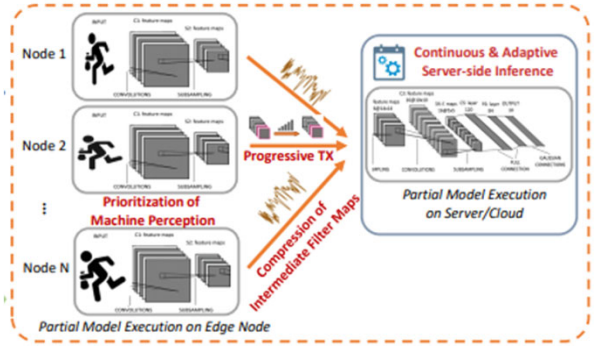


Opportunistic and Unconventional Sensing

- Exploit non-visual commodity sensors to extract information beyond their intended sensing capability
- Non-line-of-sight sensing through walls, fog, and in the dark
- Multi-modal sensing using RF (LTE, LoRa, WiFi), LiDAR, audio, motion, thermal

Adaptive, Distributed Prediction

- Distributed neural network inference across heterogeneous hardware resources in contested environments
- Enable flexible trade-offs between accuracy/latency of prediction services and resources available through run-time adaptation



Theory of Uncertainty Quantification

- URSAbench: Analysis of the relation between uncertainty, latency, and resource use
- Reduce uncertainty by separating complex scenes into regions based on the tolerable uncertainty of the object states
- Algorithms and architectures for uncertainty/cost tradeoffs

OPPORTUNISTIC AND UNCONVENTIONAL SENSING

IoBT REIGN DEVCOM ARMY RESEARCH LABORATORY

Opportunistic and unconventional sensing enables sensing/localizing/tracking capabilities using commodity radios or commodity sensors beyond their intended purpose

Army Relevance and Value Proposition

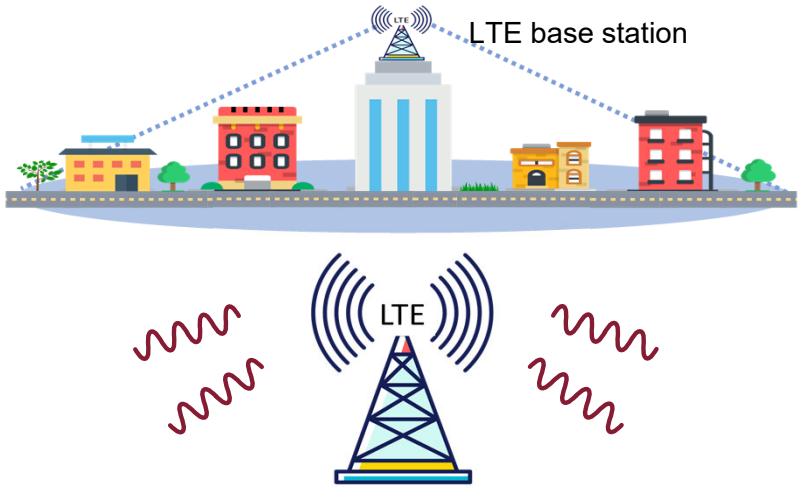
Commercially available commodity sensors will augment conventional military sensing capabilities by providing opportunistic sensing given their wide proliferation at a significantly lower cost

Prior State-of-the-Art

- Commodity sensors and IoT devices designed for single-purpose sensing need
- Commodity radios embed side-channel information that has not been exploited by sensing methods

Technical Approach

- Focus on RF emissions to penetrate objects at lower frequencies (Beyond Line of Sight)
- Detect signal reflection, amplitude and phase of object (Reasoning about the type and features of material)
- Detect phase of reflected signals offers a vibrometry capability (Detection of breathing, speech and equipment)
- Detect changes in magnetic field due to electrical current (Detect and localize electronics and their functionality)
- Reconstruct audio from accelerometer data by frequency unfolding to reconstruct high-frequency signals from low frequency signals



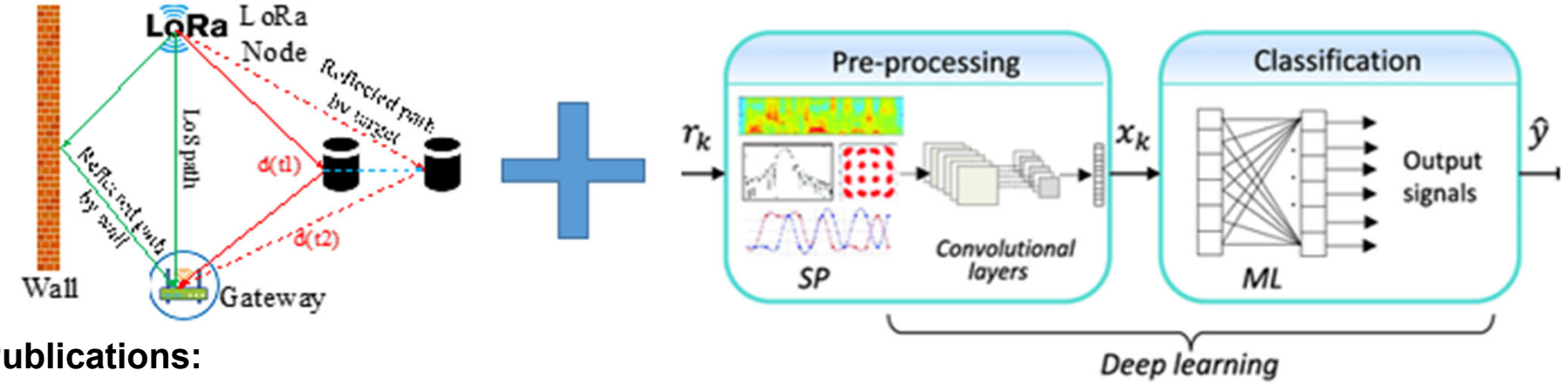
Contribution: Widely-deployed commodity radios and sensors can be exploited to perform opportunistic sensing of the environment

UNCONVENTIONAL SENSING: LTE-BASED PERVASIVE SENSING

Army Relevance: Gain situational advantage by using existing pervasive RF signals to both sense and communicate

Contribution: Designed novel signal processing schemes to combat against the severe interference and low SNR to make LTE sensing possible: first-of-its kind demonstration of pervasive sensing using ambient 4G LTE signals

- Utilize signal variations to sense the context of targets such as human respiration rate and car speed
- Use pervasive LTE signals to sense both human targets and non-human objects across indoor & outdoor environments
- Utilize the unique properties of LTE to address fundamental issues such as performance instability



Key Publications:

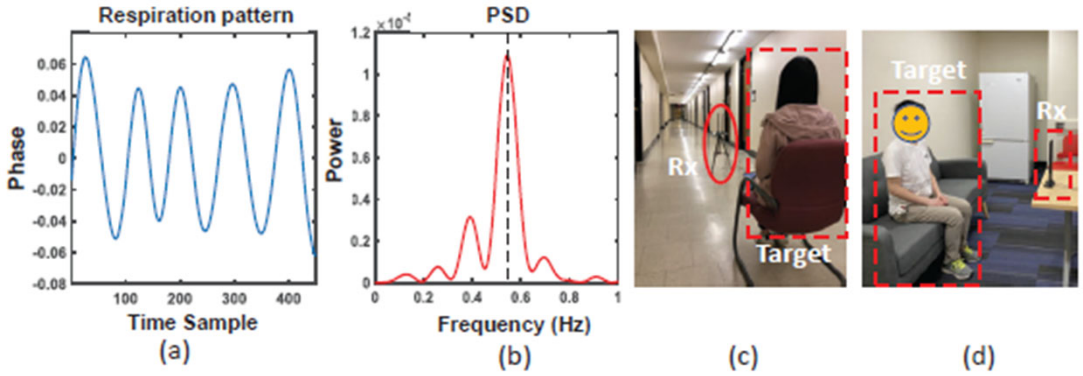
LTE-based Pervasive Sensing Across Indoor and Outdoor
F. Yuda, X. Yaxiong, D. Ganesan, X. Jie, *ACM SenSys*, 2021.

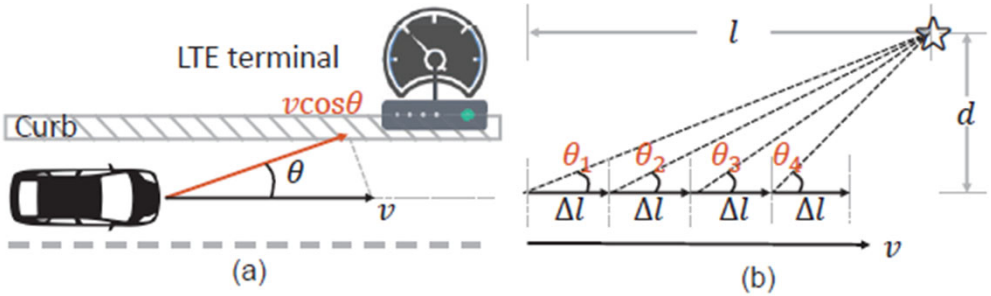
Github Code: LTScope
https://github.com/kyleatprinceton/LTEScope_release

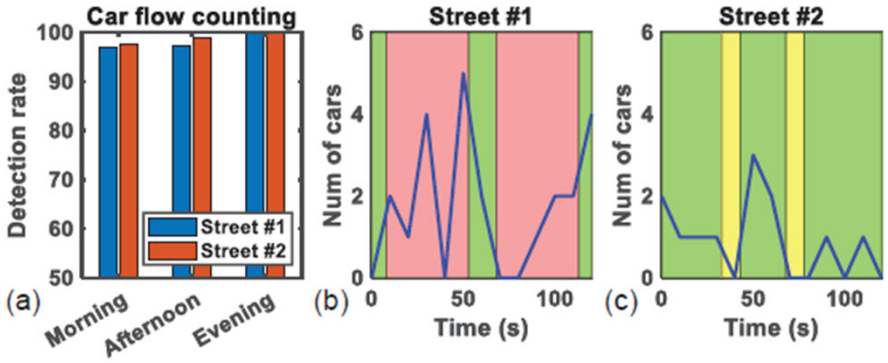
UNCONVENTIONAL SENSING: LTE-BASED PERVASIVE SENSING

Experiments and Results:

- Human presence and Respiration Sensing
  - Highly accurate with the blind spot and orientation-sensitive issues mitigated
- Vehicle speed Monitoring
  - Less than 2 mph error in car speed estimation







Summary:
Deploying LTE sensing capability can enable situational awareness in areas with 4G cell coverage

Demo video available: [Multimodal Sensing for Human Monitoring]

PARTITIONED EXECUTION OF AI MODELS

Optimize the allocation of distributed analytics by exploiting heterogeneous edge computing

Army Relevance and Value Proposition

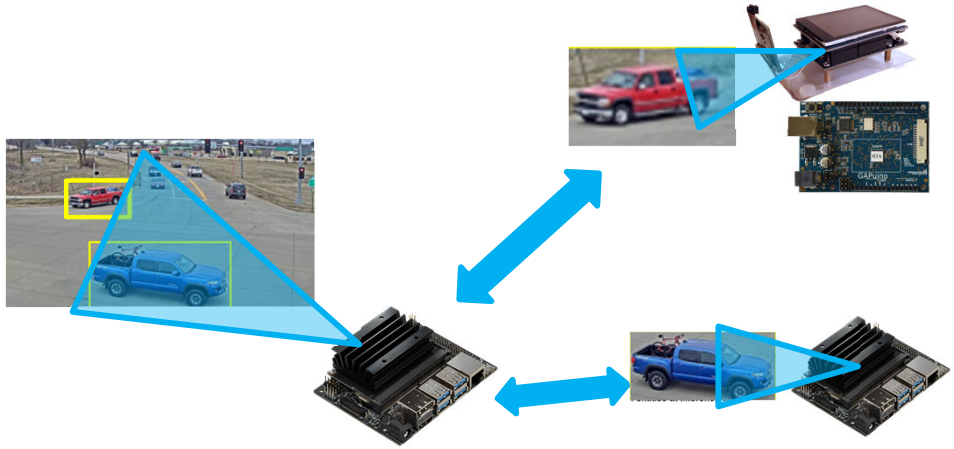
Intelligent situation-awareness at the point of need require distributed analytics that can operate over a complex IoBT network, can tolerate resource outages and bottlenecks, and are robust to adversarial agents

Prior State-of-the-Art

- Inference models are run over reliable networks connecting resource-rich homogeneous hardware
- Local computation is limited to single edge resources and is not adaptive for optimal decision making

Technical Approach

- Optimally distribute execution of AI models on heterogeneous edge devices under computation and communication constraints
- Develop prediction methods that are robust to channel dynamics or resource fluctuations at inference time
- Run-time adaptation of latency, accuracy, energy tradeoffs using novel neural network architectures



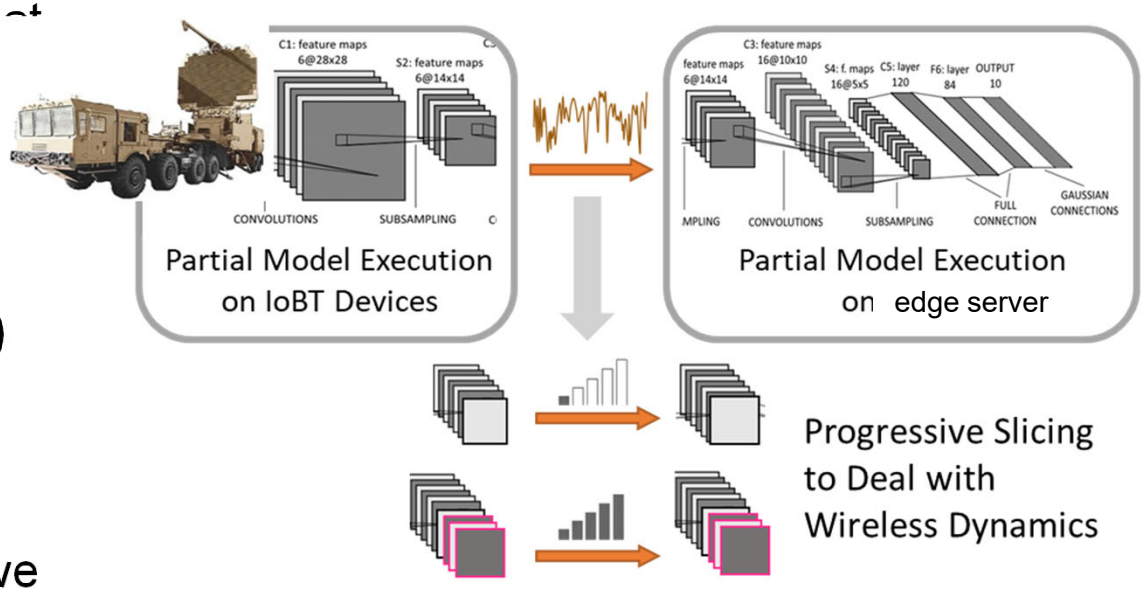
Contribution: Adaptive, distributed machine learning prediction systems can achieve maximally robust trade-offs between performance and use of available resources under varying constraints and requirements

PARTITIONED EXECUTION OF AI MODELS: CLIO

Army Relevance: Enable baseline situational awareness on resource constrained devices in contested areas; Increase situational awareness when more resources are available

Contribution: Distribute neural network models across heterogeneous hardware resources assuming a wireless network link with stochastic channel capacity

- **Adaptive partitioning**: dynamically partition how many layers are executed at device and edge server
- **Progressive slicing**: dynamically vary the amount of internal result sent over network
- Optimally choose **best (partition, slice)** to maximize accuracy for available wireless bandwidth
- Combine CLIO's resource partitioning with network compression in compressive offloading (DeepCOD)



Key Publications:

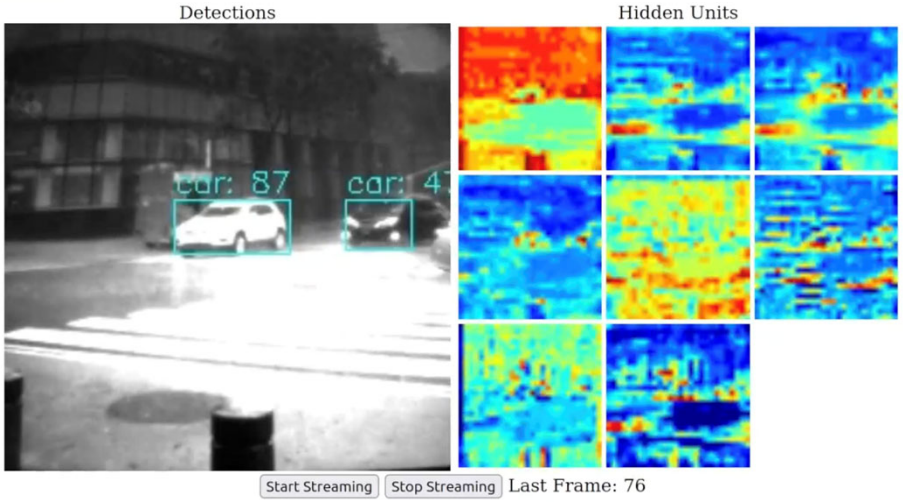
Optimizing Intelligent Edge-Clouds With Partitioning, Compression and Speculative Inference
F. Fang, J. Huang, C. Samplawski, D. Ganesan, B. Marlin, T. Abdelzaher, M. Wigness, *IEEE MILCOM*, 2021.

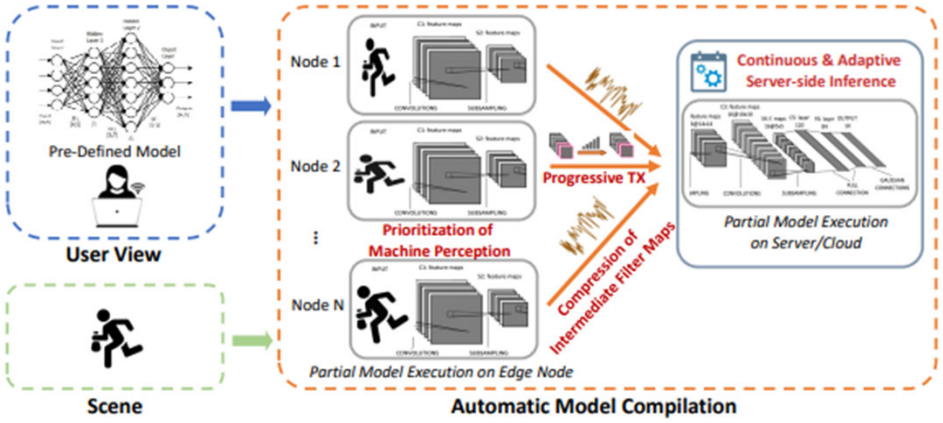
CLIO: Enabling Automatic Compilation of Deep Learning Pipelines Across IoT and Cloud
J. Huang, C. Samplawski, D. Ganesan, B. Marlin, H. Kwon, *Int. Conf on Mobile Computing and Networking*, 2020.

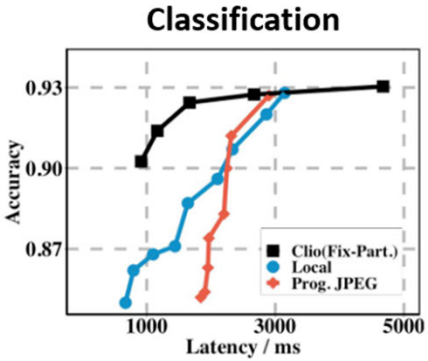
PARTITIONED EXECUTION OF AI MODELS: CLIO

Experiments and Results:

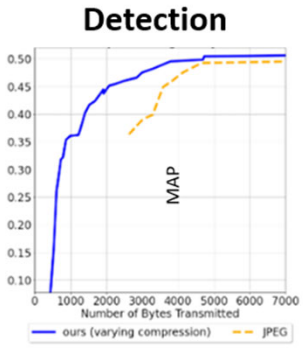
- Approach implemented using the ultra low power Gap8 neural accelerator platform
- Model partitioning and progressive slicing applied to image classification and object detection.
- 300% faster predictions with same accuracy on very resource-constrained edge devices



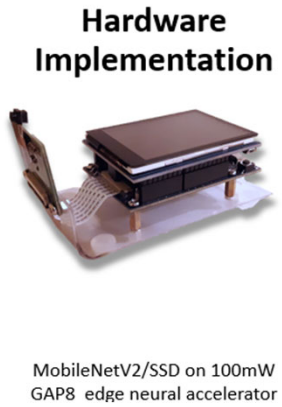




20-class ImageNet using ResNet 18



Pascal VOC with MobileNetV2/SSD



MobileNetV2/SSD on 100mW GAP8 edge neural accelerator

Demo videos available: [Distributed NN Partitioning with CLIO] and [Multimodal Sensing for Human Monitoring]

Summary:
Improve the resilience of machine prediction by expanding the space of accuracy-energy-latency trade-offs

FUNDAMENTALS OF UNCERTAINTY QUANTIFICATION

IoBT REIGN DEVCOM ARMY RESEARCH LABORATORY

Fundamental theories and analysis tools to quantify uncertainty propagation through IoBT networks in adversarial environments

Army Relevance and Value Proposition

An AI/ML system that correctly recognizes when it is “not sure” is key to the safe integration of machine intelligence into the battlefield. Machine’s assessing risk prior to action can reduce catastrophic failure and loss of trust

Prior State-of-the-Art

- Uncertainty quantification and mitigation from sensor data is well understood
- Much less understanding of the uncertainty trade-offs and performance limits when AI and deep neural networks execute in the tactical edge

URSA Bench
Uncertainty • Robustness • Speed • Accuracy
https://github.com/reml-lab/URSABench

Technical Approach

- Automated benchmarking tools to estimate uncertainty of AI models on edge devices
- Reduce uncertainty by separating complex scenes into regions based on the amount of tolerable uncertainty of the object states
- Improve robustness to considering platform uncertainty (latency) in the RL model

Contribution: Quantifying uncertainty from AI models and edge platforms enables human decision makers to better incorporate machine predictions into decision making

FUNDAMENTALS OF UNCERTAINTY QUANTIFICATION: URSABENCH

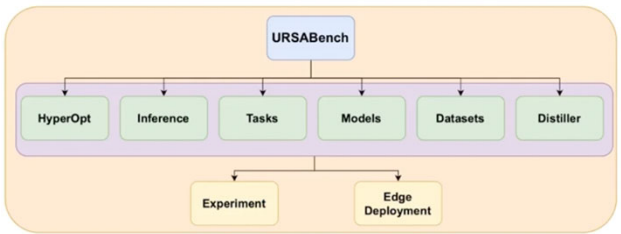
Army Relevance: Moving uncertainty quantification to the edge to provide confidence estimates to aid in tactical edge decision-making

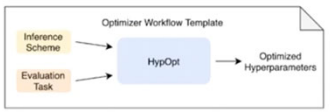
Contribution: Development and evaluation of robust methods for accelerating uncertainty quantification for deep neural network models at the edge

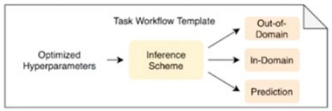
- Multi-faceted benchmarking tool for characterizing
  i. accuracy: prediction accuracy
  ii. uncertainty: log likelihood, model calibration
  iii. robustness: out-of-distribution classification
  iv. scalability: memory and execution cost in edge
- Automated model compilation and assessment for DNN models to quantify URSA metrics on edge devices
- *Potential Transition*: URSABench is being refined with collaboration from the JAIC

URSA Bench
Uncertainty • Robustness • Speed • Accuracy
https://github.com/reml-lab/URSABench

Bayesian Deep Learning with URSA Bench Uncertainty • Robustness • Speed • Accuracy







Key Publications:

A System for Comprehensive Benchmarking of Bayesian Deep Neural Network Models and Inference Models
M. Vadera, J. Li, T. Abdelzaher, B. Marlin, *Conference on Machine Learning and Systems*, 2022.

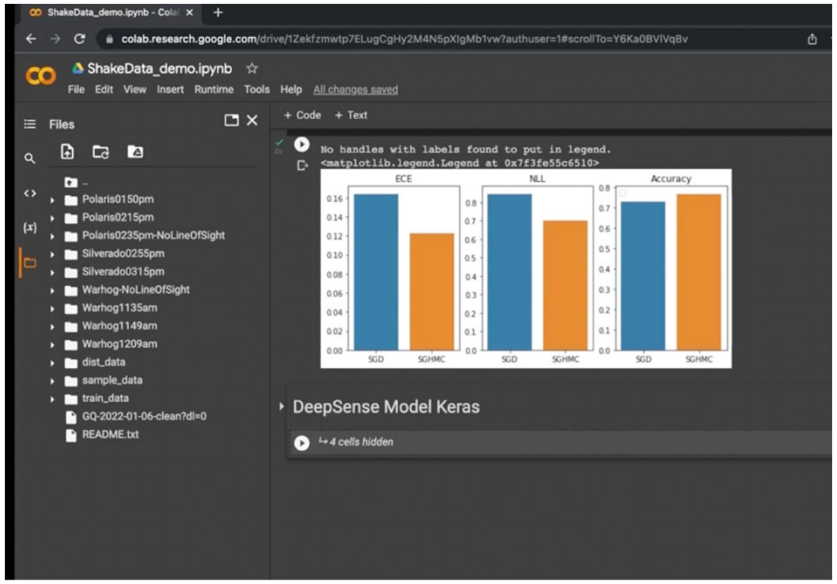
URSABench: Comprehensive Benchmarking of Approximate Bayesian Inference Methods for Deep Neural Networks
M. Vadera, A. Cobb, B. Jalaian, B. Marlin, *Workshop at Int. Conf on Machine Learning*, 2020.

FUNDAMENTALS OF UNCERTAINTY QUANTIFICATION: URSABENCH

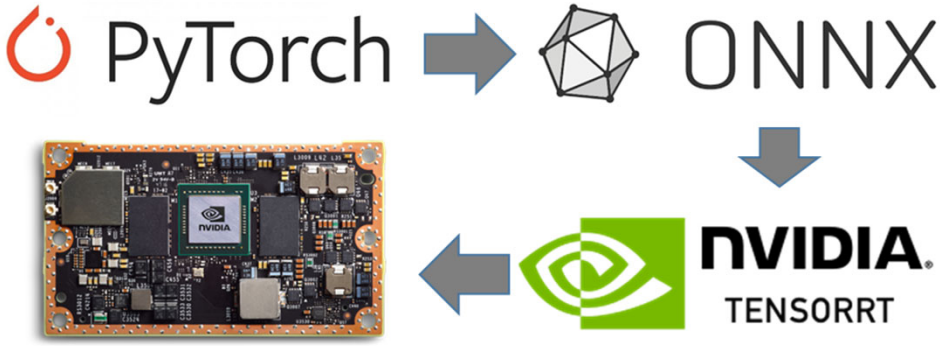
Experiments and Results

- URSABench provides modules for compiling a variety of approximate Bayesian deep learning models for deployment and benchmarking on NVIDIA Jetson edge devices.
- Initial results obtained for the Jetson TX2 platform for several models, data sets, and approximate inference approaches

Inference	Accuracy ↑	NLL ↓	Robustness ↑	Uncertainty ↑	Scalability ↓
SGLD	0.869	0.524	0.803	0.916	129.3
SGHMC	0.868	0.539	0.808	0.916	129.3
cSGLD	0.892	0.396	0.810	0.912	2103.3
cSGHMC	0.886	0.443	0.798	0.898	2114.9
SWAG	0.824	0.735	0.759	0.885	1351.7
PCA + ESS (SI)	0.869	0.482	0.804	0.901	1940.0
MC dropout	0.872	0.554	0.775	0.914	127.6
SGD	0.861	0.625	-	-	127.7



Demo video available: URSABench



Summary:
Quantitative understanding of accuracy-robustness-speed-storage tradeoffs in resource-constrained edge environments

IOBT CRA IMPACT

Before IoBT CRA		With IoBT CRA
Commodity sensors capability limited to designed purpose	**Unconventional Sensing** →	Opportunistic use of commodity radios as sensors and use of side-channel information in sensor data
Machine learning inference on resource-rich computing resources	**Distributed Edge Inference** →	Distributed ML predictions in heterogeneous edge networks by flexibly adapting to wireless and resource dynamics
Uncertainty information limited to sensed data and limited knowledge of uncertainty from AI models	**Uncertainty Quantification** →	Theory of uncertainty quantification in edge networks that run deep learning and Bayesian inference

NEXT STEPS

IoBT Cornerstone Challenges: Coordination – Heterogeneity – Scale

- How can commodity radios be designed for **coordinated joint sensing and communication**?
- How can multiple unconventional sensing modalities be fused together **flexibly based on dynamic availability** for better situational awareness?
- How can edge processing **prioritize what information to observe and report** in heterogeneous networks that produce enormous volumes of sensor data?
- What are the theoretical foundations of **multi-modal** and **multi-vantage** confirmation/verification? How can we quantify uncertainty propagation through large, multimodal, heterogenous networks?

Prospective Army Capabilities

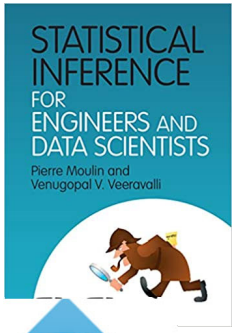
Current modernization efforts are working towards intelligence at the point of need. Better sensing capabilities empowers better decision making. Making use of unconventional sensing and model partitioning can reduce the resources required, thus pushing the technologies further to edge and providing more effective situational awareness under adversarial conditions

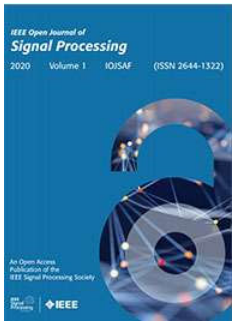
ACCOMPLISHMENT SUMMARY

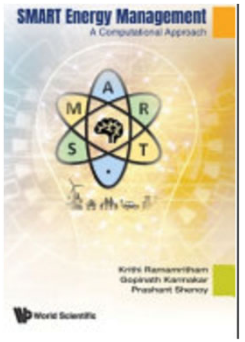
Questions?

Notable Awards/Honors:

- **Society Fellow Honors:**
  - IEEE Fellow, 2021, (T. Abdelzaher)
  - ACM Fellow, 2020, (T. Abdelzaher)
  - ACM Fellow, 2020, (P. Shenoy)
  - NAE member 2022, (K. Nahrstedt)
- **Journal and Conference Leadership**
  - PC Chair, IEEE Infocom'21 (T. Abdelzaher)
  - General Chair, IoTDI'20 (P. Shenoy)
  - Editor, IEEE Open Journal of Signal Processing (V.V. Veeravalli)
- **Books**
  - Statistical Inference for Engineers and Data Scientists, Cambridge U. Press (V.V. Veeravalli)
  - Smart Energy Management, World Scientific Press (P. Shenoy)
- **Best paper Awards:**
  - RTSS (T. Abdelzaher)
  - Sensys (T. Abdelzaher)
  - Sensys (P. Shenoy)
- **Professorships**
  - Distinguished Professor UMass, (P. Shenoy)







acm

IEEE INFOCOM

ACM SenSys 2020

IoTDI 2020

URSA Bench
Uncertainty • Robustness • Speed • Accuracy
https://github.com/reml-lab/URSABench

Notable Transitions:

- **[DoD]** URSABench refinement with JAIC for possible transition
- **[Industry]** Boeing award to develop digital twin for Navy
- **[Medical AI]** MassAITC NIH Center on mobile/wearable sensing and AI

Demos & Posters: Visit https://abdelzaher.cs.illinois.edu/RMB22-Demos.html

- Vital Sign Assessment Using Multimodal Sensing
- Distributed NN Partitioning with CLIO
- URSABench
- Unconventional Sensing: Audio Sensing with Accelerometers
- Exposing Hidden Sensors