Knowledge Product: IoBT Edge Efficiency (Time is a Weapon)

ARL POC: *Gunjan Verma, gunjan.verma.civ@army.mil, (810) 207-1098*
Consortium POC: *Christina Fragouli, christina.fragouli@ucla.edu, (310) 206-8795*

OPPORTUNITY: The proliferation of battlefield sensors and tactical data analytics enables the execution of various warfighting functions (such as target detection, identification, and localization) at an unprecedented speed and scale. A key challenge in realizing this potential is taming the combined latency of the various machine learning (ML) and signal processing algorithms, whose sum total will otherwise contribute significantly to the lag from data to decision-making.

GOAL: Significantly reduce the latencies involved in the sensor-to-decision loop by designing for timeliness and efficiency of intelligent processing on edge devices. Rapidly create IoBT networks capable of executing a variety of mission functions. Reduce the amount of data required to accurately detect events of interest. Minimize the data that need to be communicated for successful downstream inference.

NEW SCIENCE: The IoBT CRA has developed significant theoretical and empirical breakthroughs in edge efficiency thanks to: (1) advances in optimization theory that enable fast algorithms for *network creation and re-configuration* to meet objectives, (2) development of new statistical theory for rapid *networked anomaly detection* at low false-alarm rates in distribution-agnostic regimes, (3) extensions of information theory to enable highly compressed *network inference* while increasing quality, (4) advances in compressive-sensing inspired neural networks to realize data *compression* algorithms for extremely low-latency and low-bandwidth network communications, (5) introduction of a new paradigm for prioritizing ML processing to focus resources on the most important features first, and (6) advances in information theory and statistical learning theory creating the first *theory of model compression* that explains why dramatically smaller models (and hence faster *inference*) can actually *improve* accuracy.

We extend the theory of submodular optimization to produce a very general analytical framework for network synthesis which maximizes sensing quality; we are able to formulate very general communication and computation constraints in a suitable way so as to still provide theoretical guarantees on the optimality of the resulting problem and demonstrate it admits a tractable solution. We develop new (Bayesian) statistical method inspired test statistics which can rapidly detect anomalies in networks while achieving very low false alarm rates. We incorporate the theory of compressive sensing into neural network design to create a state-of-the-art offloading system capable of highly compressed, very low-bandwidth, very low-latency joint edge/server joint processing. The resulting framework consists of a highly asymmetric encoder/decoder to account for the fact the tactical edge (cloud) is heavily (lightly) resource constrained. We connect the ideas of rate-distortion theory from the field of information theory with the ideas of model complexity and generalization error of statistical learning theory to provide the first ever theoretical argument for why dramatically compressing a model can also improve its accuracy.

SCIENTIFIC CHALLENGES AND RISK: The above contributions are noteworthy in overcoming several key challenges, namely: (1) expediting operations may trade off speed for decreased accuracy – one must ensure that accuracy goals are not compromised; (2) ML techniques are not free of misclassifications – solutions must estimate correct confidence in results; (3) the relevant networking problems are rooted in combinatorial optimization, which is intractable – scalable but provably accurate approximations are needed; and (4) non-stationary environments necessitate frequent model re-training – overhead must be minimized. Risks include overheads of extreme scale, potentially poorly understood cross-modality trade-offs, uncertainty regarding data prioritization, and adversarial actions that degrade decision timeliness.

RESULTS FROM EXPERIMENTATION: All innovations have been implemented and tested on well-established datasets in the scientific community as well as in experiments at facilities at consortium institutions and ARL. Network synthesis algorithms input real map imagery and result in actual sensor deployments that have been tested in a dense urban environment (Los Angeles). Rapid change detection algorithms have been demonstrated on data collected from ARL's Multi-Sensor Array (MSA) at White Sands in which our algorithms used non-visual sensors to detect motion in less than half the time required by the state-of-the-art. Compressive offloading on real world wireless links showed 2-10x improvements over the state-of-the-art and was used in ARL's **ACROPOLIS** experiment (a side effort to **PC21**) to efficiently communicate imagery between a robot and a ground base, in which offloading was critical to enable mission success. Highly compressed distributed model training algorithms have been implemented and tested on several benchmark datasets and also on acoustic and seismic data collected from experiment at ARL's R2C2 Campus (at Graces Quarters), showing communication savings of 5x-100x.

TECHNOLOGY TRANSITION: Our edge Efficiency and fast network configuration innovations resulted in an ongoing transition (funded by Boeing Inc. at approximately $0.5M) to an IoBT Digital Twin project. The digital twin will encapsulate IoBT's rapid anomaly detection and optimization/configuration algorithms to improve agility of the computational substrate of the tactical edge. Boeing is considering Navy application.

IMPACT ON THE SCIENTIFIC COMMUNITY: The work has produced high-quality publications in the premier peer-reviewed conferences and journals including NeurIPS, ICML, Infocom, Sensys, MLSys, IEEE JSA Information Theory, IEEE Trans. Information Theory, ACM Trans. Sensor Networks, MILCOM, among many others. PIs involved in this knowledge product have been elected as fellows of the IEEE, ACM, AAAI, IFAC, and Guggenheim societies. One was inducted to the National Academy of Engineering. PIs have been recognized for their research through various research awards including the ACM Mobicom test of time award, the Adobe Data Science Research Award, and the Sandia National Labs Research Award, as well as research awards from Google, Amazon, IBM, JP Morgan, and Texas Instruments. PIs have delivered keynote lectures at various venues like IEEE COMSNETS, ACM Edge Computing, IEEE ICNP, and the NSF Workshop on Real-time Learning and Decision Making, PIs have chaired several influential conferences including ACM special interest group on energy, ACM/IEEE Conference on IoT Design and Implementation, IEEE International Symposium on Info Theory, and ACM/IEEE International Conference on Cyberphysical systems. as well as serving as president of the IEEE Information Theory Society and Claude Shannon award selection committee. Students funded under the IoBT CRA whose work has contributed to this knowledge product have graduated and been placed at Facebook, Walmart Labs, MIT, Univ. Buffalo UIUC, and the Technical University of Netherlands.

POTENTIAL ARMY CAPABILITIES: In a conflict against a near-peer adversary, connecting sensors and effects more efficiently may offer decision advantage. A recent Focused Excursion organized by DEVCOM and FCC (a collaborative investigative process that supports concept and capability development with participation from FCC DoC, DoIS, DAC, and ARL) investigated IoBT implications on select learning demands of the Army Concepts Framework 2040. Among other hypotheses, the excursion ratified the hypothesis that "*IoBT will help commanders distill prioritized information from vast amounts of data faster for decision making by providing efficient, real-time processing*". We thus envision a future in which IoBT improves tactical decision-making thanks to distributed, communications-efficient, and optimized algorithms that rapidly collect key information pertaining to Commander priorities. This information is efficiently exchanged and synthesized across networked battlefield entities, cross-validated via multi-modal and multi-vantage sensing, and passed through ML-based inference algorithms to distill insights relevant to the mission. Because there will be vastly more data available than can possibly be parsed by humans, machine learning and signal processing algorithms will play a definitive central role.

EXPERIMENTAL DEMONSTRATIONS:

See videos on IoBT Edge Efficiency Innovations at: https://abdelzaher.cs.illinois.edu/RMB22-Demos.html

KEY PUBLICATIONS*:

1. D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations." *IEEE Journal on Selected Areas in Information Theory* 1 (1), 2020.
2. Y. Bu, W. Gao, S. Zou, and V. Veeravalli, "Population Risk Improvement with Model Compression: An Information-theoretic Approach," *Entropy* 23 (10), 2021.
3. Y. Bu, S. Zou, and V. Veeravalli, "Tightening Mutual Information-based Bounds on Generalization Error," *IEEE Journal on Selected Areas in Information Theory*, 2020.
4. J. Bunton, T. Anevlavis, **G. Verma**, C. Fragouli, and P. Tabuada "Split to Win: Near-optimal Sensor Network Synthesis via Path-greedy Subproblems," In Proc. *IEEE Military Communications Conference (MILCOM)*, 2021.
5. A. Deshmukh, J. Liu, V. Veeravalli, and **G. Verma**, "Information Flow Optimization in Inference Networks," In Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
6. P. Ghosh, J. Bunton, D. Pylorof, M. Vieira, K. Chan, R. Govindan, G. Sukhatme, P. Tabuada, and **G. Verma** "Synthesis of Large-Scale Instant IoT Networks," *IEEE Transactions on Mobile Computing*, 2021.
7. P. Ghosh, J. Bunton, D. Pylorof, M. Vieira, **K. Chan**, R. Govindan, G. Sukhatme, P. Tabuada, and **G. Verma**, "Rapid Top-down Synthesis of Large-scale IoT Networks," In Proc. *29th International Conference on Computer Communications and Networks (ICCCN)*, 2020.
8. Y. Hu, S. Liu, T. Abdelzaher, **M. Wigness**, **P. David**, "On Exploring Image Resizing for Optimizing Criticality-based Machine Perception," In Proc. *IEEE RTCSA*, August 2021.
9. S. Liu, S. Yao, X. Fu, R. Tabish, S. Yu, A. Bansal, H. Yun, L. Sha, and T. Abdelzaher, "On Removing Algorithmic Priority Inversion from Mission-critical Machine Inference Pipelines," In Proc. *IEEE Real-Time Systems Symposium (RTSS)*, 2020.
10. S. Liu, X. Fu, **M. Wigness**, **P. David**, S. Yao, L. Sha, T. Abdelzaher, "Self-cueing Real-time Attention Scheduling in Criticality-driven Visual Machine Perception," In Proc. IEEE RTAS, May 2022.
11. N. Singh, D. Data, **J. George**, and S. Diggavi "Squarm-SGD: Communication-efficient Momentum SGD for Decentralized Optimization," *IEEE Journal on Selected Areas in Information Theory* 2 (3), 2021.
12. G. Rovatsos, V. Veeravalli, D. Towsley, and **A. Swami**, "Quickest Detection of Anomalies of Varying Location and Size in Sensor Networks," *IEEE Transactions on Aerospace and Electronic Systems*, 57(4), 2021.
13. G. Rovatsos, V. Veeravalli, D. Towsley, **A. Swami**, "Quickest Detection of Growing Anomalies in Networks," In Proc. *IEEE ICASSP*, February 2020.
14. J. Li, V. Veeravalli, D. Towsley, **A. Swami**, S. Zou, "Distributed Quickest Detection of Significant Events in Networks," In Proc. *ICASSP*, May 2019.
15. S. Yao, J. Li, D. Liu, T. Wang, S. Liu, H. Shao, and T. Abdelzaher. "Deep Compressive Offloading: Speeding up Neural Network Inference by Trading Edge Computation for Network Latency." In Proc. *18th Conference on Embedded Networked Sensor Systems*, 2020.

*Note: Names in **blue** are government co-authors.