Internet of Battlefield Things

COLLABORATIVE RESEARCH ALLIANCE

IoBT REIGN

U.S.ARMY

DEVCOM ARMY RESEARCH LABORATORY

Timeliness and Efficiency for Network Sensing, Communication, and Learning

Gunjan Verma, Research Area Lead
DEVCOM, Army Research Laboratory

Christina Fragouli, Research Task Lead
University of California, Los Angeles

March 18, 2022

OVERVIEW

Edge Efficiency: Time Is a Weapon

GOAL

Dramatically improve the efficiency of all facets of IoBT operations

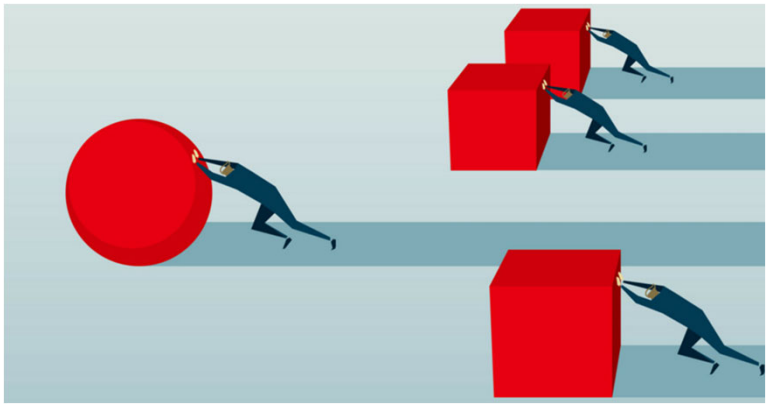
- Sensing
- Communication
- Model training
- Model inference

to enable rapid and resource-efficient realization of situational awareness

NEW SCIENCE

Fundamental advances in optimization theory, information theory, and statistical learning

- Extend theory of submodular optimization for provably near-optimal rapid network synthesis
- Information theory and nonconvex optimization for fast distributed learning
- New statistical methods for distribution-agnostic rapid change detection
- Theory of compressive sensing for very low latency tactical edge/cloud joint model inferencing



RESEARCH CONTRIBUTIONS

IoBT REIGN DEVCOM ARMY RESEARCH LABORATORY

Edge Efficiency: Time Is a Weapon

Accelerating the speed of data-to-decision-making

Rapid sensing

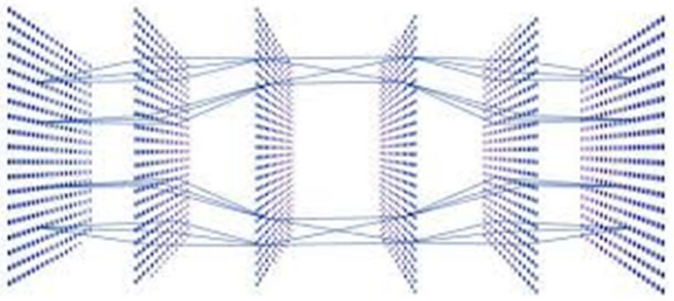
- Submodular optimization for network synthesis
- Statistical theory of rapid change detection

Rapid communication

- Communications-efficient compressed distributed learning (“compressing the data for training”)
- Compressive offloading (“compressing the data for inference”)

Rapid learning and inference

- Criticality-based attention prioritization
- Adaptive run-time execution of DNN inference (speed/energy tradeoffs)
- Algorithms for model compression
- Theory of model compression (“compressing the model parameters”)



RESEARCH CONTRIBUTIONS

Edge Efficiency: Time Is a Weapon

Accelerating the speed of data-to-decision-making

Rapid sensing

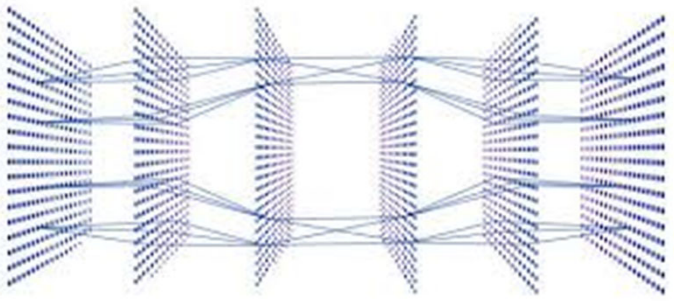
- Submodular optimization for network synthesis
- Statistical theory of rapid change detection

Rapid communication

- Communications-efficient compressed distributed learning (“compressing the data for training”)
- Compressive offloading (“compressing the data for inference”)

Rapid learning and inference

- Criticality-based attention prioritization
- Adaptive run-time execution of DNN inference (speed/energy tradeoffs)
- Algorithms for model compression
- Theory of model compression (“compressing the model parameters”)



RAPID NETWORK SYNTHESIS

IoBT REIGN

DEVCOM ARMY RESEARCH LABORATORY

Provably near-optimal algorithm to rapidly place and connect sensors into a network maximizing sensing utility while respecting communications and computation constraints

Army Relevance and Value Proposition

IoBT requires rapid (re-)synthesis of mission-capable networks from IoBT assets to enable Army operations in dense urban environments

Prior State-of-the-Art

- Network deployments in civilian context are methodically planned over months/years
- Army network designs are typically optimized for a single objective
- No techniques exist for synthesizing on-demand networks capable of optimizing for multiple goals

Technical Approach

- Frame problem in very general terms: Select a set of locations to deploy sensors that maximizes a flexible notion of utility subject to routing, communication, and data processing constraints
- Combinatorial optimization problem (NP hard) - Propose a lightweight algorithm based on submodular optimization that selects a set of locations satisfying constraints with provably near-optimal sensing utility

Contribution: A framework for the rapid design of synthesized networks satisfying a range of desired performance characteristics using modern optimization methods

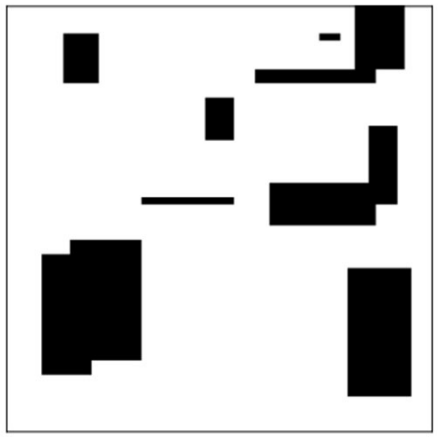
RAPID NETWORK SYNTHESIS

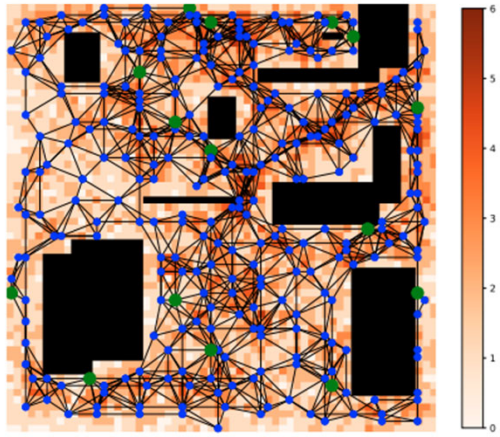
IoBT REIGN DEVCOM ARMY RESEARCH LABORATORY

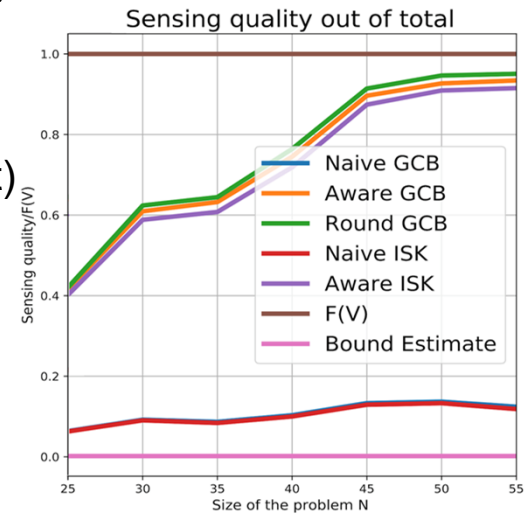
Army Relevance: Rapid reconfiguration of the tactical network given dynamic battlefield constraints and priorities will provide flexibility

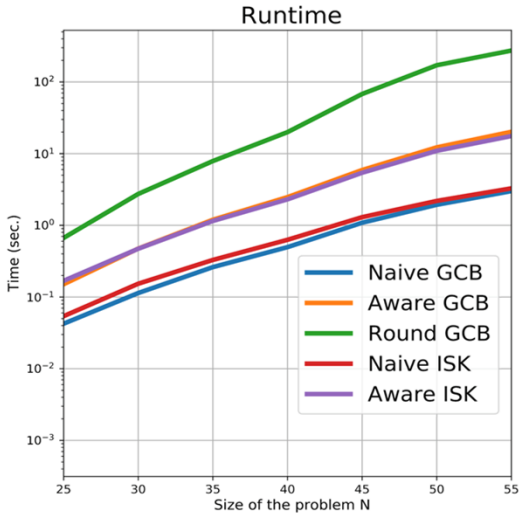
Contribution: An algorithm to select a subset of locations $A \subseteq V$ to place sensors that maximizes a sensing quality measure F(A) subject to constraints on communication and computation

- Framework is very general; uses a generic model of sensor utility (any submodular set function), communications (arbitrary weighted graph), and data processing limits (additive budget constraint)
- Lightweight algorithm with near-optimality properties from submodular maximization
- Key insight: decompose problem into clusters of sensor locations (each tied to 1 compute element)
- Reformulate problem from an arbitrary set of locations to a set of distance-constrained paths









Summary:

Fast solutions to NP-hard IoBT synthesis

Demo video available: Rapid Network Synthesis

Key Publications:

Split to Win: Near-optimal Sensor Network Synthesis via Path-greedy Subproblems
J. Bunton, T. Anevlavis, G. Verma, C. Fragouli, P. Tabuada, *IEEE MILCOM (pp. 789-794)*, 2021.

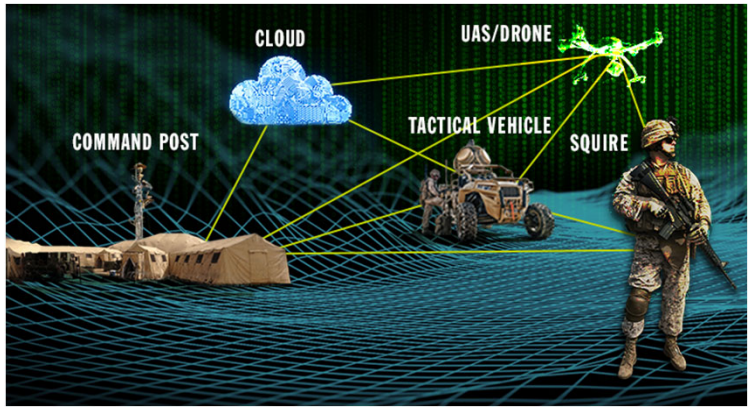
COMPRESSIVE OFFLOADING

IoBT REIGN DEVCOM ARMY RESEARCH LABORATORY

An algorithm for joint tactical edge/server computing that achieves a dramatic reduction in communications and latency

Army Relevance and Value Proposition

Execution of complex tactical ML pipelines in congested wireless environments on resource-constrained edge devices requires highly compressed edge-data offloading

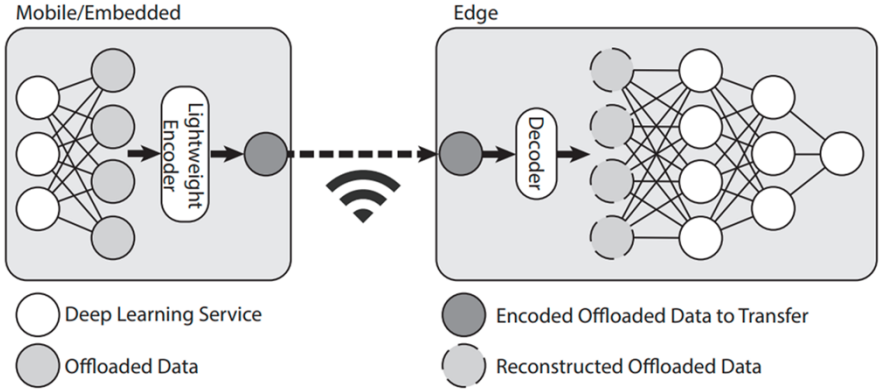


Prior State-of-the-Art

- Modern ML pipelines require high execution time and energy consumption, impeding their deployment on lower-end embedded and/or mobile sensing devices which will predominate IoBTs

Technical Approach

- Highly asymmetric encoder (edge) / decoder (server) design to account for imbalanced computational capabilities
- Theory of compressive sensing explicitly incorporated in design of objective function resulting in very small data sizes transmitted over the channel with virtually no loss in accuracy



Contribution: Design and implementation of efficient edge-based ML with significantly improved energy, latency, bandwidth, and accuracy tradeoffs

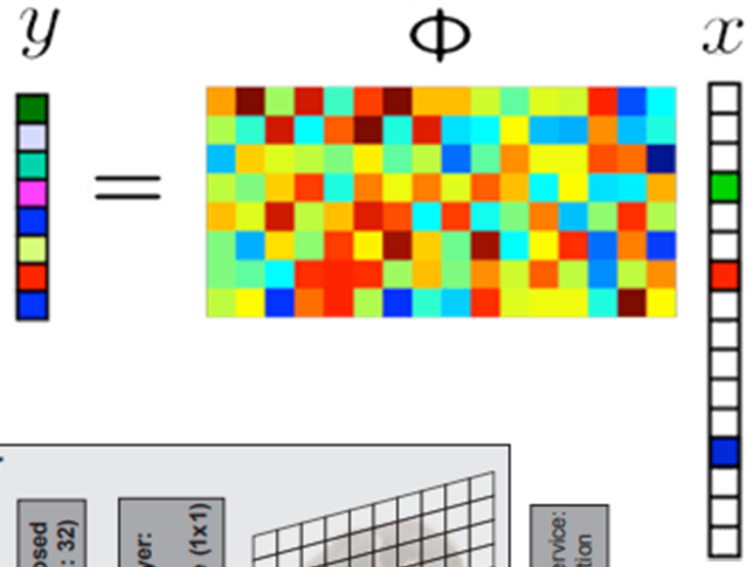
COMPRESSIVE OFFLOADING

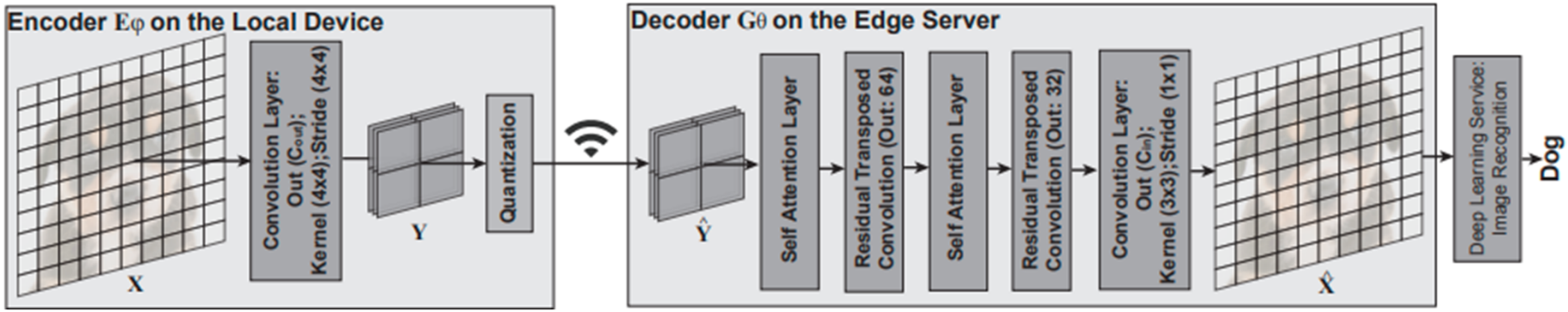
IoBT REIGN DEVCOM ARMY RESEARCH LABORATORY

Army Relevance: Enables faster operation of sensor-to-decision loops due to increased communication speed and joint analysis of sensor data

Contribution: A sensor data offload scheme which dramatically outperforms the state-of-the-art by reducing communications latency with almost no loss of accuracy (of downstream analytics)

- Theory of compressive sensing integrated into compressor design
- Reconstruct the encoded data in a manner that achieves the best inference result (not reconstruction).
- Substantially reduces offloading latency, while contributing only a negligible computational overhead on local end-devices, and incurring almost no degradation in inference accuracy
- Application-agnostic (applicable to multi-modal vision, speech, RF based ML models) and architecture-agnostic (applicable to any machine learning or signal processing algorithm)





Key Publications:

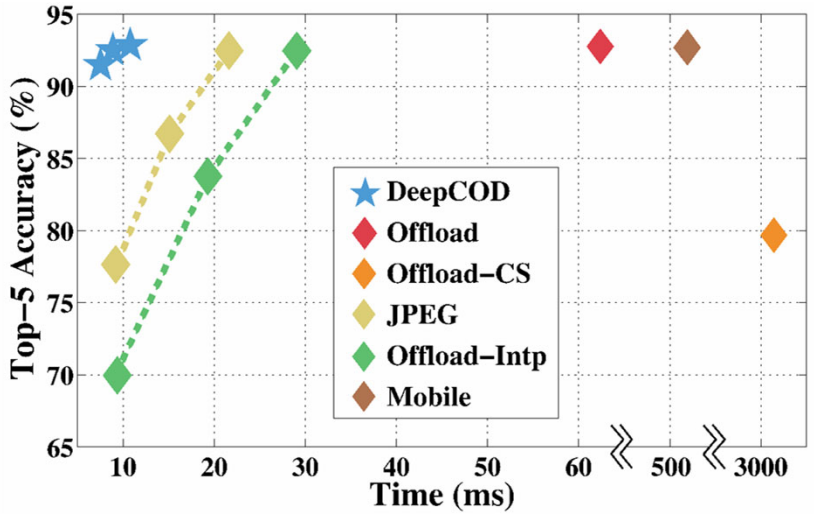
Deep Compressive Offloading: Speeding up Neural Network Inference by Trading Edge Computation for Network Latency
S. Yao, J. Li, D. Liu, T. Wang, S. Liu, H. Shao, T. Abdelzaher, *Conf. on Embedded Networked Sensor Systems* (pp. 476-488), 2022.

COMPRESSIVE OFFLOADING

Network Latency / Accuracy for Speech Recognition

	Input			Layer1			Layer2		
	Size	t_{net}	WER	Size	t_{net}	WER	Size	t_{net}	WER
DeepCOD	**17.9KB (1.5%)**	**8.8ms (8.2%)**	**0.085 (+0.003)**	**7.3KB (0.2%)**	**6.9ms (1.8%)**	**0.087 (+0.005)**	**5.5KB (0.1%)**	**4.3ms (1.1%)**	**0.085 (+0.003)**
Offload-CS	140KB (12.1%)	25.8ms (24.1%)	0.231 (+0.149)	551KB (11.5%)	50.6ms (13.4%)	0.144 (+0.062)	550KB (11.5%)	50.5ms (13.4%)	0.128 (+0.046)
Offload-Intp	142KB (12.3%)	25.9ms (24.2%)	0.262 (+0.18)	550KB (11.5%)	50.6ms (13.4%)	0.148 (+0.066)	550KB (11.5%)	50.6ms (13.4%)	0.313 (+0.231)
Offload-Lossy	144KB (12.4%)	25.9ms (24.2%)	0.264 (+0.182)	551KB (11.5%)	50.6ms (13.4%)	0.145 (+0.063)	551KB (22.4%)	50.6ms (13.4%)	0.135 (+0.053)
Offload-AE+	21.7KB (1.9%)	8.9ms (8.3%)	0.088 (+0.006)	45KB (0.9%)	23.3ms (6.2%)	0.09 (+0.008)	30KB (0.6%)	20.3ms (5.4%)	0.087 (+0.005)
Offload	1158KB	107.2ms	0.082	4800KB	377.9ms	0.082	4800KB	377.9ms	0.082

Latency/Accuracy Trade-off



Energy Consumption (mJ)

	DeepCOD	Offload-Intp	Offload-Lossy
Image	28	27	41
Speech	38	35	25

Summary:

Compressive offloading significantly improves the trade-off between offload latency and downstream classification accuracy

Demo videos available: Compressive Offloading and Improving Edge Efficiency

THEORY OF MODEL COMPRESSION

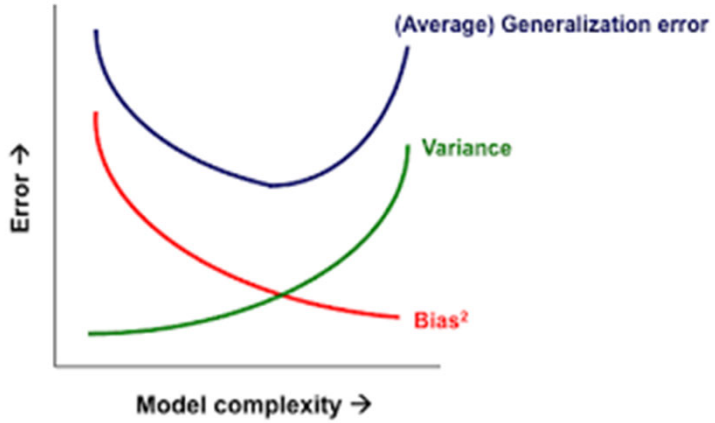
A theory for neural network compression that explains why compressed models (with dramatically reduced memory and run-time latency) can also have a lower generalization error

Army Relevance and Value Proposition

In order to capitalize on the large number of resource-limited (storage, bandwidth, energy, compute) edge devices, IoBT will require the use of highly compressed models

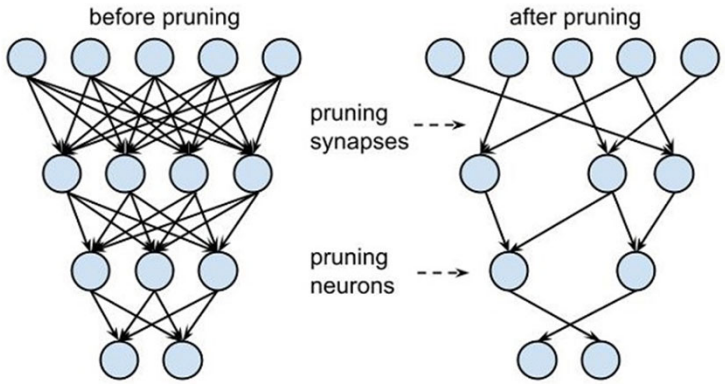
Prior State-of-the-Art

- Cutting edge neural network models are comprised of (10-100) millions of parameters, requiring >100 MB of storage and costly FLOPs during inference
- Existing techniques for NN model compression are ad-hoc and lack theoretical guarantees; based only on numerical experiments



Technical Approach

- Derive relation between model compression and generalization error of machine learning to rate-distortion theory of information theory
- Exploit this relation to motivate new compression algorithms and identify regimes where compression reduces generalization error



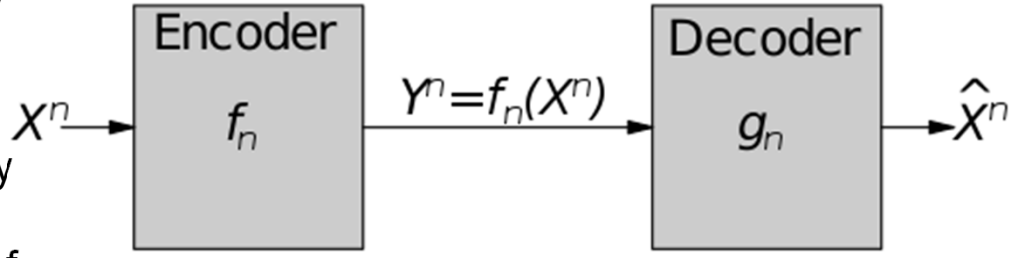
Contribution: A theoretical basis for why models can be compressed to dramatically reduce memory and computation while maintaining or increasing their accuracy

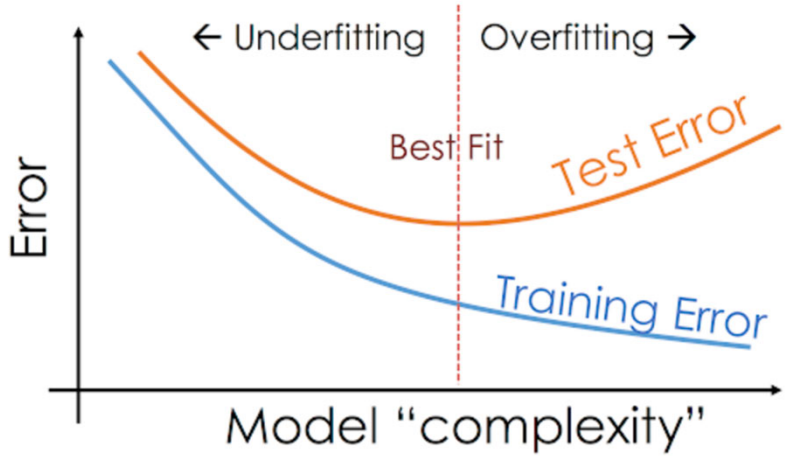
THEORY OF MODEL COMPRESSION

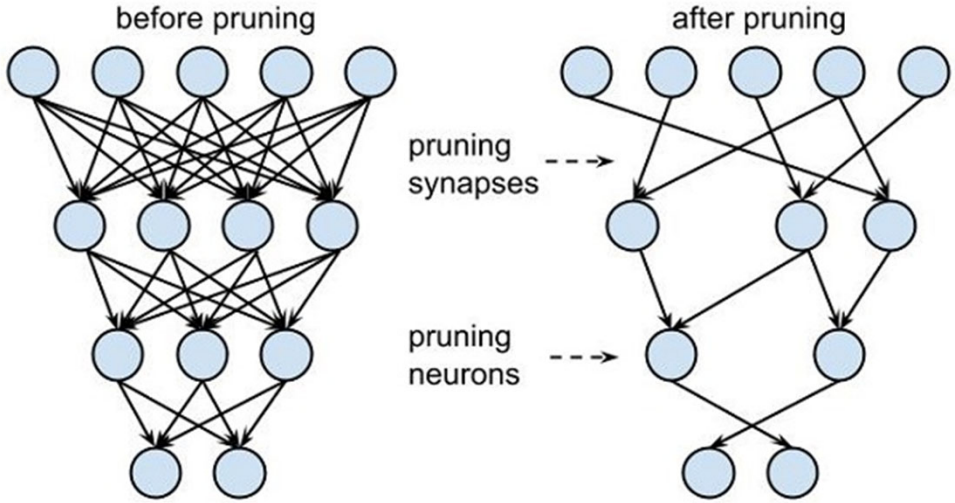
Army Relevance: Enables increased situational awareness tempo due to faster analysis of sensor data

Contribution: First theoretical explanation for why model compression can improve ML performance

- Model compression acts as a regularizer
- Key idea: rate/distortion theory from information theory
  - Rate → size of compressed model
  - Distortion → increase in empirical (training) error of compressed vs. uncompressed model
  - Higher empirical risk → lower population risk







Key Publications:

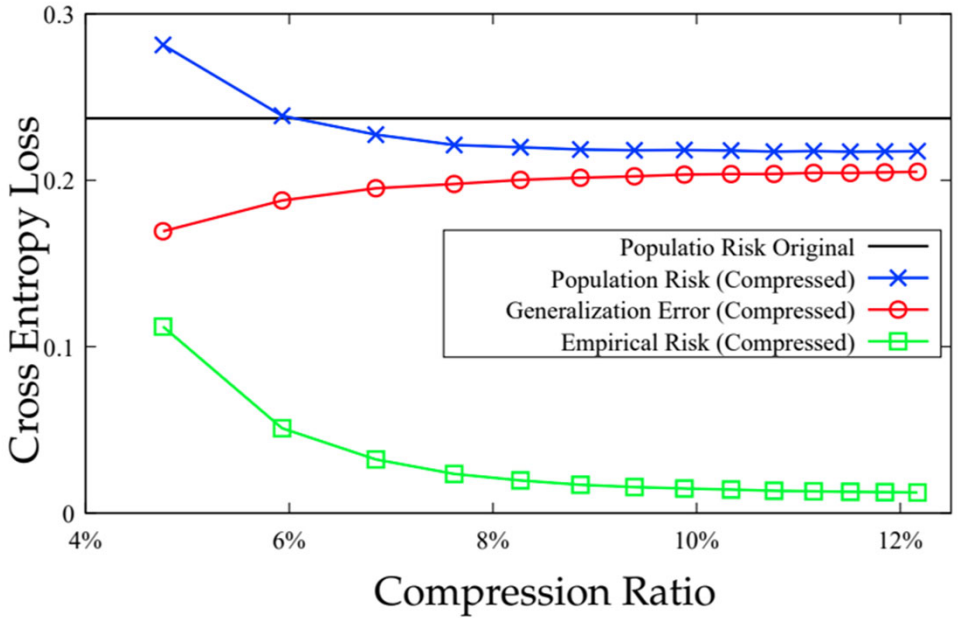
Population Risk Improvement with Model Compression: An Information-theoretic Approach
Y. Bu, W. Gao, S. Zou, V.V. Veeravalli, *Entropy,* 23(10), 2021.

THEORY OF MODEL COMPRESSION

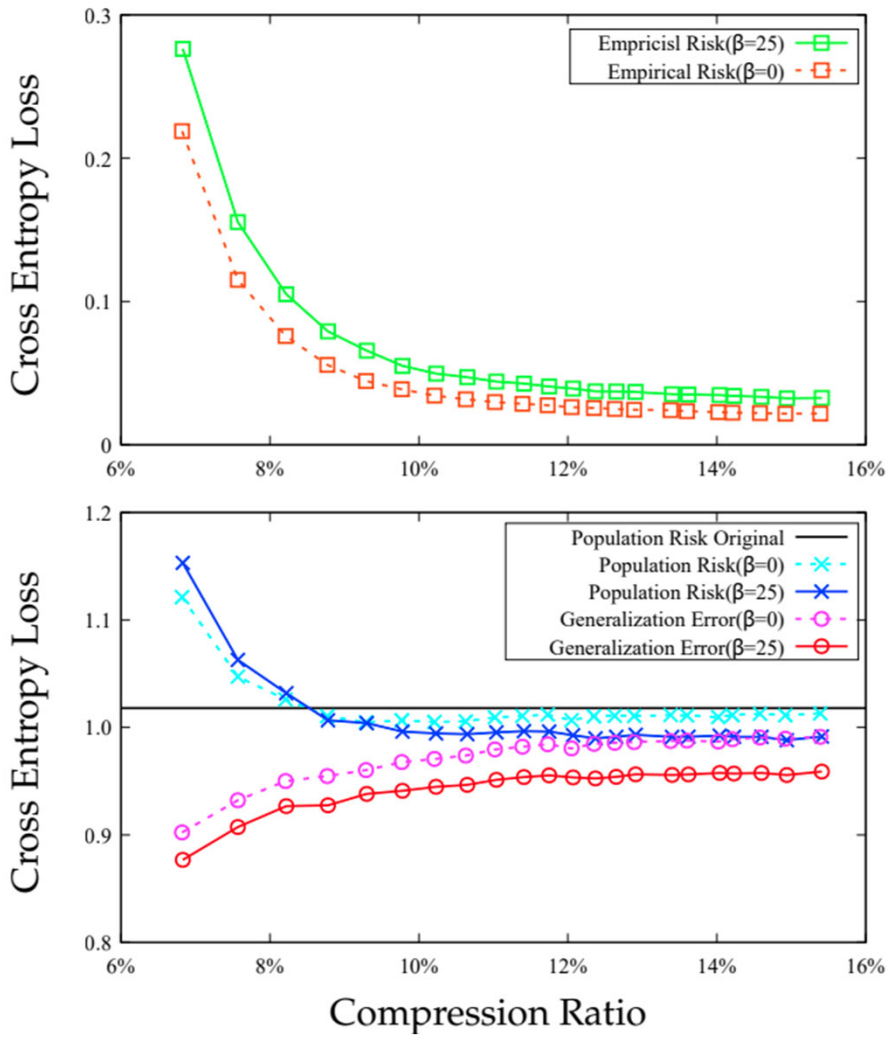
Bound for generalization error

$$\min_{P_{\hat{W}|W}:I(W;\hat{W})=R} \mathbb{E}_{S,W,\hat{W}}[L_\mu(\hat{W}) - L_S(W)] \leq \sqrt{\frac{2\sigma^2}{n}R + D(R)}$$

Empirical validation on MNIST



Empirical validation on CIFAR10



Summary:

A fundamental analytical explanation of conditions that allow for both a high compression ratio and improved accuracy over the uncompressed neural network

IOBT CRA IMPACT

Before IoBT CRA		With IoBT CRA
Very limited class of objective functions and supported constraints; limited optimality guarantees	**Network Synthesis** →	Highly flexible class of objective functions and constraints supporting network synthesis with optimality guarantees
# bits communicated: B Accuracy: A	**Distributed Model Training** →	# bits communicated: B/20 Accuracy: A
End-end latency: L Accuracy: A Energy: E	**Model Inference** →	End-end latency: L/10 to L/2 Accuracy: 0.99 A Energy: 1.05 E
Time to detection: T False alarm rate: F Pre-, post-change distributions known	**Change detection** →	Time to detection: T/2 False alarm rate: F Post-change distribution unknown

NEXT STEPS

IoBT Cornerstone Challenges: Robustness - Dynamics - Scale - Heterogeneity

- Rapid execution is one important part of a larger design space: How can ML models be designed that enable on-the-fly **scalable** optimization across **latency**, **energy**, accuracy, and **robustness** factors? How can such models be optimally trained and adapted at run-time?
- Can we extend our network synthesis framework to be **robust** to changing mission objectives and deliberate adversarial attacks and offer **scalability** even under **dynamically changing** topologies, including sensor node mobility?
- How to **prioritize data** and **adapt compression** levels (depending on data semantics) to better **meet CCIRs**?
- Can the relationship between learning performance and compression be characterized via even tighter information-theoretic bounds? Can we use the insights from these new bounds to improve upon existing compression schemes with even lower latency and memory requirements - or inspire new ones?
- How to design provably near optimal data compression for **multi-modal heterogeneous** data that **scale gracefully** with the number of agents, account for multiple inference objectives, and offer **uncertainty bounds**?
- Demonstrate a system at large scale that combines distributed network synthesis, compression, and ML processing over dynamically changing topologies

Prospective Army Capabilities

A new generation of theoretically-principled networked algorithms designed to execute on, and enable collaboration of, extremely resource-limited battlefield devices, while flexibly optimizing across accuracy, energy, latency, and robustness - in order to maximize delivery of actionable knowledge to the Commander for mission success

Questions?

Notable Awards/Honors:

- **Book:** P. Moulin and V.V. Veeravalli. Statistical Inference for Engineers and Data Scientists. Cambridge University Press, 2019
- **Area Editor:** IEEE Open Journal of Signal Processing (V.V. Veeravalli)
- **Guest Editor:** Special issue of Acta Informatica on Synthesis. IFAC Fellow (P. Tabuada)
- **President of IEEE Information Theory Society** (C. Fragouli)
- **Google Faculty Research** Award in Machine Learning (S. Diggavi)
- **Amazon Faculty Research Award** (in AI) (S. Diggavi, A. Wang)
- **IBM, Texas Instruments Faculty Research Award** (A. Wang)
- **Founding Chair**, ACM Special Interest Group on Energy (P. Shenoy)
- **ACM Fellow**, 2020 (P. Shenoy)

Tightening Mutual Information Based Bounds on Generalization Error

Yuheng Bu, *Member, IEEE*, Shaofeng Zou, *Member, IEEE*, and Venugopal V. Veeravalli *Fellow, IEEE*

IEEE Journal on Selected Areas in Information Theory, 2020
Entropy, Special Issue on Information Theory for Machine Learning, 2021,

Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations

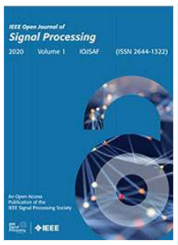
IEEE Journal on Selected Areas in Information Theory, May 2020

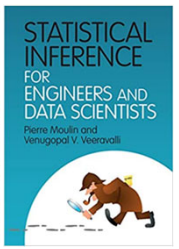
Debraj Basu *
Adobe Inc.
dbasu@adobe.com

Deepesh Data
UCLA
deepeshdata@ucla.edu

Can Karakus *
Amazon Inc.
cakarak@amazon.com

Suhas Diggavi
UCLA
suhasdiggavi@ucla.edu





IBM Faculty Award

acm

amazon research awards

IFAC
INTERNATIONAL FEDERATION OF AUTOMATIC CONTROL

Google Research

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

- Rapid Network Synthesis
- Compressive Offloading
- Fast Anomaly Detection