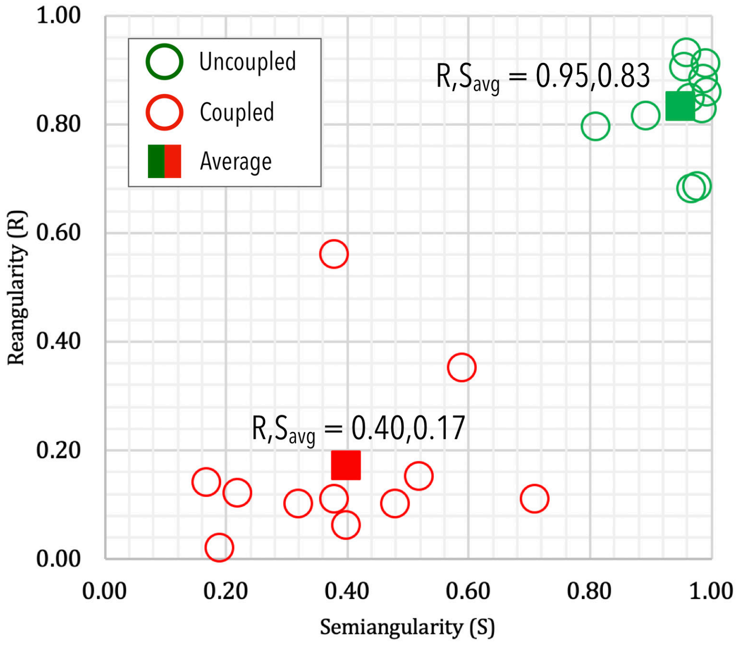
Machine Learning and Neuromorphic Computing

Hybrid Intelligence in Design

H. Akay, S.-G. Kim
Sponsorship: NSF, MIT-Sensetime

One of the greatest challenges facing society is addressing the complexities of big picture, system-level, interdisciplinary problems in a holistic way. Human designers, architects, and engineers have come to rely on steadily improving computational tools to design, model, and analyze their systems of interest. At this stage one might ask several questions: "How could we teach junior engineers, architects, and scientists to design complex systems successfully without spending years on job training? Could we also assist human experts to minimize the probability of failure by leveraging recent developments in artificial intelligence (AI) and big data?" While the resurgence of AI and machine learning suggest ways to even more fully automate downstream tasks in the design process, we propose to go up-stream of design, where all the key concepts are determined. Could machine intelligence help this early stage of designing beyond routine design and the optimization of pre-specified goals toward the generation of good, novel designs?

To capture the benefit of machine learning for design, the information and knowledge embodied in design must be represented in a method that machines can understand, memorize, and retrieve, with the goal of enhancing the practice of design. Preliminary investigation has shown how Natural Language Processing (NLP) models can be applied to accurately estimate design metrics such as functional independence based solely on descriptions of different design cases, as shown in Figure 1. With a framework for representing design knowledge, machines can effectively augment the work of human designers at the early stages of the design process.



▲ Figure 1: Descriptions of faucet design can be used to estimate functional independence.

FURTHER READING

- H. Akay and S.-G. Kim, "Measuring Functional Independence in Design with Deep-learning Language Representation Models," *Procedia CIRP*, 2020.
- S.-G. Kim, S. M. Yoon, M. Yang, J. Choi, H. Akay, and E. Burnell, "AI for Design: Virtual Design Assistant," *CIRP Annals*, 2019.
- H. Akay and S.-G. Kim, "Design Transcription: Deep Learning-based Design Feature Representation," *CIRP Annals*, 2020.

Partition WaveNet for Deep Modeling of Automated Material Handling System Traffic

D. Amirault, D. S. Boning

The throughput of a modern semiconductor fabrication plant depends greatly on the performance of its automated material handling system. Spatiotemporal modeling of the dynamics of a material handling system can lead to a multi-purpose model capable of generalizing to many tasks, including dynamic route optimization, traffic prediction, and anomaly detection. Graph-based deep learning methods have enjoyed considerable success in other traffic modeling domains, but semiconductor fabrication plants are out of reach because of their prohibitively large transport graphs. In this report, we consider a novel neural network architecture, Partition WaveNet, for spatiotemporal modeling on large graphs. Partition WaveNet uses a learned graph partition as an encoder to reduce the input size combined with a WaveNet-based stacked dilated 1D convolution component. The adjacency structure from the original graph is propagated to the induced partition graph. We discuss the motivation for framing our problem as a supervised learning task instead of a reinforcement learning task, as well as the benefits of Partition WaveNet over alternative neural network architectures. We evaluate Partition WaveNet on data from a simulated and a real semiconductor fabrication plant. We find that Partition WaveNet outperforms other spatiotemporal networks using network embeddings or graph partitions for dimensionality reduction.

Efficient AutoML with Once-for-all Network

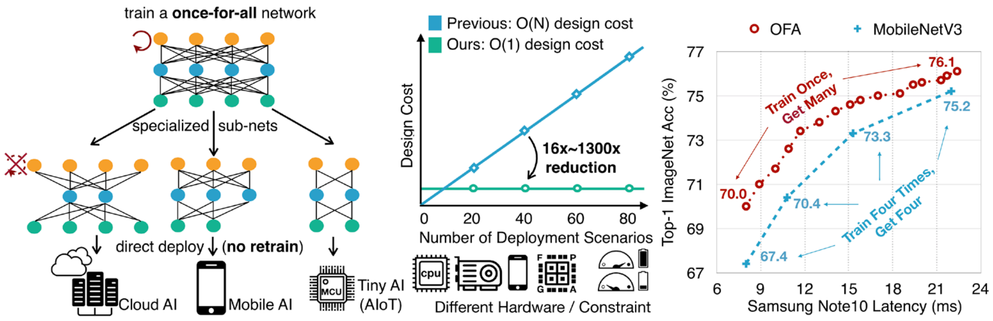
H. Cai, C. Gan, T. Wang, Z. Zhang, S. Han

Sponsorship: NSF, MIT-IBM Watson AI Lab, Google Research Award, AWS Machine Learning Research Award

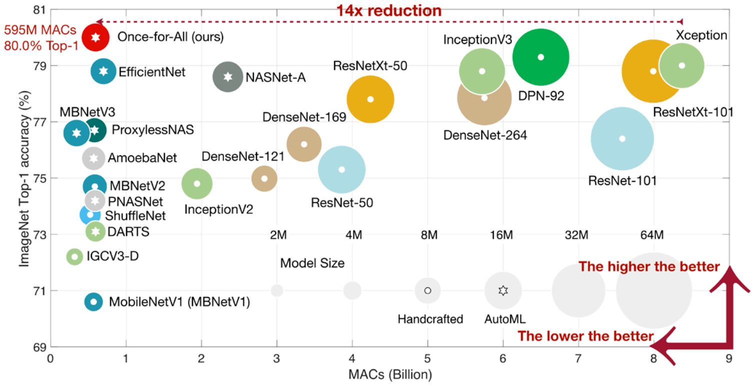
We address the challenging problem of efficient inference across many devices and resource constraints, especially on edge devices. Conventional approaches either manually design or use neural architecture search (NAS) to find a specialized neural network and train it from scratch for each case, which is computationally prohibitive (causing CO2 emission as much as 5 cars' lifetime) and thus unscalable. In this work, we propose to train a once-for-all (OFA) network that supports diverse architectural settings by decoupling training and search, to reduce the cost. We can quickly get a specialized sub-network by selecting from the OFA network without additional training. To efficiently train OFA networks, we also propose a novel progressive shrinking algorithm, a generalized pruning method that reduces the model size across many more dimensions than pruning (depth, width, kernel size, and resolution). It can obtain a surprisingly large number of sub-networks that can fit different hardware platforms and latency constraints while maintaining the same level of accuracy as training independently.

On diverse edge devices, OFA consistently outperforms state-of-the-art (SOTA) NAS methods (up to 4.0% ImageNet top1 accuracy improvement over MobileNetV3, or same accuracy but 1.5x faster than MobileNetV3, and 2.6x faster than EfficientNet w.r.t measured latency) while reducing GPU hours and CO2 emission by many orders of magnitude. In particular, OFA achieves a new SOTA 80.0% ImageNet top1 accuracy under the mobile setting (<600M MACs).

OFA is the winning solution for the 3rd Low Power Computer Vision Challenge (LPCVC, classification DSP track) and the 4th LPCVC (both classification track and detection track).



▲Figure 1: The OFA network can produce diverse specialized sub-networks without retraining. It removes the need for repeated architecture design and model training, saving orders-of-magnitude GPU training cost, and also produces efficient models for fast inference on mobile devices.



◀Figure 2: OFA network achieves high accuracy at low computation cost, being at the top-left corner of the accuracy-computation trade-off curve.

FURTHER READING

- H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-All: Train One Network and Specialize It for Efficient Deployment," *ICLR*, 2020.
- H. Cai, L. Zhu, and S. Han, "Proxyless NAS: Direct Neural Architecture Search on Target Task and Hardware," *ICLR*, 2019.

Robustness Verification and Defense for Tree-based Machine Learning Models

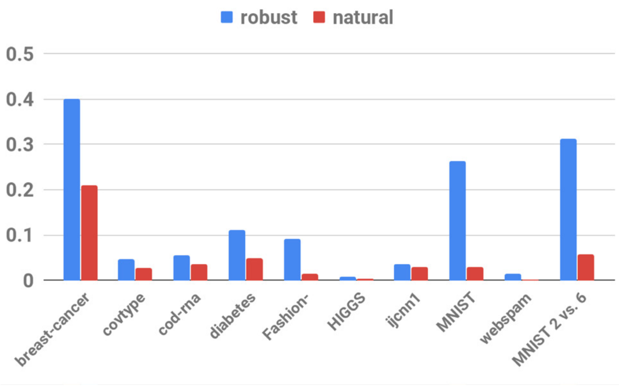
H. Chen, D. S. Boning

Sponsorship: MIT-SenseTime Alliance (MIT Quest for Intelligence)

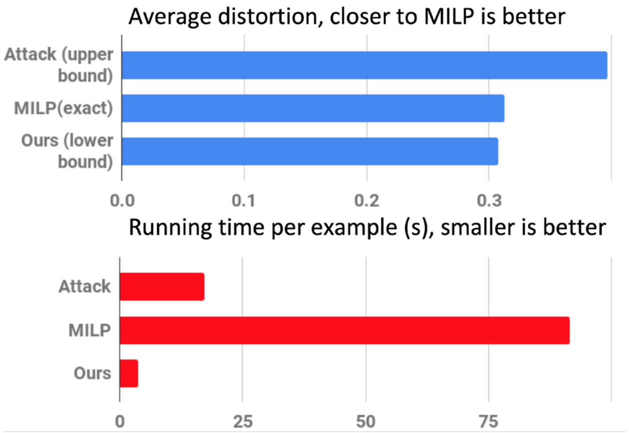
Although adversarial examples and model robustness have been extensively studied in the context of linear models and neural networks, research on this issue in tree-based models is still limited, despite the prevalence of tree-based models in manufacturing and other domains. In this work, we develop a novel algorithm to learn robust trees, as well as an efficient algorithm to evaluate the robustness of a tree-based model.

Our first algorithm aims to optimize the performance under the worst-case perturbation of input features, which leads to a max-min saddle point problem. Incorporating this saddle point objective into the decision tree building procedure is nontrivial due to the discrete nature of trees—a naive approach to finding the best split according to this saddle point objective will take exponential time. To make our approach practical and scalable, we approximate the inner minimizer in this saddle point problem and present implementations for classical information gain-based trees as well as state-of-the-art tree boosting models such as XGBoost. As demonstrated in Figure 1, experimental results on real world datasets demonstrate that the proposed algorithms can substantially improve the robustness of tree-based models against adversarial examples.

Formal robustness verification of decision tree ensembles involves finding the exact minimal adversarial perturbation or a guaranteed lower bound, which is NP-complete in general. We show that for tree ensembles, the verification problem can be cast as a max-clique problem on a multipartite graph with bounded boxicity. For low dimensional problems when boxicity can be viewed as constant, this reformulation leads to a polynomial time algorithm. For general problems, by exploiting the boxicity of the graph, we develop an efficient multi-level verification algorithm that can give tight lower bounds on the robustness of decision tree ensembles while allowing iterative improvement and anytime termination. As in Figure 2, our algorithm is much faster than a previous approach that requires solving mixed integer linear programming (MILP) and can give tight robustness verification bounds on large models with one thousand deep trees.



▲ Figure 1: Average distortion of the minimum adversarial examples for robust tree-based models in this work and natural tree-based models. Larger values mean better robustness.



▲ Figure 2: Average distortion and per example running time of our method on a 1000-tree robust GBDT model trained with MNIST 2 vs. 6.

FURTHER READING

- H. Chen, H. Zhang, D. S. Boning, and C.-J. Hsieh, "Robust Decision Trees Against Adversarial Examples," *International Conference on Machine Learning*, pp. 1122-1131, Jun. 2019.
- H. Chen, H. Zhang, S. Si, Y. Li, D. S. Boning, and C.-J. Hsieh. "Robustness Verification of Tree-based Models," Advances in Neural Information *Processing Systems*, pp. 12317-12328, Dec. 2019.

An Efficient and Continuous Approach to Information-theoretic Exploration

T. Henderson, V. Sze, S. Karaman

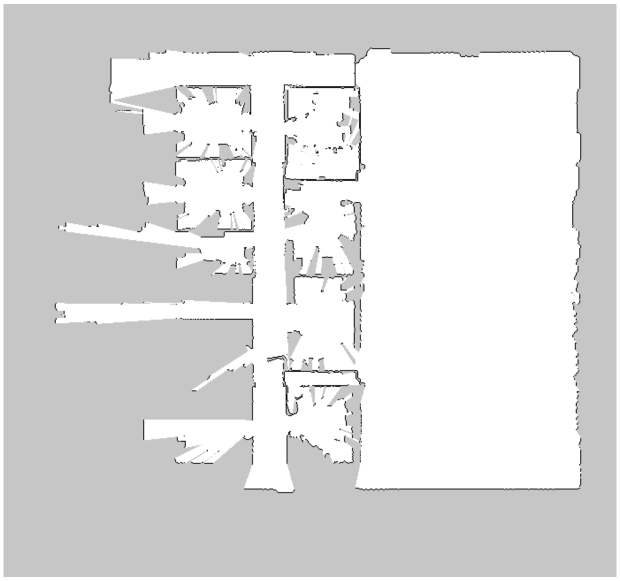
Sponsorship: NSF Cyber-Physical Systems (CPS) Program

Exploration of unknown environments is embedded in many robotics applications: search and rescue, crop survey, space exploration, etc. The central problem an exploring robot must answer is "where should I move next?" The answer should balance travel cost with the amount of information expected to be gained about the environment. Traditionally, this question has been answered by a variety of heuristics that provide no guarantees on their exploration efficiency. Information-theoretic methods can produce an optimal solution, but until now they were thought to be computationally intractable.

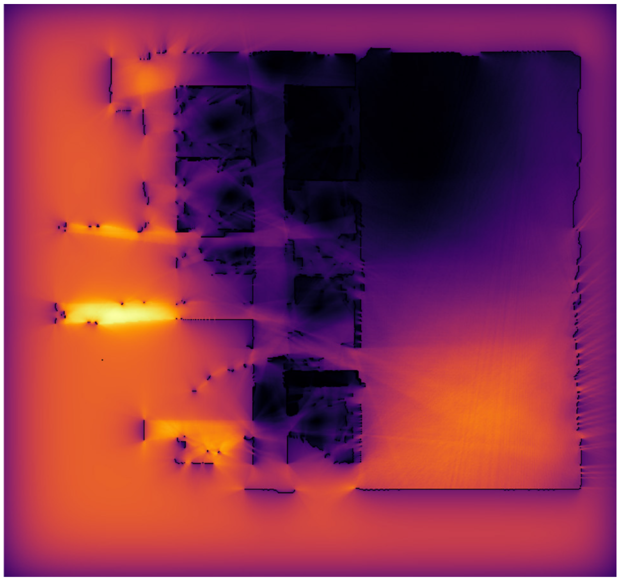
In our recent work we describe the Fast Continuous Mutual Information (FCMI) algorithm, which computes the information-theoretic exploration metric efficiently. FCMI takes as input an incomplete occupancy map like the one shown in Figure 1, where white pixels indicate free space, black pixels indicate occupied space, and gray pixels indicate unknown space. It then returns an information surface as shown in Figure 2, where the brightness of each pixel indicates how much information is expected to be gained by exploring at that location. The algorithm also works on multi-resolution or 3-dimensional maps. FCMI has a lower asymptotic complexity than existing methods and our experiments demonstrate that it is hundreds of times faster than the state-of-the-art for practical inputs.

The key insight that enables FCMI is to consider the occupancy map as a continuous random field rather than a discrete collection of cells. This reveals a nested information structure that makes it possible to recursively reuse computation from one map location in adjacent locations. The continuous structure also provides more general insights that are relevant to any occupancy mapping system.

For practical map sizes, FCMI runs in seconds on a single threaded laptop CPU which is well within the timing constraints for most robotic applications. It provides considerable savings to energy constrained systems by reducing both the exploration travel cost and the computation cost. FCMI is also highly parallelizable and suited for a rapid, low energy, embedded implementation.



▲ Figure 1: An incomplete occupancy grid map of MIT's building 31.



▲ Figure 2: An information surface produced by the FCMI algorithm.

FURTHER READING

- T. Henderson, V. Sze, S. Karaman "An Efficient and Continuous Approach to Information-Theoretic Exploration," Proceedings of the the *2020 Informational Conference on Robotics and Automation (ICRA)*, 2020.

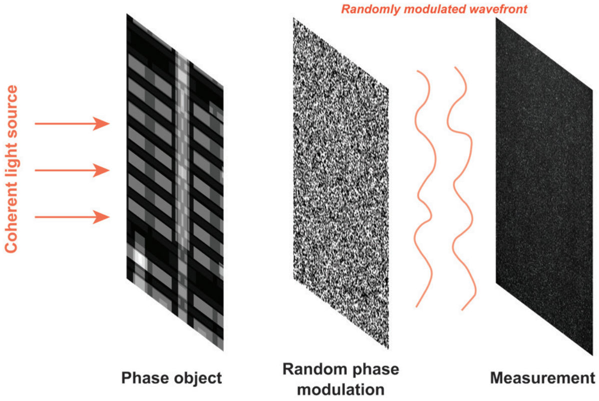
On the Use of Deep Learning for Retrieving Phase from Noisy Inputs in the Coherent Modulation Imaging Scheme

I. Kang, F. Zhang, G. Barbastathis
Sponsorship: IARPA RAVEN, MIT-SUSTech, NSF China, KFAS

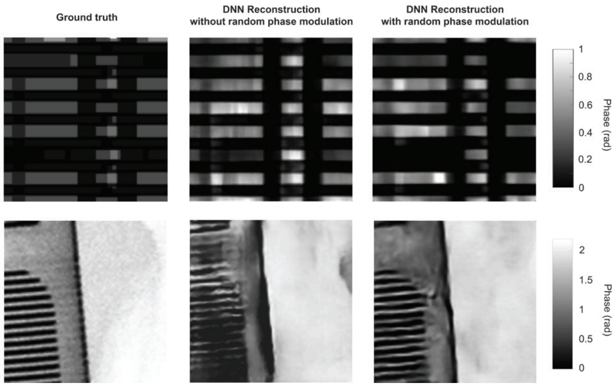
Low-dose light imaging is of significance in many cases when minimal radiation exposure of samples is desired. In biological imaging, high-dose light may induce phototoxic effects at the cost of larger signal-to-noise ratio (SNR). In particle imaging, for instance, imaging integrated circuits (IC) with high-power beam leads to destructive side-effects, e.g., heat-induced deformation. However, quantum nature of photon detection influences and degrades the quality of intensity measurements, and on top of the Poisson statistics, other types of noise sources, e.g., thermal or readout noise, add up.

Deep neural networks (DNNs) have been used for retrieving phase information from noisy intensity measurements. Nonetheless, the ill-posedness of the inversion problem, governed by a physical system design, could not be sufficiently addressed when the DNN alone was used. Due to the ill-posedness of the system, residual artifacts remained in reconstructions, thus a decrease in image fidelity. Therefore, we suggest the application of random phase modulation on an optical field, also known as a coherent modulation imaging (CMI) scheme, along with the DNNs as a method of reconstruction.

In this work, we provide both quantitative and qualitative results that unwanted artifacts in reconstructions are largely removed in the coherent modulation imaging scheme under low-light conditions in conjunction with the DNNs. Here, phase extraction neural network (PhENN), which is an encoder-decoder DNN architecture based on ResNet specifically optimized for phase retrieval tasks, was used as a design of the DNN.



▲Figure 1: Coherent modulation imaging scheme involves three planes: object plane, modulation plane, and detector plane. Random phase modulation applies random phase on optical field.



▲Figure 2: Display of DNN reconstructions with and without random phase modulation of some images from ImageNet and IC layout database. Modulation largely removes residual artifacts.

FURTHER READING

- A. Goy, K. Arthur, S. Li, and G. Barbastathis, "Low Photon Count Phase Retrieval Using Deep Learning," *Physical Review Letters*, vol. 121, p. 243902, 2018.
- F. Zhang, B. Chen, G. R. Morrison, J. Vila-Comamala, M. Guizar-Sicairos, and I. K. Robinson, "Phase Retrieval by Coherent Modulation Imaging," *Nature Communications*, vol. 7, pp. 1-8, 2016.
- A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless Computational Imaging Through Deep Learning," *Optica*, vol. 4, pp. 1117-1125, 2017.

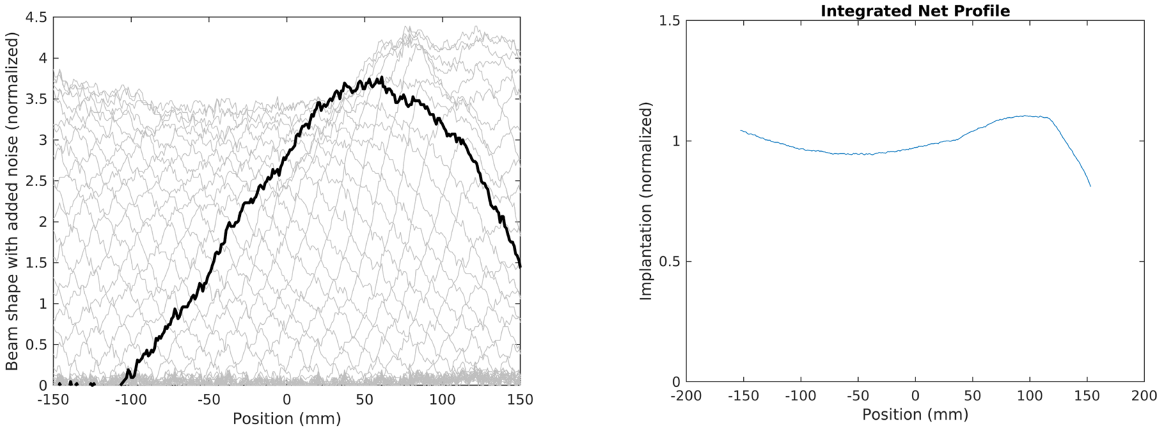
Rapid Uniformity Tuning in Ion Implantation Systems Using Bayesian Optimization

C. I. Lang, D. S. Boning

Sponsorship: Applied Materials

As the size of integrated circuits continue to shrink, variations in their fabrication processes become more significant, hindering their electrical performances and yields. One such wafer-scale variation occurs in ion implantation processes, where an ion beam implants charged particles into a substrate. As the beam is scanned across the wafer, its shape and intensity often change, resulting in a non-uniform implantation. This effect can be compensated for by adjusting the speed of the ion beam as it moves across the wafer; however, in order to do so, the dynamics of the ion beam shape must be known.

Our work focuses on using Bayesian optimization, a form of reinforcement learning, to rapidly learn how the beam shape changes, and to optimize the beam speeds in order to reduce non-uniformities. Here, we capture our knowledge of the beam shapes by treating its intensities as multivariate, normally distributed, random variables. After observing new implantations, we then use this framework to update our belief of the beam shapes, then solve for a new set of scan speeds which result in our desired profile under this updated model. We then continue this process until we converge to our desired profile. After this initial tuning, the same tuning algorithm continues to run during normal operation. Implantation measurements are periodically made, the model is updated using these measurements, and any corrections to the scan speeds are made in order to maximize uniformity. This process allows us to both quickly tune a new implantation recipe, while also allowing us to learn and compensate for any changing conditions in the tool.



▲Figure 1: Image showing how the beam shape changes as a function of wafer position (A) and resulting implantation using constant beam speed (B).

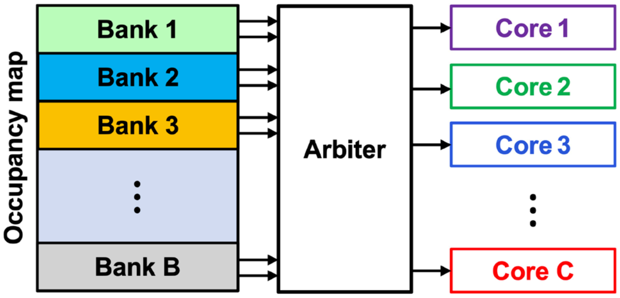
A Mutual Information Accelerator for Autonomous Robot Exploration

P. Z. X. Li, S. Karaman, V. Sze
Sponsorship: AFOSR YIP, NSF

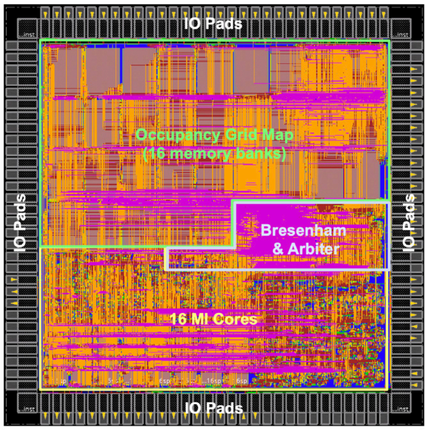
Robotic exploration problems arise in various contexts, ranging from search and rescue missions to underwater and space exploration. In these domains, exploration algorithms that allow the robot to rapidly create the map of the unknown environment can reduce the time and energy for the robot to complete its mission. Shannon mutual information (MI) at a given location is a measure of how much new information of the unknown environment the robot will obtain given what the robot already know from its incomplete understanding of the environment. In a typical exploration pipeline, robot starts with an incomplete map of the environment. At every step, the robot computes the MI across the entire map. Then, the robot can select the location with the highest mutual information for exploration in order to gain the most information about the unknown environment.

However, on the CPUs and GPUs typically found on mobile robotic platforms, computing MI using the state-of-the-art Fast Shannon Mutual Information (FSMI) algorithm across the entire map takes more than one second, which is too slow for enabling fast autonomous exploration. As a result, the emerging literature considers approximation techniques, and many practitioners rely on heuristics that often fail to provide any theoretical guarantees.

To eliminate the bottleneck associated with the computation of MI across the entire map, we propose a novel multicore hardware architecture (Figure 1) with a memory subsystem that efficiently organizes the storage of the occupancy grid map and an arbiter that effectively resolves memory access conflicts among MI cores so that the entire system achieves high throughput. In addition, we provide rigorous analysis of memory subsystem and arbiter in order to justify our design decisions and provide provable performance guarantees. Finally, we thoroughly validated the entire hardware architecture by implementing it using a commercial 65nm ASIC technology (Figure 2).



▲ Figure 1: Proposed multi-core hardware architecture that provides sufficient memory bandwidth so that the computation cores are active.



▲ Figure 2: Layout of the proposed hardware architecture with 16 cores using a commercial 65nm technology.

FURTHER READING

- P. Z. X. Li*, Z. Zhang*, S. Karaman, V. Sze, “High-throughput Computation of Shannon Mutual Information on Chip,” *Robotics: Science and Systems (RSS)*, Jun. 2019.
- Z. Zhang, T. Henderson, V. Sze, S. Karaman, “FSMI: Fast Computation of Shannon Mutual Information for Information-theoretic Mapping,” *IEEE International Conference on Robotics and Automation (ICRA)*, May 2019.

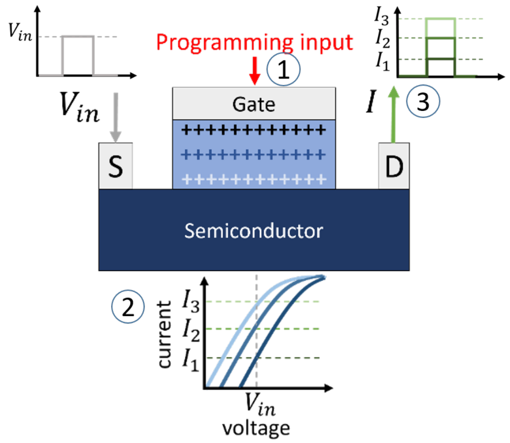
Ionic Analog Synapses for Deep Learning Accelerators

K. Limanta, A. Zubair, T. Palacios

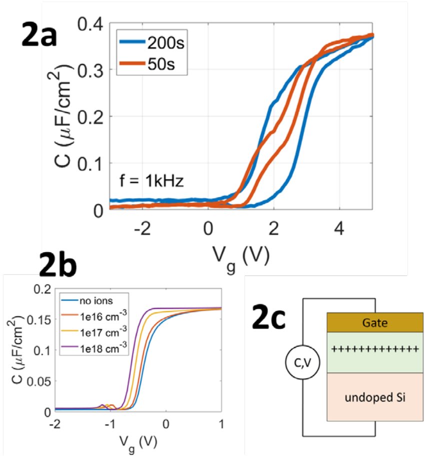
Sponsorship: Advanced Micro Devices Undergraduate Research and Innovation Scholar

The recent progress in novel hardware/software co-optimizations for machine learning has led to tremendous improvement of the efficiency of neural networks. Nevertheless, the energy efficiency is still orders of magnitude lower than biological counterpart – the brain. Digital CMOS architecture has inherent limitations for deep learning applications due to their volatile memory, spatially separated memory and computation, and the lack of connectivity between nodes. Crossbar arrays of non-volatile memory devices, able to perform simple operations (e.g. bit multiplication), can potentially achieve a 30000× improvement in energy efficiency. State-of-the-art analog "synaptic" devices based on resistive memories suffer from stochastic, asymmetric, and non-linear weight updates, detrimental to training accuracy. Electrochemical ionic devices have been shown to be fast, energy efficient, and exhibit symmetric, linear weight updates. However, electrolytes used for the electrochemical reaction are often CMOS incompatible and suffers from scalability.

Here we propose a new transistor-based analog synapse, consisting of a proton-doped SiO2 gate oxide which electrostatically modifies the threshold voltage of the semiconductor channel, tuning the channel conductance (Figure 1). Non-volatility is maintained by trapping of protons in the oxide. Due to electrostatics, we expect to observe a symmetric and linear shift in threshold voltage, leading to linear weight updates. We study the proton diffusion and electrostatic effects through device simulation via Silvaco Atlas and analytical modeling. Simulations show a threshold voltage shift of the MOS gate stack due to the presence of ions in the gate oxide (Figure 2). We fabricate n-Si/ALD SiO2/Al MOS capacitor and to demonstrate the feasibility of our ionic device. We observe that the MOS gate stack exhibits hysteretic behavior below 2V, indicating non-volatility and low-voltage operation. The results of this work will shed light on the feasibility of simple CMOS-compatible ionic devices for the next generation of neural network hardware accelerators.



▲Figure 1: All inputs are based on simple square pulses. 1. Programming input pulse pushes protons toward semiconductor, changing device to a new state. 2. Threshold voltage VT shifts. 3. For a given input voltage Vin, the conductance of the device (measured by output current) changes based on state of device.



▲Figure 2: a. Capacitance-voltage measured on fabricated MOS gate stack. b. Device simulation of MOS gate stack doped with protons shows a threshold voltage shift. Effect is stronger with increasing implantation dose. c. Diagram of MOS structure.

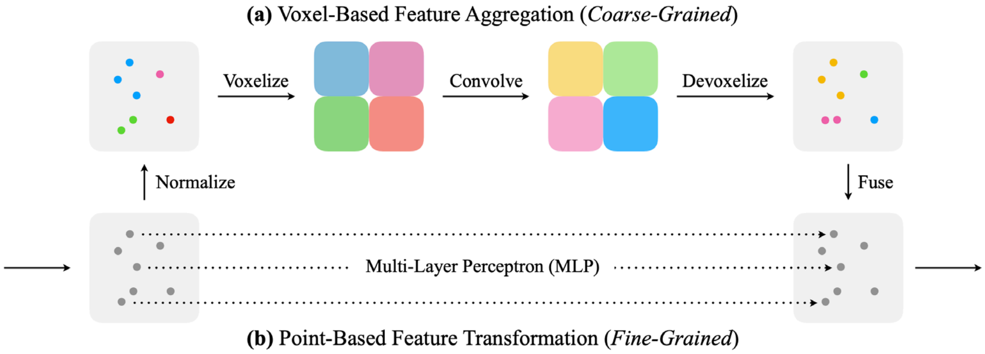
Efficient 3D Deep Learning with Point-voxel CNN

Z. Liu, H. Tang, Y. Lin, S. Han

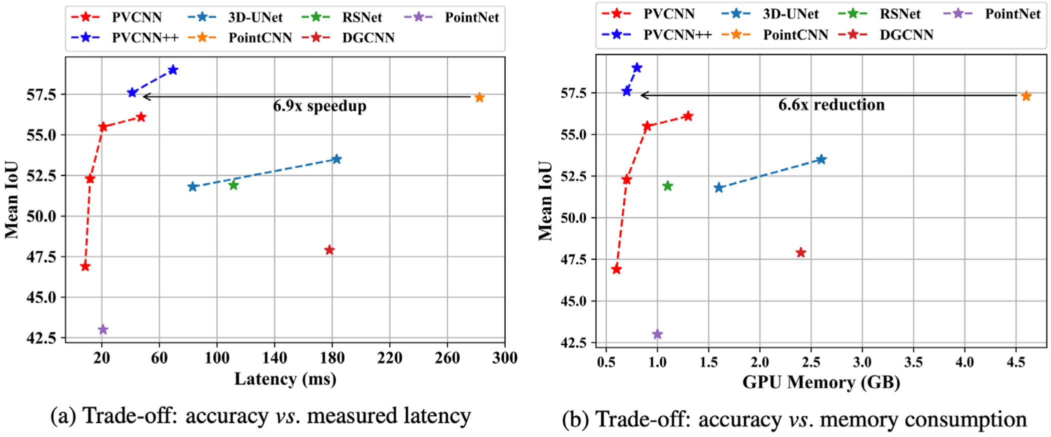
Sponsorship: MIT Quest for Intelligence, MIT-IBM Watson AI Lab, Samsung, Facebook, SONY

3D deep learning has received increased attention thanks to its wide applications: e.g., AR/VR and autonomous driving. These applications need to interact with people in real time and therefore require low latency. However, edge devices (such as AR/VR headsets and self-driving cars) are tightly constrained by hardware resources and battery. Previous work processes 3D data using either voxel-based or point-based NN models. However, both approaches are computationally inefficient. The computation cost and memory footprints of the voxel-based models grow cubically with the input resolution, making it memory-prohibitive to scale up the resolution. As for point-based networks, up to 80% of the time is wasted on structuring the sparse data which have rather poor memory locality, not on the actual feature extraction.

To this end, we propose Point-Voxel CNN (PVCNN) that represents the 3D input data as point clouds to take advantage of the sparsity to reduce the memory footprint, and leverages the voxel-based convolution to obtain the contiguous memory access pattern (Figure 1). Evaluated on semantic and part segmentation datasets, it achieves a much higher accuracy than the voxel-based baseline with 10× GPU memory reduction; it also outperforms the state-of-the-art point-based models with 7× measured speedup on average (Figure 2). We validate its general effectiveness on 3D object detection: Frustrum PVCNN outperforms Frustrum PointNet++ by up to 2.4% mAP with 1.8× measured speedup and 1.4× GPU memory reduction.


(a) Voxel-Based Feature Aggregation (Coarse-Grained)
Voxelize
Convolve
Devoxelize
Normalize
Fuse
Multi-Layer Perceptron (MLP)
(b) Point-Based Feature Transformation (Fine-Grained)

▲Figure 1: PVCNN is composed of several PVConv's, each of which has a low-resolution voxel-based branch and a high-resolution point-based branch. The voxel-based branch extracts coarse-grained neighborhood information, which is supplemented by the fine-grained individual point features extracted from the point-based branch.



(a) Trade-off: accuracy *vs.* measured latency

(b) Trade-off: accuracy *vs.* memory consumption

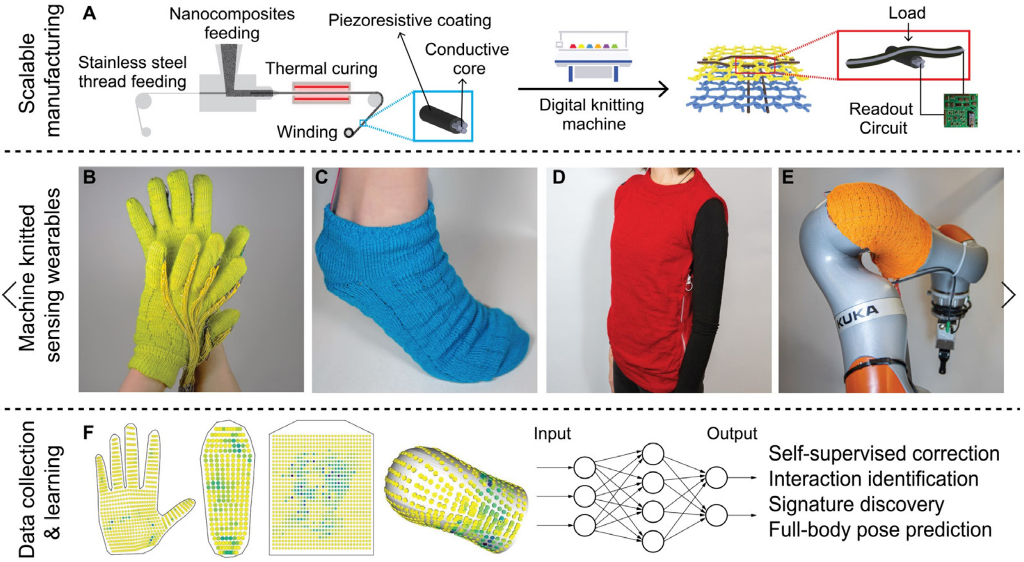
▲Figure 2: Results of indoor scene segmentation on S3DIS. On average, our PVCNN and PVCNN++ outperform the point-based models with 8× measured speedup and 3× memory reduction, and outperform the voxel-based baseline with 14× measured speedup and 10× memory reduction.

Learning Human-environment Interactions Using Scalable Functional Textiles

Y. Luo, W. Matusik, T. Palacios

Living organisms extract information and learn from the surroundings through constant physical interactions. For example, humans are particularly receptive to tactile cues (on hands, limbs, and torso), which enable the performing of complex tasks like dexterous grasp and locomotion. Observing and modeling interactions between humans and the physical world are fundamental for the study of human behavior, healthcare, robotics, and human-computer interactions. However, many studies of human-environment interactions rely on more easily observable visual or audible datasets because it is challenging to obtain tactile data in a scalable manner. Recently, Sundaram et al. coupled tactile-sensing gloves and machine learning to uncover signatures of the human grasp. However, the recording and analysis of whole-body interactions remain elusive, as this would require large-scale wearable sensors with low cost, dense coverage, conformal fit, and minimal presence to permit natural human activities.

We present a textile-based tactile learning platform that enables researchers to record, monitor, and learn human activities and the associated interactions. Realized with inexpensive piezoresistive fibers (0.2 USD/m) and automated machine knitting, our functional textiles offer dense coverage (> 1000 sensors) over large complex surfaces (> 2000 cm2). Further, we leverage machine learning for sensing correction, ensuring that our system is robust against potential variations from individual receptors. To validate the capability of our sensor, we capture diverse human-environment interactions (> 1,000,000 tactile frames) and demonstrate that machine learning techniques can be used with our data to classify human activities, predict whole-body poses, and discover novel motion signatures. This work opens new possibilities in wearable electronics, healthcare, manufacturing, and robotics.



▲ Figure 1: Schematic of textile-based tactile learning platform. (A) Scalable manufacturing of tactile sensing textiles using customized coaxial piezoresistive fiber fabrication and digital machine knitting. Commercial conductive stainless steel thread is coated with piezoresistive nanocomposite (composed of polydimethylsiloxane (PDMS) elastomer as matrix and graphite/copper nanoparticles as conductive fillers). Knitted full-sized tactile sensing: (B) gloves, (C) sock, (D) vest, and (E) robot arm sleeve. (F) Examples of tactile frames collected during human-environment interactions and applications explored using machine learning techniques.

FURTHER READING

- R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile Sensing—From Humans to Humanoids," *IEEE Transactions on Robotics*, vol. 26, p. 1, 2009.
- S. Sundaram, P. Kellnhofer, J. Y. Zhu, A. Torralba, and W. Matusik, *Nature*, vol. 569, p. 698, 2019.

Automated Fault Detection in Manufacturing Equipment Using Semi-supervised Deep Learning

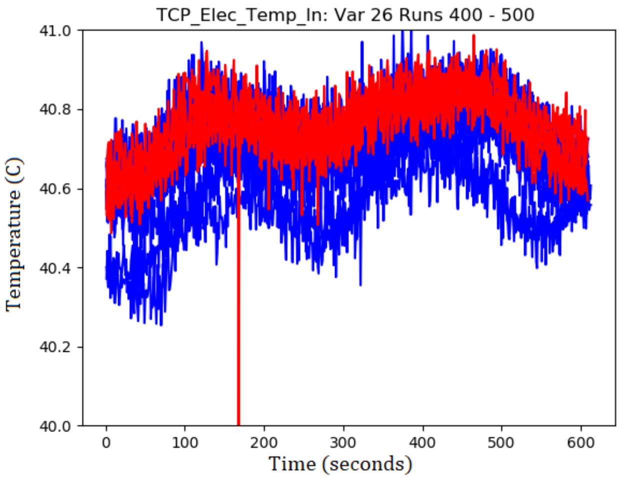
D. Martin, D. Boning, J. Lang

Sponsorship: Harting Technology Group, Lam Research

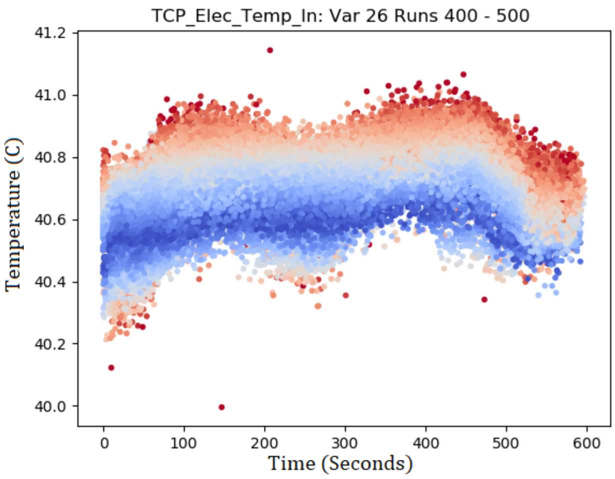
Our project investigates the use of semi-supervised deep learning systems for automated fault detection and predictive maintenance of manufacturing equipment. Unexpected equipment faults can be highly costly to manufacturing lines, but data-driven fault detection systems often require a high level of domain-specific expertise to implement as well as continued human oversight. To this end, we are developing and testing general-purpose fault detection systems that require minimal labeled data.

Our system trains deep autoencoders to function as a non-linear compression algorithm for sensor readings from manufacturing equipment. The compressed sensor signals are used as a proxy for the equipment's hidden state, and the reconstruction error is used to detect unexpected behavior. The compressed representation and reconstruction error are combined to provide a robust anomaly score. Instances in time with the highest anomaly score are then flagged to be labeled by a human operator as faulty or nominal. With sparsely labeled faults, the system then uses Gaussian mixture models to classify different types of errors and predicts future faults by monitoring parameter drift towards known fault states. Our system is currently being trained to detect failed runs on a plasma etcher (used for integrated circuit fabrication) using internal sensors that take voltage, current, pressure, and temperature readings. In preliminary tests, the system was able to correctly detect 88% of failed etching runs and identify specific markers in different signals indicative of faults. For example, a failure mode of the plasma etcher involves an abnormally high temperature (Figure 1). Without any labeled errors, the system flagged the higher temperatures as possibly indicative of faults (Figure 2).

We are currently testing the system on a wider range of applications, including estimating wear of milling machine cutting tools and predicting the risk of breakage. We are also developing prototypes of contactless voltage/current sensors that can easily be retrofitted onto older machinery to test the efficacy of fault detection systems using only external power draw.



▲ Figure 1: Temperature sensor reading over 100 etching runs. Failed etches are shown in red, and successful runs are shown in blue.



▲ Figure 2: System score for nominal temperature sensor readings with predicted anomalies shown in red.

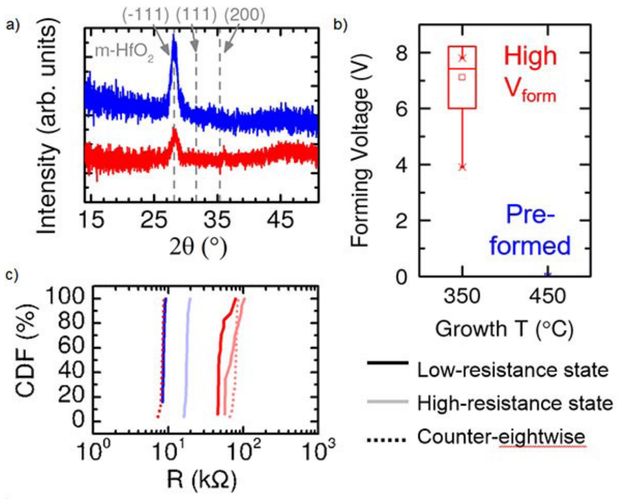
Control of Conductive Filaments in Resistive Switching Oxides

K. J. May, Y. R. Zhou, T. Ando, V. Narayanan, H. L. Tuller, B. Yildiz

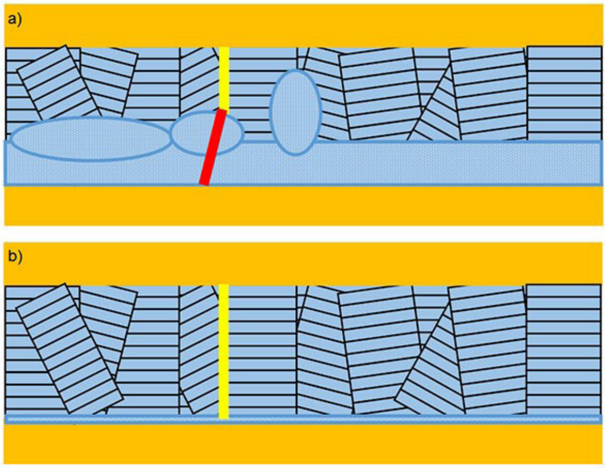
Sponsorship: IBM

There is a growing interest in using specialized neuromorphic hardware for artificial neural network applications such as image and speech processing, which require significant computational resources. These neuromorphic devices show promise for reducing the demands of such applications by increasing speed and decreasing power consumption compared to current software-based methods. One approach to achieving this goal is through oxide thin film resistive switching devices arranged in a crossbar array configuration. Resistive switching can mimic several aspects of neural networks, such as short- and long-term plasticity, via the dynamics of switching between multiple analog conductance states-dominated by the creation, annihilation, and movement of defects within the film (such as oxygen vacancies). These processes can be stochastic in nature and contribute significantly to device variability, both within and between individual devices.

Our research focuses on reducing the variability of the set/reset voltages and enhancing control of the conductance state with voltage pulsing using model systems of HfO2 grown on Nb:SrTiO3 substrates through the control of film growth and processing parameters. We show that depending on the growth temperature, substrate orientation, and substrate surface treatment, devices can exhibit forming-free switching or forming voltages ranging from 4 to 7 V. Forming-free devices show lower variability in the high and low conductance states but have a lower on/off conductance ratio. We rationalize these results using film microstructure information obtained from 2D X-ray diffraction and cross-sectional transmission electron microscopy. This work provides a significant step towards controlling the mechanisms behind device variability and achieving devices that meet the strict requirements of neuromorphic computing.



▲Figure 1: Comparison of HfO2 films grown on Nb:SrTiO3 at 350°C (red) and 450°C (blue): a) X-ray diffraction patterns showing higher crystallinity at higher growth temperature; b) forming voltage distribution box plot vs. growth temperature; c) cumulative distribution function (CDF) for the low (dark shade) and high (light shade) resistance states, where eightwise switching devices are solid lines, and counter-eightwise switching devices are dotted lines.



▲Figure 2: Schematic showing hypothetical filament formation pathways for films with a) low growth temperature and b) high growth temperature.

FURTHER READING

- R. Dittmann and J. P. Strachan, "Redox-based Memristive Devices for New Computing Paradigm," *APL Materials*, vol. 7, no. 11, p. 110903, Nov. 2019, doi: 10.1063/1.5129101.
- D. S. Jeong and C. S. Hwang, "Nonvolatile Memory Materials for Neuromorphic Intelligent Machines," *Advanced Materials*, vol. 30, no. 42, p. 1704729, 2018, doi: 10.1002/adma.201704729

Variational Inference for Model-free Simulation of Dynamic Systems with Unknown Parameters

K. Yeo, D. E. C. Grullon, F.-K. Sun, D. S. Boning, J. R. Kalagnanam

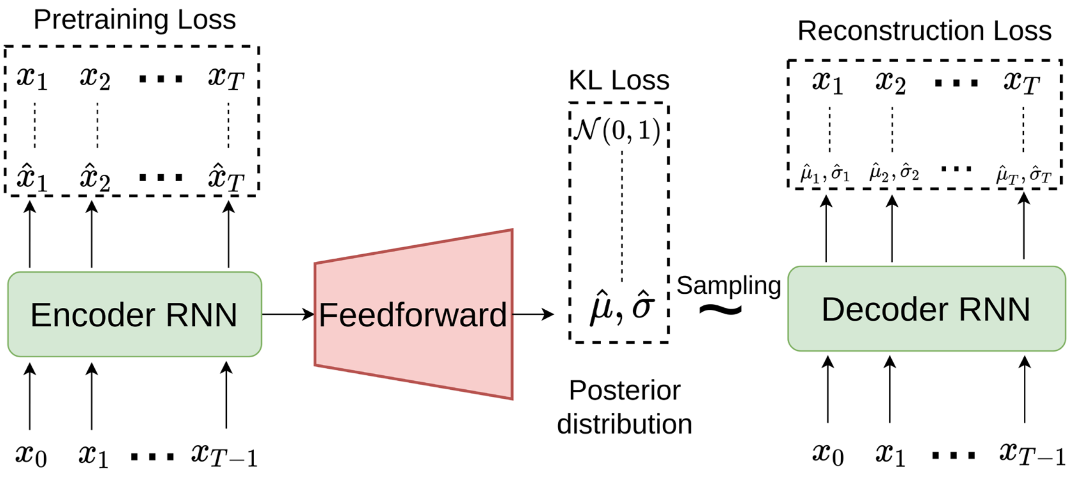
Sponsorship: MIT-IBM Watson AI Lab (MIT Quest for Intelligence)

Complex physical, biological, and engineering processes can be modelled using dynamic systems with few parameters. However, in real-world applications including manufacturing, it is possible to encounter systems for which the dynamics are not well understood and identifying the parameters is challenging.

"Model-free" approaches aim to learn the dynamics of the system from data. Classical statistical models assume the dynamics are linear to make the inference analytically tractable. Extension to nonlinearity usually requires partial knowledge about the system. Our goal is to achieve modeling of nonlinear dynamic systems purely by using data with the strength of deep learning.

In this work, we formulate the learning task as variational inference by considering the unknown parameters as random variables. Then, we use two recurrent neural networks and a feedforward network as the variational autoencoder to learn an approximate posterior distribution. The first recurrent neural network is a pre-trained encoder that encodes the input into a dense representation. Then, the feedforward network transforms the representation into the posterior distribution. Finally, the second recurrent neural network receives samples from the posterior distribution to predict the mean and variance of the output. Loss functions include pretraining loss, reconstruction loss, and KL divergence loss with regard to the prior. Figure 1 gives an overview of our model.

The numerical experiments show that the proposed model produces a more accurate simulation than the standard recurrent neural networks, especially when the Monte Carlo method is applied to perform multiple-step simulations. In addition, by analyzing the learned posterior distribution, we show that our approach can correctly identify the number of underlying parameters.


Pretraining Loss
x_1 x_2 ... x_T
$\hat{x}_1$ $\hat{x}_2$ ... $\hat{x}_T$
Encoder RNN
x_0 x_1 ... x_{T-1}
Feedforward
KL Loss
$\mathcal{N}(0,1)$
$\hat{\mu}, \hat{\sigma}$
Posterior distribution
Sampling
~
Reconstruction Loss
x_1 x_2 ... x_T
$\hat{\mu}_1, \hat{\sigma}_1$ $\hat{\mu}_2, \hat{\sigma}_2$ ... $\hat{\mu}_T, \hat{\sigma}_T$
Decoder RNN
x_0 x_1 ... x_{T-1}

▲Figure 1: Overview of our variational autoencoder. x_i is the input, and symbols with hats are outputs of the model. The posterior distribution represents the random variables of the underlying parameters.

FURTHER READING

- K. Yeo, D. E. C. Grullon, F.-K. Sun, D. S. Boning, and J. R. Kalagnanam, "Variational Inference Formulation for a Model-free Simulation of a Dynamical System with Unknown Parameters by a Recurrent Neural Network," arXiv preprint arXiv:2003.01184, 2020.

SpArch: Efficient Architecture for Sparse Matrix Multiplication

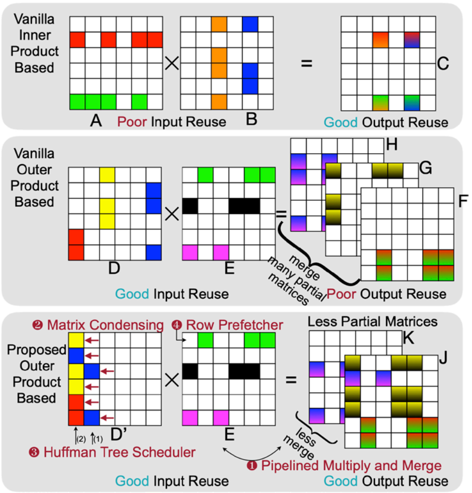
Z. Zhang*, H. Wang*, S. Han, W. J. Dally
(*Equal Contributions)
Sponsorship: NSF, DARPA

Generalized sparse matrix-matrix multiplication (SpGEMM) is the key computing kernel for many algorithms such as compressed deep neural networks. However, the performance of SpGEMM is memory-bounded on the traditional general-purpose computing platforms (CPU, GPU) because of the irregular memory access pattern and poor locality brought by the extremely sparse matrices. For instance, the density of Twitter's adjacency matrix is as low as 0.000214%. Previous accelerator OuterSPACE proposed an outer product method that has perfect input reuse but poor output reuse due to enormous partial matrices, thus achieving only 10.4% of the theoretical peak.

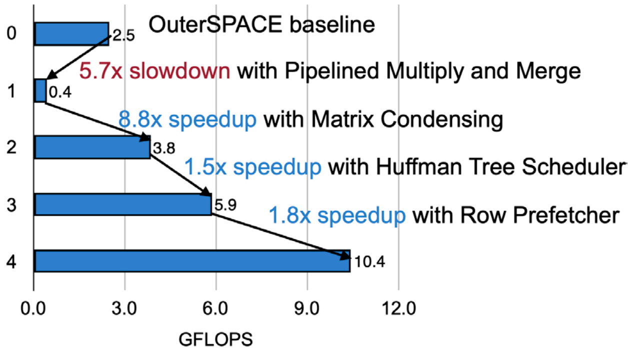
Therefore, we propose SpArch (HPCA'2020) to jointly optimize input and output data reuse. We obtain input reuse by using the outer product and output reuse by on-chip partial matrix merging (Figure 1).

We first design a highly parallelized merger to pipeline the two computing stages, Multiply and Merge. However, the number of partial matrices can easily exceed the on-chip merger's parallelism and incurs even larger DRAM access. We thus propose a condensed matrix representation for the left input matrix, where all non-zero elements are pushed to the left, forming much denser columns and fewer partial matrices. Unfortunately, the condensed representation can still produce more partial matrices than the merger's parallelism. Since the merge order impacts DRAM access, we should merge matrices with fewer non-zeros first. To this end, we design a Huffman tree scheduler to decide the near-optimal merge order of the partial matrices. Finally, we propose a row prefetcher to prefetch rows of the right matrix and store to a row buffer, thus improving the input reuse.

We evaluate SpArch on real-world datasets from SuiteSparse, SNAP, and rMAT, achieving 4×, 19×, 18×, 17×, and 1285× speedup and 6×, 164×, 435×, 307×, and 62× energy saving over OuterSPACE, MKL, cuSPARSE, CUSP and ARM Armadillo, respectively. Figure 2 shows the speedup breakdown of SpArch over OuterSPACE.



▲Figure 1: Four Innovations in SpArch.



▲Figure 2: SpArch Speedup Breakdown.

FURTHER READING

- Z. Zhang*, H. Wang*, S. Han, and W. J. Dally, "SpArch: Efficient Architecture for Sparse Matrix Multiplication," *HPCA 2020*, pp. 261-274, 2020.
- S. Pal, J. Beaumont, D. H. Park, A. Amarnath, S. Feng, C. Chakrabarti, C., et, al, "OuterSPACE: An Outer Product Based Sparse Matrix Multiplication Accelerator," *HPCA 2018*, pp. 724-736, 2018.
- J. Cong, Z. Fang, M. Lo, H. Wang, J. Xu, and S. Zhang, "Understanding Performance Differences of FPGAs and GPUs," *FCCM 2018*, pp. 93-96, 2018.

Efficient Natural Language Processing with Hardware-aware Transformers (HAT)

H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, S. Han

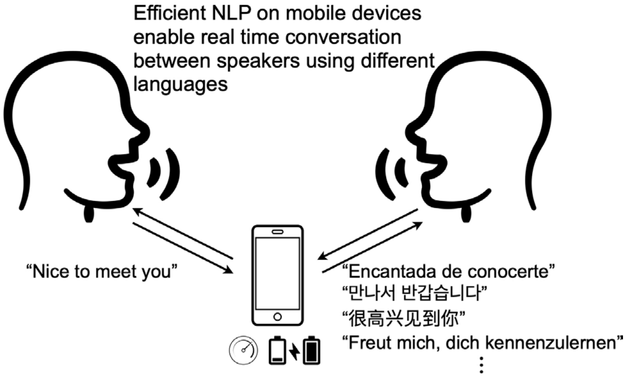
Sponsorship: Intel, TI, Facebook

Transformers have been widely used in Natural Language Processing (NLP) tasks, providing a significant performance improvement over previous convolutional and recurrent models. Nevertheless, transformers cannot be easily deployed in mobile/edge devices due to their extremely high cost of computation. For instance, to translate a sentence with only 30 words, a Transformer-Big model executes 13G Mult-Adds and takes 20 seconds on Raspberry Pi 4, making real-time NLP impossible.

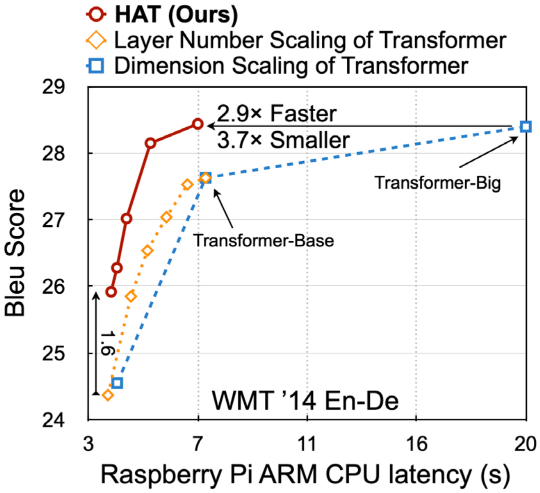
We found two critical phenomena that impact the transformer's efficiency: (1) FLOPs cannot reflect real latency and (2) efficient model architecture varies for different hardware. The reason is that for different hardware, the latency influencing factors differ a lot. For example, the embedding size has a large impact on Raspberry Pi but can hardly influence GPU latency.

Inspired by the success of neural architecture search (NAS), we propose to search for hardware-aware transformers (HAT, ACL'2020) by directly involving hardware latency feedback in the design loop (Figure 1). Hence, we do not need FLOPs as a latency proxy and can search hardware-specific models. We first construct a large search space with two features: (1) arbitrary encoder-decoder attention to allow all decoder layers to attend to multiple and different encoder layers and (2) heterogeneous layers to let different layers have different architectures. To conduct a low-cost search, we first train a SuperTransformer, which contains many Sub-Transformers with weight-sharing. Then we perform an evolutionary search in the SuperTransformer to find the best SubTransformers under hardware latency constraints.

We evaluate our HAT with three translation tasks on Raspberry pi ARM CPU, Intel CPU, and Nvidia GPU. HAT achieves up to 3× speedup and 3.7× smaller size over the conventional Transformer-Big model (Figure 2). With 10000× less search cost, HAT outperforms the Evolved Transformer with 2.7× speedup and 3.6× smaller size. Therefore, HAT enables efficient NLP on mobile devices.


Efficient NLP on mobile devices
enable real time conversation
between speakers using different
languages
"Nice to meet you"
"Encantada de conocerte"
"만나서 반갑습니다"
"很高兴见到你"
"Freut mich, dich kennenzulernen"
⋮

▲ Figure 1: HAT enables efficient NLP.



▲ Figure 2: DHAT achieves 2.9× speedup and 3.7× smaller size than conventional models.

FURTHER READING

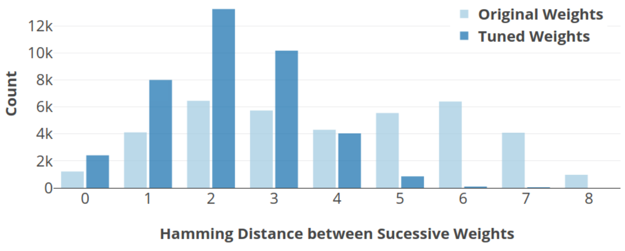
- H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, and S. Han, "HAT: Hardware-Aware Transformers for Efficient Neural Machine Translation," *ACL 2020*, 2020.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need," *NeurIPS 2017*, pp. 5998-6008, 2017.
- Y. He, J. Lin, Z. Liu, H. Wang, L. J. Li, and S. Han, "Amc: Automl for Model Compression and Acceleration on Mobile Devices," *ECCV 2018*, pp. 784-800, 2018.

Flexible Low Power CNN Accelerator for Edge Computing with Weight Tuning

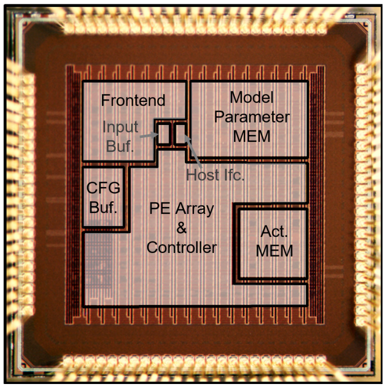
M. Wang, A. P. Chandrakasan
Sponsorship: Foxconn Technology Group

Smart edge devices that support efficient neural network (NN) processing have recently gained public attention. With algorithm development, previous work has proposed small-footprint NNs achieving high performance in various medium complexity tasks, e.g. speech keyword spotting (KWS), human activity recognition (HAR), etc. Among them, convolutional NNs (CNNs) perform well, which gives rise to the deployment of CNNs on edge devices. A hardware platform for edge devices should be (1) flexible to support various NN structures optimized for different applications; (2) energy efficient to operate within the power budget; (3) achieving high accuracy to minimize spurious triggering of power-hungry downstream processing, since it is often part of a large system.

This work proposes a weight tuning algorithm to improve the energy efficiency by lowering the switching activity of weight-related components, e.g. weight buses and multipliers. To achieve that, the algorithm reduces the Hamming distance between successive weights as shown in Figure 1. A flexible and runtime-reconfigurable CNN accelerator is co-designed with the algorithm. The system is fully self-contained for small CNNs. Speech keyword spotting is shown as an example with an integrated feature extraction frontend. As shown in Figure 2, a fully integrated custom ASIC is fabricated for this system. Based on post place-and-route simulation of the ASIC, the weight tuning algorithm reduces the energy consumption of weight delivery and computation by 1.70x and 1.20x respectively with little loss in accuracy.



▲Figure 1: The histogram of the distribution of Hamming distance between successive weights.



▲Figure 2: Chip micrograph.

FURTHER READING

- M. Wang, and A. P. Chandrakasan. "Flexible Low Power CNN Accelerator for Edge Computing with Weight Tuning." *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 209-212, Nov., 2019.

Protonic Solid-state Electrochemical Synapse for Physical Neural Networks

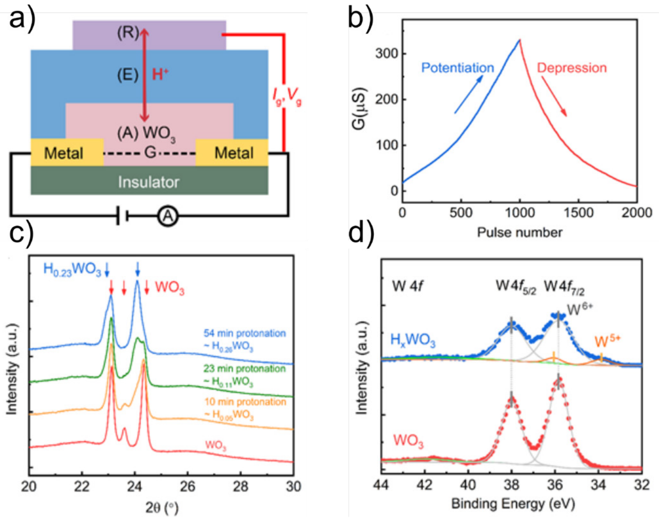
X. Yao, K. Klyukin, W. Lu, M. Onen, S. Ryu, D. Kim, N. Emond, I. Waluyo, A. Hunt, J. A. del Alamo, J. Li, B. Yildiz

Sponsorship: MIT Skoltech Program, by the SenseTime Group, Ltd. through MIT Quest for Intelligence, MRSEC Program of NSF

Physical neural networks made of analog resistive switching processors are promising platforms for analog computing and for emulating biological synapses. State-of-the-art resistive switches rely on either conductive filament formation or phase change, processes that suffer from poor reproducibility or high energy consumption, respectively. To avoid such shortcomings, we establish an alternative synapse design (Figure 1a) that relies on a deterministic charge-controlled mechanism, modulated electrochemically in solid state, that consists of shuffling the smallest cation, the proton.

This proof-of-concept, protonic solid-state electrochemical synapse is a three-terminal configuration and has a channel of active material (A), here taken as WO_3. By protonation/deprotonation, we modulate the electronic conductivity of the channel over seven orders of magnitude, obtaining a continuum of resistance states (Figure 1b). A solid proton reservoir layer (R), PdHx, serves as the gate terminal. A proton conducting solid electrolyte (E), Nafion, separates the channel and the reservoir. By probing the atomic, electronic, and crystal structures (Figure 1c-d) involved during proton intercalating, we reveal an increase in the electronic conductivity of WO_3 resulting from the increase of both the carrier density and the mobility. This switching mechanism has several key advantages over other switching mechanisms, including low energy dissipation and good reversibility and symmetry in programming.

We are also working to improve device properties and integrability of this protonic synapse by exploring alternative materials for both the active channel and the solid-state electrolyte. On one hand, promising host materials for the intercalation of protons and multivalent ions, such as vanadium pentaoxide, graphene oxide, and tantalum pentaoxide, are being investigated as potential active materials. On the other hand, nanocrystalline yttrium-doped barium zirconate and gadolinium-doped cerium oxide are being studied as possible room-temperature fast proton conductor ceramics.



▲Figure 1: (a) Schematic of the protonic electrochemical synapse device structure and (b) its potentiation/depression behavior. (c) Tungsten oxidation state (XPS of W 4*f* peak) and (d) crystal structure (XRD) change with protonation, from WO_3 to H_xWO_3.

FURTHER READING

- E. J. Fuller, F. El Gabaly, F. Léonard, S. Agarwal, S. J. Plimpton, R. B. Jacobs-Gedrim, C. D. James, M. J. Marinella, and A. A. Talin, "Li-ion Synaptic Transistor for Low Power Analog Computing," *Advanced Materials*, vol. 29, no. 1604310, 2017.
- Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. A. Talin and A. Salleo, "A Non-volatile Organic Electrochemical Device as a Low-voltage Artificial Synapse for Neuromorphic Computing," *Nature Materials*, vol. 16, pp. 414-419, 2017.