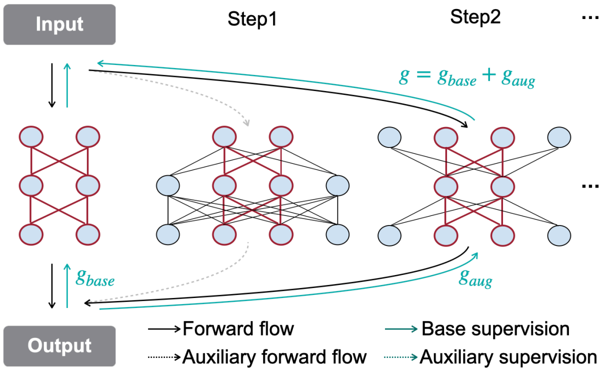
Machine Learning, Neurmorphic Computing, and AI

Network Augmentation for Tiny Deep Learning

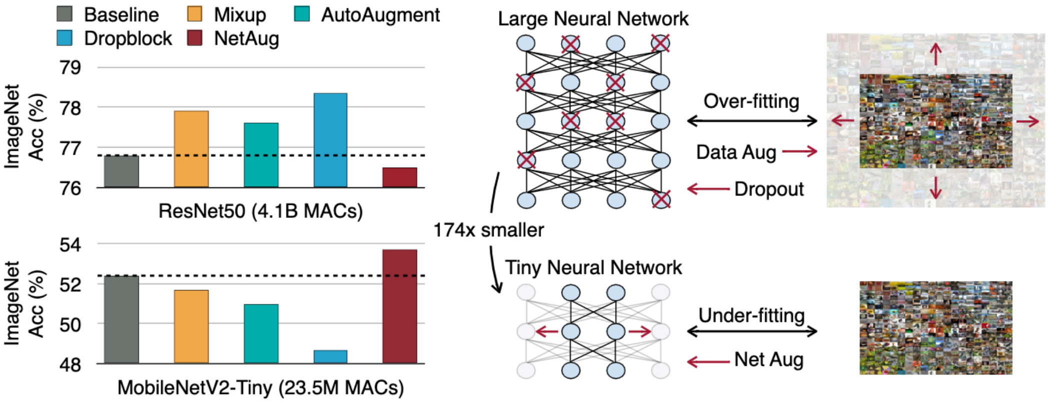
H. Cai, C. Gan, J. Lin, S. Han

Sponsorship: MIT-IBM Watson AI Lab, NSF, Hyundai, Ford, Intel, Amazon

We introduce Network Augmentation (NetAug), a new training method for improving the performance of tiny neural networks. Existing regularization techniques (e.g., data augmentation, dropout) have shown much success on large neural networks by adding noise to overcome over-fitting. However, we found that these techniques hurt the performance of tiny neural networks. We argue that training tiny models differ from large models: rather than augmenting the data, we should augment the model, since tiny models tend to suffer from under-fitting rather than over-fitting due to limited capacity. To alleviate this issue, NetAug augments the network (reverse dropout) instead of inserting noise into the dataset or the network. NetAug puts the tiny model into larger models and encourages it to work as a sub-model of larger models to get extra supervision, in addition to functioning as an independent model. At test time, only the tiny model is used for inference, incurring zero inference overhead. We demonstrate the effectiveness of NetAug on image classification and object detection. NetAug consistently improves the performance of tiny models, achieving up to 2.2% accuracy improvement on ImageNet. On object detection, achieving the same level of performance, NetAug requires 41% fewer MACs on Pascal VOC and 38% fewer MACs on COCO than the baseline.



◀ Figure 1: NetAug encourages the target tiny model to work as a sub-model of larger models to get extra supervision.



▲ Figure 2: NetAug improves the accuracy of the tiny model while regularization methods hurt its accuracy.

FURTHER READING

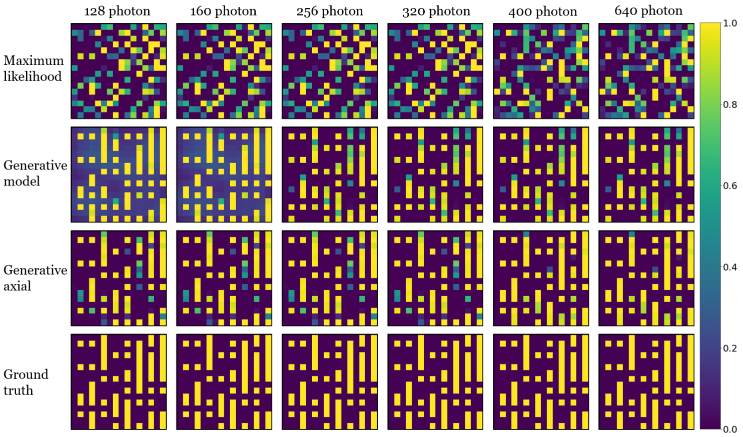
- H. Cai, C. Gan, J., Lin, and S. Han, "Network Augmentation for Tiny Deep Learning," *ICLR*, 2022.
- H. Cai, et al. "TinyTL: Reduce Activations, Not Trainable Parameters for Efficient On-Device Learning," *Advances in Neural Information Processing Systems*, vol. 33, p. ?, 2020.
- H. Cai, et al. "Once-for-all: Train One Network and Specialize it for Efficient Deployment," *ICLR*, 2020.

Physics-assisted Generative Adversarial Network for X-Ray Tomography

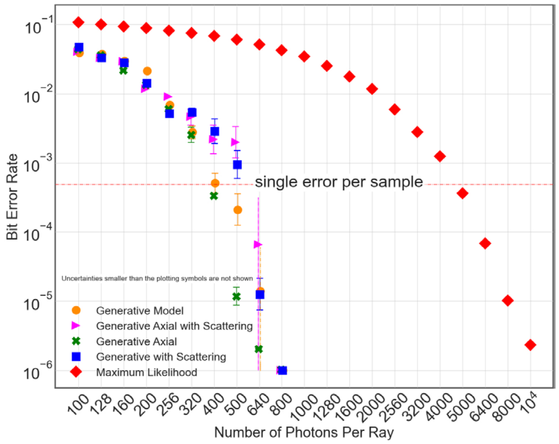
Z. Guo, J. K. Song, G. Barbastathis, M. E. Glinsky, C. T. Vaughan, K. W. Larson, B. K. Alpert, Z. H. Levine

Sponsorship: Intelligence Advanced Research Projects Activity, Office of the Director of National Intelligence, U.S. Department of Energy's National Nuclear Security Administration, Singapore National Research Foundation

X-ray tomography has applications in biomedical imaging, material study, electronic inspection, and more. The technique is capable of imaging the internal of objects in three dimensions non-invasively but may require prior regularization to obtain satisfactory reconstruction. In this work, we developed a physics-assisted generative adversarial network (PGAN) to determine and apply a learned prior in the reconstruction process. In contrast to previous efforts, our PGAN utilizes the maximum likelihood estimation to regularize the reconstruction with both physical and learned priors. The objects are synthetic integrated-circuits (ICs) from a proposed model dubbed CircuitFaker. Compared with maximum likelihood estimation, our PGAN can dramatically improve the synthetic IC reconstruction quality when the projection angles and photon budgets are limited. The advantages of using learned priors from deep learning in X-ray tomography may further enable its applications in low-photon nanoscale imaging.



▲Figure 1: Selected examples of IC reconstructions for different methods. Black stands for copper, white stands for silicon.



▲Figure 2: Bit-Error-Rate comparison between reconstruction methods at different imaging conditions.

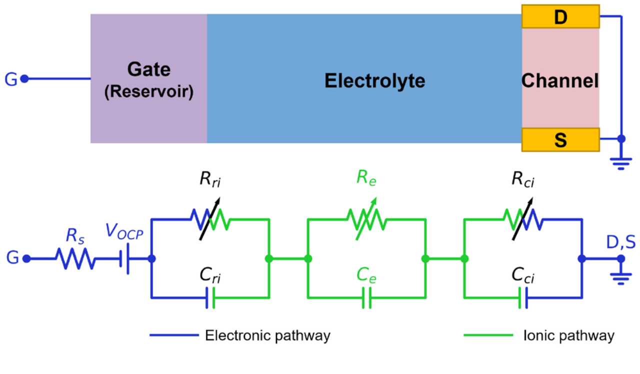
An Equivalent Circuit Model of an Electrochemical Artificial Synapse for Neuromorphic Computing

M. Huang, M. Onen, J. A. del Alamo, J. Li, B. Yildiz

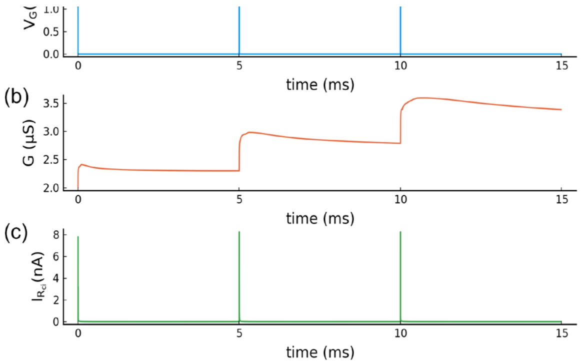
Sponsorship: MIT Quest for Intelligence

Deep learning based on artificial neural networks has achieved outstanding performance in a wide range of artificial intelligence applications. However, such computations are energy intensive to perform on conventional digital computers. A promising energy efficient approach to performing deep learning is to use neuromorphic computing hardware based on analog nonvolatile resistive switching devices in crossbar arrays. Among different resistive switching devices, electrochemical artificial synapses are a promising candidate as they have shown uniform and deterministic switching with low energy consumption. Electrochemical artificial synapses are programmable resistors in a three-terminal configuration where the conductance of a channel is controlled by reversible ion intercalation driven by voltage or current applied to a gate terminal that also serves as ion reservoir. Understanding the physical processes and the behavior of these devices is critical for applying them in brain-inspired computing systems.

In this work, we propose a 1D equivalent circuit model to describe the electrochemical processes of the electrochemical artificial synapse, including ionic transport, charge transfer, and diffusion processes. The model aims to predict the behavior of devices with different geometries and materials properties under various gate voltage or current waveforms. The model provides insight into processes such as the change of channel conductance and its relaxation after electrical pulses are applied to the gate and the interactions between successive pulses. In addition, the model can potentially guide the design of material properties and the optimization of device performance for achieving lower operating voltage, faster operation speed, and improved energy efficiency.



▲Figure 1: Schematic of electrochemical artificial synapses and circuit diagram of the equivalent circuit model.



▲Figure 2: Effect of voltage pulse application and post-pulse relaxation showing (a) gate voltage, (b) channel conductance, and (c) Faraday current at the channel electrode as a function of time.

FURTHER READING

- X. Yao, K. Klyukin, W. Lu, M. Onen, S. Ryu, D. Kim, N. Emond, I. Waluyo, A. Hunt, J. A. del Alamo, J. Li, and B. Yildiz, "Protonic Solid-State Electrochemical Synapse for Physical Neural Networks," *Nature Communications*, vol. 11, p. 3134, Jun. 2020.
- M. Onen, N. Emond, J. Li, B. Yildiz, and J. A. del Alamo, "CMOS-Compatible Protonic Programmable Resistor Based on Phosphosilicate Glass Electrolyte for Analog Deep Learning," *Nano Letts*., vol. 21, pp. 6111–6116, Jul. 2021.

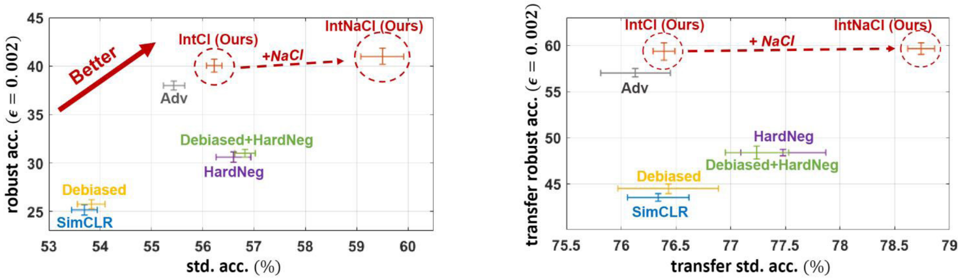
Revisiting Contrastive Learning through the Lens of Neighborhood Component Analysis: An Integrated Framework

C.-Y. Ko, J. Mohapatra, S. Liu, P.-Y. Chen, L. Daniel, T.-W. Weng

Sponsorship: MIT-IBM Watson AI Lab

Contrastive learning has drawn much attention and has become one of the most effective representation learning techniques recently. In essence, contrastive learning aims to leverage pairs of positive and negative samples for representation learning; however, positive/negative pairs are hard to define without the knowledge of downstream tasks. To provide a surrogate of measuring similarity, Current mainstream contrastive learning algorithms build up and optimize over a surrogate of the ideal contrastive loss. Although this formulation seems to put no assumptions on the downstream task classes, we find that there are in fact implicit assumptions on the class probability prior of the downstream tasks

In this project, we formally establish the connection between the neighborhood component analysis (NCA) and the unsupervised contrastive learning. Inspired by this interesting relationship to NCA, we further propose a new contrastive loss (named NaCl) which outperform existing paradigm. Furthermore, by inspecting the robust accuracy of several existing methods (e.g., Figure 1's y-axis, the classification accuracy when inputs are corrupted by crafted perturbations), one can see the insufficiency of existing methods in addressing robustness. Thus, we propose a new integrated contrastive framework (named IntNaCI and IntCl) that accounts for *both* the standard accuracy and adversarial cases: our proposed method's performance remains in the desired upper-right region (circled) as shown in Figure 1. A conceptual illustration of our proposals is given in Figure 2.

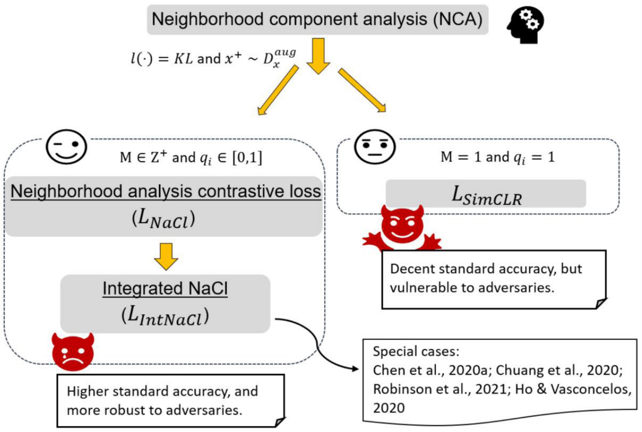


(a) CIFAR100

(b) CIFAR10

▲Figure 1. The performance of existing methods and our proposal (IntNaCl & IntCl) in terms of their standard accuracy (x-axis) and robust accuracy under Fast Gradient Sign Method attacks **ε** = 0.002 (y-axis). The transfer performance refers to fine-tuning a linear layer for CIFAR10 with representation networks trained on CIFAR100.

▶ Figure 2. A conceptual illustration of our proposals.


Neighborhood component analysis (NCA)
$l(\cdot) = KL$ and $x^+ \sim D_x^{aug}$
$M \in Z^+$ and $q_i \in [0,1]$
Neighborhood analysis contrastive loss
(L_{NaCl})
Integrated NaCl
($L_{IntNaCl}$)
Higher standard accuracy, and more robust to adversaries.
$M = 1$ and $q_i = 1$
L_{SimCLR}
Decent standard accuracy, but vulnerable to adversaries.
Special cases:
Chen et al., 2020a; Chuang et al., 2020; Robinson et al., 2021; Ho & Vasconcelos, 2020

FURTHER READING

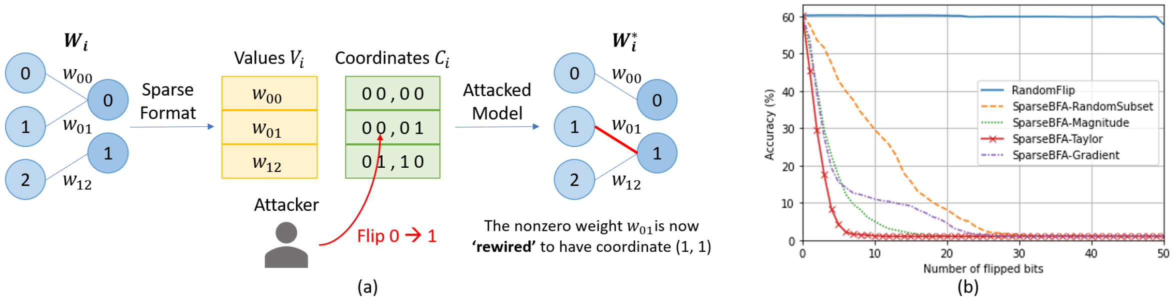
- T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *Proc. of the International Conference on Machine Learning*, pp. 1597-1607, 2020.
- J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood Components Analysis," *Advances in Neural Information Processing Systems*, pp. 513-520, 2004.

SparseBFA: Attacking Sparse Deep Neural Networks with the Worst-case Bit Flips on Coordinates

K. Lee, A. P. Chandrakasan

Sponsorship: Facebook, Korea Foundation for Advanced Studies

Deep neural networks (DNNs) are shown to be vulnerable to a few carefully chosen bit flips in their parameters, and bit flip attacks (BFAs) exploit such vulnerability to degrade the performance of DNNs. In this work, we show that DNNs with high sparsity that typically result from weight pruning have a unique source of vulnerability to bit flips when their coordinates of nonzero weights are attacked. We propose SparseBFA, an algorithm that searches for a small number of bits among the coordinates of nonzero weights when the parameters of DNNs are stored using sparse matrix formats. Using SparseBFA, we find that the performance of DNNs drops to the random-guess level by flipping less than 0.00005% (1 in 2 million) of the total bits.



▲ Figure 1: (a) When an attacker flips a bit in the coordinates representing the location of nonzero weights, the connection between neurons is rewired. (b) Accuracy of the ResNet50 model as bits in the coordinate list are flipped using SparseBFA.

Memory-efficient Gaussian Fitting for Depth Images in Real Time

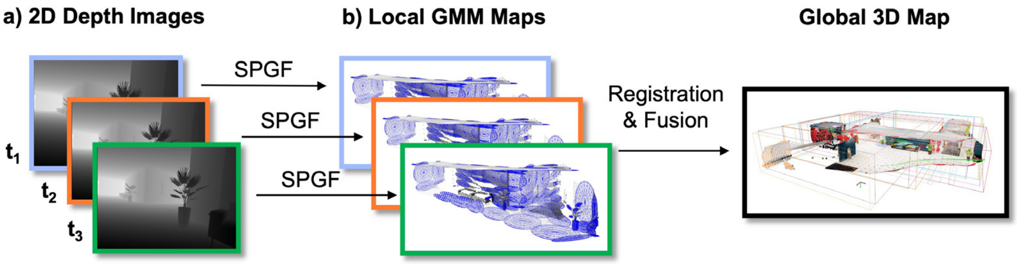
P. Z. X. Li, S. Karaman, V. Sze

Sponsorship: NSF RTML 1937501, NSF CPS 1837212

Energy-constrained microrobots, such as insect-sized flapping wing robots and palm-sized drones, are expected to be deployed for search and rescue missions in dangerous and unknown environments. These robots have very limited battery capacity, which limits the energy available for computation. Since the energy cost of memory access can be significant, algorithms designed for these robots should reduce memory overhead so that most data and variables used during computation can be efficiently stored in and accessed from lower-level caches (KBs in storage) instead of a larger off-chip dynamic random-access memory (DRAM).

Constructing a compact representation for 3D environments is essential for enabling autonomy for tasks such as navigation, localization, and exploration. From a sequence of depth images, many existing algorithms convert each image into a compact Gaussian mixture model (GMM) where each Gaussian models a surface in the environment. Then, GMMs across all images are fused together into a coherent global 3D map (Figure 1). While existing algorithms focus on reducing the size of each GMM, they require significant memory overhead due to the storage of the entire depth image or its intermediate representation in memory for multi-pass processing.

In this work, we present the Single-Pass Gaussian Fitting (SPGF) algorithm that incrementally constructs a GMM one pixel at a time in a single pass through a depth image. Since only one pixel is stored in memory at any time, SPGF achieves orders-of-magnitude lower memory overhead than prior approaches. By processing each depth image row-by-row, SPGF can efficiently and accurately infer surface geometries, which leads to higher precision than prior multi-pass approaches while maintaining the same compactness of the GMM. Using a low-power ARM Cortex-A57 CPU, SPGF operates at 32 fps, requires 43 KB of memory overhead, and consumes only 0.11 J per image. Thus, SPGF enables real-time mapping of large 3D environments on energy-constrained robots.



▲ Figure 1: (a) A depth image from a depth camera, and (b) a GMM (blue) generated using the proposed SPGF algorithm with a root-mean-square error of 9 cm, a memory overhead of 43 KB, a throughput of 32 fps, and an energy consumption of 0.11 J per frame using the low-power ARM Cortex-A57 CPU.

FURTHER READING

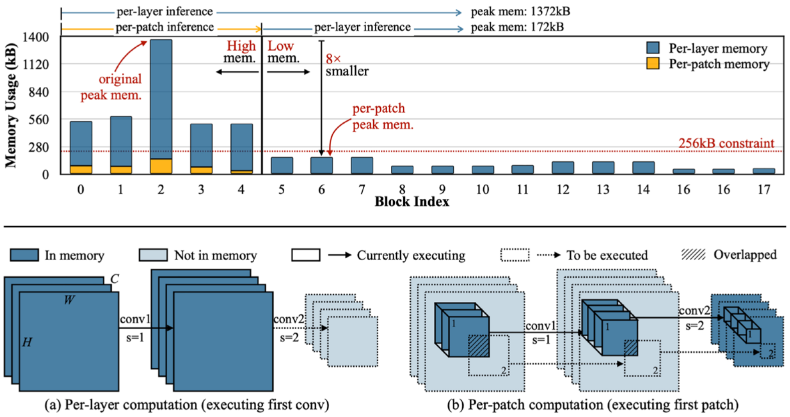
- P. Z. X. Li, S. Karaman, V. Sze, “Memory-Efficient Gaussian Fitting for Depth Images in Real Time,” *IEEE International Conference on Robotics and Automation (ICRA)*, May 2022.

MCUNetV2: Memory-efficient Patch-based Inference for Tiny Deep Learning

J. Lin, W. Chen, H. Cai, C. Gan, S. Han

Sponsorship: MIT-IBM Watson AI Lab, Samsung, Woodside Energy, NSF CAREER Award #1943349

Tiny deep learning on microcontroller units (MCUs) is challenging due to the limited memory size. We find that the memory bottleneck is due to the imbalanced memory distribution in convolutional neural network (CNN) designs: the first several blocks have an order-of-magnitude larger memory usage than the rest of the network. To alleviate this issue, we propose a generic patch-by-patch inference scheduling, which operates only on a small spatial region of the feature map and significantly cuts down the peak memory. However, naive implementation brings overlapping patches and computation overhead. We further propose network redistribution to shift the receptive field and floating-point operations (FLOPs) to the later stage and reduce the computation overhead. Manually redistributing the receptive field is difficult. We automate the process with neural architecture search to jointly optimize the neural architecture and inference scheduling, leading to MCUNetV2. Patch-based inference effectively reduces the peak memory usage of existing networks by 4-8x. Co-designed with neural networks, MCUNetV2 sets a record ImageNet accuracy on MCU (71.8%), and achieves >90% accuracy on the visual wake words dataset under only 32kB static random access memory (SRAM). MCUNetV2 also unblocks object detection on tiny devices, achieving 16.9% higher mean Average Precision (mAP) on Pascal VOC compared to the state-of-the-art result. Our study largely addresses the memory bottleneck in tinyML and paves the way for various vision applications beyond image classification.



▲ Figure 1: MobileNetV2 has a very imbalanced memory usage distribution: the peak memory is determined by the first 5 blocks with high peak memory, while the later blocks all share a small memory usage. By using per-patch inference, we are able to significantly reduce the peak memory by 8x, fitting MCUs with a 256 kB memory budget.

FURTHER READING

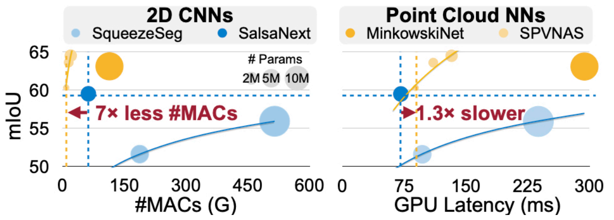
- J. Lin, W. M. Chen, Y. Lin, C. Gan, and S. Han, "MCUNet: Tiny Deep Learning on Iot Devices," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11711-11722, 2020.

PointAcc: Efficient Point Cloud Deep Learning Accelerator

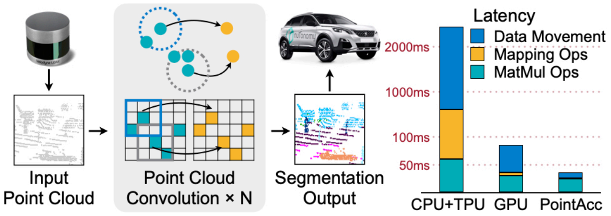
Y. Lin, Z. Zhang, H. Tang, H. Wang, S. Han
Sponsorship: NSF, Hyundai, Qualcomm, MIT-IBM Watson AI Lab

Deep learning on point clouds plays a vital role in a wide range of applications such as autonomous driving and augmented reality (AR) and virtual reality (VR). These applications interact with people in real time on edge devices and thus require low latency and low energy. Compared to projecting the point cloud to 2D space, directly processing 3D point cloud yields higher accuracy and lower number of multiply-accumulations (#MACs). However, the extremely sparse nature of point cloud poses challenges to hardware acceleration. For example, we need to explicitly determine the nonzero outputs and search for the nonzero neighbors (mapping operation), which is unsupported in existing accelerators. Furthermore, explicit gathering and scattering of sparse features are required, resulting in large data movement overhead.

In this work, we comprehensively analyze the performance bottleneck of modern point cloud networks on central processing, graphics processing, and tensor processing units (CPU/GPU/TPU). To address the challenges, we then present PointAcc, a novel point cloud deep learning accelerator. PointAcc maps diverse mapping operations onto one versatile ranking-based kernel, streams the sparse computation with configurable caching, and temporally fuses consecutive dense layers to reduce the memory footprint. Evaluated on 8 point cloud models across 4 applications, PointAcc achieves 3.7× speedup and 22× energy savings over RTX 2080Ti GPU. Co-designed with light-weight neural networks, PointAcc rivals the prior accelerator Mesorasi by 100× speedup with 9.1% higher accuracy running segmentation on the S3DIS dataset. PointAcc paves the way for efficient point cloud recognition.



▲Figure 1: Compared to 2D CNNs, point cloud networks have higher accuracy and lower #MACs, but higher GPU latency due to low utilization brought by sparsity and irregularity.



▲Figure 2: Point cloud deep learning is crucial for real-time AI applications. PointAcc accelerates point cloud computations by resolving sparsity and data movement bottlenecks.

FURTHER READING

- Y. Lin, Z. Zhang, H. Tang, H. Wang, and S. Han, “PointAcc: Efficient Point Cloud Accelerator,” *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 449-461, Oct. 2021.
- Y. Feng, B. Tian, T. Xu, P. Whatmough, and Y. Zhu, “Mesorasi: Architecture Support for Point Cloud Analytics via Delayed-aggregation,” *MICRO-53: 53rd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 1037-1050, Oct. 2020.

Algorithm-system Co-design for Efficient Calorimetry Clustering

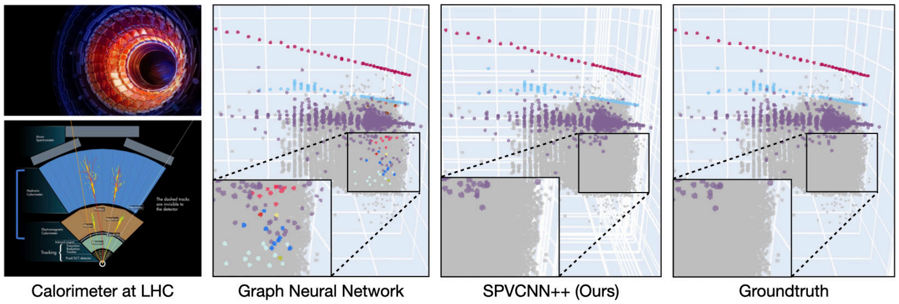
Z. Liu, X. Yang, S. Han
Collaborators: A. Schuy (UW), S-C. Hsu (UW), J. Krupa (MIT), P. Harris (MIT)
Sponsorship: NSF

The content management system (CMS) detector at the Large Hadron Collider (LHC) reconstructs high-energy proton-proton collisions to understand physics beyond the standard model. A key part of the CMS detector is the calorimeter, which reconstructs particle energies by clustering 3D energy deposits from particle showers. The LHC observes ~1 billion collisions per second and must decide within ~1 millisecond which collisions to keep; this imposes a strict throughput/latency requirement. Furthermore, the LHC data flow will increase tenfold by 2027. The corresponding increase in computing requirements using traditional algorithms is beyond our capabilities. Therefore, there is an urgent need to develop accurate algorithms capable of scaling under resource and latency constraints.

3D point cloud neural networks are very suitable for calorimetry clustering. However, they are ten times more computationally expensive than 2D convoluted neural networks (CNNs). Moreover, the sparse and irregular nature of the point cloud makes them less favored by general-purpose hardware (such as CPU, GPU, and TPU). We approach these challenges with algorithm-system co-design.

From the algorithm side, we have developed SPVCNN++, which brings together the best from point-based and voxel-based models. SPVCNN++ is composed of a fine-grained point-based branch that keeps the 3D data in high resolution without large memory footprints and a coarse-grained voxel-based branch that aggregates the neighboring features without many random memory accesses. Compared with the GNN-based approach, our SPVCNN++ achieves a 4% higher panoptic quality on the particle physics benchmark.

From the system side, we have developed TorchSparse, a specialized high-performance GPU computing library for 3D sparse computations. TorchSparse directly optimizes the two bottlenecks of sparse convolution: irregular computation and data movement. As a result, our TorchSparse achieves more than 1.5x measured end-to-end speedup over the state of the art.


Calorimeter at LHC
Graph Neural Network
SPVCNN++ (Ours)
Groundtruth

▲ Figure 1: Results of our algorithm-system co-design solution for efficient calorimetry clustering. Compared with conventional GNN-based approach, our SPVCNN++ provides much more accurate clustering results.

FURTHER READING

- H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution," *European Conference on Computer Vision (ECCV)*, Aug. 2020.
- H. Tang, Z. Liu, X. Li, Y. Lin, and S. Han, "TorchSparse: Efficient Point Cloud Inference Engine," to be presented at *Conference on Machine Learning and Systems (MLSys)*, August 2022.

Fabrication of Electrochemical Artificial Synapses Based on Intercalation of Mg^{2+} Ions

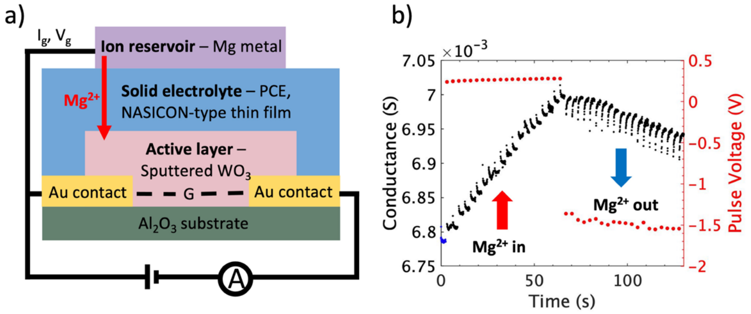
M. Schwacke, J. A. del Alamo, J. Li, B. Yildiz

Sponsorship: SRC, MIT Quest

Deep learning based on neural networks has gained much attention due to its success in a wide range of applications. However, running neural networks on traditional computer systems requires large amounts of power and memory due to the von-Neumann structure which separates the central processing unit (CPU) and memory. Alternatively, crossbar architectures with two-terminal resistive switches can imitate neurons and synapses, allowing for the integration of memory and computation.

Electrochemical artificial synapses (EAS) are a promising, emerging resistive switching mechanism. Ions are shuttled between the reservoir and active layer, changing the ion concentration and thus the conductivity of the channel, which allows for storage of an analog state (Figure 1a). Several studies have demonstrated successful EAS based on the intercalation of protons or Li+ ions. However, Li is incompatible with complementary metal-oxide-semiconductor (CMOS) fabrication, and protons can diffuse out of the channel after insertion, creating problems for the long-term retention of stored states and compromising endurance. This research focuses on EAS that function by the intercalation of Mg2+ ions. Mg was chosen for its abundance, CMOS compatibility, and presence in battery literature.

Current devices based on a radio-frequency sputtered WO3 channel, succinonitrile/Mg(TFSI)2 phase convertible electrolyte (PCE), and Mg metal reservoir show successful modulation in channel conductance with applied current pulses (Figure 1b). However, formation of resistive interfacial phases and electronic conductance through the electrolyte have proven problematic for device performance, in terms of the repeatability, symmetry, and energy consumption. The former problem has been solved by formation of a MgI2 artificial solid-electrolyte interphase (SEI) on the Mg prior to device assembly. To resolve the latter, we are currently developing a thin film, Sodium (Na) Super Ionic Conductor (NASICON)-type electrolyte which has lower electronic conductivity than the PCE and is also compatible with CMOS processing. This could allow these devices to be used as fast, enduring, and energy-efficient computing elements.



▲Figure 1: (a) Schematic of EAS device based on movement of Mg2+ ions between reservoir and active layer. (b) Modulation of active layer conductance with applied current pulses to gate (15 0.5-s pulses of +20 uA, followed by 15 0.5-s pulses of -60 uA with 2.5-s read between each pulse).

FURTHER READING

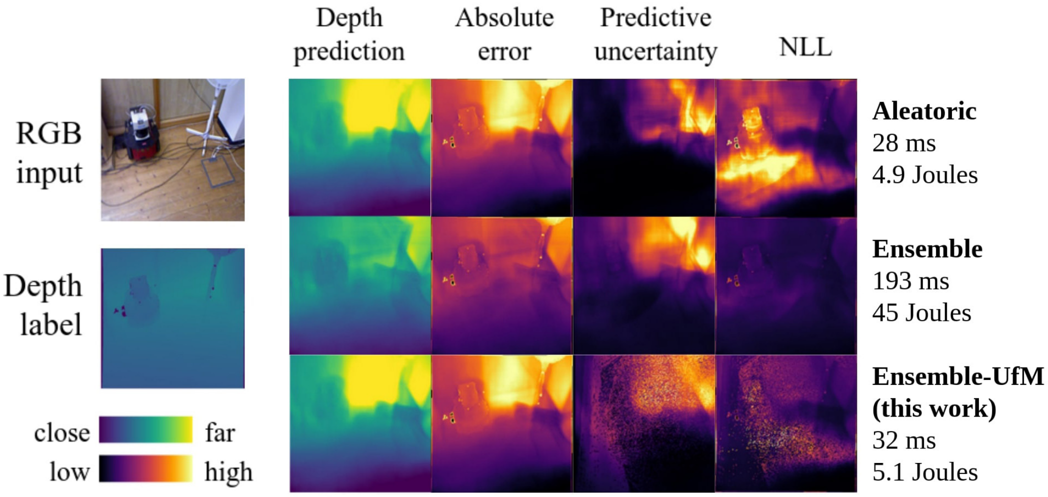
- Yao, X., Klyukin, K., Lu, W., Onen, M., Ryu, S., Kim, D., Edmond, N., Waluyo, I. et al. "Protonic Solid-state Electrochemical Synapse for Physical Neural Networks," *Nature Communications*, vol. 11, p. 3134, Jun. 2020.

Uncertainty from Motion for DNN Monocular Depth Estimation

S. Sudhakar, V. Sze, S. Karaman

Sponsorship: NSF Cyber-Physical Systems Program Grant no. 1837212, NSF Real-Time Machine Learning Program Grant no. 1937501, MIT-Accenture Fellowship

Deployment of deep neural networks (DNNs) for monocular depth estimation in safety-critical scenarios on resource-constrained platforms requires well-calibrated and efficient uncertainty estimates. However, uncertainty estimates from state-of-the-art ensembles are computationally expensive, requiring multiple inferences per input. We propose a new algorithm, called Uncertainty from Motion (UfM), that runs only one inference per input by exploiting the temporal redundancy in video inputs to merge incrementally the per-pixel depth prediction and per-pixel uncertainty over a sequence of frames. In a set of experiments using a DenseNet-based autoencoder on a single GPU, UfM offers near ensemble uncertainty quality while consuming on average 5.1 Joules with a latency of 32 ms per frame, which is 8.8x less energy and 6.4x faster than the ensemble. In Figure 1, we compare the results of a DNN that predicts only its data (aleatoric) uncertainty, an ensemble that predicts its overall uncertainty, and a DNN with UfM. We see that UfM retains the uncertainty quality of ensembles at a fraction of the energy and latency, enabling uncertainty estimation for resource-constrained, real-time scenarios.



▲Figure 1: Uncertainty estimation comparison for an aleatoric network, ensemble, and UfM applied to ensembles on an out-of-distribution example from the TUM RGBD dataset. Lower negative log-likelihood (NLL) indicates better uncertainty quality.

FURTHER READING

- S. Sudhakar, S. Karaman, and V. Sze, "Uncertainty from Motion for DNN Monocular Depth Estimation," to be presented at *2022 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2022.

Unsupervised Anomaly Detection on High-frequency Time Series in the Frequency Domain

F.-K. Sun, J. H. Lang, D. S. Boning

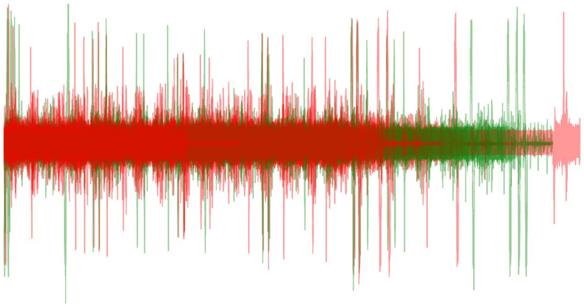
Sponsorship: HARTING Technology Group, Lam Research

Unplanned downtime caused by machine faults is costly. By installing sensors and an automated fault detection system, organizations can monitor process and machine sensor data and flag anomalies before problems become serious. However, anomalies are inherently rare, and detecting anomalies typically requires domain expertise.

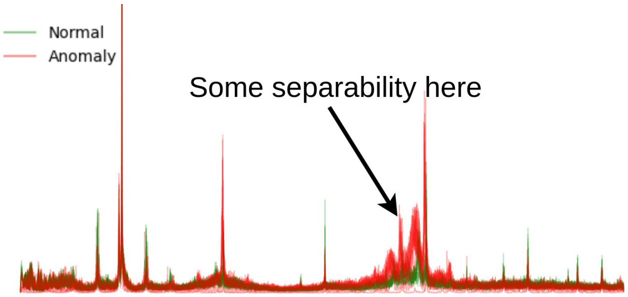
To address these issues, we formulate the problem as unsupervised anomaly detection on time series. That is, we use only known good data, so our method is applicable even before anomalies are observed. Furthermore, we propose an autoencoder model in combination with several techniques to automatically learn from the data without domain expertise. The autoencoder model is small, so it is suitable when only a small amount of data is available and requires relatively modest computational resources.

The input is the time series representing the sensor values of a manufacturing run. Given the input, we first apply the Fourier Transform to the whole series because frequency domain representation is particularly useful for high-frequency series. Secondly, we propose fractional average pooling to normalize all series to the same number of frequency components. Next, we train the autoencoder on only known good runs, with dropout as data augmentation. Finally, we assign anomaly scores to runs based on the reconstruction error and set two standard deviations as the threshold to classify runs.

We evaluate our method on two vibration datasets about milling machines: one considers worn tools and another considers switching off the cooling system as anomalies. On both datasets, we achieve area-under-curves (AUC) of 99%+ and an average accuracy of 90% on classifying anomaly runs vs. normal runs in the testing set. Our method is applicable to manufacturing industries where high-frequency signals are accessible and has the following advantages: (1) only good run data used, (2) no domain expertise required, and (3) a small and simple model.



▲Figure 1: Normal and anomalous series in the time domain are difficult to classify.



▲Figure 2: Normal and anomalous series are easier to classify in the frequency domain.

FURTHER READING

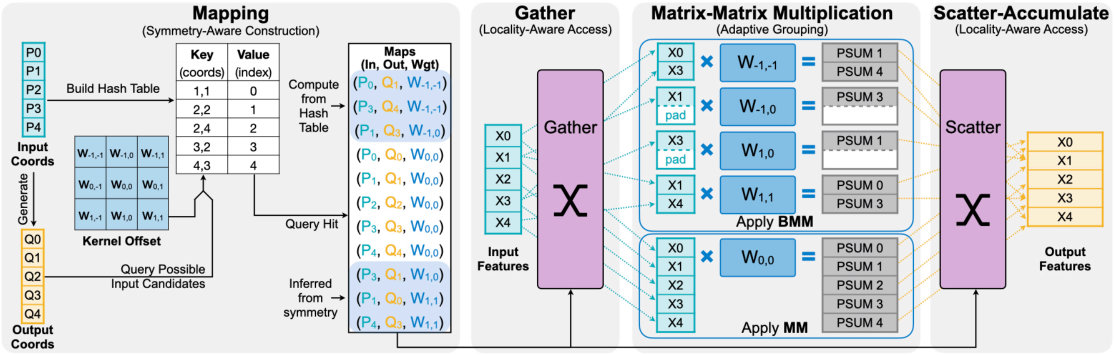
- D. W. Martin, "Fault Detection in Manufacturing Equipment Using Unsupervised Deep Learning," *MIT EECS Meng Thesis*, Feb. 2021.

TorchSparse: Efficient Point Cloud Inference Engine

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

Sponsorship: NSF CAREER Award, Ford, Hyundai, Qualcomm Innovation Fellowship

Deep learning on point clouds has received increased attention thanks to its wide applications in augmented and virtual reality and autonomous driving. These applications require low latency and high accuracy to provide real-time user experience and ensure user safety. Unlike conventional dense workloads, the sparse and irregular nature of point clouds poses severe challenges to running sparse convoluted neural networks efficiently on general-purpose hardware. Furthermore, existing sparse acceleration techniques for 2D images do not translate to 3D point clouds. In this paper, we introduce TorchSparse, a high-performance point cloud inference engine that accelerates sparse convolution computation on graphics processing units. TorchSparse directly optimizes the two bottlenecks of sparse convolution: irregular computation and data movement. It applies adaptive matrix multiplication grouping to trade computation for better regularity, achieving 1.4-1.5x speedup for matrix multiplication. It also optimizes the data movement by adopting vectorized, quantized, and fused locality-aware memory access, reducing the memory movement cost by 2.7x. Evaluated on seven representative models across three benchmark datasets, TorchSparse achieves 1.6x and 1.5x measured end-to-end speedup over the state-of-the-art MinkowskiEngine and SpConv, respectively.



▲ Figure 1: TorchSparse aims at accelerating sparse convolution, which consists of four stages: mapping, gathering, matrix multiplication. and scatter-accumulation. We follow two general principles: (1) memory footprint should be reduced, and (2) computation regularity should be increased to optimize these four components with quantized, vectorized, row-major scatter/gather (Principle 1); adaptively batched MM (Principle 2); and mapping kernel fusion (Principle 1).

FURTHER READING

- H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution," *European Conference on Computer Vision (ECCV)*, Aug. 2020.
- H. Tang, Z. Liu, X. Li, Y. Lin, and S. Han, "TorchSparse: Efficient Point Cloud Inference Engine," *Conference on Machine Learning and Systems (MLSys)*, Aug. 2022.

QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits

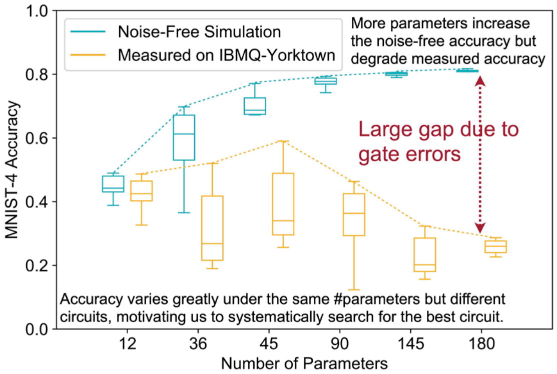
H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T. Chong, S. Han

Sponsorship: MIT-IBM Watson AI Lab, NSF CAREER Award, Qualcomm Innovation Fellowship

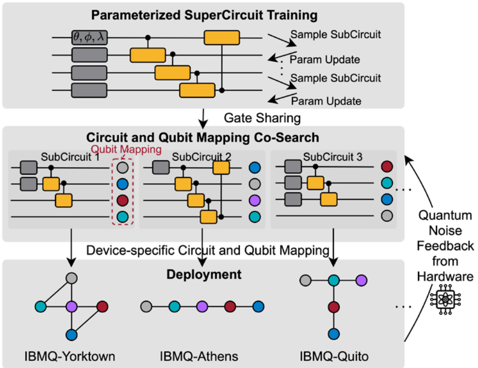
Quantum noise is the key challenge in noisy intermediate-scale quantum (NISQ) computers. Previous work on mitigating noise has primarily focused on gate-level or pulse-level noise-adaptive compilation. However, few research efforts have explored a *higher level of optimization* by making the quantum circuits themselves resilient to noise.

We propose QuantumNAS, a comprehensive framework for noise-adaptive co-search of the variational circuit and qubit mapping. Variational quantum circuits are a promising approach for performing quantum machine learning (QML) and simulation. However, finding the best variational circuit and its optimal parameters is challenging due to the large design space and parameter training cost. We propose to decouple the circuit search and parameter training by introducing a novel *SuperCircuit*. The SuperCircuit is constructed with multiple layers of pre-defined parameterized gates and trained by iteratively sampling and updating the parameter subsets (SubCircuits) of it. It provides an accurate estimation of SubCircuits performance trained from scratch. Then we perform an evolutionary co-search of SubCircuit and its qubit mapping. The SubCircuit performance is estimated with parameters inherited from SuperCircuit and simulated with real device noise models. Finally, we perform iterative gate pruning and finetuning to remove redundant gates.

Extensively evaluated with 12 QML and Variational Quantum Eigensolver (VQE) benchmarks on 14 quantum computers, QuantumNAS significantly outperforms baselines. For QML, QuantumNAS is the first to demonstrate over 95% 2-class, 85% 4-class, and 32% 10-class classification accuracy on real quantum machines. It also achieves the lowest eigenvalue for VQE tasks on H_2, H_2O, LiH, CH_4, and BeH_2 compared with UCCSD. We also open-source TorchQuantum (https://github.com/mit-han-lab/torchquantum) for fast training of parameterized quantum circuits to facilitate future research.



▲Figure 1: MNIST-4 on noise-free simulator / real QC. More parameters increase the noise-free accuracy but degrade measured accuracy due to larger gate errors. Accuracy gap is large.



▲Figure 2: Noise-adaptive circuit and qubit mapping co-search improves the robustness on real machines.

FURTHER READING

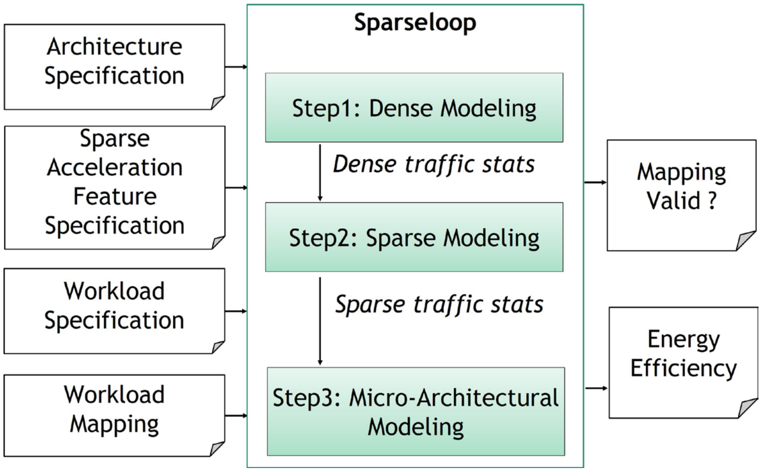
- H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T., Chong, and S. Han, "QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits," *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2022, pp. 692-708, doi: 10.1109/HPCA53966.2022.00057..
- H. Wang, J. Gu, Y. Ding, Z. Li, F. T. Chong, D. Z. Pan, and S. Han, "QuantumNAT: Quantum Noise-Aware Training with Noise Injection, Quantization and Normalization," *2022 Design Automation Conference*, 2022.
- H. Wang, Z. Li, J. Gu, Y. Ding, D. Z. Pan, & S. Han, "QOC: Quantum On-Chip Training with Parameter Shift and Gradient Pruning," *2022 Design Automation Conference*, 2022.

Sparseloop: An Analytical Approach to Sparse Tensor Accelerator Modeling

Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, J. S. Emer
Sponsorship: DARPA (HR0011-18-3-0007), Ericsson

In recent years, a myriad of accelerators has been proposed to efficiently process sparse tensor algebra applications (e.g., neural networks), leading to a large and diverse design space. However, the lack of systematic description and modeling support for these sparse tensor accelerators prevents hardware designers from efficient design space exploration.

To solve the problem, we present Sparseloop, the first fast, accurate, and flexible analytical modeling framework for sparse tensor accelerators. Figure 1 shows Sparseloop's high-level framework. Based on a unified taxonomy to describe the diverse designs, Sparseloop comprehends a wide set of architecture specifications and calculates designs' performance based on stochastic tensor density models. Across a representative set of accelerators and workloads, Sparseloop achieves >600x faster modeling speed than cycle-level simulations, <1% error compared to a custom accelerator model with statistical data modeling, and <8% error compared to simulations with real data.



▲ Figure 1: MNIST-4 on noise-free simulator / real QC. More parameters increase the noise-free accuracy but degrade measured accuracy due to larger gate errors. Accuracy gap is large.

FURTHER READING

- Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, and J. S. Emer, "Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators," presented at *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Mar., 2021.
- Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," presented at *2019 International Conference on Computer Aided Design (ICCAD)*, Nov. 2019.
- A. Parashar et al., "Timeloop: A Systematic Approach to DNN Accelerator Evaluation," presented at *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, Mar., 2019.

Fast Convergence of Unstable Reinforcement Learning Problems

W. Zhang

For many of the reinforcement learning applications, the system is assumed inherently stable and with bounded reward, state, and action space. These are key requirements for the optimization convergence of classical reinforcement learning reward function with discount factors. Unfortunately, these assumptions are no longer valid for many real-world problems such as an unstable linear–quadratic regulator (LQR). In this work, we propose new methods to stabilize and speed up the convergence of unstable reinforcement learning problems with the policy gradient methods. We provide theoretical insights on the efficiency of our methods. In practice, we achieve good experimental results over multiple examples where the vanilla method could hardly fail to converge due to system instability.

Latency-tolerant On-device Learning

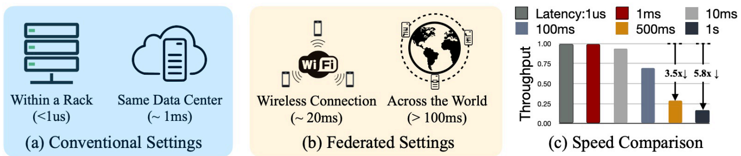
L. Zhu, S. Han

Sponsorship: MIT-IBM Watson AI Lab, Samsung, Woodside Energy, NSF, Amazon

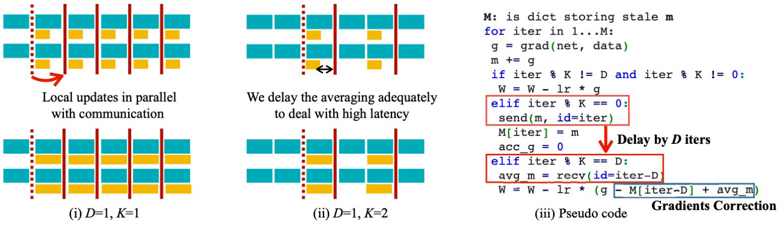
Much new and sensitive data are generated and collected by intelligent edge devices with rich sensors every day. On-device federated learning is an emerging direction that enables jointly training a model without sharing the data. Since the data is distributed across many edge devices through wireless / long-distance connections, federated learning suffers from inevitable high communication latency. However, the latency issues are undermined in the current literature and existing approaches such as FedAvg become less efficient when the latency increases.

To overcome the problem, we propose delayed gradient averaging (DGA) to address the latency bottleneck. The key idea is to delay the gradient averaging to a future iteration; thus the communication can be pipelined with computation (as shown in Figure 2). By accepting stale average gradients for model updates, DGA allows the communication to execute in parallel with the computation and become scalable even under extreme latency.

We theoretically prove that DGA attains a similar convergence rate as FedAvg and empirically show that our algorithm can tolerate high network latency without compromising accuracy. Specifically, we benchmark the training speed on various vision (CIFAR, ImageNet) and language tasks (Shakespeare), with both independent and identically distributed (IID) and non-IID partitions, and show that DGA can bring 2.55× to 4.07× speedup. Moreover, we built a 16-node Raspberry Pi cluster and show that DGA can consistently speed up real-world federated learning applications

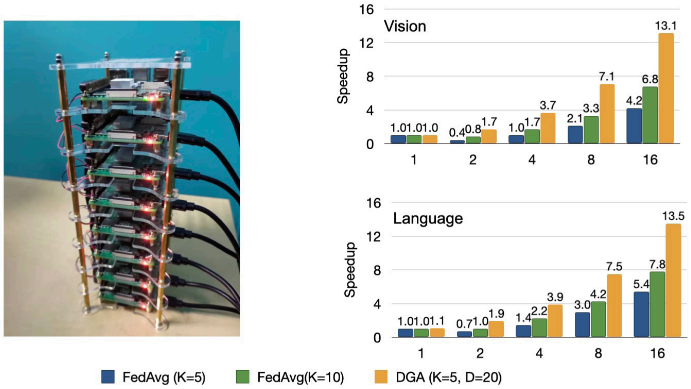


▲Figure 1: Training settings of conventional distributed training and federated learning differ greatly. High latency cost greatly degrades FedAvg's performance, posing a severe challenge to scale up the training.



```
M: is dict storing stale m
for iter in 1...M:
 g = grad(net, data)
 m += g
 if iter % K != D and iter % K != 0:
  W = W - lr * g
 elif iter % K == 0:
  send(m, id=iter)
  M[iter] = m
  acc_g = 0
 elif iter % K == D:
  avg_m = recv(id=iter-D)
  W = W - lr * (g - M[iter-D] + avg_m)
```

▲Figure 2: Overview of our proposed DGA.



▲Figure 3: Our benchmark Pi Farm setup and speedup comparison.

FURTHER READING

- H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *arXiv,* 2016.
- L. Zhu, H. Lin, Y. Lu, Y. Lin, and S. Han, "Delayed Gradient Averaging: Tolerate the Communication Latency for Federated Learning," *NeurIPS* 2021.