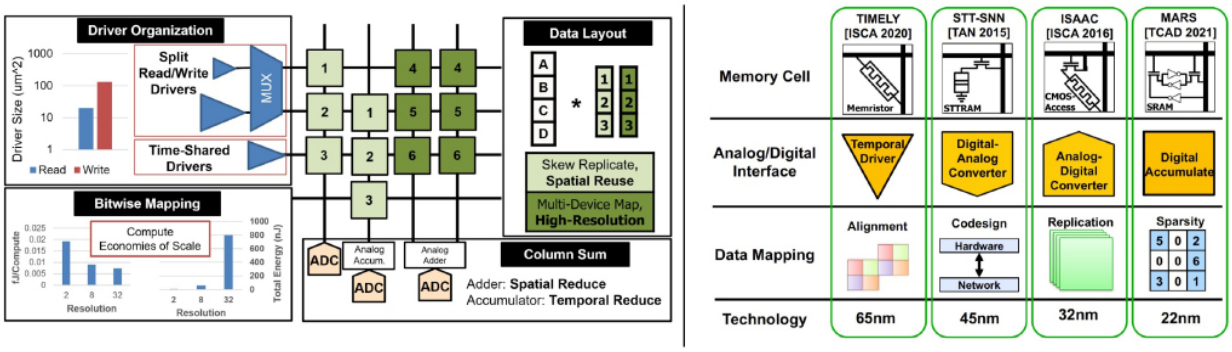
Machine Learning and Other Accelerators

Architectural Evaluation of Processing-in-Memory Accelerators

T. Andrulis, J. S. Emer, V. Sze
Sponsorship: Ericsson, TSMC, MIT Artificial Intelligence Hardware Program

Processing-in-memory (PIM) accelerators are a promising approach to efficiently run deep neural networks (DNNs) as they move compute into memory and reduce high DNN data movement costs. Unfortunately, research has mainly focused on devices (e.g., memristors), circuits (e.g., analog converters), or architecture (e.g., dataflow) in isolation. It is desirable to see how innovation at any level, such as new devices, may change the efficiency and performance of whole accelerators. This would enable fair comparison of innovations and yield insight into the vast number of ways to combine them. We present a framework that models PIM at an architectural level. With fast simulation and easy-to-change PIM device, circuit, and architecture models, our framework enables researchers to see how innovations affect the efficiency and performance of PIM accelerators. Further, we simulate up to 10,000x faster, enabling fast evaluation of different PIM accelerators and exploration of the vast design space.

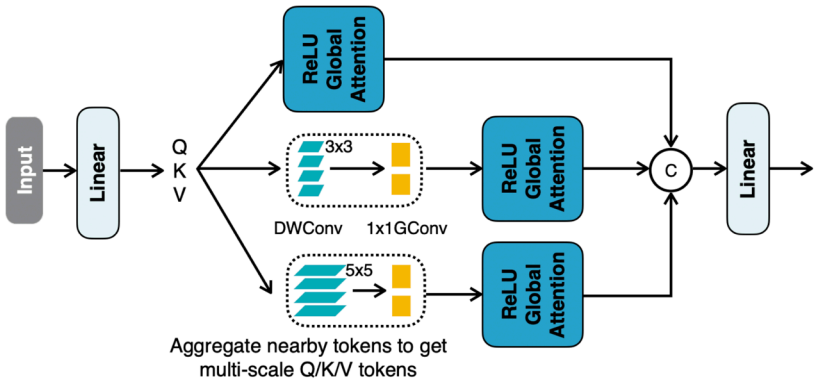


▲ Figure 1: (Left) PIM analog crossbar design showing options in drivers/column sum (Circuit) and in data layout/bit mapping (Architecture) (Right) Previous works' design choices. The number of combinations is large and needs a rapid framework to explore.

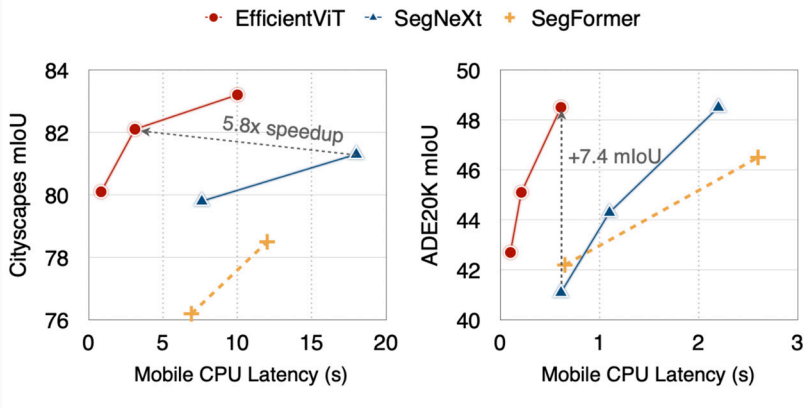
EfficientViT: Lightweight Multi-scale Attention for On-device Semantic Segmentation

H. Cai, J. Li, M. Hu, C. Gan, S. Han
Sponsorship: NSF, MIT-IBM Watson AI Lab, Ford, Intel, Qualcomm

Semantic segmentation enables many appealing real-world applications, such as computational photography, autonomous driving, etc. However, the vast computational cost makes deploying state-of-the-art semantic segmentation models on edge devices with limited hardware resources difficult. This work presents EfficientViT, a new family of semantic segmentation models with a novel lightweight multi-scale attention for on-device semantic segmentation. Unlike prior semantic segmentation models that rely on heavy self-attention, hardware-inefficient large-kernel convolution, or complicated topology structure to obtain good performances, our lightweight multi-scale attention achieves a global receptive field and multi-scale learning (two critical features for semantic segmentation models) with only lightweight and hardware-efficient operations. As such, EfficientViT delivers remarkable performance gains over previous state-of-the-art (SOTA) semantic segmentation models across popular benchmark datasets with significant speedup on the mobile platform. Without performance loss on Cityscapes, our EfficientViT provides up to 15x and 9.3x mobile latency reduction over SegFormer and SegNeXt, respectively. Maintaining the same mobile latency, EfficientViT provides +7.4 mIoU gain on ADE20K over SegNeXt.



◀ Figure 1: EfficientViT achieves a global receptive field and multi-scale learning with only efficient operations.



▶ Figure 2: EfficientViT provides significant performance boosts compared with prior state-of-the-art semantic segmentation models.

FURTHER READING

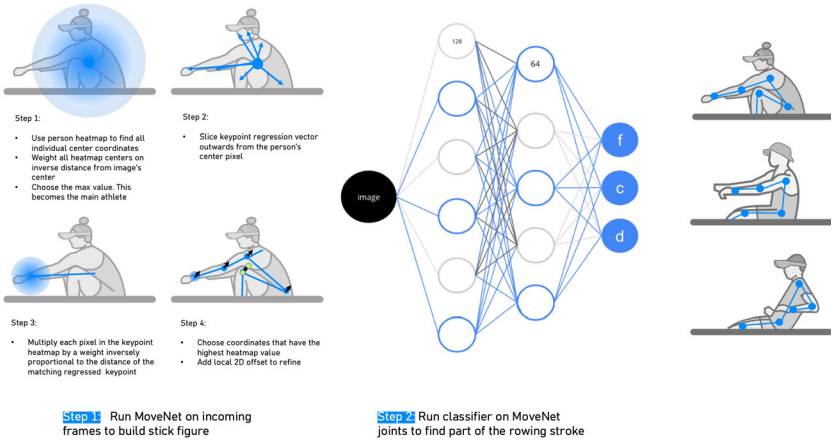
- Cai, Han, Li, Junyan, Hu, Muyan, Gan, Chuang, Han, Song, " EfficientViT: Lightweight Multi-scale Attention for On-device Semantic Segmentation," *arXiv preprint arXiv:2205.14756* (2022).

AI-Row: A Real-time Mobile ML Analytics Application for Para- and Non-para Rowers

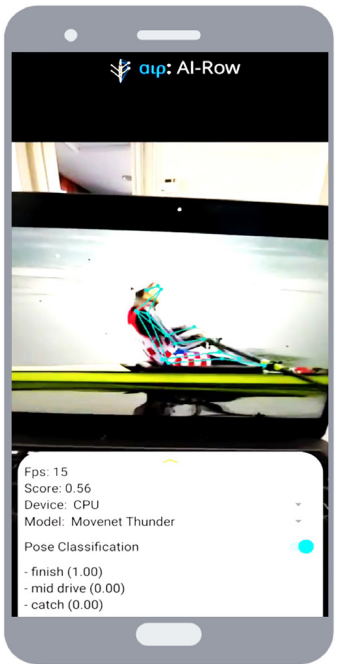
E. Eldracher, V. Muriga, S. Rodríguez, Y. Lin, Z. Liu, S. Han

Lightweight machine learning (ML) models can learn how to deliver elite rowing coaching to anyone, regardless of that individual's skill or physical ability. Despite the sport's predictability and angle-based technique, few tools exist to deliver data analytics on performance. Those that do are often expensive, not frequently optimized for both on-the-water and ergometer video, and trained without para inclusivity. This work harnesses the power of artificial intelligence to deliver technical insights in real time on a mobile device. This mobile application is the first use of mobile pose estimation ML for para-optimized rowing. It works for all athletes: those who use only their arms, those who use their arms and upper body, and those who use their legs, body, and arms.

By utilizing TensorFlow's MoveNet for mobile two-dimensional (2D) pose estimation combined with a simple classifier, we categorized three specific rowing poses both on the water and on the rowing machine. After locating an athlete's joints through MoveNet, our 26kb classifier predicts what part of the stroke a rower is in. At 18 frames/second, this classifier achieves around 89% accuracy. Because we built our own dataset of images (935 training, 235 testing, YouTube and USRowing images), this model is diverse and inclusive of both para and non-para athletes.



▲ Figure 1: Rendering of how our model predicts parts of the rowing stroke. Step 1 demonstrates MoveNet, Step 2 details our classifier, and the right-most side shows the three potential rowing positions categorized.



▲ Figure 2: Real-time screen recording of AI-Row. An android device correctly classifies a rower in the layback finish position.

FURTHER READING

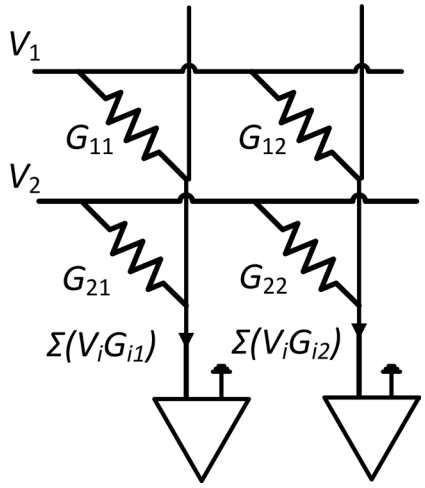
- R. Votel and N. Li, "Next-Generation Pose Detection with MoveNet and TensorFlow.js," *https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html* (May 2021).
- Y. Wang, M. Li, H. Cai, W. Chen, and S. Han, "Lite pose: Efficient Architecture Design for 2D Human Pose Estimation," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13126-13136, 2022.

ADCs for Analog Neural Nets

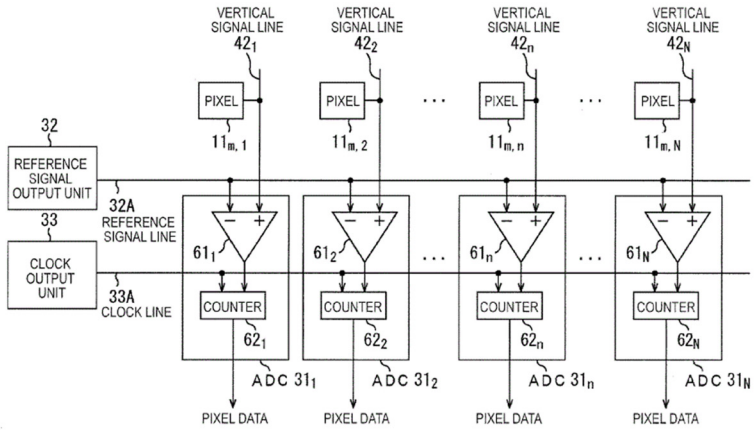
M. A. G. Elsheikh, H.-S. Lee
Sponsorship: MIT/MTL Samsung Semiconductor Research Fund

The meteoric rise of neural networks in recent years has been fueled by applications in several domains such as image recognition, self-driving cars, signal processing, and drug discovery. For more widespread deployment in portable applications, speed and power consumption must be improved. Employing specialized accelerator architectures offers improvements on both fronts. One accelerator topology, analog neural networks (ANNs), shown in Figure 1, perform matrix-vector-multiplications (MVMs) in the analog domain by encoding the inputs as the voltages and the weights as conductances. Summing up the currents in a common, virtual ground node is equivalent to performing the MVM process in one cycle, thereby potentially saving energy and time.

An analog-to-digital converter (ADC) architecture that is proposed for this application is the single slope ADC (SS-ADC), shown in Figure 2. The SS-ADC has become the standard architecture for complementary metal-oxide semiconductor (CMOS) image sensors as it is suitable for a medium number of bits, it is highly reconfigurable, and the peripheral circuits can fit the column pitch of the sensor elements. The column-parallel SS-ADC consists of a comparator and a counter for each column and a central ramp generator. At the beginning of the quantization process, the ramp and the counter are reset. The ramp starts increasing as the counter starts incrementing, and each column comparator compares between its analog column voltage and the ramp voltage. When the ramp value exceeds the column value, the comparator trips, and the counter value is held. This value is the digital representation of the column signal. In this research several innovations can be introduced to the SS-ADC to tailor it for ANNs to improve them beyond the state of the art in terms of speed and power consumption.



▲ Figure 1: Analog neural network accelerator.



▲ Figure 2: The column parallel architecture SS-ADC.

FURTHER READING

- Y. N. Wu, V. Sze and J. S. Emer, "An Architecture-Level Energy and Area Estimator for Processing-In-Memory Accelerator Designs," *2020 IEEE International Symposium on Performance Analysis of Systems and Software* (ISPASS), pp. 116-118, 2020.

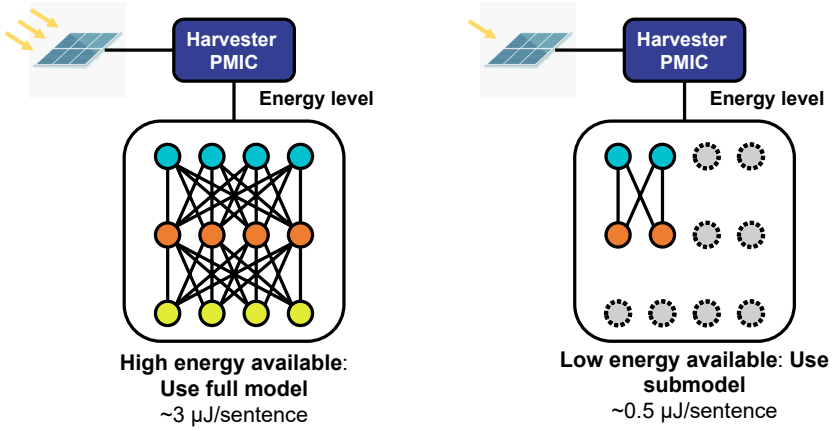
A Fully-integrated Energy-scalable Transformer Accelerator Supporting Adaptive Model Configuration and Word Elimination for Language Understanding on Edge Devices

A. Ji, H. Wang, M. Wang, S. Han, A. P. Chandrakasan
Sponsorship: TSMC

Efficient natural language processing (NLP) on the edge is needed to interpret voice commands, which have become an increasingly common way to interact with devices around us. Attention-based transformer models have replaced recurrent neural networks as the predominant model for NLP applications due to parallel input processing and the attention mechanism being able to capture both short and long-range relations. However, existing mainstream models (e.g., BERT, GPT) are way too large for edge devices. For simple NLP tasks on the edge, tiny custom transformer models can achieve good accuracy while being much more suitable for constrained hardware.

There are two main challenges when deploying lightweight NLP models on edge devices. Firstly, hardware constraints can fluctuate based on battery level, latency requirements, availability of compute resources, and accuracy tolerance. Adapting to these conditions typically requires multiple models of different sizes. For instance, when the device is less constrained, we may use a large model while under more constrained conditions, we may opt for a small model. But storing multiple models incurs a significant memory overhead. Secondly, sentences usually contain redundant words that contribute little to the overall understanding and may potentially be skipped during the majority of the processing. Conventional models spend an equal amount of time processing each word, leading to unnecessary computation.

Our work addresses these challenges with an energy-scalable transformer accelerator targeting small Internet of Things devices with two key features: 1) adaptive model configuration using a custom SuperTransformer model to generate models of various sizes while taking up only the memory footprint of a single full model; and 2) a comparator-based word elimination unit to progressively remove unimportant words from the sentence, reducing computation. We achieve 5.8× scalability in the network energy and latency. Word elimination can reduce network energy by 16% with some accuracy loss.



▲ Figure 1: Adaptive model configuration based on energy level of energy harvester using a single SuperTransformer model.

FURTHER READING

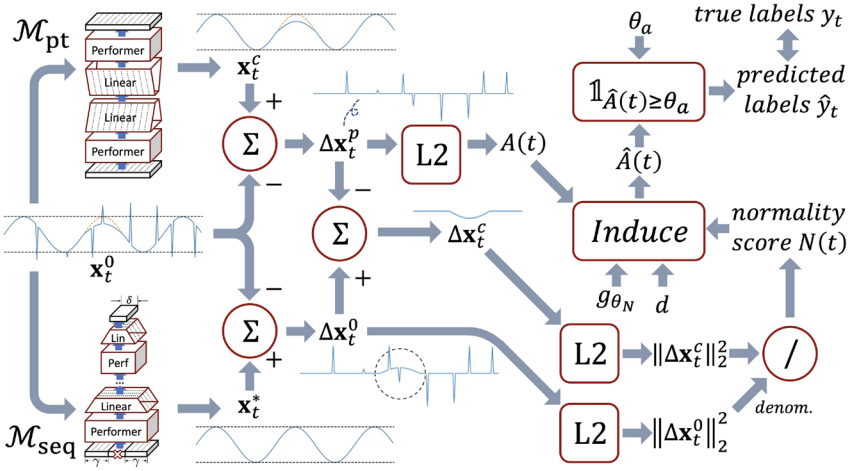
- H. Wang, Z. Zhang, and S. Han, "SpAtten: Efficient Sparse Architecture with Cascade Token Pruning and Head Pruning," *HPCA*, 2021.
- H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, and S. Han, "HAT: Hardware-Aware Transformers for Efficient Natural Language Processing," *ACL*, 2020.

Unsupervised Time Series Anomaly Detection via Point/Sequential Reconstruction

C.-Y. Lai, F.-K. Sun, Z. Gao, J. H. Lang, D. S. Boning
Sponsorship: Turntide Technologies

Fortune Global 500 manufacturing and industrial firms lose around 3.3 million hours of production time due to machine failure, resulting in an economic impact of $864 billion of their annual revenue. By identifying anomalies in time series data for manufacturing and operations, one can reduce downtime and prevent financial setbacks. However, time series anomaly detection is challenging due to the complexity and variety of patterns that can occur. One major difficulty arises from modeling time-dependent relationships to find contextual anomalies while maintaining detection accuracy for point anomalies.

In this work, we propose a novel framework--normality score conditioned time series anomaly detection by point/sequential reconstruction (NPSR)--for unsupervised time series anomaly detection that utilizes point-based and sequence-based reconstruction models. The general scheme is shown in Figure 1. The point-based model attempts to quantify point anomalies, and the sequence-based model attempts to quantify both point and contextual anomalies. Under the formulation that the observed time point is a two-stage deviated value from a nominal time point, we introduce a normality score calculated from the ratio of a combined value of the reconstruction errors. We derive an induced anomaly score by further integrating the normality score and anomaly score, and then theoretically prove the superiority of the induced anomaly score over the original anomaly score under certain conditions. Extensive studies conducted on several public datasets show that the proposed framework outperforms most state-of-the-art baselines for time series anomaly detection. It also possesses the potential to decrease labor needs for fault monitoring and correspondingly accelerate decision making and can contribute to artificial intelligence (AI) sustainability by preventing energy waste or system failure.



▲ Figure 1: General scheme for NPSR.

FURTHER READING

- A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–33, 2021.
- K. Choi, J. Yi, C. Park, and S. Yoon, "Deep Learning for Anomaly Detection in Time-series Data," *Review, Analysis, and Guidelines. IEEE Access*, vol. 9, pp. 120043–120065, 2021.

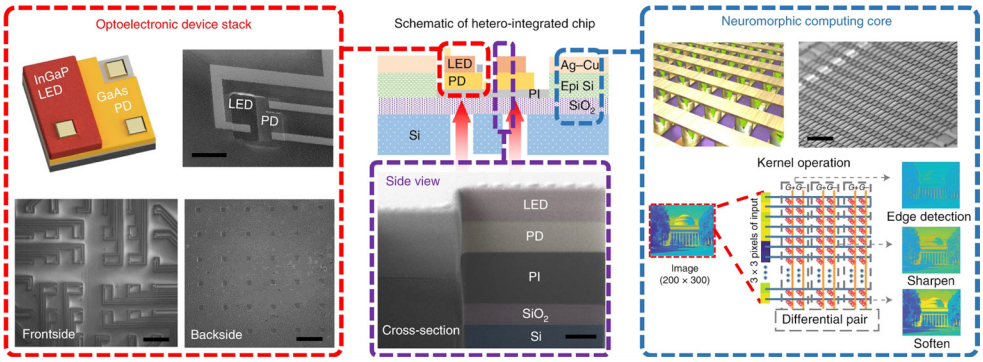
Lego-like Reconfigurable Sensor Computing System

G. Lee, C. Choi, H. Kim, J.-H. Kang, M.-K. Song, H. Leon, J. Kim
Sponsorship: Ministry of Trade, Industry, and Energy, Korea Institute of Science and Technology, Samsung Global Research Outreach Program

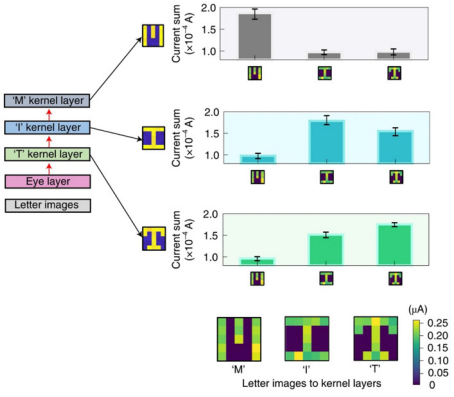
The emergence of artificial intelligence applications has transformed computer design, leading to a quest for hardware architecture that can process large volumes of data with high power, area, and time efficiency. To improve data communication bandwidth among sensors, memory, and processors, three-dimensional (3D) heterogeneous integration combined with advanced packaging technologies is a promising solution. However, these systems have limitations such as a lack of hardware reconfigurability and reliance on conventional von Neumann architecture.

In this project, we address these issues by introducing stackable heater-integrated chips that employ optoelectronic device arrays for inter-chip communication and neuromorphic cores built with memristor crossbar arrays for parallel data processing (Figure 1). With these stackable and replaceable chips, we created a system that can directly classify information from a light-based image source. First, we showed that three different preprogrammed neuromorphic core layers can be stacked and share the light inputs, illustrating the robustness of light-signal-based communication (Figure 2). Further, we showed that an additional noise reduction layer inserted after the sensor layer successfully improves the letter recognition performance in a noisy environment (Figure 3).

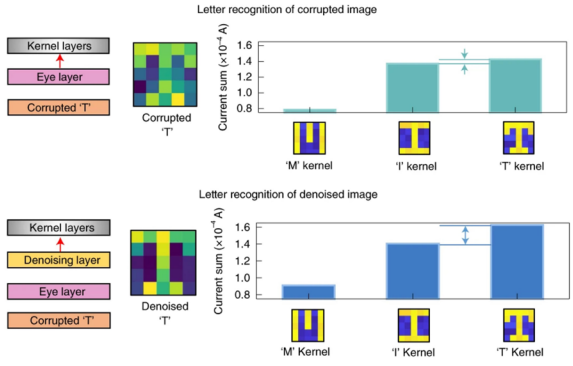
This project provides a reconfigurable 3D hetero-integrated platform that enables vertical staking of various functional layers. This could provide an energy-efficient data communication and processing solution to sensor computing or edge computing applications.



◀ Figure 1: Stackable hetero-integrated neuromorphic chips.



▲ Figure 2: Robust kernel operation of stackable heater-integrated neuromorphic chips.



▲ Figure 3: Noise reduction using additional layer.

FURTHER READING

- C. Choi, H. Kim, J.-H. Kang, M.-K. Song, H. Yeon, C. S. Chang, J. Suh, J. Shin, et al., "Reconfigurable Heterogeneous Integration Using Stackable Chips withEmbedded Artificial Intelligence," *Nat. Electron.* vol. 5, no. 6, pp. 386-393, Jun. 2022.
- H. Leon, P. Lin, C. Choi, S. H. Tan, Y. Park, D. Lee, J. Lee, F. Xu, et al., "Alloying Conducting Channels for Reliable Neuromorphic Computing," *Nat. Nanotechnol.*, vol. 15, no. 6, pp. 574-579, Jun. 2020.

LEGO: Spatial Accelerator Generation and Optimization for Tensor Applications

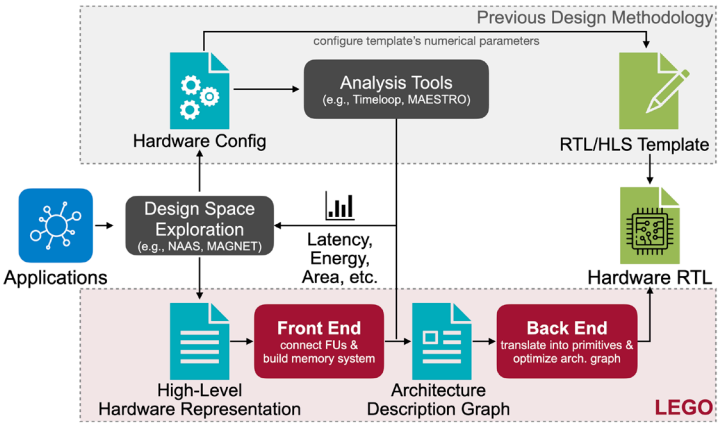
Y. Lin, Z. Zhang, S. Han
Sponsorship: SRC

The proliferation of tensor applications, such as deep neural networks, has led to an unprecedented demand for efficient and high-performing solutions. Particularly, modern foundation models and generative artificial intelligence (AI) applications require multiple input modalities (both vision and language), which increases the demand for flexible accelerator architecture. Existing frameworks suffer from the trade-off between design flexibility and productivity of register transfer language (RTL) generation: either limited to very few hand-written templates or unable to automatically generate the RTL.

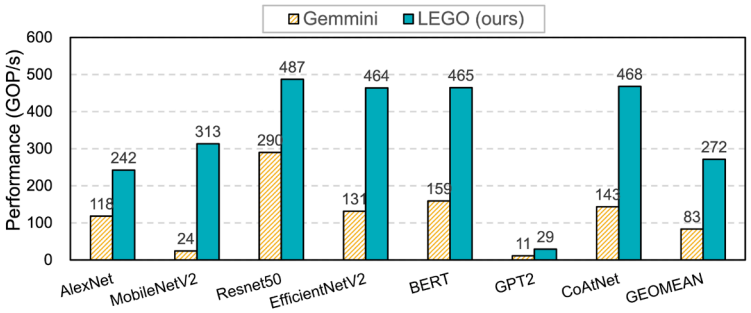
To address this challenge, we propose the LEGO framework, which automatically generates and optimizes spatial architecture design in the front end and outputs synthesizable RTL code in the back end without RTL templates. LEGO front end finds all possible interconnections between function units and determines the memory system shape by solving the integer linear equations and establishes the connections by a minimum-spanning-tree-based algorithm and a breadth-first-search-based heuristic algorithm for merging different spatial dataflow designs.

LEGO back end then translates the hardware in a primitive-level graph to perform lower-level optimizations and applies a set of linear-programming algorithms to optimally insert pipeline registers and reduce the overhead of unused logic when switching spatial dataflows.

Our evaluation demonstrates that LEGO can achieve 3.3× speedup and 2.1× energy efficiency compared to previous work by Gemmini and can generate one architecture for diverse modern foundation models in generative AI applications.



◀ Figure 1: Instead of configuring the sizing parameters in the hardware template, LEGO directly generates spatial architecture design and outputs RTL code from high-level hardware description.



▶ Figure 2: Performance comparison of Gemmini and LEGO. LEGO achieved an average 3.3× speedup over Gemmini. Both Gemmini and LEGO are bounded by memory bandwidth on GPT2. LEGO performs much better on MobileNetV2 due to its efficient support of depthwise convolution by dataflow switching.

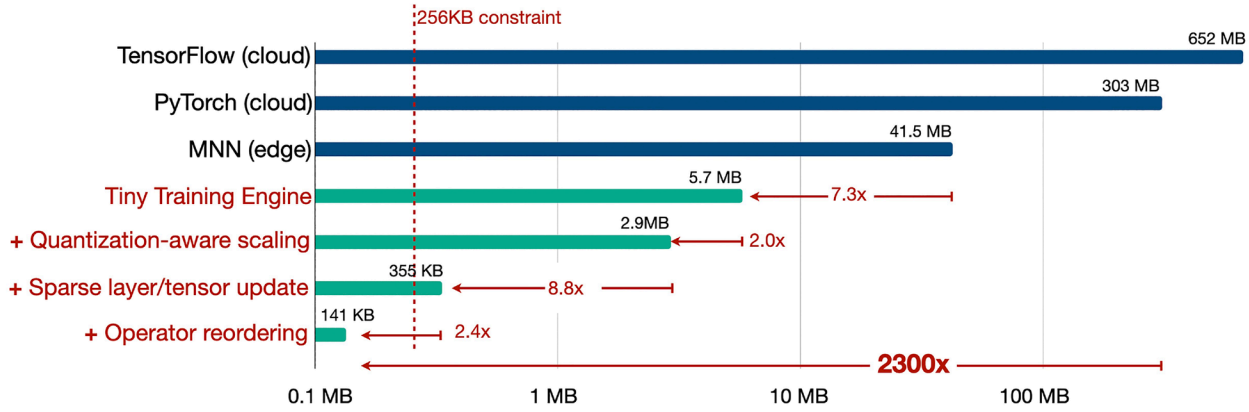
FURTHER READING

- Y. Lin, Z. Zhang, and S. Han, "LEGO: Spatial Accelerator Generation and Optimization for Tensor Applications," *https://naas.mit.edu.*
- H. Genc, S. Kim, A. Amid., A. Haj-Ali, V. Iyer, P. Prakash, J. Zhao, D. Grubb, H. Liew, H. Mao and A. Ou, "Gemmini: Enabling Systematic Deep-learning Architecture Evaluation via Full-stack Integration," *58th ACM/IEEE Design Automation Conference (DAC)* (pp. 769-774). IEEE, Dec. 2021.

On-device Training Under 256KB Memory

J. Lin, L. Zhu, W. M. Chen, W. C. Wang, C. Gan, S. Han
Sponsorship: NSF, MIT-IBM Watson AI Lab, MIT AI Hardware Program, Amazon, Intel, Qualcomm, Ford, Google

On-device training enables the model to adapt to new data collected from the sensors by fine-tuning a pre-trained model. Users can benefit from customized artificial intelligence (AI) models without having to transfer the data to the cloud, protecting privacy. However, the training memory consumption is prohibitive for Internet of Things (IoT) devices that have tiny memory resources. We propose an algorithm-system co-design framework to make on-device training possible with only 256KB of memory. On-device training faces two unique challenges: (1) the quantized graphs of neural networks are hard to optimize due to low bit-precision and the lack of normalization, and (2) the limited hardware resource does not allow full back-propagation. To cope with the optimization difficulty, we propose quantization-aware scaling to calibrate the gradient scales and stabilize 8-bit quantized training. To reduce the memory footprint, we propose sparse update to skip the gradient computation of less important layers and sub-tensors. The algorithm innovation is implemented by a lightweight training system, Tiny Training Engine, which prunes the backward computation graph to support sparse updates and offload the runtime auto-differentiation to compile time. Our framework is the first solution to enable tiny on-device training of convolutional neural networks under 256KB static random-access memory (SRAM) and 1MB Flash without auxiliary memory, using less than 1/1000 of the memory of PyTorch and TensorFlow while matching the accuracy on tinyML application VWW. Our study enables IoT devices not only to perform inference but also to continuously adapt to new data for on-device lifelong learning.



▲ Figure 1: Algorithm and system co-design reduces the training memory from 303MB (PyTorch) to 141KB with the same transfer learning accuracy, leading to 2300× reduction.

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*, pp.11711-11722, 2020.
- J. Lin, W. M. Chen, H. Cai, C. Gan, and S. Han, "MCUNetV2: Memory-Efficient Patch-Based Inference for Tiny Deep Learning," *Advances in Neural Information Processing Systems*, pp.2346-2358, 2021.

Efficient Camera-radar Fusion for 3D Perception

Z. Liu, H. Tang, K. Shao, X. Chen, S. Han
Sponsors: National Science Foundation, Hyundai Motor, Qualcomm, NVIDIA, Apple

The development of three-dimensional (3D) perception systems is crucial to the widespread adoption of autonomous driving. However, many studies tend to overlook practical considerations, resulting in relatively little focus on the efficient and sensible deployment of 3D perception models in the real world. A successful practical deployment requires consideration of several factors simultaneously, such as accuracy, speed, cost, and deployability. Given the high cost of light detection and ranging (LiDAR) sensors compared to cameras or radars, we propose an efficient camera-radar fusion approach for 3D perception.

Various studies have explored the question of how to fuse information from different modalities (see Figure 1). In our approach, we propose performing the fusion in bird's eye view (BEV) space, as it retains semantic and spatial information from each modality. We apply a modality-specific encoder to each input, followed by the BEV projection and a two-dimensional (2D) decoder. To further improve the model, we designed a novel view transformer module that is responsible for transforming image features from the camera view to 3D space. By fusing radar points onto the image plane, our model is capable of more accurate depth estimation, leading to better spatial alignment and improved performance. Additionally, we modify previous architectures by ensuring that all operations in our model are capable of hardware acceleration.

Our architecture design results in a camera-radar fusion model that improves on the previous state-of-the-art single-frame models. Our model achieves 52.6% NDS on the nuScenes detection dataset. Moreover, our model can leverage TensorRT acceleration to achieve a much greater speedup than competing methods (see Figure 2). Ultimately, our model achieves real-time latencies on NVIDIA Jetson AGX Orin.

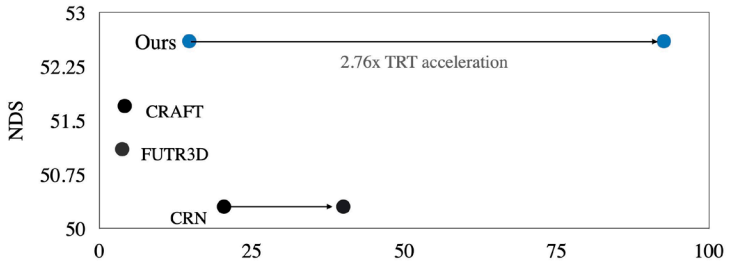


Figure 1: Visualization of radar points projected onto corresponding image. A principal challenge of utilizing radar data is its sparsity, as plainly seen here. However, radar returns also carry semantically useful information, such as radar cross section and radial velocity, which are helpful for 3D object detection models.

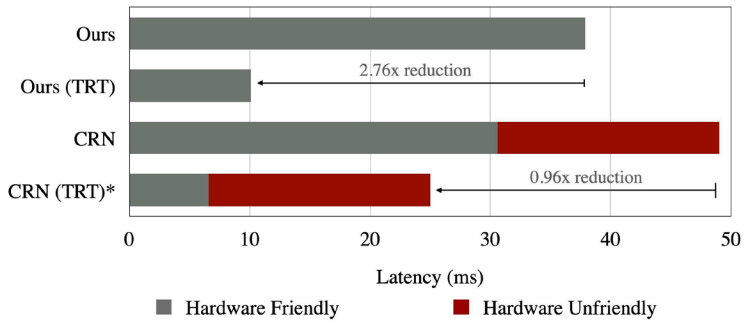


Figure 2: Comparison between our models and previous camera radar fusion methods. Due to our design, our model can better leverage hardware acceleration to achieve stronger accuracy at faster speeds than CRN, the currently leading approach.

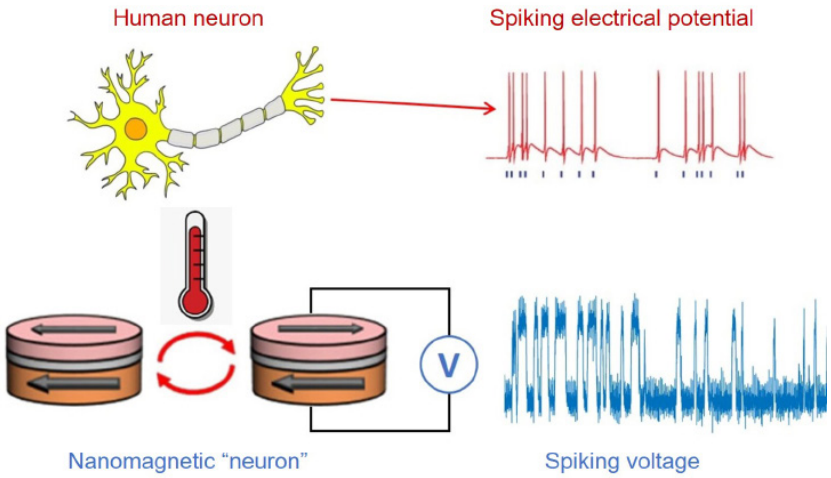
FURTHER READING

- Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," to be presented at *International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, May 2023.

Neuromorphic Computing with Probabilistic Nanomagnets

B. C. McGoldrick, M. A. Baldo, L. Liu
Sponsorship: EECS Mathworks Fellowship

The human brain is capable of performing complex tasks such as object recognition and inference, all while operating at low power. In contrast, performing the same tasks on digital computers requires significant energy, time, and hardware. While the brain exhibits random, noisy behavior, digital computers are designed oppositely to be deterministic and low-noise. By emulating the brain's probabilistic nature in hardware, we can potentially achieve superior speed and energy-efficiency on solving the aforementioned problems. We develop a probabilistic bit (p-bit) based on a nanomagnetic device that produces a random bipolar voltage signal driven by ambient thermal noise. We can control the p-bit's probability and fluctuation rate by applying magnetic fields or charge currents. Finally, we elucidate a plan to integrate our tunable p-bits with traditional CMOS circuits to realize a probabilistic inference system with potentially greater speed and energy efficiency than existing approaches.



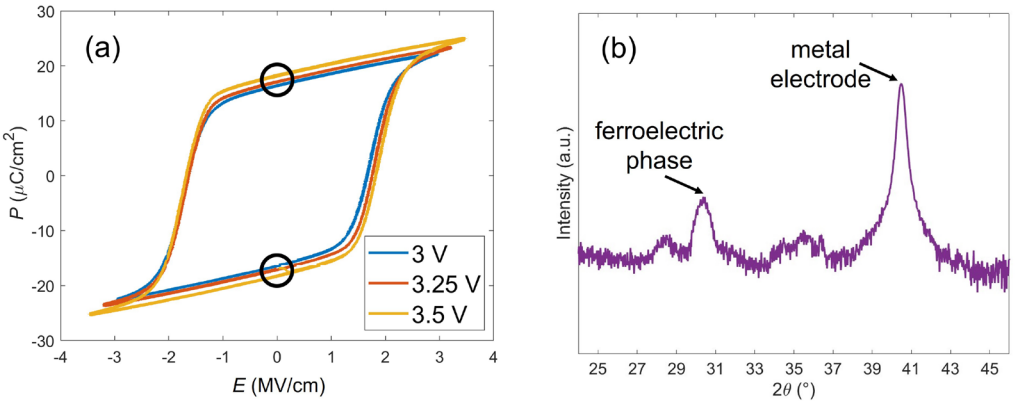
▲ Figure 1: Nanomagnetic "neuron" that converts ambient thermal fluctuations at room temperature to a random, fluctuating bipolar voltage signal. This behavior can be likened to the random spiking behavior of a neuron in the human brain.

CMOS-Compatible Ferroelectric Materials and Structures

E. Rafie Borujeny, Y. Shao, T. Kim, J. A. del Alamo
Sponsorship: MIT Quest for Intelligence, SRC

Ferroelectrics are a class of materials that exhibit a nonlinear relationship between the externally applied electric field (E) and the electric polarization (P) formed inside them. In addition, they show a spontaneous non-zero polarization even when no external E is applied (Figure 1a). Moreover, the value of P in ferroelectric materials depends not only on the value of E but also on its history (for example, note in Figure 1 that at E=0, P can have two values depending on whether we reach E=0 from E>0 or from E<0). Having a spontaneous and history-dependent polarization means that ferroelectric materials can be incorporated into electronic devices to act as non-volatile memory elements.

Our research focuses on high-quality complementary metal-oxide semiconductor- (CMOS) compatible ferroelectrics that can be incorporated into nanometer-scale electronic devices. We specifically focus on developing ferroelectric structures based on hafnium zirconium oxide (HZO) thin (~10 nm) films fabricated at low temperature (at or below 400˚C). We investigate how process variations influence the composition, structure and electrical behavior of these films. We also investigate how the presence of other materials in contact with these ferroelectrics, as it occurs in real-world electronic devices, influences their performance. As a result of these investigations, we have successfully developed high-quality HZO films that contain the desired ferroelectric crystalline phase responsible for the ferroelectric properties (Figure 1b) and show clear and symmetric polarization-voltage characteristics even when the fabrication process temperature is limited to 400˚C (Figure 1a). These investigations pave the way for the incorporation of ferroelectric materials into standard CMOS technology to enhance the functionality and performance of future electronics.



▲ Figure 1: (a) Relationship between P and applied E in a ferroelectric HZO film annealed at 400°C- -note the hollow circles that show two values of P at E=0; (b) X-ray diffraction pattern of the corresponding film depicting the presence of the desired orthorhombic ferroelectric structure inside the film.

FURTHER READING

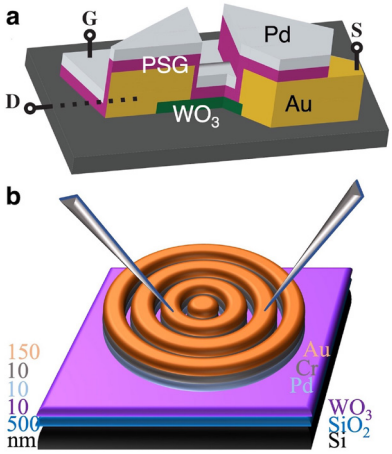
- T. Kim, J. A. del Alamo, and D. A. Antoniadis, "Switching Dynamics in Metal–Ferroelectric $HfZrO_2$–Metal Structures," *IEEE Trans. Electron Devices*, vol. 69, no. 7, pp. 4016–4021, Jul. 2022.
- T. Kim, J. A. del Alamo, and D. A. Antoniadis, "Dynamics of HfZrO2 Ferroelectric Structures: Experiments and Models," *2020 IEEE International Electron Devices Meeting* (IEDM), vol. 21, no. 4, pp. 1-4, 2020.

Circular TLM Characteristics of WO_3 for Protonic Programmable Resistors

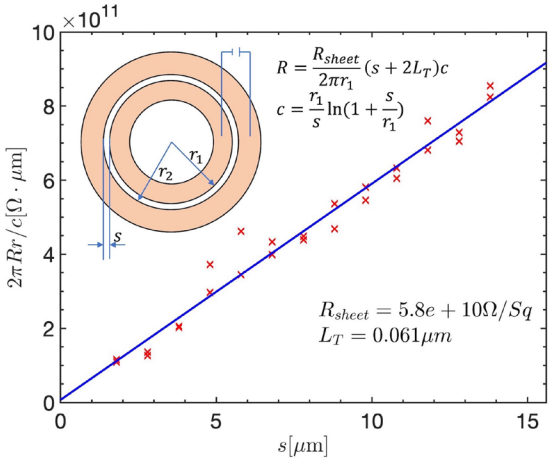
D. Shen, J. A. del Alamo
Sponsorship: MIT-IBM Watson AI Lab

Analog computing offers a potential solution for overcoming computational bottlenecks in traditional digital systems utilized for deep learning. The fundamental concept of analog deep learning accelerators involves processing information locally by leveraging the physical properties of devices, rather than conventional Boolean arithmetic—specifically, using Ohm's and Kirchhoff's laws for matrix inner product calculations and threshold-based updating for the outer product. Among various physical principles, electrochemical ion-intercalation makes possible a three-terminal device with a channel resistance that is modulated by ionic exchange between the channel and a gate reservoir via an electrolyte. This study focuses on such ionic programmable resistors featuring WO_3 as the channel and protons as the ions, aiming to provide information processing with increased energy savings, efficiency, non-volatility, and low latency. Our group's previous work, with a device structure shown in Figure 1a, has demonstrated silicon-compatible nanoscale devices that are 1,000x smaller than biological neurons, enabling channel conductance modulation of an over 20x range with nanosecond operation at room temperature.

To further examine the properties of WO_3 as the channel material and efficiently understand the impact of different fabrication process conditions, we have developed a straightforward and effective test structure, depicted in Figure 1b. The test structure employs the conventional circular transfer length method (TLM) to measure sheet resistance and contact resistance via linear fitting of resistance data collected from a series of devices, as shown in Figure 2. By obtaining resistance information before, during, or after protonation, we gleaned valuable WO_3 characteristics—such as a 104x conductance modulation range, low proton diffusion coefficient, and high resistance recovery ability under heating. Additionally, we will compare different protonation methods such as metal diffusion method with Pd and hydrogen spillover method with HCl. These insights help us optimize the fabrication process for improved programmable resistors.



▲Figure 1: WO_3 protonic device structures: (a) Three-dimensional illustration of the device studied in our previous work, WO_3 as channel, nano-porous phosphosilicate glass (PSG) as the electrolyte, and Pd as hydrogen reservoir and controlling gate. (b) Circular TLM structure for rapid test, Pd absorbing hydrogen and shuttling protons.



▲Figure 2: Typical circular TLM fitting demonstration. The inset shows the parameters of concentric metal rings and the near-linear relation between cross-ring resistance and gap-spacing. Main figure shows results for WO_3 deposited with 90W RF sputtering and 400℃ annealing, with r_1 = 100 μm and s ranging from 1.8 μm to 13.8 μm.

FURTHER READING

- O. Murat, N. Emond, B. Wang, D. Zhang, F. M. Ross, J. Li, B. Yildiz, and J. A. del Alamo, "Nanosecond Protonic Programmable Resistors for Analog Deep Learning." *Science*, vol. 377, no. 6605, pp. 539-543, Jul. 2022.
- J. H. Klootwijk and C. E. Timmering, "Merits and Limitations of Circular TLM Structures for Contact Resistance Determination for Novel III-V HBTs," *2004 International Conference on Microelectronic Test Structures* (IEEE Cat. No. 04CH37516), pp. 247-252, 2004.

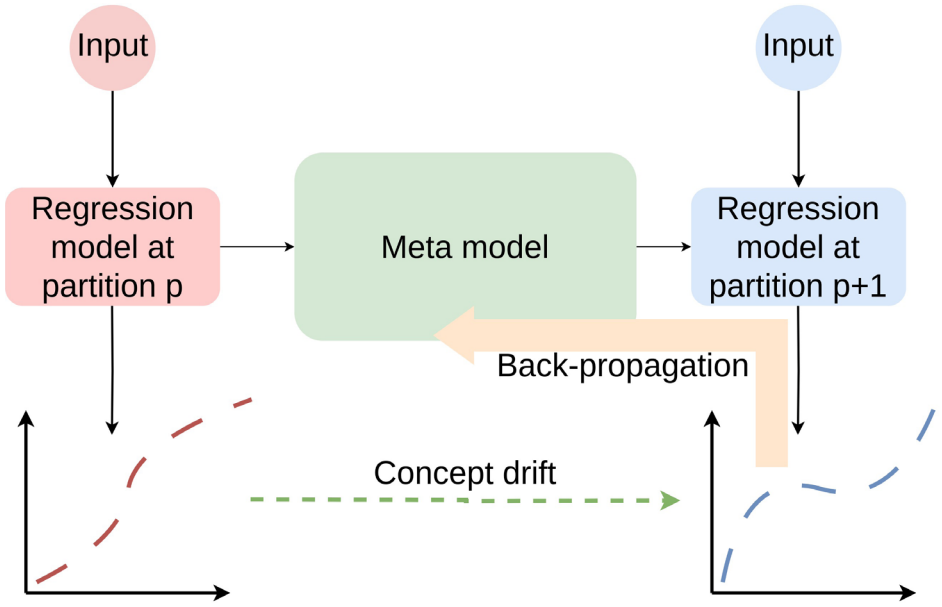
Training Meta Neural Networks for Concept Drift Adaptation in Time Series

F.-K. Sun, D. S. Boning
Sponsorship: Lam Research

Time series analysis and modeling play crucial roles in various applications such as forecasting and anomaly detection. However, one common challenge is the occurrence of concept drift, where the dynamics of the underlying system change over time due to factors like wear, tear, and environmental variations. Dealing with concept drift typically involves retraining the model whenever new data points are observed, but this process can be time-consuming and computationally intensive.

To better understand the impact of concept drift, we start by synthesizing a multivariate linear dataset with linear drift and training a regression model using it. We divide the dataset into 10 sets, training the regression model up to the p-th set and evaluating its performance on the subsequent (p+1)-th set. As expected, we observe a degradation in the model's performance over time.

To address this issue, we propose a solution in the form of a "meta model" designed to learn and predict the drift dynamics of the regression models. The underlying assumption is that the drift dynamics are predictable. The "meta model" takes the p-th regression model as input and predicts the (p+1)-th regression model. Notably, we have also developed a technique enabling end-to-end training of the meta model. Consequently, even if the regression model is trained on data only up to the p-th set, we can utilize the meta model to predict the (p+1)-th model and evaluate it on the corresponding (p+1)-th set of data. This approach is valuable in real-world scenarios where we want to assess new batches of data but possess only an "outdated" model. Our experimental results demonstrate that the meta model effectively learns the drift dynamics, resulting in a performance degradation reduction ranging from 2x to 10x compared to not adapting to the drift.



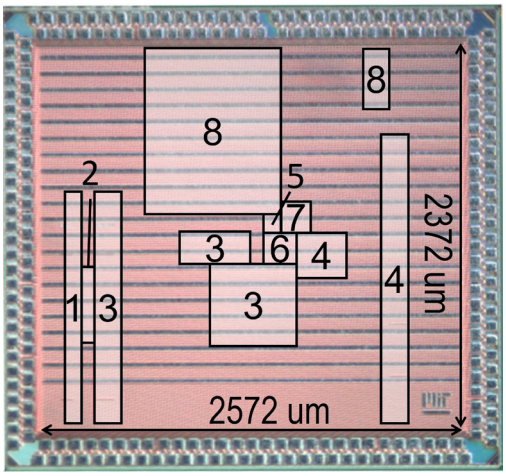
▲ Figure 1: The algorithm flow of training the meta model. Notably, the gradient back-propagation is directly propagated from the output to the meta model, enabling an end-to-end training approach.

Algorithm and Hardware Co-design for Efficient Video Understanding on the Edge

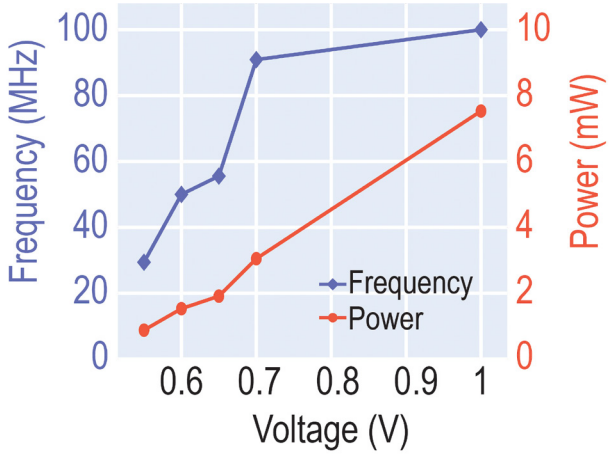
M. Wang, Y. Lin, Z. Zhang, J. Lin, S. Han, A. P. Chandrakasan
Sponsorship: Qualcomm Incorporated, TSMC University Shuttle Plan

With the rise of various applications including augmented reality/virtual reality, autonomous driving, object tracking for unmanned aerial vehicles, etc., there is an increasing need for accurate and energy-efficient video understanding on the edge. Although many deep learning chips are designed for images, little work has been done for videos. Video understanding on the edge has three major challenges. First, video understanding requires temporal modeling. For example, it identifies the difference between opening and closing a box, which is distinguishable only with temporal information considered. Second, many applications are delay-critical, such as self-driving cars and artificial intelligence drones. Third, high energy efficiency is important for edge devices with a tight power budget. Due to temporal continuity, consecutive frames might share a lot of common information, providing the potential to improve processing efficiency. However, an image-based processing system cannot utilize that since each frame is processed individually.

In this project, we co-design algorithms and hardware for energy-efficient video processing for delay-critical applications. We propose a real-time DiffFrame convolution achieving 2.2x dynamic random-access memory (DRAM) access reduction compared to conventional convolution at single-frame latency, design a sorter-free architecture for efficient utilization of temporal similarities between video frames, enable temporal modeling capability achieving high accuracy on video understanding applications, and optimize data buffering to remove DRAM traffic overhead for temporal modeling and reduce 55%--79% input activation DRAM traffic in depth-wise convolution layers. The chip consumes 40uJ/frame with 38 frames/second at 0.6V in 28nm TSMC 28-nm complementary metal-oxide-semiconductor (CMOS) process. Figure 1 shows the chip photograph; Figure 2 presents the frequency and power measurement results.Our demonstration of ferroelectricity in stacking-engineered TMD bilayers consolidates the feasibility of engineering 2D ferroelectric semiconductors and opens up a broad way of engineering various functional heterostructures out of non-ferroelectrics.



▲ Figure 1: Die micrograph (1: 16kB weight buffer; 2: 8x8 multiplier-and-accumulator array; 3: 32kB input activation buffer; 4: 44kB output activation and RefFrame unit; 5: DiffFrame generator; 6: DiffFrame pruning; 7: ConvMap buffer; 8: convolution map generator & coordinate buffer).

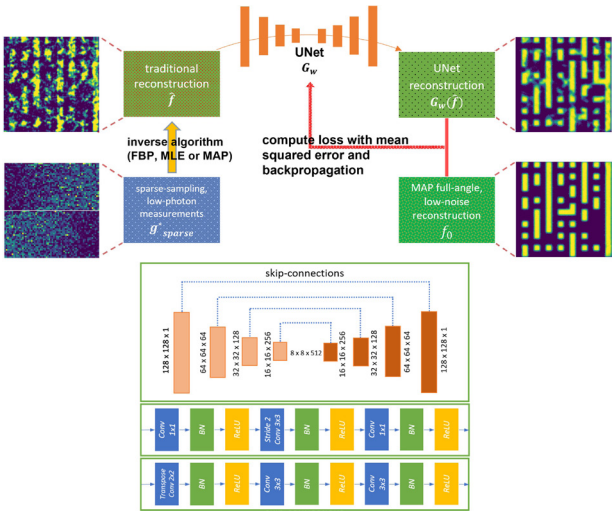


▲ Figure 2: Frequency and power measurements.

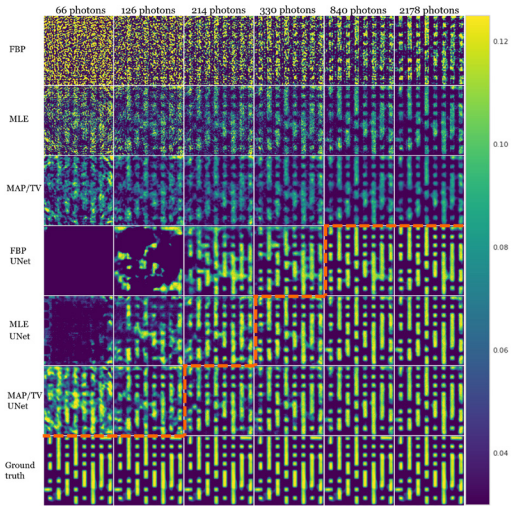
Noise Resilience Deep Reconstruction for X-ray Tomography

Z. Guo, Z. Liu, Q. Zhang, G. Barbastathis
Sponsorship: Singapore's National Research Foundation

X-ray tomography is a non-destructive imaging technique that visualizes the interior features of solid objects, with applications in biomedical imaging, materials science, manufacturing inspection, and other disciplines. Under limited-angle and low-photon sampling, a regularization prior is required to retrieve a high-fidelity reconstruction. Recently, deep learning has been used in X-ray tomography. The prior learned from training data replaces the general-purpose priors in iterative algorithms, achieving high-quality reconstructions with a neural network. Previous studies typically assume the noise statistics of testing data is acquired a priori from training data, leaving the network susceptible to a change in the noise characteristics under practical imaging conditions. In this work, we propose a noise-resilient deep-reconstruction algorithm for X-ray tomography. Our approach improves the noise resilience of the learned prior by using noise-resilient maximum a posteriori (MAP) reconstructions as the input to the neural network. Unlike previous efforts, we focus on the generalization of the deep learning algorithms to test data with different noise levels than the training data, which is critical in practical applications. Without training samples from different photon statistics, the MAP+UNet approach can produce acceptable reconstruction down to 50 photons per ray in simulations and 214 per ray in experiments, whereas the filtered back projection (FBP)+UNet approach requires around 10x more photons per ray in simulations and 2.5x more in experiments.



▲Figure 1: A conceptual diagram for the learning-based algorithms.



▲Figure 2: Selected 2D reconstruction for algorithms using experimental data. Each row represents a reconstruction algorithm. Each column represents an intensity of the photon rays. Dotted orange line is the boundary between acceptable and unacceptable performance as determined by the MST metric.

FURTHER READING

- Z. Guo, Z. Liu, G. Barbastathis, Q. Zhang, M. E. Glinsky, B. K. Alpert, and Z. H. Levine, "Noise-resilient Deep Learning for Integrated Circuit Tomography," *Optics Express*, vol. 31, no. 10, pp. 15355-15371, 2023.
- Z. Guo, "Noise Resilience Deep Reconstruction for X ray Tomography," (Version 1.0.0) [Computer software]. https://github.com/zguo0525/Noise-resilience-deep-reconstruction-for-X-ray-Tomography (2022).