Edge Intelligence

Montréal-Québec Workshop 2022

September 19-20, 2022

Book of Abstracts

PLATINUM SPONSOR

HUAWEI

Editorial

There have been a few impactful technological revolutions that impacted human life to a great extent. Such inventions include the wheel, steam engine, combustion engine, electricity, and the computer. The early inventions transferred work from animals to machines while computing technology and artificial intelligence (AI) aim to off-load decisions from humans to machines. It is not surprising to see that the global workforce needs to undergo the next industrial revolution to make this transition and AI is the key to this major shift. While deep learning is taking over almost all predictive models nowadays, deep models require large computational resources for training and inference.

To make artificial intelligence accessible widely, we need to focus on adapting these models for low-resource devices, especially edge devices. In this workshop we will focus on issues related to efficient computation of artificial intelligence on resource-constrained devices.

The objective of this workshop is to bring experts from academia and industry together to discuss avenues for bringing artificial intelligence to edge devices. Artificial intelligence is having significant environmental impact that will increase through time as more industries embed AI in their products. We hope that such discussions open new horizons for a more efficient and more responsible AI worldwide. We take advantage of the vibrant academic culture of Montreal, which is also known as the international hub for deep learning, to tackle deep models as the main focus of this workshop.

Warren Gross and Vahid Partovi Nia

Platinum Sponsor

It is my great pleasure to sponsor the edge intelligence workshop for the second time. The first workshop in 2020 focused mostly on optimization and mathematical challenges, run by the internationally renowned Montreal-based research Centre GERAD. I am happy to see that the second workshop in 2022 has shifted its focus to hardware and algorithm directions. Huawei Noah's Ark lab's mission is to push boundaries related to machine learning and artificial intelligence; as the director of Noah's Ark Canada, I am privileged to play a role in bringing researchers from academia and industry together on this important topic. Huawei's mission is to serve customers as one of its main cultural key values. We strongly believe that bringing artificial intelligence closer to the edge empowers people against companies. This creates new room for customer-centric innovation and technological democratization. I want to thank the organizers, the scientific committee, invited speakers, and poster contributions which helped enrich this scientific event. I hope that such a friendly ecosystem endures for a long time in Montreal.

Wulong Liu,
Director of the Montreal Research Center, and Huawei Noah's Ark Canada.

Conference Chairs

- Warren Gross — McGill University
- Vahid Partovi Nia — Huawei Noah's Ark Lab, Polytechnique Montréal

Organizing Committee

- Xi Chen — McGill University
- James Clark — McGill University
- Negin Firouzian — McGill University
- Ghouti Boukli Hacene — Sony
- Mehdi Rezagholizadeh — Huawei Noah's Ark Lab
- Brett Meyer — McGill University
- Mohammadreza Tayaranian — McGill University

Technical Committee

- Masoud Asgharian — McGill University
- Charles Audet — GERAD and Polytechnique Montréal
- Tiago Falk — INRS
- Ali Ghodsi — University of Waterloo
- Sebastien Le Digabel — Polytechnique Montreal
- Alejandro Murua — Université de Montréal
- Dominique Orban — GERAD and Polytechnique Montréal
- Yvon Savaria — Polytechnique Montreal
- Yaoliang Yu — University of Waterloo
- Chao Xing — Huawei Noah's Ark Lab

Table of Contents

Scientific Program

Monday, September 19

	Session 1	
08:30-09:30	Diana Marculescu	Keynote: "When Sustainability Meets Machine Learning: Efficient Learning from Cloud to Edge"
09:30-10:00	James Clark	What is lost when networks are compressed?
	Session 2	
10:30-11:00	Vahid Partovi Nia	Edge implementation of deep models
11:00-11:30	Mark Coates	Efficient Bayesian Network Architecture Search for Graph Neural Networks
11:30-12:00	Ehsan Saboori	Running 2 bit quantized CNN models on ARM CPUs
	Session 3	
13:00-15:00	Evgeni Gousev	Keynote: "tinyML: ultimate energy efficient machine learning solution for edgeAI"
14:00-14:30	Muthucumaru Maheswaran	JAMScript: A Programming Language for Edge Oriented Mobile Internet of Things
14:30-15:00	Shahrokh Valaee	Cooperative Location Estimation using Federated Learning
	Session 4	
15:30-16:00	Rachel E. Bouserhal	Hearables and their potential as a tool for early disease detection
16:00-16:30	Dounia Lakhmiri	A Stochastic Proximal Method for Nonsmooth Regularized Finite Sum Optimization
16:30-17:00	Masoud Asgharian	Causal Discovery, Independence of Mechanism and Input Assumption and Selection Bias
17:00-17:30	Michael Rabbat	Asynchronous Federated Learning at Scale

Tuesday, September 20

	Session 1	
08:30-09:30	Wen Tong	Keynote: Machine Learning Based Post-Shannon Cognition Communications
09:30-10:00	Brett Meyer	Transforming Intelligence for the Edge: Challenges and Opportunities in Modeling, Optimization, and Deployment
	Session 2	
10:30-11:00	Yunaho Yu	Challenges for Edge Device Machine Learning Platform
11:00-11:30	Naoya Onizawa	Fast-Converging Simulated Annealing for Ising Models Based on Integral Stochastic Computing
11:30-12:00	Ghouthi Boukli Hacene	DNN Quantization and acceleration for training and inference
	Session 3	
13:00-14:00	Song Han	Keynote: Efficient AI Computing with Sparsity
14:00-14:30	Christophe Dubach	Very High-Level Synthesis of Neural Networks Accelerators for FPGAs
14:30-15:00	Francois Leduc-Primeau	Building Energy-Efficient AI Chips by Exploiting Energy-Reliability Tradeoffs
	Session 4	
15:30-16:00	Yvon Savaria	Applications of Edge Intelligence, Applications, Lessons Learned and Platforms
16:00-16:30	Andreas Moshovos	Boosting Machine Learning Innovation: Computing Systems that Learn and Adapt
16:30-17:00	Pascal Poupart	Uncertainty Aware Federated Learning
17:00-17:30	Sarath Chandar	TBD

In what follows the abstracts for all the keynotes speakers, invited speakers are provided in their order of presentation. The abstracts for poster presentations are provided in alphabetical order of their title for easier searchability.

Keynote Speakers

When Sustainability Meets Machine Learning: Efficient Learning from Cloud to Edge

- Diana Marculescu University of Texas at Austin

Abstract:
A large portion of current cloud and edge workloads feature Machine Learning (ML) tasks, thereby requiring a deep understanding of their energy efficiency. While the holy grail for judging the quality of a ML model has largely been testing accuracy, and only recently its resource usage and training efficiency, neither of these metrics translate directly to energy efficiency, runtime, or mobile device battery lifetime. This talk uncovers the need for building accurate, platform-specific power and latency models for ML and efficient hardware-aware ML design methodologies, thus allowing machine learners and hardware designers to identify not just the best accuracy ML model configuration, but also those that satisfy given hardware constraints and are likely to have a low carbon footprint. I will discuss our early supernet-based Single-Path Neural Architecture Search (NAS) approach which finds the final model configuration up to 5,000× faster compared to prior work, translating in only 0.75 CO_2 for finding a constraint-satisfying ML model, and continue with approaches that allow efficient transfer of large models to the edge with little to no accuracy loss. First, we use a supernet ML model to transfer object detection on edge by Adapting Networks over Time (ANT) and show that top object detection accuracy can be achieved at almost 50% less cost than existing work. Furthermore, we identify which parts of ML model should be transferred and how, thereby resulting in superior accuracy at a fraction of the cost compared to existing work. To this end, we employ selective fine-tuning and reinforcement learning for on-device transfer learning that can achieve top image classification accuracy at almost 10x less cost than state of the art approaches. Finally, we underscore the importance of thermal effects on edge ML performance and identify a supernet-based approach for mitigating their impact with little loss in overall accuracy, suggesting ways to reduce the carbon footprint for on-edge learning under thermal constraints.

tinyML: ultimate energy efficient machine learning solution for edgeAI

- Evgeni Gousev

Senior Director, Qualcomm AI Research, Qualcomm Technologies, Inc. and tinyML Foundation, Chairman, Board of Directors

Abstract:

Recent progress in computing hardware, machine learning algorithms and networks and availability of large datasets for model training have created a strong momentum in development and wide deployment of game changing AI applications. Dedicated hardware becomes tiny and very energy efficient (with mW or less power consumption), algorithms and models - smaller (down to 10s of kB of memory requirements), software – lighter down to deployment on deeply embedded platforms. This enormous technology innovative wave and fast growing ecosystem create a strong momentum towards new applications and business opportunities. As we are in the midst of the digital transformation revolution, tinyML offers ultimate benefits of extreme energy savings of performing on-device machine intelligence and data analytics at low cost, combined with inherent privacy features. Within the decade, we are going to witness a major impact that tinyML phenomenon is going to create in the society at different levels. At the commercial level, there will be an explosive growth of tinyML applications in all verticals around us: smart city, smart home, smart manufacturing, smart agriculture, etc.. By giving this tiny artificial technology back to people, there will be many applications of tinyML for Good, such as in the sustainability, STEM and healthcare areas. And, tinyML opens up a path to democratize AI via enabling data ownership by those who produce these data streams and make them actionable, locally, at the very edge. As a result, we see a new world with trillions of intelligent devices enabled by tinyML technologies that sense, analyze and autonomously act together to create a healthier and more sustainable environment for all.

Machine Learning Based Post-Shannon Cognition Communications

- Wen Tong Huawei Technologies Canada

Abstract:
In this talk, a new machine learning based communications architecture is presented; the Type-1 is a direct-communications framework; where we use machine learning technique to identify the specific object in the physical world and to use the extreme compression with ultra-low data-rate to communicate the object scenery information in real-time; the Type-2 is a hierarchical-communications framework; which consists of the partition of the System-1 and System-2, such that the communication entropy is distilled into System-2, this enables so-called intelligence communications. Both Type-1 and Type-2 communications can be developed as foundational technologies for 6G to enable the new paradigm for machine-to-machine, and human-to-machine communications.

Efficient AI Computing with Sparsity

- Song Han Massachusetts Institute of Technology

Abstract:
Modern deep learning requires a massive amount of computational resources, energy, and engineering efforts, making on-device machine learning challenging; retraining the model on-device is even more difficult. We make machine learning efficient by utilizing sparsity. We'll first present neural architecture search techniques by searching sparsely activated sub-networks from the once-for-all network (OFA), and MCUNet that brings AI to micro-controllers. Then I'll describe TinyTL and on-device training that enables the model to adapt to new data collected from the sensors using only 256KB memory, 1000x less than PyTorch. Next I'll talk about improving the efficiency by utilizing the temporal sparsity for videos, spatial sparsity for point cloud, and token level sparsity for NLP. I'll conclude by hardware and system support for sparsity (TorchSparse, SpAtten, SpArch, PointAcc). The presentation will highlight full-stack optimizations, including the neural network topology, inference library, and the hardware architecture, which allows a larger design space to unearth the underlying principles for sparsity and efficient AI.

Invited Speakers

What is lost when networks are compressed?

- James Clark McGill University

Abstract:
Network compression, which is the process of taking a large neural network and reducing its computational load and memory footprint, is a key step in developing neural network solutions for implementation on edge devices.
In this talk I look at what is lost when networks are compressed. It is well known that when networks are heavily compressed, their performance suffers, such as reduced accuracy on classification tasks. However, accuracy is not the only aspect of network performance which designers of edge devices may care about. In this talk, I consider two other aspects – uncertainty calibration and domain adaptability. Network calibration refers to the correlation between the network's uncertainty measure and its accuracy. In practice, you would like the uncertainty measure to track the classifier accuracy, so that, for example, inputs yielding an uncertainty of 0.75 means that the network classifies these inputs correctly 75% of the time. Domain adaptability means how well a network performs on a new domain, when domain adaptation processes are carried out to shift the operation of the network trained in one domain to work on a new domain. I will consider the relative effect of compression rates and compression technique on network calibration and domain adaptation.

Edge implementation of deep models

- Vahid Partovi Nia Huawei Noah's Ark Lab, Polytechnique Montréal

Abstract:
Neural networks are not new to statisticians. Brian Ripley wrote a seminal textbook on the topic in the 1990s, and the projection pursuit regression of Friedman and Tuckey is closely related to one hidden layer fully connected neural network, but it stopped there. There has been little contribution by statisticians in recent decades to neural networks, however, most statistical machine learning books now include a chapter on this topic. Statistics departments should take this topic seriously because of its impact on technological products and embed it in their applied coursework. Deep neural networks can provide a new nonparametric modeling horizon to build regression models on the output of some other regression models, hierarchically. Perhaps with this definition we must classify nonparametric statistics experts into i) "shallower", or ii) "deeper", which most statisticians fall into the first category. I aim at promoting the second category while focusing on the implementation of such models on edge devices.

Efficient Bayesian Network Architecture Search for Graph Neural Networks

- Mark Coates McGill University

Abstract:
Real life data often arises from relational structures that are best modeled by graphs. Bayesian learning on graphs has emerged as a framework which allows us to model prior beliefs about network data in a mathematically principled way. The approach provides uncertainty estimates and can perform very well on a small sample size when provided with an informative prior. Much of the work on Bayesian graph neural networks (GNNs) has focused on inferring the structure of the underlying graph and the model weights. An important factor that can strongly influence the performance of the network is the choice of the architecture. In a GNN this includes the number of layers, the number of active neurons, the aggregators, and pooling procedures. Searching over candidate architectures can be extremely demanding in terms of computation. It is essential that we can perform the search efficiently. In this talk we will describe search strategies that employ proxy scoring functions based on neural network Gaussian processes. We also introduce a Bayesian approach that allows us to specify a posterior over the performance rather than a point estimate. This allows us to make architectural decisions that are framed around performance robustness as well as accuracy and complexity.

Running 2bit quantized CNN models on Arm CPUs

- Ehsan Saboori Deeplite

Abstract:
The emergence of Deep Neural Networks (DNNs) on embedded and low-end devices holds tremendous potential to expand the adoption of AI technologies to wider audiences. However, making DNNs applicable for inference on such devices using techniques such as quantization and model compression, while maintaining model accuracy, remains a challenge for production. Ehsan will be presenting a novel inference engine (DeepliteRT) and compression framework (Deeplite Neutrino), developed by Deeplite Inc., that automatically quantizes and runs PyTorch deep learning models at real 2bit and 1bit precision.

JAMScript: A Programming Language for Edge Oriented Mobile Internet of Things

- Muthucumaru Maheswaran McGill University

Abstract:
Mobile Internet of Things (IoT) is becoming very common place with many different realizations of them including smart vehicles, wearables, drones, and other mobile sensors. Mobile IoT is getting a tremendous boost from 5G and edge computing. Mobile IoT has many applications such as e-bike tracking, home appliances monitoring, assets tracking, and environmental pollution detection. With the introduction of edge computing, mobile IoT becomes even more powerful and suitable for hosting more demanding applications. However, there are many outstanding issues including the following that need to be addressed before edge oriented mobile IoT can reach its full potential: (a) efficiently and quickly discover other mobile IoT and edge servers, (b) collaborate opportunistically with available mobile IoT and edge servers, (c) create shared data at the edge and use them as needed in time and location, and (d) efficiently deal with unexpected movements and failures in a graceful and safe manner. We have developed a novel programming language called JAMScript that addresses these challenges for edge oriented mobile IoT. JAMScript supports three types of tasks: batch tasks, real-time tasks, and synchronous tasks. With synchronous tasks, we can time align the task executions across different worker nodes. A program written in JAMScript can be deployed to run either in a single device, many devices and an edge server, or many devices and many edge servers and the cloud. The language runtime is responsible for managing the task executions as a program is deployed under different configurations. That is, a device can operate independently or under the control of an edge server and the runtime is responsible for detecting the active configuration and routing the task execution requests and data transfers.

Cooperative Location Estimation using Federated Learning

- Shahrokh Valaee University of Toronto

Abstract:
Channel State Information (CSI) based fingerprinting is surfacing as an accurate and robust method of indoor localization. However, the high-dimensional nature of CSI data impedes its adoption in multi access point (AP) systems. To reap the rewards of cooperative localization with privacy and limited system complexity in mind, we propose a federated learning (FL) architecture. Each AP has an individual model and a shared model, where the individual model parameters are unique to each AP and the shared model parameters are communicated to a central server for aggregation. The server averages the models and sends them back to each AP, which uses this joint model as a regularization term. To capture the spatiotemporal characteristics of CSI, we propose a convolutional neural network (CNN) as each AP's individual model and a multi-layer perceptron (MLP) as the shared model. Extensive experimental studies verify the superiority of the proposed edge computing approach compared to the exiting methods in the literature. We use commercial off-the-shelf APs collecting CSI data in multiple indoor environments and compare the proposed system to the state-of-the-art deep learning models.

Hearables and their potential as a tool for early disease detection

- Rachel E. Bouserhal École de technologie supérieure

Abstract:
In-ear wearables, or hearables, have become increasingly popular over recent years. This is mainly due to two reasons: people have become accustomed to continuously wearing in-ear devices and, more importantly, when occluded, the ear becomes a portal of access to a plethora of human-produced events ranging from speech to a blink of an eye. Access to such a variety of signals coupled with advancements in artificial intelligence pave the way for automatic disease detection with hearables. This is particularly interesting for degenerative diseases such as Parkinson's and Alzheimer's disease because intervention at the early stages could slow down the decline of motor and cognitive abilities. The continuous individual use of hearables as well as simultaneous tracking of a diverse set of signals provides an opportunity for multimodal prediction models for early disease detection. This talk focuses on current advancements on classification models for in-ear signals and future applications of such algorithms.

A Stochastic Proximal Method for Nonsmooth Regularized Finite Sum Optimization

- Dounia Lakhmiri Polytechnique Montréal

Abstract:
We consider the problem of training a deep neural network with non-smooth regularization to retrieve a sparse and efficient sub-structure. Our regularizer is only assumed to be lower semi-continuous and prox-bounded. We combine an adaptive quadratic regularization approach with proximal stochastic gradient principles to derive a new solver, called SR2, whose convergence and worst-case complexity are established without knowledge or approximation of the gradient's Lipschitz constant. We formulate a stopping criteria that ensures an appropriate first-order stationarity measure converges to zero under certain conditions. We establish a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-2})$ that matches those of related methods like ProxGEN, where the learning rate is assumed to be related to the Lipschitz constant. Our experiments on network instances trained on CIFAR-10 and CIFAR-100 with l_1 and l_0 regularizations show that SR2 consistently achieves higher sparsity and accuracy than related methods such as ProxGEN and ProxSGD.

Asynchronous Federated Learning at Scale

- Michael Rabbat Facebook AI Research

Abstract:
Federated Learning (FL) trains a model across distributed devices without the training data ever leaving the device. When combined with appropriate privacy-enhancing technologies, such as secure aggregation and differential privacy, this enables distributed privacy-preserving machine learning. Most FL systems described in the literature are synchronous - they perform a synchronized aggregation of model updates from individual clients. Scaling synchronous FL is challenging since increasing the number of clients training in parallel leads to diminishing returns in training speed, analogous to large-batch training. Moreover, stragglers hinder synchronous FL training. This talk will discuss experience designing and deploying a production asynchronous FL system. Our work tackles the aforementioned issues, sketches of some of the system design challenges and their solutions, and touches upon principles that emerged from building a production FL system for millions of clients. Empirically, we demonstrate that asynchronous FL converges significantly faster than synchronous FL, while incurring less communication overhead, when training across nearly one hundred million devices.

Transforming Intelligence for the Edge: Challenges and Opportunities in Modeling, Optimization, and Deployment

- Brett Meyer McGill University

Abstract:

The state-of-the-art in deep learning has advanced rapidly, with especially significant improvements in algorithm performance on diverse computer vision and natural language processing. Our capacity to deploy such models at the edge lags considerably, however, for a variety of reasons. First, models that achieve SOTA performance are often large and complex, and far beyond what resource-constrained edge devices can manage. Second, the software environments available for deployment to edge devices generally cannot keep pace with the evolution of models and their components. Case in point: though quantization has been a standard practice for complexity reduction for years, support for edge devices remains remarkably uneven, and even varies across different frameworks for the same hardware.

In this talk, I will explore recent work supporting the optimization and deployment of transformer-based models to resourced-constrained heterogeneous systems at the edge. Heterogeneous multiprocessors offer a wide variety of opportunities for improving edge inference but navigating the multitude of options is time consuming. Efficient decision making must be supported by power and performance modeling and estimation. Along the way, I will also present our recent experiences addressing the various challenges that emerge when attempting to make SOTA models run on edge hardware, and what it means to practically optimize such systems.

Challenges for Edge Device Machine Learning Platform

- Yuanhao Yu Huawei Technologies

Abstract:

In this talk, an abstract or concept example of industrial machine learning platform oriented for edge device will be first presented. Based on this setting, a number of challenges shall be further introduced from different perspective, such as robustness, efficiency and maintenance.

From efficiency point of view, given a number of “prototype” models without considering numerous edge device constraint like computational resource, an auto-distillation mechanism of the platform to build edge device deployable model will be introduced. Within this module, the automation and generalization remain unsolved and shall be studied.

To enhance robustness of user experience, a platform functionality called model adaptation will be discussed as well, which aims to take advantage of user’s data to boost model capability. Similar ideas have been also investigated in multiple fields, including federated learning, meta learning, transfer learning and edge device efficient training. However, in order to productize the technique, some issues still remain.

In terms of maintenance, an online automatic way to maintain the training dataset for further model major update on the cloud end has drawn significant attention in industry, as usually the new “stream”-like data constantly coming in and it would not be economical to keep it all. On the other side, there is a strong motivation to make use of the incremental data and deal with their noisy labels in order to strengthen the model. How to find an optimal or trade-off solution is still beyond answered.

Fast-Converging Simulated Annealing for Ising Models Based on Integral

- Naoya Onizawa Tohoku University

Abstract:
Probabilistic bits (p-bits) have recently been presented as a spin (basic computing element) for the simulated annealing (SA) of Ising models. In this brief, we introduce fast-converging SA based on p-bits designed using integral stochastic computing. The stochastic implementation approximates a p-bit function, which can search for a solution to a combinatorial optimization problem at lower energy than conventional p-bits. Searching around the global minimum energy can increase the probability of finding a solution. The proposed stochastic computing-based SA method is compared with conventional SA and quantum annealing (QA) with a D- Wave Two quantum annealer on the traveling salesman, maximum cut (MAX-CUT), and graph isomorphism (GI) problems. The proposed method achieves a convergence speed a few orders of magnitude faster while dealing with an order of magnitude larger number of spins than the other methods.

DNN Quantization and acceleration for training and inference

- Ghouti Boukli Hacene Sony

Abstract:
Efficient deep neural network (DNN) inference on mobile or embedded devices typically involves quantization of the network parameters and activations. In this talk we will see how we can quantize efficiently a neural network using existing quantization technics and improve them using a learnable linear combination during training that allows us to quantize DNNs gradually and converge to a quantized DNN with better performance than other counter-part methods. We will also see in this talk how can compression methods to reduce training complexity. We show that using decomposition methods and dimension reductions methods help us to reduce the training complexity of DNNs and achieve better performance while reducing the number of operations required to train DNNs.

Very High-Level Synthesis of Neural Networks Accelerators for FPGAs

- Christophe Dubach McGill University

Abstract:
FPGAs (Field Programmable Gate Arrays) have become the substrate of choice to implement accelerators. They deliver high performance with low power consumption, while offering the flexibility of being re-programmable. But they are notoriously hard to program directly using HDLs (Hardware Description Languages). Traditional HLS (High-Level Synthesis) are far from being perfect as programmers are still required to write hardware-specific code and existing HLS tools often produce suboptimal designs.
This talk will present current efforts to address these shortcomings, using a multi-level functional IR (Intermediate Representation). As we will see, a functional IR makes optimizations via rewrite rules simple to express, and abstract away the hardware details. This approach has the advantage of generating high performance designs in a predictable way, drastically reducing design time. This talk will show how neural networks are easily represented using functional hardware-agnostic constructs. The resulting FPGA synthesized designs achieve near-peak performance and are competitive with the output produced by current HSL tools.

Building Energy-Efficient AI Chips by Exploiting Energy-Reliability Tradeoffs

- Francois Leduc-Primeau Polytechnique Montreal

Abstract:
For the last ten years or so, the energy efficiency of CMOS integrated circuits has been improving only slowly, even though the size of transistors on the chip continues to decrease. To continue improving the capabilities of digital systems, it is crucial to improve their energy efficiency. Many approaches have been proposed, such as near-threshold CMOS circuits or processing-in-memory architectures, but these approaches all have in common that they make it more difficult to control the reliability of the fabricated circuits. In communications, we know since the seminal work of Claude Shannon that adding redundancy to a message allows to greatly reduce the energy needed to communicate it across a noisy channel. In this talk, we will discuss ways in which this approach can be applied to deep neural networks, with the objective of building artificial intelligence systems that achieve energy efficiency comparable to human intelligence.

Applications of Edge Intelligence, Applications, Lessons Learned and Platforms

- Yvon Savaria Polytechnique Montreal

Abstract:
Machine learning based data processing methods find uses and provide competitive solutions in countless applications. A team under his supervision was recently exposed to a wide range of applications related to intelligence at the edge including: epileptic seizure prediction, pose estimation, medical image segmentation, human activity detection at a bus stop in a public transportation system, and cyber-attack detection. Solutions providing excellent processing quality could be found for all these very diverse applications. Edge intelligence was a winning solution for most of these applications, but not for all. The talk will report on solutions that were proposed and will share lessons learned on the features that influence the degree of success of edge intelligence for various applications. Quantization on use of low-resolution models is a fruitful solution to obtain compact models that facilitate the use of machine learning at the edge. The tradeoffs associated with processing quality in low complexity models is directly linked to effective learning. Successful deployment of intelligence at the edge also depends on the use of suitable hardware platforms Experience and efforts towards the deployment and use of these platform will also be reported.

Boosting Machine Learning Innovation: Computing Systems that Learn and Adapt

- Andreas Moshovos University of Toronto

Abstract:
Machine learning enabled computing devices to learn to “think”, to “see”, to “hear”, to “read”, to “write” and in general interact with the physical world in ways that we typically associate only with humans and high intelligence. These computing systems can rival human abilities and promise to enhance our ability to discover, learn and benefit from information. The application areas are enormous covering science, medicine, health, commerce, civil planning, security, and so on. Further enabling innovation in machine learning however requires computing systems that can store even larger amounts of information. We have been exploring hardware/software techniques that can enable further innovation in machine learning. We target methods that are designed from the ground up to take advantage of behaviors that emerge when machine learning workloads execute. This talk will overview some of those methods: APack a lossless compression method for inference with fixed-point values, Mokey a quantization method for transformers, and Schrödinger's Floating-Point a compressor to accelerate training. APack takes advantage of the lopsided values, Mokey rethinks dictionary-based quantization to enable storage and computation of value indexes, and Schrödinger's Floating-Point dynamically adjusts floating-point containers to reduce memory overheads.

Uncertainty Aware Federated Learning

- Pascal Poupart University of Waterloo

Abstract:
Federated Learning (FL) has gained popularity for combining models trained on edge devices without the data leaving its owner's device. Since each edge device may not have a lot of data, several questions arise: How do we avoid overfitting? How do we quantify model uncertainty? How can the uncertainty of local models be taken into account during their aggregation into a global model? How can we calibrate the predictive uncertainty? In this talk, I will describe two Bayesian FL techniques that quantify uncertainty and then use this uncertainty to obtain robust global models with improved predictions. The first technique is based on Gaussian processes, which express a distribution over the parameters of the last layer of a predictor. The second technique directly computes a distribution over predictions while reducing the amount of communication between edge devices and the server to a single round of messages.

Contributed Posters

A Decomposition Method Supporting Many Factorization Structures

- Marawan Gamal Abdel Hameed
  Huawei Technologies Canada, University of Waterloo
- Ali Mosleh — Huawei Technologies Canada
- Marzieh S. Tahaei — Huawei Technologies Canada
- Vahid Partovi Nia — Huawei Technologies Canada

Abstract:
While convolutional neural networks (CNNs) have become the de facto standard for most image processing and com- puter vision applications, their deployment on edge devices remains challenging. Tensor decomposition methods provide a means of compressing CNNs to meet the wide range of de- vice constraints by imposing certain factorization structures on their convolution tensors. However, being limited to the small set of factorization structures presented by state-of-the- art decomposition approaches can lead to sub-optimal performance. We propose SeKron, a novel tensor decomposition method that offers a wide variety of factorization structures, using sequences of Kronecker products. By recursively finding approximating Kronecker factors, we arrive at optimal decompositions for each of the factorization structures. We show that SeKron is a flexible decomposition that generalizes widely used methods, such as Tensor-Train (TT), Tensor- Ring (TR), Canonical Polyadic (CP) and Tucker decompositions. Furthermore, we derive an efficient convolution projection algorithm shared by all possible factorization structures of SeKron to obtain the compressed version of a given CNN. We validate SeKron for model compression on both high-level and low-level computer vision tasks and find that it outperforms state-of-the-art decomposition methods.

A Short Study on Compressing Decoder-Based Language Models

- Tianda Li — Huawei Technologies Canada
- Yassir El Mesbahi — Huawei Technologies Canada
- Ivan Kobyzev — Huawei Technologies Canada
- Ahmad Rashid — Huawei Technologies Canada
- Atif Mahmud — Huawei Technologies Canada
- Nithin Anchuri — Huawei Technologies Canada
- Mehdi Rezagholizadeh — Huawei Technologies Canada

Abstract:

Pre-trained Language Models (PLMs) have been successful for a wide range of natural language processing (NLP) tasks. The state-of-the-art of PLMs, however, are extremely large to be used on edge devices. As a result, the topic of model compression has attracted increasing attention in the NLP community.

Most of the existing works focus on compressing encoder-based models (tiny-BERT, distilBERT, distilRoBERTa, etc), however, to the best of our knowledge, the compression of decoder-based models (such as GPT-2) has not been investigated much. Our paper aims to fill this gap. Specifically, we explore the pre-training of a compressed GPT-2 model using layer truncation and compare it against the distillation-based method (DistilGPT2). The training time of our compressed model is significantly less than DistilGPT-2, but it can achieve better performance when fine-tuned on downstream tasks. We also demonstrate the impact of data cleaning on model performance.

An Exploration into the Performance of Unsupervised Cross-Task Speech Representations for ''In the Wild'' Edge Applications

- Heitor Guimarães INRS
- Arthur Pimentel INRS
- Anderson Avila Huawei Technologies Canada
- Mehdi Rezagholizadeh Huawei Technologies Canada
- Tiago Falk INRS

Abstract:
Unsupervised speech models are becoming ubiquitous in the speech and machine learning communities. Upstream models are responsible for learning meaningful representations from raw audio. Later, these representations serve as input to downstream models to solve a number of tasks, such as keyword spotting or emotion recognition. As edge speech applications start to emerge, it is important to gauge how robust these cross-task representations are on edge devices with limited resources and different noise levels. To this end, in this study we evaluate the robustness of four different versions of HuBERT, namely: base, large, and extra-large versions, as well as a recent version termed Robust-HuBERT. Tests are conducted under different additive and convolutive noise conditions for three downstream tasks: keyword spotting, intent classification, and emotion recognition. Our results show that while larger models can provide some important robustness to environmental factors, they may not be applicable to edge applications. Smaller models, on the other hand, showed substantial accuracy drops in noisy conditions, especially in the presence of room reverberation. These findings suggest that cross-task speech representations are not yet ready for edge applications and innovations are still needed.

ARMCL BERT: Novel Quantizable BERT Implementation for ARM SoCs

- Murray Kornelsen McGill University
- Seyyed Mozafari McGill University
- James Clark McGill University
- Brett Meyer McGill University
- Warren Gross McGill University

Abstract:
One challenge in moving BERT inference onto edge devices is a lack of support in popular inference frameworks. Furthermore, in order to lower latency and decrease memory usage, fp16 and quantized int8 operations are preferable. In this regard, we developed an ARMCL BERT framework that runs on CPU and GPU and supports fp16 and int8 data types. To enable this implementation and optimize latency, we extend ARMCL with new quantizable GELU and LayerNorm implementations. We measure BERT latency on two popular edge hardware platforms, showing up to 30% improvement moving from fp32 to fp16. Comparing mobile GPU to CPU, we find the ARM Mali GPU could reduce latency by over 50%, while the tested Qualcomm Adreno GPU presented major difficulties, with poor performance and driver errors.

BERT Inference Energy Predictor for Efficient Hardware-aware NAS

- Chuning Li McGill University
- Seyyad Mozafari McGill University
- James Clark McGill University
- Warren Gross McGill University
- Brett Meyer McGill University

Abstract:
Deploying large language models, such as BERT, to energy-constrained devices is challenging. One way to optimize these complex models is to utilize hardware-aware neural architecture search (NAS). However, hardware-aware NAS needs to incorporate hardware performance metrics, such as energy consumption. Although on-device measurements provide accurate feedback, the overhead is huge. For example, the on-device energy measurement of a design space with the size of 10^5 models would take 49 weeks on an embedded system, such as Hikey970 platform.

To address this problem, we propose an energy modelling framework on the Hikey970 ARM big.LITTLE CPU to predict the energy consumption of BERT models. Our deep neural network (DNN) based energy model is evaluated on the DynaBERT design space of 240 models. It achieves a mean absolute percentage error (MAPE) of 0.14. Combining with the real accuracy values on the QNLI task, we evaluate our energy model's ability to predict Pareto-optimal front in the 2D accuracy-energy design space. Our model correctly predicts 14 out of the 17 true Pareto-optimal models.

Dyadic Integer Only BERT

- Charles Le McGill University
- Arash Ardakani McGill University
- James Clark McGill University
- Brett Meyer McGill University
- Warren Gross McGill University

Abstract:
Transformer-based models such as BERT have seen successes in various Natural Language Processing (NLP) tasks, but their huge computational costs from Layer Normalization, GeLU, and softmax, and also their huge model size prevent deployment on edge device for reduced inference latency and low power inference. Quantization has been applied to make BERT's inference more energy efficient and faster, but most works that focus on quantization, quantize the matrix multiplication and dequantize or use floating-point arithmetic for non linear operations such as Layer Normalization, GeLU, and softmax. As a result of that Integer-only units such as Turing TensorCores or integer-only ARM processors cannot be efficiently utilized. In this work, we propose approximation functions that make BERT more integer friendly than previous works on integer-based BERT, without floating point arithmetic and having less division operations compared to previous works. We also propose a training pipeline to allow BERT with the non-linear approximation functions to maintain the same accuracy as its full precision counterpart.

Faster Attention Is What You Need: A Fast Self-Attention Neural Network Backbone Architecture for the Edge via Double-Condensing Attention Condensers

- Alexander Wong University of Waterloo
- Mohammad Javad Shafiee University of Waterloo
- Saad Abbasi University of Waterloo
- Saeejith Nair University of Waterloo
- Mahmoud Famouri University of Waterloo

Abstract:
With the growing adoption of deep learning for on-device TinyML applications, there has been an ever-increasing demand for more efficient neural network backbones optimized for the edge. Recently, the introduction of attention condenser networks have resulted in low-footprint, highly-efficient, self-attention neural networks that strike a strong balance between accuracy and speed. In this study, we introduce a new faster attention condenser design called double-condensing attention condensers that enable more condensed feature embedding. We further employ a machine-driven design exploration strategy that imposes best practices design constraints for greater efficiency and robustness to produce the macro-micro architecture constructs of the backbone. The resulting backbone (which we name AttendNeXt) achieves significantly higher inference throughput on an embedded ARM processor when compared to several other state-of-the-art efficient backbones (>10X faster than FB-Net C at higher accuracy and speed) while having a small model size (>1.47X smaller than OFA-62 at higher speed and similar accuracy) and strong accuracy (1.1% higher top-1 accuracy than MobileViT XS on ImageNet at higher speed). These promising results demonstrate that exploring different efficient architecture designs and self-attention mechanisms can lead to interesting new building blocks for TinyML applications.

Generalizing ProxConnect on Vision Transformer Binarization

- Yiwei Lu — University of Waterloo
- Yaoliang Yu — University of Waterloo
- Vahid Partovi Nia — Huawei Noah's Ark Lab
- Xinlin Li — Huawei Noah's Ark Lab
- Ali Mosleh — Huawei Noah's Ark Lab
- Arash Ardakani — Huawei Noah's Ark Lab

Abstract:
In neural network binarization, BinaryConnect (BC) and its variants are considered as the gold standard. Such methods usually apply sign function in the forward pass of the neural network and perform backpropagation using straight through estimator (STE) to avoid zero gradients. Although such implementation works well in practice, the inner workings of the inconsistent forward and backward pass is not clear. In this work, we aim at closing this gap by generalizing the broader BC family, i.e., ProxConnect (PC): (1)
We identify that considering the weight transformation in PC, we can naturally introduce forward and backward proximal quantizers, (2) we refine the true (regularized) objective function of BC family and thus entitle their empirical success from the optimization justification. (3) We apply the generalized PC algorithm with different forward-backward proximal quantizers to the popular vision transformer architecture on CIFAR-10/100 and ImageNet, which also serves as the first exploration of binarizing vision transformers.

GHN-Q: Parameter Prediction for Unseen Quantized Convolutional Architectures via Graph Hypernetworks

- Stone Yun University of Waterloo
- Alexander Wong University of Waterloo

Abstract:
Deep convolutional neural network (CNN) training via iterative optimization has had incredible success in finding optimal parameters. However, modern CNN architectures often contain millions of parameters. Thus, any given model for a single architecture resides in a massive parameter space. Models with similar loss could have drastically different characteristics such as adversarial robustness, generalizability, and quantization robustness. For deep learning on the edge, quantization robustness is often crucial. Finding a model that is quantization-robust can sometimes require significant efforts. Recent works using Graph Hypernetworks (GHN) have shown remarkable performance predicting high-performant parameters of varying CNN architectures. Inspired by these successes, we wonder if the graph representations of GHN-2 can be leveraged to predict quantization-robust parameters as well, which we call GHN-Q. We conduct the first-ever study exploring the use of graph hypernetworks for predicting parameters of unseen quantized CNN architectures. We focus on a reduced CNN search space and find that GHN-Q can in fact predict quantization-robust parameters for various 8-bit quantized CNNs. Decent quantized accuracies are observed even with 4-bit quantization despite GHN-Q not being trained on it. Quantized finetuning of GHN-Q at lower bitwidths may bring further improvements and is currently being explored.

Gradient Distribution Theory for Exploding and Vanishing Gradient Problem

- Justin Yu Huawei Noah's Ark Lab

Abstract:
While quantization has been shown to be successful in practice with minimal accuracy degradation, there has not been any theoretical analysis on the effects of quantization. With the goal to find the theoretical optimal number format, we investigate the effects of quantizing gradients for the backward pass. Instead of gradients, or neural gradients, we study the neural network's Jacobian matrix's column norms, a related but more stable quantity. This norm is provably approximately log-normally distributed on ReLu linear networks with no skip connections. Experiments on networks with skip connections such as Resnet-18 and Shufflenet on CIFAR-10 demonstrate a near perfect log-normal distribution for these norms on all layers except those with skip connection. For layers with skip connection, it has been observed that applying a square root transformation normalizes the distribution. From this result, we derive bounds on the size of the gradient with respect to the accumulator bit width to minimize the probability of exploding and vanishing gradient. This bound gives a theoretical and quantitative proof for quantizing gradient descent with high stability.

How Robust is Robust wav2vec 2.0 for Edge Applications?: An Exploration into the Effects of Quantization and Model Pruning on “In-the-Wild” Speech Recognition

- Arthur Pimentel — INRS
- Heitor Guimarães — INRS
- Anderson Avila — Noah's Ark Lab, Huawei Technologies
- Mehdi Rezagholizadeh — Noah's Ark Lab, Huawei Technologies
- Tiago Falk — INRS

Abstract:
Recent advances on self-supervised learning have allowed speech recognition systems to achieve state-of-the-art (SOTA) word error rates (WER) while requiring only a fraction of the labeled training data needed by its predecessors (e.g., the wav2vec 2.0 model). Notwithstanding, while such models have shown to achieve SOTA performance on matched conditions, their performance has shown to degrade in unmatched conditions, which is typically the case in edge applications. To overcome this problem, strategies such as data augmentation and/or multi-condition training have been explored and a robust version of wav2vec 2.0 has been implemented. It is argued here, however, that such models are still too large to be considered for edge applications on resource-constrained devices, which justifies why model compression tools are needed. In this paper, we report findings on the effects of quantization and model pruning on speech recognition tasks in noisy conditions. Our results show that model compression has minimal impact on final WER, while signal-to-noise ratio (SNR) has a stronger impact.

Inspecting the Role of Pretrained Transformers in Federated Learning

- Ankur Agarwal — Huawei Noah's Ark Lab
- Mehdi Rezagholizadeh — Huawei Noah's Ark Lab
- Prasanna Parthasarathi — Huawei Noah's Ark Lab

Abstract:
Real world applications of language models entail data privacy constraints when learning from diverse data domains. Federated learning with pretrained language models for language tasks has been gaining attention lately but there are definite confounders that warrants a careful study towards its limitations. The variables we explored are heterogeneity, the trade-off between training time and performance, downstream task, client distribution and sensitivity of the shared model to learning local distributions. Studying these variables is necessary to evaluate whether language models indeed learn to generalize by adapting to the different domains. Towards that, we elaborate different hypotheses over the components in federated NLP architectures and study them in detail with relevant experiments over three tasks: Stanford Sentiment Treebank-2, OntoNotes-5.0 and GigaWord. The experiments with different Transformer inductive biases on the variety of tasks allow us in having a profound understanding of federated learning in NLP. The analysis suggests that regularization due to the ensembling effect may be masquerading as domain adaptation of federated learning in NLP with pre-trained language models.

iRNN: Integer-only Recurrent Neural Network

- Vanessa Courville Huawei Technologies
- Vahid Partovi Nia Huawei Technologies

Abstract :
Recurrent neural networks (RNN), found in many text and speech applications, are made up of complex computational components which make them diffcult to deploy on edge devices. We present a quantization-aware training method for obtaining a highly accurate integer-only recurrent neural network (iRNN). The proposed method enables RNN-based language models to run on edge devices with 2× improvement in runtime, and 4× reduction in model size while maintaining similar accuracy as its full-precision counterpart.

Kronecker Decomposition for GPT Compression

- Ali Edalati — McGill University
- Marzieh Tahaei — Huawei Noah's Ark Laboratory
- Ahmad Rashid — Huawei Noah's Ark Laboratory
- Vahid Partovi Nia — Huawei Noah's Ark Laboratory
- James J. Clark — McGill University
- Mehdi Rezagholizadeh — Huawei Noah's Ark Laboratory

Abstract:

GPT is an auto-regressive Transformer-based pre-trained language model which has attracted a lot of attention in the natural language processing (NLP) domain. The success of GPT is mostly attributed to its pre-training on huge amount of data and its large number of parameters. Despite the superior performance of GPT, this overparameterized nature of GPT can be very prohibitive for deploying this model on devices with limited computational power or memory.

This problem can be mitigated using model compression techniques; however, compressing GPT models has not been investigated much in the literature.

In this work, we use Kronecker decomposition to compress the linear mappings of the GPT-2 model. Our Kronecker GPT-2 model (KnGPT2) is initialized based on the Kronecker decomposed version of the GPT-2 model and then is undergone a very light pre-training on only a small portion of the training data with intermediate layer knowledge distillation (ILKD). Finally, our KnGPT2 is fine-tuned on downstream tasks using ILKD as well.

We evaluate our model on both language modeling and General Language Understanding Evaluation benchmark tasks and show that with more efficient pre-training and similar number of parameters, our KnGPT2 outperforms the existing DistilGPT2 model significantly.

Latency and Accuracy Predictors for Efficient BERT Hardware-aware NAS

- Negin Firouzian McGill University
- Seyyed Hasan Mozafari McGill University
- James J. Clark McGill University
- Warren J. Gross McGill University
- Brett H. Meyer McGill University

Abstract:
With the increased size and complexity of state-of-the-art language models such as BERT, deploying them on resource-constrained devices has become challenging.
Latency-aware Neural Architecture Search (NAS) is an effective solution for finding an efficient implementation of complex models that satisfy hardware limitations.
However, collecting on-device latency and measuring accuracy feedback would significantly slow down the search process, making NAS impractical.
To address this, we propose a low-cost method that models the latency of BERT-based models on a target embedded device, NVIDIA Jetson TX2, and also, it predicts models' accuracy.
As a result, our method removes accuracy measurement and hardware-related delays from the search loop of NAS.
Using a Gradient Boosting regression, our accuracy and latency predictors outperform the state-of-the-art and achieve up to 57x speedup while finding a set of near-optimal models.

Learning Gaussian Restricted Boltzmann Machine using tensorial decompositions

- Bruno Monsia Universite de Montreal
- Alejandro Murua Universite de Montreal

Abstract :
The Boltzmann machine is a probabilistic graphical model used in several practical cases. It involves several parameters and takes a long time to learn them; which does not facilitate their use, in particular in devices with low memory capacities. In this paper, we proposed a tensor formulation for learning Gaussian Restricted Boltzmann machine (GRBM). We precisely used various tensorial decompositions to represent the weight matrix in GRBM. Tensor decompositions (TD) represent the elements of a tensor as tensorial product and allow to obtain fewer elements compared to the initial tensor form. This formulation made it possible to obtain fewer parameters in GRBM without deteriorating its qualities. The results obtained show that the matrix product operator (MPO) outperforms among the range of tensor decompositions used.

Limited-Memory Stochastic Partitioned Quasi-Newton Training

- Paul Raynaud GERAD
- Dominique Orban GERAD

Abstract:
In unconstrained smooth optimization, quasi-Newton methods construct Hessian approximations and are known to perform better than first-order methods.
Among them are partitioned methods, which exploit an objective's partially separable structure, i.e., as a sum of element functions, each depending on a small subset of variables.
Their advantages include parallelism, finer Hessian approximations, and faster convergence than quasi-Newton methods that ignore structure.
In particular, they aggregate the element-function Hessian approximations to produce an accurate Hessian approximation of the loss.
We exploit the weighted sum structure to induce partial separability in training problems.
We propose a partially separable loss function and describe how network architecture impacts partial separability of the problem with the concept of separable layers.
Network partial-separable structure usually leads to smaller neural networks where element functions only touch a fraction of the neural network.
In classification networks, the more classes there are, the smaller the fraction is, which theoretically permits element-function deployment on several computational units.

Mixed representation integer fine-tuning of transformer-based models

- Mohammadreza Tayaranian — Huawei Noah's Ark Lab, McGill University
- Alireza Ghaffari — Huawei Noah's Ark Lab
- Marzieh S. Tahaei — Huawei Noah's Ark Lab
- Mehdi Rezagholizadeh — Huawei Noah's Ark Lab
- Vahid Partovi Nia — Huawei Noah's Ark Lab
- Masoud Asgharian — McGill University

Abstract:

Transformer based models are used to achieve state-of-the-art performance on various deep learning tasks.

Since transformer-based models have large numbers of parameters, fine-tuning them on downstream tasks is computationally intensive and energy hungry.

Automatic mixed-precision FP32/FP16 fine-tuning of such models has been previously used to lower the compute resource requirements. However, with the recent advances in the low-bit integer back-propagation, it is possible to further reduce the computation and memory footprint.

In this work, we explore a novel integer training method that uses integer arithmetic for both forward propagation and gradient computation of linear, convolutional, layer-norm, and embedding layers in transformer-based models.

Furthermore, we study the effect of various integer bit-widths to find the minimum required bit-width for integer fine-tuning of transformer-based models.

We fine-tune BERT and ViT models on popular downstream tasks using integer layers. We show that 16-bit integer models match the floating-point baseline performance. Reducing the bit-width to 10, we observe 0.5 average score drop. Finally, further reduction of the bit-width to 8 provides an average score drop of 1.7 points.

NAS plus Pipeline for High Throughput Edge Inference BERT

- Hung-Yang Chang McGill University
- Seyyed Hasan Mozafari McGill University
- James Clark McGill University
- Brett Meyer McGill University
- Warren Gross McGill University

Abstract:
To meet BERT inference throughput requirements on resource-constrained edge devices, we incorporate pipelining for BERT inference. Also, to find optimal solutions in the 2D design space of accuracy and throughput, we should perform Neural Architecture Search (NAS) over different BERT configurations. In this paper, we incorporate NAS with pipeline inference for DynaBERT models. We show that performing NAS with homogeneous core inference first and then apply pipeline inference (NAS-then-Pipeline) achieves, on average, 56% higher throughput, compared to NAS with homogeneous inference core only (NAS-only).

However, incorporating NAS with the pipeline raises the question of whether to do (1) NAS-then-Pipeline or (2) Pipeline-then-NAS (apply pipeline inference and then perform NAS). NAS-then-Pipeline would help designers to save time since implementing and profiling hardware metrics for pipelined models in a large design space would be complex. However, we show that even though this convention saves time, it results in non-optimal design choices. The result shows that 50% of found POF sets in Pipeline-then-NAS are different from what is found in NAS-then-Pipeline in the throughput-accuracy design space. Moreover, in terms of the relative distance of POF sets in the throughput-accuracy design space, Pipeline-then-NAS shows up to 17% better results than NAS-then-Pipeline. This shows the necessity of applying pipeline inference before NAS for finding optimal throughput-accuracy solutions.

On the Importance of Integrating Curriculum Design for Teacher Assistant-based Knowledge Distillation

- Ibtihel Amara — McGill University
- Maryam Ziaeefard — McGill University
- Brett H. Meyer — McGill University
- Warren Gross — McGill University
- James J. Clark — McGill University

Abstract:
Knowledge distillation (KD), a teacher-student training paradigm has gained a lot of attention thanks to its versatility and easy usability. The intuition that large models are the best teachers is not always guaranteed. In fact, researches has shown that as the capacity gap between the teacher and student networks increases, the harder it is to train the student. Teacher assistant based KD, a sequential and gradual distillation training, has been proposed to mitigate this problem. For classification tasks, understanding the complexity of the data by level of difficulty provides additional guidance to the student when performing distillation. A student network learning gradually form easy to difficult samples can benefit from the known advantages of curriculum learning.
In this work, we show how integrating a curriculum design when performing distillation could contribute to a better training process for the compact student.
We specifically apply the proposed method onto the teacher-assistant -based knowledge distillation. We mainly perform our experimentation with CIFAR-10, CIFAR-100, and ImageNet datasets and show improved accuracy on various architectures.

Persona Controlled Dialogue Prompting

- Runcheng Liu — University of Waterloo
- Ahmad Rashid — University of Waterloo
- Ivan Kobyzev — Huawei Noah's Ark Lab
- Mehdi Rezagholizadeh — Huawei Noah's Ark Lab
- Pascal Poupart — University of Waterloo

Abstract:
We present a novel persona-driven prompt tuning algorithm for dialogue generation. Specifically, we fine-tune a lightweight module (only 5% of the total number of parameters) to generate prompts conditioned on sentences describing a persona, rather than the conversation history. This allows prompting modules fine-tuned for different domains to be stored without too much overhead on constrained edge devices. Experiments on the FoCus open-domain dialogue dataset demonstrate that our method achieves superior performance in terms of both automated metrics and human evaluation.

Quadratic Regularization Optimizer in Low Precision for Deep Neural Networks: Implementation and Numerical Experience

- Farhad Rahbarnia — GERAD, Polytechnique Montréal
- Dominique Orban — GERAD, Polytechnique Montréal
- Nathan Allaire — GERAD, Polytechnique Montréal
- Dominique Monnet — GERAD, Polytechnique Montréal

Abstract:
We consider the training of deep neural networks (DNNs) from an implementation and floating-point arithmetic point of view. First, we propose a Julia framework in which a deterministic method for traditional optimization is automatically ported to the stochastic setting in view of training DNNs. We illustrate the process on the quadratic regularization method R2, a variant of the gradient method with adaptive step size. Secondly, we report numerical experience with stochastic R2 on classification networks using low-precision arithmetic, edge weight scaling, and stochastic rounding.

Quantized One-dimensional Stacked CNN for Seizure Forecasting with Wearables

- Yang Zhang — Polytechnique Montreal
- Yvon Savaria — Polytechnique Montreal
- Shiqi Zhao — Westlake University
- Gonçalo Mordido — Polytechnique Montreal
- Mohamad Sawan — Westlake University
- François Leduc-Primeau — Polytechnique Montreal

Abstract:
Epilepsy is a life-threatening disease affecting millions of people all over the world. Artificial intelligence epileptic predictors offer excellent potential to improve epilepsy therapy. Particularly, deep learning models such as convolutional neural networks (CNN) can be used to accurately detect ictogenesis through deep structured learning representations. In this work, a tiny one-dimensional stacked convolutional neural network (1DSCNN) is proposed based on short-time Fourier transform (STFT) to predict epileptic seizure. The results demonstrate that the proposed method obtains better performance compared to recent state-of-the-art methods, achieving an average sensitivity of 94.44%, average false prediction rate (FPR) of 0.011/h and average area under the curve (AUC) of 0.979 on the test set of the American Epilepsy Society Seizure Prediction Challenge dataset, while featuring a model size of only 21.32kB. Furthermore, after adapting the model to 4-bit quantization, its size is significantly decreased by 7.08x with only 0.51% AUC score precision loss, which shows excellent potential for hardware-friendly wearable implementation.

Quasi-convex floating points optimization

- Matteo Cacciola Huawei Noah's Ark Lab

Abstract:
The properties of gradient descent algorithms in the convex case have been widely studied and several special cases have been analyzed (Polyak 1967), included but not limited to the inexact context (when perturbation are present) and the stochastic setting (Schmidt, Roux, and Bach 2011), (Ram, Nedic, and Veeravalli 2009). However, many real applications involves functions that are not convex, so this hypothesis needs to be relaxed.

Retention of Domain Adaptability in Compressed Neural Networks

- Lulan Shen — McGill University
- Brett Meyer — McGill University
- Warren Gross — McGill University
- James J. Clark — McGill University

Abstract:
It is necessary to develop efficient deep neural networks which can be deployed on edge devices with limited computation resources. However, the compressed networks often execute new tasks in the target domain, which is different from the source domain where the original network is trained, and the current deep domain adaptation methods for computer vision are used to minimize the distribution difference between the two domains do not consider network compression. Hence, we investigate the retained domain adaptability of compressed networks in various domain shifts and discover that the compressed networks lose certain domain adaptability compared with their original networks. In addition, while obtaining a similar size to the compressed networks, the original larger network retains less domain adaptability than the original smaller one.

S^3 Sign-Sparse-Shift Reparametrization for Effective Training of Low-bit Shift Networks

- Xinlin Li — Huawei Noah's Ark Lab
- Vahid Partovi Nia — Huawei Noah's Ark Lab

Abstract:
Shift neural networks reduce computation complexity by removing expensive multiplication operations and quantizing continuous weights into low-bit discrete values, which are fast and energy-efficient compared to conventional neural networks. However, existing shift networks are sensitive to the weight initialization and yield a degraded performance caused by vanishing gradient and weight sign freezing problem. To address these issues, we propose S^3 re-parameterization, a novel technique for training low-bit shift networks. Our method decomposes a discrete parameter in a sign-sparse-shift 3-fold manner. This way, it efficiently learns a low-bit network with weight dynamics similar to full-precision networks and insensitive to weight initialization. Our proposed training method pushes the boundaries of shift neural networks and shows 3-bit shift networks compete with their full-precision counterparts in terms of top-1 accuracy on ImageNet.

Sharpness-Aware Training for Accurate Inference on Noisy DNN Accelerators

- Gonçalo Mordido — Mila & Polytechnique Montreal
- Sarath Chandar — Mila & Polytechnique Montreal
- François Leduc-Primeau — Polytechnique Montreal

Abstract:
Energy-efficient deep neural network (DNN) accelerators are prone to non-idealities which degrade DNN performance at inference time. To mitigate such degradation, existing methods typically add random noise to DNN weights during training to simulate inference on noisy hardware. However, this increases training complexity and often requires knowledge about the target hardware. We demonstrate that applying sharpness-aware training by optimizing for both the loss value and the loss sharpness significantly improves robustness to noisy hardware at inference time. We show superior performance compared to injecting noise during training on multiple architectures, without relying on target hardware measurements or increasing training complexity.

Speeding up Resnet Architecture with Layers Targeted Low Rank Decomposition

- Walid Ahmed — Huawei Technologies
- Habib Hajimolahoseini — Huawei Technologies
- Austin Wen — Huawei Technologies
- Yang Liu — Huawei Technologies

Abstract:
Compression of a neural network can help in speeding up both the training and the inference. In this research, we study applying compression using low rank decomposition on network layers. Our research demonstrates that to acquire a speed up, the compression methodology should be aware of the underlying hardware as analysis should be done to choose which layers to compress. The advantage of our approach is demonstrated via a case study of compressing ResNet50 trained on full ImageNet-ILSVRC2012. We tested on two different hardware systems Nvidia V100 and Huawei Ascend910. With hardware targeted compression, results on Ascend910 showed 5.36% training speedup and 15.79% inference speed on Ascend310 with only 1% drop in accuracy compared to the original uncompressed model.

Standard Deviation-Based Quantization for Deep Neural Networks

- Amir Ardakani — McGill University
- Arash Ardakani — McGill University
- Brett Meyer — McGill University
- James J. Clark — McGill University
- Warren J. Gross — McGill University

Abstract:
Quantization of deep neural networks is a promising approach that reduces the inference cost, making it feasible to run deep networks on resource-restricted devices. Inspired by existing methods, we propose a new framework to learn the quantization intervals (discrete values) using the knowledge of the network's weight and activation distributions, i.e., standard deviation. Furthermore, we propose a novel base-2 logarithmic quantization scheme to quantize weights to power-of-two discrete values. Our proposed scheme allows us to replace resource-hungry high-precision multipliers with simple shift-add operations. According to our evaluations, our method outperforms existing work on CIFAR10 and ImageNet datasets and even achieves better accuracy performance with 3-bit weights and activations when compared to the full-precision models.

Toward Training Neural Networks with a Multi-Precision Quadratic Regularization Algorithm

- Dominique Monnet GERAD, Polytechnique Montreal
- Dominique Orban GERAD, Polytechnique Montreal

Abstract:
We introduce SMPR2, a stochastic multi-precision quadratic regularization algorithm for training neural networks. SMPR2 is a stochastic gradient-descent algorithm with adaptive step size that dynamically selects the floating-point format used to evaluate the sampled objective and gradient in hopes to perform as many operations and evaluations as possible in low precision, thereby saving computational time and energy expended.

Towards Finding Efficient Students via Blockwise Neural Architecture Search and Knowledge Distillation

- Hang Zhang McGill University
- Seyyed Hasan Mozafari McGill University
- James Clark McGill University
- Brett Meyer McGill University
- Warren Gross McGill University

Abstract:
The inference of pre-trained attention-based structures, such as BERT and DistilBERT, requires huge resources, and thus, is difficult to be deployed on edge devices. Neural Architecture Search (NAS) is an efficient solution to find compressed sub-networks in multi-objective optimization. However, the solution is not accurate if not all the candidates are fairly and fully trained. Therefore, we propose a novel Differentiable Blockwise Neural Architecture Search (DBNAS) with Knowledge Distillation (KD), to automatically find the optimal student models under different constraints. We implement our method (DBNAS) to compress BERT and DistilBERT. Afterward, we report the accuracy and sizes of the compressed models DBNAS_BERT and DBNAS_DistilBERT. Compared with the corresponding teacher models, they reduces the model size by 90%, while maintaining more than 95% performance on the SST-2 task.

Training Acceleration of Low-Rank Decomposed Networks using Sequential Freezing and Rank Quantization

- Habib Hajimolahoseini — Huawei Technologies
- Walid Ahmed — Huawei Technologies
- Yang Liu — Huawei Technologies

Abstract:
Low Rank Decomposition (LRD) is a technique used for reducing the computational complexity of deep learning models. However, it will make the models deeper as each layer is decomposed into a sequence of smaller layers, which may not lead to training acceleration if the decomposition ranks are not small enough. In this paper, we propose two techniques that accelerate the training of the LRD decomposed models, including rank quantization and sequential freezing. Experiments on both convolutional and transformer-based models show that these techniques can improve the model throughput up to 60% during training when combined together while preserving the accuracy close to that of the original models.

Weighted Group L0-norm Constraint for Sparse Training

- Michael Metel Huawei Noah's Ark Lab

Abstract:
Motivated by structured sparsity for neural network training, we study a weighted group l0-norm constraint and present the Euclidean projection operator and normal cone of this set.