EXPO 2019

VANCOUVER

DEC 8th, 2019

NEURAL INFORMATION
PROCESSING SYSTEMS

2019 SPONSOR EXPO

Presented by Neural Information Processing Systems

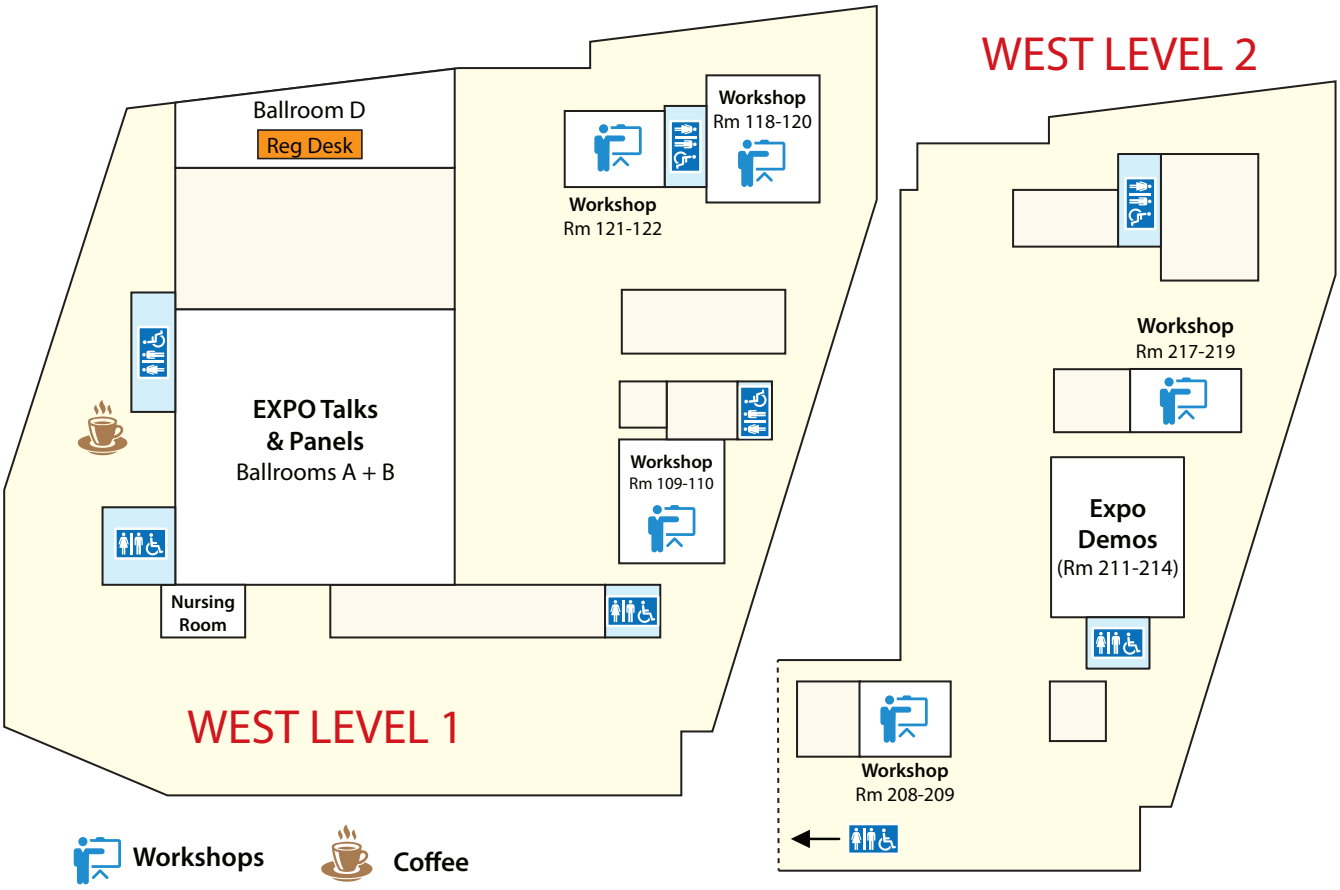
Now in its second year, The NeurIPS EXPO presents work in AI ML done in an industrial setting. The EXPO offers a unique opportunity for a firsthand perspective on how industry is translating the power of the tools developed in this field. As a bridge between academia and industry, the NeurIPS EXPO is the premier forum for issues in artificial intelligence in a real-world setting.

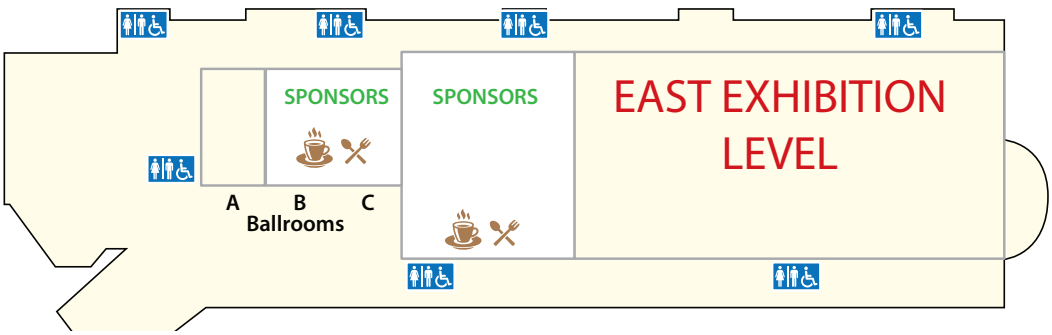
This year our curatorial team focused on bringing a diversity of topics to the EXPO with a focus on promoting the intellectual currency that is at the heart of all research, both academic and industrial.

We hope you will find a lot to explore and even more that inspires.

Neil Lawrence
Organizing Committee Chair

Table Of Contents





WORKSHOPS - 9:00 AM - 1:00 PM

West Level 1 - Rm 109 - 110

Causal Inference & Reinforcement Learning: Making The Right intervention

The successes of predictive modelling encourage many organizations to leverage machine learning models in their decision making. Examples include the design of marketing, and user relationship strategies. Such analyses typically pose the following challenges: (1) the need to distinguish between events that cause outcomes from those that merely correlate; (2) the need to optimize the medium to long term results, instead of focussing on short term gains alone. The two areas of ML research that are particularly relevant here are Causal Inference (CI) and Reinforcement learning (RL).

There have been astonishing successes in the fields of RL and CI over recent years. However, applications in industry are lagging behind. One key challenge is to choose the right analytical framing to make the approach tractable, whilst also solving the real-world problem. Moreover, existing applications often assume the ability to perform experiments, which is infeasible for many questions of interest in industry. Models have to rely on historically observed data alone. Finally, the lack of mature open source software makes attempts to apply RL and CI in fast-paced environments risky. In this workshop, we will provide an overview of QB's experiences implementing these methods. Our objective is to facilitate a discussion about practical challenges in real-world applications: 1. Introduction 2. Framework for shaping real-world strategy problem into ML problem 3. CI in practice with focus on Bayesian Networks 4. RL in practice with focus on offline learning 5. Closing remarks and discussion.

At QB, we're passionate about diversity. We are actively working to achieve more equality in tech in all dimensions including gender, race, nationality, and sexual orientation. It is important to us that the presenters of this workshop will reflect our diversity. Some recent articles we wrote on gender diversity are tiny.cc/1zlucz and tiny.cc/p1lucz, and we are proud sponsors of WiML.

West Level 1 - Rm 118 - 120

Multi-modal Research To Production

The content at Facebook and more broadly continues to increase in diversity and is made up of a number of modalities (text, audio, video, etc..). For example, an Ad may contain multiple components including image, body text, title, video and landing pages. Even an individual component may bear multimodal traits, for instance, a video contains visual and audio signals, a landing page is composed of images, texts, HTML sources, etc. This workshop will dive into a number modalities such as computer vision (large scale image classification and instance segmentation) and Translation and Speech (seq-to-seq Transformers) from the lens of taking cutting edge research to production. Lastly, we will also walk through how to use the latest APIs in PyTorch to take eager mode developed models into graph mode via Torchscript and quantize them for scale production deployment on servers or mobile devices. Libraries used include:

- PyTorch - a popular deep learning framework for research to production.
- Classy Vision - a newly open sourced PyTorch framework developed by Facebook AI for research on large-scale image and video classification. Classy Vision allows researchers to quickly prototype and iterate on large distributed training jobs. Models built on Classy Vision can be seamlessly deployed to production, and Classy Vision powers the next generation of classification models in production at Facebook.
- Detectron2 - the recently released object detection library built by the FAIR computer vision team. We will articulate the improvements over the previous version including: 1) Support for latest models and new tasks; 2) Increased flexibility, to enable new computer vision research; 3) Maintainable and scalable, to support production use cases.
- Fairseq - general purpose sequence-to-sequence library, can be used in many applications, including (unsupervised) translation, summarization, dialog and speech recognition.

West Level 2 - Rm 217 - 219

Challenges, Advances and Opportunities for Machine Learning Algorithms and Infrastructure at Alibaba

The last decade has witnessed great success of machine learning in multiple domains, including speech recognition, image classification, video content analysis, search, computational advertisement and finance. Despite the amazing progress, we also run into many challenges when coming to practical applications of machine learning technologies.

In this workshop, we would like to share some of the key developments of machine learning at Alibaba Group that explicitly address limitations of the existing machine learning techniques, both on algorithms and infrastructure. More specifically, we will discuss and share our practical experiences of (a) A Data-driven Approach for GPU Resource Optimization over A Large-scale AI Cluster; (b) FusionStitching: Boosting Execution Efficiency of Memory Intensive Computations for DL Workloads; (c) AliGraph: A Comprehensive Graph Neural Network Platform; (d) Shared Learning: Cross-Organization Joint Machine Learning without Compromising Privacy or Data Security; (e) Machine Learning for All-Inclusive Finance; (f) Extremely Large Scale Image Classification with Long- tailed Noisy Data. These talks almost cover the biggest challenges and most recent advancements in our core departments and we hope that this workshop will nourish further development in these areas.

WORKSHOPS - 9:00 AM - 1:00 PM

West Level 1 - Rm 121 - 122

From Research To Industrial NLP At Baidu

This workshop will propose recent advances of NLP research and industrial applications at Baidu. We will first give an introductory talk on Baidu's NLP combined with our open source deep learning platform – PaddlePaddle. The following talks will go depth into specific NLP topics. 1. Overview of NLP at Baidu. This talk will give an overview of NLP at Baidu, including researches and products. 2. PaddleNLP. This talk will introduce our open source deep learning platform - PaddlePaddle, followed by frameworks and models customized for NLP tasks. 3. ERNIE (Enhanced Representation through kNowledge IntEgration), a brand-new natural language understanding framework. Based on this framework, Baidu also open sourced a pre-trained language understanding model which achieved state-of-the-art results and outperformed BERT and the recent XLNet in 16 NLP tasks in both Chinese and English. 4. Machine Reading comprehension. This talk will describe our research efforts on machine reading comprehension (MRC) to deal with the real challenges when applying MRC technologies to the production of open-domain question answering in Baidu Search, including multi-passage MRC, knowledge-enhanced MRC and improving the robustness and generalization of MRC models, that is the winner solution at MRQA 2019. 5. Machine Translation. This talk will introduce the technologies we used in this year's WMT evaluation campaign, in which our system ranked the 1st in Chinese-English newswire translation. We will also introduce our efforts on simultaneous machine translation. Recently, we released a speech-to-speech simultaneous machine translation system, achieving comparable performance to human interpreters in delivering high-quality simultaneous speech translation with low latency.

West Level 2 - Rm 208 - 209

Real world reinforcement learning with Vowpal Wabbit

Vowpal Wabbit (https://vowpalwabbit.org) is an open source machine learning library, extensively used by industry, and is the first public terascale learning system (https://arxiv.org/abs/1110.4198). It provides fast, scalable machine learning and has unique capabilities such as learning to search, active learning, contextual memory, and extreme multiclass learning. It has a focus on reinforcement learning and provides production ready implementations of Contextual Bandit algorithms. Vowpal Wabbit sees significant innovation as a research to production vehicle for Microsoft Research.

Come and learn about reinforcement learning, Vowpal Wabbit, and applying contextual bandits to problems using Vowpal Wabbit.

WORKSHOPS - 2 PM - 6 PM

West Level 2 - Rm 208 - 209

Hands-On Workshop: High-Performance Inference with Habana Goya AI Processor

Recently, the ML community has seen interest spike in the area of AI processors (AIPs), computational engines specifically designed for Machine Learning workloads. The workshop addresses practical aspects of implementing AIPs.

Part One: 2.5-hour hands-on session provides attendees with in-depth information on how to implement a high-performance inference system employing the Habana Goya Inference Processor. After sharing the foundations of the Goya AI processor hardware and software, we will demonstrate how to use the AI processor to solve the most common and computationally extensive inference tasks. We will demonstrate the entire inference process:

1. Starting from a trained model from any of the common frameworks;
2. Converting the trained model into Habana IR (Intermediate Representation) model --Graph Optimization techniques – encourage/discourage fusion, typecasting, etc.
3. Quantization: explain the concept, discuss best known methods and considerations;
4. Optimizing for performance or accuracy: identifying best compromise;
5. Debugging and profiling for bottlenecks;
6. Demonstrating how the Habana TPC Kernel library can be extended with User Specific Kernels.

Part Two: 30-minute talk by a Facebook AI hardware engineer on the industry's new AI hardware standards.

Audience/Attendee Experience: Provide attendees with practical "how to" information for implementing AI workloads with AIPs. We will conduct a live demo and tutorial on a Goya-equipped desktop (lab machine), review alternative approaches to different inference steps, pros and cons of each approach. We will provide attendees with access to the lab machine; participants will run inferences on supplied models, or they may run workloads on models from the internet. Advanced users will write their own kernels and run them on the platform. Users should bring their laptop with an SSH or VNC or NoMachine Client installed, to be able to connect to the lab machine.

WORKSHOPS - 2 PM - 6 PM

West Level 1 - Rm 109 - 110

AI & Machine Learning in Finance: Applications and Opportunities

In this workshop three companies, Two Sigma, Hudson River Trading, and JP Morgan, will discuss how artificial intelligence is transforming the world of financial services, markets, and investing. Finance is data-rich field, presenting a variety of opportunities and challenges for the AI practitioner. The presenters will cover the nuances of using this data from their unique perspectives, including the process of building models, understanding them, deploying them, and the ethical issues to address when doing so. In addition to predictive modelling, finance has rich control and optimization problems where risk, return, and compliance must be carefully balanced. Topics we will cover include portfolio optimization and reinforcement learning.

Each company will give a separate 45-60 min talk, followed by a joint panel discussion with representatives from each company.

West Level 1 - Rm 121 - 122

Tensor Networks with TensorNetwork

Tensor networks are high-powered tools developed in physics and now finding new use in machine learning. Here we'll showcase TensorNetwork, a new open source library built on TensorFlow for doing tensor network computations. As tensor networks grow in popularity in the ML world, this library will be an invaluable resource for researchers and practitioners alike.

For more than fifteen years, physicists have been using tensor networks to study the properties of complex quantum systems. Fundamentally, these tensor networks are all about efficient storage and manipulation of correlations in a sparse representation of the data, built using nothing more complicated than matrix multiplication. The machine learning community is starting to catch on, and a growing community of researchers are exploring the applications of tensor networks to deep learning. A PhD in quantum physics is not required to understand these concepts.

Until now, there has not been a software library for tensor networks that has a low barrier to entry and is also powerful enough for state-of-the-art computations. That role is being filled by our new open-source library, TensorNetwork.

In this workshop, you'll:

- **Learn what a tensor network is.**
- **See examples of a tensor network in action.**
- **Get introduced to an easy-to-use open-source API for tensor networks using Python, with NumPy, TensorFlow, JAX, and PyTorch backends.**
- **Leave excited and ready to use tensor networks for your own applications!**

West Level 1 - Rm 118 - 120

Responsible and Reproducible AI

This workshop on Responsible and Reproducible AI will dive into important areas that are shaping the future of how we interpret, reproduce research, and build AI with privacy in mind. We will cover major challenges, walk through solutions, and finish each talk with a hands on tutorial.

Reproducibility: As the number of research papers submitted to arXiv and conferences skyrockets, scaling reproducibility becomes difficult. We must address the following challenges: aid extensibility by standardizing code bases, democratize paper implementation by writing hardware agnostic code, facilitate results validation by documenting "tricks" authors use to make their complex systems function. To offer solutions, we will dive into tools PyTorch Hub and PyTorch Lightning which are used by some of the top researchers in the world to reproduce the state of the art.

Interpretability: With the increase in model complexity and the resulting lack of transparency, model interpretability methods have become increasingly important. Model understanding is both an active area of research as well as an area of focus for practical applications across industries using machine learning. To get hands on, we will use the recently released Captum library that provides state-of-the-art algorithms to provide researchers and developers with an easy way to understand the importance of neurons/layers and the predictions made by our models.`

Private AI: Practical applications of ML via cloud-based or machine-learning-as-a-service platforms pose a range of security and privacy challenges. There are a number of technical approaches being studied including: homomorphic encryption, secure multi-party computation, trusted execution environments, on-device computation, and differential privacy. To provide an immersive understanding of how some of these technologies are applied, we will use the CrypTen project which provides a community based research platform to take the field of Private AI forward.

DEMONSTRATIONS - 9 AM - 6 PM
WEST LEVEL 1 - ROOM 211-214

#1 AliGraph: A Comprehensive Graph Neural Network Platform

An increasing number of machine learning tasks require dealing with large graph datasets, which capture rich and complex relation- ship among potentially billions of elements. Graph Neural Network (GNN) becomes an effective way to address the graph learning problem by converting the graph data into a low dimensional space while keeping both the structural and property information to the maximum extent and constructing a neural network for training and referencing. However, it is challenging to provide an efficient graph storage and computation capabilities to facilitate GNN training and enable development of new GNN algorithms. In this paper, we present a comprehensive graph neural network system, namely AliGraph, which consists of distributed graph storage, optimized sampling operators and runtime to efficiently support not only existing popular GNNs but also a series of in- house developed ones for different scenarios.

The system is currently deployed at Alibaba to support a variety of business scenarios, including product recommendation and personalized search at Alibaba's E-Commerce platform. By conducting extensive experiments on a real-world dataset with 492.90 million vertices, 6.82 billion edges and rich attributes, Ali- Graph performs an order of magnitude faster in terms of graph building (5 minutes vs hours reported from the state-of- the-art PowerGraph platform). At training, AliGraph runs 40%- 50% faster with the novel caching strategy and demonstrates around 12 times speed up with the improved runtime. In addition, our in-house developed GNN models all showcase their statistically significant superiorities in terms of both effectiveness and efficiency (e.g., 4.12%–17.19% lift by F1 scores)

#2 PaddlePaddle: An industry-Grade End-to-End Deep Learning Platform

PaddlePaddle is a deep learning platform developed at Baidu. We will demonstrate the core technology behind PaddlePaddle and a range of applications built on its top. The main features to be demonstrated include its "easy to use" APIs, powerful inference engine and tool chain for fast inference and deployment, its one-stop learning environment, and EZDL Pro, an integrated user interface for industrial deep learning applications. We will also demonstrate our semantic representation model ERNIE and simultaneous translation developed using PaddlePaddle.

#3 Deep Learning-Based End-to-end Automatic Contouring and Automated Radiation Therapy Treatment Planning System

For patients suffering from cancer, surgery may not be a feasible solution and radiation therapy is a common alternative treatment method. Radiation therapy is the treatment of cancers via the use of targeted high-energy radiation, which may be delivered externally or internally. Performing such treatment involves creating a treatment plan for implementing the radiation dose onto the patient, which requires radiologists to manually delineate the boundaries of patients' tumors and organs-at-risks, and iteratively adjust the constraints and weights of the dose objective functions relating to the delineated tumors and organs-at-risks. However, the process is mostly manual, making it time-consuming and inconsistent. This is undesirable considering that a poorly designed treatment plan could adversely affect the function of healthy organs if the organs are overdosed, or unsuccessfully remove tumors if the tumors are underdosed.

Therefore, this demonstration shows how deep learning can be applied in the field of radiation therapy, specifically the creation of a radiation therapy treatment plan from only patient medical images and prescription information from healthcare professionals. Attendees will be able to interact with the system through an easy-to-use interface, allowing them to witness the treatment planning process and observing and interacting with the automated contouring, treatment plan and patient radiation dose obtained via the treatment planning process. The purpose of the demonstration is to show attendees that deep learning is ready to be deployed for radiation therapy, creating a positive difference by saving lives and freeing up time for radiologists to enhance their ability to provide healthcare for patients.

#4 Tensor Networks with TensorNetwork

Tensor networks are high-powered tools developed in physics and now finding new use in machine learning. Here we'll showcase TensorNetwork, a new open source library built on TensorFlow for doing tensor network computations. As tensor networks grow in popularity in the ML world, this library will be an invaluable resource for researchers and practitioners alike.

#5 Showcasing How Extremely Large Language Models Can Be Trained Using New AI Processor Technology From Graphcore

Language model pre-training has been shown to be effective for improving many natural language processing tasks. In the last two years we have seen major new breakthroughs in extremely large langauge models including Google's BERT - Deep Bidirectional Transformers for Language Understanding (Devlin Chang Lee Toutanova 2018) which applies the bidirectional training of Transformer, a popular attention model, to language modelling and OpenAI's GPT2 (Radford Wu Child Luan Amodei

Sutskever 2019) a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

In this demo, Graphcore team will showcase how recent, extremely large language models can be trained using new AI processor technology from Graphcore, the techniques employed and the state of the art results achieved, bringing further progress to this field.

#6 Industry Leading Performance Per Watt With Intel® Nervana™ Neural Network Processor for Inference

Language model pre-training has been shown to be effective for improving many natural language processing tasks. In the last two years we have seen major new breakthroughs in extremely large langauge models including Google's BERT - Deep Bidirectional Transformers for Language Understanding (Devlin Chang Lee Toutanova 2018) which applies the bidirectional training of Transformer, a popular attention model, to language modelling and OpenAI's GPT2 (Radford Wu Child Luan Amodei Sutskever 2019) a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

In this demo, Graphcore team will showcase how recent, extremely large language models can be trained using new AI processor technology from Graphcore, the techniques employed and the state of the art results achieved, bringing further progress to this field.

#7 Efficient Deep Learning computing with Intel®Nervana™ Neural Network Processor for Training

Intel® Nervana Neural Network Processor for Training™ (Intel® Nervana™ NNP-T) is designed to maximize efficiency in power usage, memory and communication. The NNP-T focuses on increasing compute utilization for AI training needs instead of just peak TOPS. The NNP-T allocates the die area judiciously between MACs, local and off-die memory, and communication (both on-die and off-die) in order to create a device with a large amount of compute that could be kept fed with data on problem sizes both large and small. The NNP-T keeps the compute fed and maximizes power efficiency by retaining data locally and reusing it as much as possible.We will demonstrate end-to-end training of an image classification workload, ResNet50, using a popular deep learning framework.

#8 Advanced Hyperparameter Optimization Methods

Modelers typically introduce hyperparameter optimization (HPO) toward the end of their model development process. Confining HPO to this "last mile" misses a variety of opportunities to applying HPO techniques throughout the modeling process that can boost modeling outcomes. In this demo, we will provide a prescriptive guide for HPO that includes techniques for using HPO throughout the modeling process and real-world examples of the impact of this approach. In particular, this demo will include ways to apply advanced HPO methods to address:

- Metric definition, selection and optimization
- Data and feature transformation parameter tuning
- Model search and selection with conditionality of parameters
- Deep learning model convergence with automated early stopping
- Long training cycles by running more efficient algorithms in parallel

For each of these use cases, the team will share more detail around the research underpinning for these HPO techniques as well as an applied use case that contextualizes how to incorporate it into the modeling process. The goal is for this combination to offer useful information for modelers, modeling team leaders and modeling platform engineers.

#9 A.I.-assisted music generation

The demo will demonstrate a selection of Sony tools that allow music creators to compose new songs within seconds using a suite of interactive generative models for music. By covering the whole spectrum of music generation, from score-based composition to explorative sound design, these various models aim at providing tools to make music production accessible and playful for a wide audience. These tools are integrated into modern music production software in otder to naturally enrich artists' workflows.

The demo will show the following plugins and discuss their interconnection: 1) Nonoto - a model-agnostic web interface for interactive music composition by inpainting, 2) DrumNet - a conditional drum pattern generator, 3) PlanetDrums - a kick drum sound generation tool, 4) DeepBach - a model able to learn and generate music in the style of J.S. Bach.

TALKS & PANELS - MORNING 9 AM - 12:05 PM

WEST BALLROOMS A + B

9:10 am
Intel® Nervana™ NNP: Domain Specific Architectures For Inference & Training

Domain specific accelerators provide more efficient computation for increasingly complex models. With the Intel® Nervana™ NNP for Inference (NNP-I), we've designed inference compute engines and coupled them with general purpose Intel CPU cores on a single die to enable fast in-lining of deep learning and non-deep learning compute. This can unlock opportunities for heterogeneous algorithms for researchers. As the field moves towards training larger models, the NNP for Training (NNP-T) is designed with 4x 2D-mesh networks to connect our tensor cores and scale models across systems. This talk will cover how we designed (1) flexibility without sacrificing performance with NNP-I, (2) scalability with NNP-T for the most complex models, and (3) software stacks to enable programmability through standard frameworks.

9:35 am
Innovative Approaches In Training Large Scale Language Models

Language model pre-training has been shown to be effective for improving many natural language processing tasks. In the last two years we have seen major new breakthroughs in extremely large langauge models including Google's BERT - Deep Bidirectional Transformers for Language Understanding (Devlin Chang Lee Toutanova 2018) which applies the bidirectional training of Transformer, a popular attention model, to language modelling and OpenAI's GPT2 (Radford Wu Child Luan Amodei Sutskever 2019) a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

In this talk, Graphcore SVP Ola Torudbakken will explain how recent, extremely large language models can be trained using new AI processor technology from Graphcore, the techniques employed and the state of the art results achieved, bringing further progress to this field.

10:00 am
Accelerating Deep Learning At Wafer Scale

Presented by Cerebras Systems

Long training times are the biggest single factor constraining innovation in Deep Learning. To speed up training, significant progress has been made in improving the performance at the single device level (via optimized kernels for common models) and in scaling out training jobs (via new methods for large-batch training). But the former makes it harder to explore innovative research ideas, as performance becomes brittle to even minor changes in models, because those optimized kernels are rigid. And the latter requires extensive model tuning from machine learning researchers and hard to reproduce.

In this talk we will discuss Cerebras approach to speed up training and reduce time to solution with the Cerebras Wafer Scale Engine - the largest chip in the world. It provides cluster-scale resources on a single chip with full utilization for tensors of any shapes, fat, square and thin, dense and sparse, enabling researchers to explore novel network architectures and optimization techniques at any batch sizes.

10:55 am
Interpretability - Now What?

Presented by Google

In this talk, we hope to reflect on some of the progress made in the field of interpretable machine learning. We will reflect on where we are going as a field, and what are the things we need to be aware and be careful as we make progress. With that perspective, We will then discuss some recent work:

1) sanity checking popular methods and
2) developing more lay person-friendly interpretability methods.

INSIGHTS & QUESTIONS

GRAPHCORE

TALKS & PANELS - MORNING 9 AM - 12:05 PM

WEST BALLROOMS A + B

11:20 am

Machine Learning for Perception at Cruise

This talk takes an exciting look into how Cruise uses Machine Learning to solve one of the hardest problems of our generation, perception for autonomous vehicles. We will explore some of the problems machine learning is well suited for as well as some of the extraordinarily difficult challenges still left unsolved. We will also take a walk through the challenges and advantages of driving in San Francisco, and how Cruise leverages this data experience to accelerate learning and build a competitive advantage in this highly competitive industry.

11:45 am

Federated Learning in Healthcare

Machine learning promises to revolutionize many industries, but its application is restricted to areas where there is enough data to train useful models. Often, the barriers to data acquisition are not technological but issues such as privacy, trust, regulatory compliance, and intellectual property. This is especially the case in healthcare, where patients and consumers expect privacy with respect to personal information and where organizations want to protect the value of their data and are also required to follow regulatory laws such as HIPAA in the United States and the GDRP in the Eurozone. Federated learning, which provides the ability to share a model without sharing the data used to train it, has the potential to address these concerns. This talk will explore the application of Federated Learning to problems in healthcare. We'll examine two applications specifically: Federated Mobile Learning, which takes place in the consumer space where data is located on a user's personal device, and Federated Cloud Learning, which focuses on business applications in which internal company data cannot be shared with other entities or even within an organization itself. We will discuss the challenges faced by Federated Learning, such as the privacy guarantees a Federated Learning approach can make, and we will look at two of the most common techniques: differential privacy and homomorphic encryption. We will also address some of the engineering and theoretical challenges of Federated Learning. Finally, we will conclude that Federated Learning is a viable approach to machine learning in the healthcare space that can address patient, business, and regulatory concerns with the application of privacy-preserving techniques.

TALKS & PANELS - AFTERNOON 1:45 - 6:00 PM

WEST BALLROOMS A + B

1:40 pm

Explainable Reinforcement Learning

XAI refers to the application of AI in such a manner that any automated process or output can be well-understood and trusted by a human expert. RL is a sequential decision-making framework for modelling agents who learn to act in an uncertain environment. Learning is based on environmental feedback in the form of a numerical control signal intended to direct desired behaviour, in contrast to other machine learning methods that explicitly label correct / incorrect behaviours. Robust industrial application has found sporadic success, primarily in decision support, e.g. expert recommendation in healthcare, media, etc. In order to facilitate wider industrial adoption of RL techniques, it is vital that RL models be easily explained in the language of clients and domain experts.

INSIGHTS & QUESTIONS

cruise

SPORT LOGiQ

TALKS & PANELS - AFTERNOON 1:45 - 6:00 PM

WEST BALLROOMS A + B

2:15 pm
Private Federated Learning

Access to significant amounts of data and computational hardware in the cloud has enabled fitting accurate models for image classification, text translation, sentiment classification, and other predictive tasks possible with accuracy that was previously completely infeasible. These models are typically deployed on embedded devices with limited computation and communication ability — remote sensors, fitness monitors — making fitting large scale predictive models both computationally and statistically challenging.

Federated Learning is a new approach that is picking up steam in the machine learning community as a way to improve global models in the cloud by leveraging on-device training on user data. At WWDC 2019, Apple announced Private Federated Learning by combining federated learning with differential privacy. We have started to use this technology in iOS 13 for a variety of use cases including QuickType keyboard, FoundIn Apps, and Personalized "Hey Siri". In this talk, we will provide more details around this.

2:40 pm
How the Gaming Industry is Driving Advances in AI Research

Many recent advances in deep reinforcement learning and artificial intelligence have stemmed from video games. In this session, we'll explore a brief history of this relationship and how Unity is pushing the boundaries of AI research with the Obstacle Tower Challenge and Animal-AI Olympics Competition. We'll also show how Unity is leveraging cutting-edge research to solve gaming's biggest challenges with the Unity Machine Learning Agents Toolkit, one of the most popular open source toolkits for deep learning.

3:35 pm
Challenges in Analyzing Two-Sided Market and Its Application on Ride Sharing Platform

In this talk, we will introduce a general analytical framework for large scale data obtained from two-sided markets, especially ride-sharing platforms like DiDi. This framework integrates classical methods including Experiment Design, Causal Inference and Reinforcement Learning, with modern machine learning methods, such as Graph Convolutional Models, Deep Learning, Transfer Learning and Generative Adversarial Network. We aim to develop fast and efficient approaches to address five major challenges for ride-sharing platform, ranging from demand-supply forecasting, demand-supply diagnosis, MDP-based policy optimization, A-B testing, to business operation simulation. Each challenge requires substantial methodological developments and inspires many researchers from both industry and academia to participate in this endeavor. Based on our preliminary results for the policy optimization challenge, we receive the Daniel Wagner Prize for Excellent in Operations Research Practice in 2019. All the research accomplishments presented in this talk are joint work by a group of researchers at Didi Chuxing and our collaborators.

4:00 pm
XPRIZE & AI for Good: Using AI in Incentivized Prizes to Solve Today's Challenges

XPRIZE runs large-scale incentive competitions to enable breakthroughs and accelerate the future. The initial Ansari XPRIZE helped launch today's private space industry and over the last 20 years XPRIZE has run & awarded prizes in ocean mapping, education, and transforming carbon emissions. Since 2016 XPRIZE has operated the IBM Watson AI XPRIZE competition. This is a $5 million global competition challenging teams to apply AI (ML, DL, NLP, etc.) to different challenges facing humanity. Currently the competition fields 30+ teams from 10 different countries working on a wide-array of AI-based solutions attempting to solve issues in mental health, education, environment, etc. XPRIZE's panel + talk will focus on a select few of the 10 semifinalists (no more than 3 + moderator) in the AI XPRIZE competition, who will be publicly announced and showcased during the week of NeurIPS. This will be the first opportunity to see who the semifinalists are and understand the groundbreaking work they are doing.

The purpose of the talk will center around AI for Good (including the future of it), the challenges around AI for Good (structured data), what it takes to operate a start-up in this new space, and the novel approaches of the companies in the AI XPRIZE. In short - this panel will help the audience understand how AI for Good solutions can move from theoretical to practical. XPRIZE's talk/panel will also feature a preview of how XPRIZE plans to make AI the center of its future competitions, including those related to solving the challenges facing the Amazon, as well as combating wildfires.

INSIGHTS & QUESTIONS

4:35 pm
Trust And AI – Addressing AI Risks And Governance

While there are conversations around the dangers of unethical AI, there has not been enough progress made in translating these concerns to an actionable approach. The ecosystem is still focused on "what" the ethics are, and not making enough progress on "how" to put checks and balances in place to know if an AI system is abiding by those ethics and can be trusted.

If AI is to be trusted by the stakeholders it interacts with, then clear and comprehensive oversight procedures need to be developed. Regulatory compliance testing, performance auditing, quality management standards, and trusted third-party assessment need to be established.

Developing these requires an interdisciplinary approach, bringing together technical, legal, assurance and audit expertise. EY is proactively addressing this 'how' by working with a wide range of stakeholders to define practical guidance and strategies for building technologies that will advance society.

This talk will cover our latest data on the topic of ethical AI, thinking on addressing what questions need to be asked and methodologies put in place for AI practitioners to design and implement a trusted system, and how to monitor if it is acting ethically.

5:00 pm
Bridging the Gap Between
Effective Algorithms and Medical Adoption

Deep learning in healthcare presents unparalleled opportunities to serve humanity through streamlining medical workflows, predicting patient outcomes, and guiding treatment decisions. But unlike enterprise or consumer products, clinical AI solutions face enormous regulatory, ethical, and legislative challenges.

We will discuss the challenges of deploying healthcare algorithms with respect to existing privacy laws and redefining what constitutes medical data. Hear the perspective from industry, academia, and government healthcare experts as they examine integration of social determinants, genomics, and lifestyle habits into models. These new opportunities raise legitimate concerns about trust, adoption, and unanticipated consequences.

5:35 pm
Computational Creativity
for Music and Gastronomy Applications

Sony R&D has been creating novel Machine Learning tools for music creators for a long time. This has led to the first music album composed with artificial intelligence -- the ground-breaking album "Daddy's Car" (2015). ML tools for style imitation under user constraints allow music creators to capture their own style or the style of somebody else (such as the Beatles) and use this to create novel melodies and lead sheets. Similar systems have been applied to model and extend the style of Bach (DeepBach) and rhythm generation. These technologies are increasingly deployed in the real world in Digital Audio workstation tools that allow creators to interact with AI generated melodies and rhythms and thereby create new music in line with the overall vision of Sony for AI as an augmentation tool for human creativity. Recently, Sony R&D is applying similar principles to a new domain: gastronomy. Gastronomy is a challenging domain for Machine Learning because it requires to deal with data from various sources including Natural Language but also images and olfactory sensor data. Moreover, there is a strong connection with Robotics and Machine Learning for manipulation of food.

The talk will introduce the long-standing research on Machine Learning for music production and composition. It will also discuss how Sony R&D is extending these concepts and applies them to the domain of gastronomy.

INSIGHTS & QUESTIONS

Anthem.a

SONY®

EXPO SPONSORS