NEURAL INFORMATION PROCESSING SYSTEMS

Online Virtual Sponsor Expo

Sunday December 6th

Amazon | Science

Alibaba

Neural Magic

Facebook

Scale AI

IBM

Sony

Apple

Quantum Black

Benevolent AI

Zalando

Kuaishou

Cruise

Microsoft

Ant Group | Alipay

Wild Me

Deep Genomics

Netflix Research

Google Research

CausaLens

Hudson River Trading

Scikit-learn and Fairness, Tools and Challenges

5 AM PST (1 PM UTC)

Speaker: Adrin Jalali

Fairness, accountability, and transparency in machine learning have become a major part of the ML discourse. Since these issues have attracted attention from the public, and certain legislations are being put in place regulating the usage of machine learning in certain domains, the industry has been catching up with the topic and a few groups have been developing toolboxes to allow practitioners incorporate fairness constraints into their pipelines and make their models more transparent and accountable. AIF360 and fairlearn are just two examples available in Python.

On the machine learning side, scikit-learn has been one of the most commonly used libraries which has been extended by third party libraries such as XGBoost and imbalanced-learn. However, when it comes to incorporating fairness constraints in a usual scikit-learn pipeline, there are challenges and limitations related to the API, which has made developing a scikit-learn compatible fairness focused package challenging and hampering the adoption of these tools in the industry.

In this talk, we start with a common classification pipeline, then we assess fairness/bias of the data/outputs using disparate impact ratio as an example metric, and finally mitigate the unfair outputs and search for hyperparameters which give the best accuracy while satisfying fairness constraints.

This workflow will expose the limitations of the API related to passing around feature names and/or sample metadata in a pipeline down to the scorers. We discuss certain workarounds and then talk about the work being done to address these issues and show how the final solution would look like. After this talk, you will be able to follow the related discussions happening in these open source communities and know where to look for them.

The code and the presentation will be publicly available on github. The speaker, Adrin Jalali, is a core maintainer of scikit-learn and a contributor to both fairlearn and aif360.

The challenges and latest advances in the field of causal AI

5 AM PST (1 PM UTC)

Speaker: Darko Matovski

The current state of the art in machine learning relies on past patterns and correlations to make predictions of the future. This approach can work in static environments and for closed problems with fixed rules. However it does not perform for financial time-series and other dynamic systems. In order to make consistently accurate predictions about the future and to achieve true artificial intelligence, the development of new science that enables machines to understand cause and effect is required. Understanding true causal drivers enables causal AI to navigate complex and dynamic systems, being able to perform as its environment changes. In addition, causal AI is capable of 'imagining' scenarios it has not encountered in the past, allowing it to simulate counterfactual worlds to learn from, instead of relying solely on 'training' data. Perhaps most interestingly, understanding causality gives an AI the ability to interact with humans more deeply, being able to explain its 'thought process' and integrate human knowledge.

At causaLens we are building the world's largest causal AI research lab to accelerate progress in this powerful science. This talk will present the current challenges causal AI must overcome to unleash its full potential, as well as the latest progress made towards achieving those goals. Finally, some examples of the positive impact this science is already having in the field will be shared.

How we leverage machine learning and AI to develop life-changing medicines - a case study with COVID-19

6 AM PST (2 PM UTC)

Speakers: Daniel Neill, Sia Togia, Saee Paliwal, Hamish Tomlinson

The model for drug discovery and development is failing patients. It is expensive and high risk, with long research and development cycles. This has a societal cost, with 9,000 diseases being untreated - in addition to the disappointing reality that the top ten best-selling drugs only are effective in 30-50% of patients. Tackling this challenge is very complex. While many companies focus on one component of the drug discovery process, BenevolentAI chooses to apply data and machine-learning driven methods across drug discovery, from the processing of scientific literature, to knowledge completion, to precision medicine, to chemistry optimization, each leveraging domain expert knowledge and state-of-the-art research.

In this talk, we will discuss the peculiarities of machine learning for the drug discovery domain. In this field, there exist many unique challenges, including tradeoffs between novelty and accuracy; questions of quality and reliability, both in extracted data and in the underlying ground-truth; how best to learn from small volumes of data; and methods to best combine human experts and ML methods. As we discuss the tools and methods that BenevolentAI has developed, we will explore these themes and walk through approaches.

Finally, to give a real example of how we apply machine learning and AI in our day-to-day work, we will showcase the application of our technology to repurpose existing drugs, using our tools and internal clinical experts, as a potential treatment for COVID-19. Baricitinib, the top drug we identified is currently being investigated in a Phase 3 clinical trial.

Accelerated Training with ML Compute on M1-Powered Mac

7 AM PST (3 PM UTC)

Speaker: Cibele Montez Halasz

Last month, Apple announced Mac powered by the M1 chip, featuring a powerful machine learning accelerator and high-performance GPU. ML Compute, a new framework available in macOS Big Sur, enables developers to accelerate the training of neural networks using the CPU and GPU.

In this talk, we discuss how we use ML Compute to speed up the training of ML models on M1-powered Mac with popular deep learning frameworks such as TensorFlow. We show how to replace the TensorFlow ops in graph and eager mode with an ML Compute graph. We also present the performance and watt improvements when training neural networks on Mac with M1. Finally, we examine how unified memory and other memory optimizations on M1-powered Mac allow us to minimize the memory footprint when training neural networks.

Drifting Efficiently Through the Stratosphere Using Deep Reinforcement Learning

7 AM PST (3 PM UTC)

Speaker: Sal Candido

Loon's mission is connecting people everywhere using audacious new technology. We use a synthesis of machine learning and automation with a superpressure balloon-based aircraft to create a unique high altitude pseudo-satellite (HAPS) that can provide connectivity, earth observation, collect weather data to improve forecasts, and perform other tasks from the stratosphere. This technology would not be possible without software automation, and as a result research-level machine learning has a great deal of applicability at Loon. A prime example of this is our Nature publication, released earlier this week, that describes how we have used deep reinforcement learning to improve station keeping for our HAPS system in real flight through the stratosphere and across our production fleet. In this talk I'll discuss some of the areas and applications where machine learning can further improve Loon, and dive into the technology described in the Nature paper as an example of successful collaboration between research and the core Loon technology stack that led to deep reinforcement learning taking flight with Loon.

Making boats fly by scaling Reinforcement Learning with Software 2.0

8 AM PST (4 PM UTC)

Speakers: Nic Hohn, Jacomo M Corbo

The increasing prevalence of high-fidelity industrial digital twins is providing a range of opportunities to apply Reinforcement Learning techniques outside of traditional academic examples and video games. While this trend is now well-established, most RL developments and deployments in the real world are done on an ad-hoc basis with little consideration given to how to repeat and scale similar initiatives in an efficient way.

In this session we will address these shortcomings and illustrate them through our experience in optimising the design of a state-of-the-art sailing boat for a prominent competition with RL. Building an agent to control the boat is a very complex RL task for several reasons: imperfect information, loosely defined goals with delayed rewards, highly dynamic state and action spaces. In racing conditions, it takes a team of Olympic-level athletes to sail the boat and make it "fly" thanks to its underwater wings (read hydro-foiling).

Beyond describing solutions to traditional RL considerations such as the learning algorithm construct and reward function, we will focus on the underlying workflows and technology stack required to carry out a project of this technical complexity in a scalable way. We will use facets of Software 2.0, such as higher-level APIs and the automation of end-to-end model development tasks, to highlight our iterative choices and the optimisation opportunities along the machine learning pipeline and ultimately the production system.

Automating Wildlife Conservation for Cetaceans

9 AM PST (5 PM UTC)

Speaker: Jason R Parham

Wild Me is a not-for-profit based in Portland, OR, USA that works directly with ecologists around the world to automate wildlife conservation. This talk covers the concepts of wildlife conservation and how it uses statistics to monitor an animal population, presents a motivating use-case for how machine learning can be used for social good, and details some of the specific machine learning algorithms and approaches that are used in the field. One of our premier ID platforms, Flukebook (flukebook.org), has helped to power state-of-the-art mark-recapture research and publications; Flukebook now supports 9 cetacean species (humpback whale, 2 right whale species, sperm whale, orca, and 4 dolphin species), has over 220,000 reported sightings, serves hundreds of collaborators, and exposes 7 unique computer vision ID algorithms (HotSpotter, CurvRank for flukes and dorsal fins, the winning 2016 Kaggle competition ID algorithm for right whales by Deepsense.ai, two learned triplet-loss embedding ID algorithms, and more). We will do a deep dive into our deep learning stack, detection pipeline, and ID algorithms, including: image classification, bounding box regression, instance classification, class segmentation, object of interest (AoI) classification, triplet-loss embedding computations, and more. Our deep learning stack utilizes Theano and PyTorch, the NVIDIA's CUDA, CNMeM, and CuDNN deep learning stack, and employs NVIDIA GPU hardware. The Flukebook platform also integrates "citizen science" input into conservation research through the help of an intelligent agent. Our agent automatically ingest video data from YouTube using NLP and OCR, plus image sightings reported via Twitter, and feeds them into our machine learning pipeline. Join our session and listen about the social good that machine learning can achieve as Wild Me helps to modernize wildlife conservation as a data-driven science. All code is available and open-source at github.com/wildbookorg.

AI against COVID-19

9 AM PST (5 PM UTC)

Speakers: Divya Pathak, Payel Das, Michal Rosen-Zvi, Salim Roukos

The fight against COVID-19 has seen a call to action to all facets of scientific discovery. This includes AI which stands to transform how we react to the pandemic and empower our scientists and public policymakers with newer and more powerful tools. As part of our corporate responsibility and AI for Good initiatives, IBM has been at the forefront of this movement. This 1-hr talk will highlight members of our community who have risen to this challenge, in the context of AI research and the NeurIPS audience. This includes data analysis for non-pharmaceutical interventions (NPIs) world-wide to help in policy decisions, research in natural language processing, and molecule discovery to advance scientific research. The WNTRAC Open Challenge deals with AI-based data generation and analysis of non-pharmaceutical interventions to track COVID-19.

The Worldwide Non-Pharmaceutical Interventions Tracker for COVID-19 (WNTRAC) is a publicly available comprehensive dataset consisting of more than 6,000 NPIs implemented worldwide since the start of the pandemic. IBM Research has created a system that leverages DL technologies applied to Wikipedia pages for generating the data. The team has illustrated various ways to leverage the data for predicting disease spread and offers mechanisms to explore the causal effect of different interventions on the pandemic. Researchers are invited to explore the data.

Natural Language Processing for COVID Literature Search and Q&A - This part of the talk will focus on the use of natural language processing to help scientists accelerate their discoveries as well as answer your COVID-19 questions from the scientific literature. It will demonstrate how to derive insights from a large corpus of papers and open datasets through Deep Search and Q&A.

Molecule Explorer using Generative AI - This part of the talk will show how generative AI frameworks have been used to help researchers generate potential new drug candidates for COVID-19, applied to three COVID-19 targets to produce>3500 novel molecules and their attributes in the molecule explorer platform under an open license. We will talk about how generative AI techniques required to be "controllable" and be able to learn from limited labels and generalize to unseen contexts such as a novel viral protein. Such AI methods lead a path toward faster generation and comprehensive virtual screening of new and optimal candidate molecules and show promise for accelerating molecule and material discovery, which is critical for responding to unprecedented crises like the COVID-19 pandemic.

Fairness, Explainability, and Privacy in AI/ML Systems

10 AM PST (6 PM UTC)

Speakers: Vidya Sagar Ravipati, Erika A Pelaez Coyotl, Ujjwal Ratan, Kenthapadi, Krishnaram

How do we develop machine learning models and systems taking fairness, accuracy, explainability, and transparency into account? How do we protect the privacy of users when building large-scale AI based systems? Model fairness and explainability and protection of user privacy are considered prerequisites for building trust and adoption of AI systems in high stakes domains such as lending and healthcare requiring reliability, safety, and fairness.

We will first motivate the need for adopting a "fairness, explainability, and privacy by design" approach when developing AI/ML models and systems for different consumer and enterprise applications from the societal, regulatory, customer, end-user, and model developer perspectives. We will then focus on the application of fairness-aware ML, explainable AI, and privacy-preserving AI techniques in practice through industry case studies. We will discuss the sociotechnical dimensions and practical challenges, and conclude with the key takeaways and open challenges.

Challenges in the adoption of Machine Learning in Health Care

11 AM PST (7 PM UTC)

Speakers: Vidya Sagar Ravipati, Ujjwal Ratan, Erika A Pelaez Coyotl, Bhatia, Parminder

Adoption of Machine Learning (ML) outside the field of research has been one of the key factors to fuel disruption in several industry verticals. ML is a powerful tool for learning to solve complex problems and is generating interest in the health care life science space (HCLS), which poses compelling questions to this technique. In particular, Deep Learning (DL) is amassing more and more interest due to the high performance it has delivered in other fields . However, Deep Learning is yet to become a comprehensive tool for all HCLS applications. This is primarily because of the peculiar nature of Healthcare. Unique challenges that clinical data bring to ML include but are not limited to multimodal and sparse nature, lack of pre-trained models, and compatibility with conventional statistics in a field that heavily relies on p-values.

At Amazon we are well aware of the difficulties in designing ML products for medical applications . For instance, NLP services as Comprehend and Transcribe have branched off their main products to create Comprehend Medical and Transcribe Medical to overcome the limitations of their generic counterparts and abide to laws like HIPAA and COPPA . The Alexa team has solutions dedicated specifically for HealthCare.. Examples of this efforts are Alexa Health, which is a trusted health assistant for patients and clinicians, Comprehend Medical, which is able to deal with Electronic medical Records, and various other initiatives.

In this talk, we would like to address the following questions:

- How different is to build a product for HCLS than any other domain ?
- What are the biggest areas of interest for ML in HCLS ?
- What are the challenges around data privacy and regulations when it comes to the adoption of ML use cases in HCLS?
- How can companies build AI applications in HCLS and maintain customer trust?
- What are some of the risks with data collection in HCLS ?
- What is the importance of explainable models in HCLS?

AI4Code @ IBM and Red Hat

11 AM PST (7 PM UTC)

Speakers: Kartik Talamadupula, Julian Dolby, Kavitha Srinivas, Fridolín Pokorný, Maja Vukovic, Anup Kalia, Alessandro Morari

The AI4Code at IBM and Red Hat talk aims to showcase the latest in AI being applied to the code and programming lifecycle as relevant to enterprise applications. The session will feature live demos and perspective of research from IBM and Red Hat that currently use AI and ML techniques to make the entire code lifecycle more efficient, scalable, creative, and secure from an industry research perspective.

CodeBreaker is a coding assistant for data science code. It performs domain specific knowledge extraction over corpora such as GitHub, data science ontologies and Wikipedia, and documentation about data science code and tutorials to create a knowledge graph (KG) related to data science code. This KG is then used in downstream products like IDEs, etc. in order to make writing ML code easier.

Red Hat's Project Thoth focuses on analyzing and recommending software stacks for AI applications. The team will demonstrate how Thoth learns new knowledge and uses that knowledge with reinforcement learning techniques to recommend application stacks.

Code and app modernization is a core problem for the software industry, with a large amount of legacy code running critical systems world over. The "Intelligent Application Insights" tool builds and customizes models that generate containerization strategies for a given application; and constructs customized cost models for cost/benefit and risk assessments based on dynamic Bayesian learning.

AI for Vulnerability Analysis (AI4VA) models source code as a graph neural network in order to map that code to potential software vulnerabilities. Specifically, it learns whether signatures of the vulnerabilities in source code can be learned from their graph representation.

Human-Centered AI @ IBM Research –Automation versus Collaboration in the Age of AI

1 PM PST (9 PM UTC)

Speakers: Werner Geyer and Casey Dugan

Advances in Artificial Intelligence create unprecedented opportunities for automating tasks that previously could only be done by humans -- for example, driving cars, writing essays, or creating novel drugs –but also impacts the workers doing those tasks today (i.e. rideshare drivers, authors, and chemists). As AI is infused into more and more systems, there is increasing need to study and understand the impact automation has on our workforce, but also on how we interact with intelligent systems and how such systems need to be designed to most effectively support their human users. Our Human-Centered AI agenda at IBM Research has explored a different perspective on Human-AI interaction, investigating the transformation of the interaction to a more collaborative relationship in which humans and AI systems work hand-in-hand to create a desired outcome.

In this talk we will explore the theme of collaboration versus automation with AI through a number of research projects and scientific studies we have conducted in the context of AI Lifecycle Management, Automated Model Generation and Exploration for Data Scientists, Human-in-Loop Data Labeling, Explainability and Trust in AI systems, and AI-Infused Process Automation, as well as Generative Models and how they fundamentally change how humans will interact with AI systems in the creative industries and for content generation. This talk will give a unique industry perspective on designing and building AI systems with users in mind.

Modern ML Meets Financial Markets: Insights and Challenges

1 PM PST (9 PM UTC)

Speaker: Iain Dunning

Hudson River Trading (HRT) is a quantitative automated trading company that trades hundreds of millions of shares each day broken up into over a million trades and spread across thousands of symbols. It trades on about 100 markets worldwide, and accounts for over 5% of US equities volume. To provide price discovery and market making services for public markets, HRT employs state-of-the-art techniques from machine learning and optimization to understand market data.

In this talk we'll discuss some of the challenges that come from applying modern ML techniques to financial markets. We will provide some background about the massive, heterogeneous, unevenly spaced, noisy, and bursty tick-by-tick datasets that many of our machine learning algorithms process. We'll explore applications of contrastive learning, which seeks to learn feature embeddings in which similar inputs have similar feature representations, for learning feature representations of related assets for downstream tasks. Next, we'll talk about applications of meta-learning for financial timeseries models, such as quickly adapting to new financial products. Finally, we'll discuss some of the issues that need to be overcome to apply state-of-the-art NLP techniques to problems in forecasting.

The Unpaved Path of Deploying Reliable and Human-Centered Machine Learning Systems

2 PM PST (10 PM UTC)

Microsoft

Speaker: Besmira Nushi

As Machine Learning systems are increasingly becoming part of user-facing applications, their reliability and robustness are key to building and maintaining trust with users, especially for high-stake domains such as healthcare. While advances in learning are continuously improving model performance in expectation and in isolation, there is an emergent need for identifying, understanding, and mitigating cases where models may fail in unexpected ways and therefore break human trust or dependencies with other larger software ecosystems. Current development infrastructures and methodologies often designed with traditional software in mind, still provide very little support to enable practitioners debug and troubleshoot systems over time. This discussion will look at such problems from two different stakeholder lenses: machine learning practitioners and end user decision makers. From a practitioner perspective, it will summarize some of the current gaps in tooling for responsible ML development and evaluation, and present ongoing work that can enable in-depth error analysis and careful model versioning. Next, from an end user perspective it will propose rethinking the optimization of machine learning models such that it takes into consideration human-centered properties of human-machine collaboration and partnership. While both these lenses pose both research and engineering practices, they also require close collaboration with domain experts who are using machine learning in the open field to ensure that deployed systems meet real-world expectations.

Building Neural Interfaces: When Real and Artificial Neurons Meet

2 PM PST (10 PM UTC)

FACEBOOK

Speaker: Ricardo Monti, Nathalie T.H Gayraud, Jeffrey Seely, Zhuo Wang, Tugce Tasci

In this talk, the team from Facebook Reality Labs Research will outline the fundamental elements of machine learning and neuroscience required to build all-day wearable, non-invasive neural interfaces to power interaction for future computing platforms.

The talk will focus on the specific challenges of building machine learning models in a wearable device from biological signals, such as managing model stability across a range of contexts and across the global population. The team has a unique combination of computational neuroscientists and machine learning engineers that are necessary to both uncover the hidden engineering of our nervous system, and design the future of our relationship with computers.

Scaling Data Labeling with Machine Learning

4 PM PST (12 AM UTC Monday)

scale

Speakers: Yuri Maruyama, Nishant Subramani, Felix Lau

At Scale AI, we label on the order of 10MM annotations per week. Our data is diverse both in image space (e.g. cameras, weather conditions, driving surface) and label space (e.g. object and attributes categories). To fully leverage the vast amount of labeled data, we want to continuously fine-tune a base model across all datasets as tasks are completed. The base model allows us to easily fine-tune it further for downstream tasks such as quality assurance, prelabeling, and image similarity search. We present a multi-task continuous learning approach that can be trained efficiently as more labeled data becomes available. In conjunction to this, we also provide real-time annotation solutions with our models as a service offering for document intelligence. We utilize our labeling platform, our machine learning expertise, and our customer's domain knowledge to provide a customer-specific, problem-specific, real-time data labeling solution.

Hypotheses Generation for Applications in Biomedicine and Gastronomy

5 PM PST (1 AM UTC Monday)

Speaker: Michael Spranger

Hypothesis generation is the problem of discovering meaningful, implicit connections in a particular domain. We focus on two application areas 1) biomedicine and the discovery of new connections between scientific terms such as diseases, chemicals, drugs, genes, 2) food pairing for discovering new connections between ingredients, taste and flavor molecules. Sony AI and its academtic partners have developed a variety of models that explore representation learning and novel link prediction models for these tasks. In the biomedical domain, we developed models able to leverage temporal data about how connections between concepts have emerged over the last 80 years. In the food domain, we deal with multi-partite graphs that link ingredients with molecule information and health aspects of ingredients. The talk will introduce hypothesis generation as a graph embedding representation learning and link prediction task. We'll present recently published models that integrate 1) variational inference for estimating priors, 2) graph embedding learning regimes and 3) application of embeddings in training ranking models.

Visually Debugging ML Models With Scale Nucleus

6 PM PST (2 AM UTC Monday)

Speakers: Elliot Branson, Srikanth Srinivas, Chun Jiang

We will show how you can achieve the concept of "Operation Vacation" for the models you create, and make sure that the model is testing the subsets that you actually care about with Scale's latest product, Nucleus. Using nuScenes 2.0, a multimodal dataset for autonomous driving, we will demonstrate how you can easily debug model performance and automatically refine your model. In the process, we'll also dive into Nucleus's features to show how to curate sub-datasets and edge cases easily with custom metrics, image similarity search, and auto-pivot, automatically augmenting the data collection to accelerate machine learning training process.

Driving New Frontiers of Machine Learning with Cruise

7 PM PST (3 AM UTC Monday)

Speaker: Edgar Molina

This panel session will share an inside look at some of the most surprising and outstanding challenges in self driving, and Cruise's ML-first approach to solving them. Leaders from Cruise AI will discuss where ML is replacing traditional methods for tracking, planning, perception and prediction and how self driving is pushing new frontiers for AI. As part of this, we will cover uncertainty modeling and the critical role this plays in enabling autonomous vehicles to make more robust and safer decisions in complex driving environments like San Francisco. We will also walk through the homegrown ML tools and the ML infrastructure we've built to foster fast experimentation and scale.

Accelerating Eye Movement Research Via Smartphone Gaze

8 PM PST (4 AM UTC Monday)

Speakers: Vidhya Navalpakkam

Eye movements have been widely studied in vision research, language and usability, yet progress has been limited since eye trackers are expensive (>$10K) and do not scale. In this talk, we'll present findings from our recent paper at Nature communications, which shows that smartphone's selfie cameras+ML can achieve high gaze accuracy comparable to SOTA eye trackers that are 100x more expensive. We demonstrate that this smartphone technology can help replicate key findings from prior eye movement research in Neuroscience/Psychology, that earlier required bulky/expensive desktop eye trackers in highly controlled settings (e.g., chin rest).

These findings offer the potential for orders-of-magnitude scaling of basic eye-movement research in Neuroscience/Psychology (with explicit user consent) and unlock new applications for improved accessibility, usability and screening of health conditions.

Building AI with Security and Privacy in mind

5 AM PST (1 PM UTC)

Speakers: Joe Spisak, Andrew Trask, Geeta Chauhan, Laurens van der Maaten, Davide Testuggine

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, federated learning, on-device computation, and differential privacy. This tutorial will dive into some of the important areas that are shaping the future of how we interpret our models and build AI with security and privacy in mind. We will cover the major challenges, walk through some solutions and finish each talk with a hands on tutorial. The material will be presented in the following talks:

- PPML 101 & Introduction - Geeta Chauhan
- Secure Computation using CrypTen (https://crypten.ai/); - Laurens van der Maaten
- Training models differentially private at scale using Opacus (https://ai.facebook.com/blog/introducing-opacus-a-high-speed-library-for-training-pytorch-models-with-differential-privacy/); - Davide Testuggine
- Training models across multiple organizations privately with federated learning and PySyft from OpenMined (https://www.openmined.org/) - Andrew Trask

The tutorial will start with basic concepts and will proceed into more advanced topics following a chronological order of the presentations. The audience is expected to have some basic understanding of deep learning frameworks, security and privacy concepts that will be supplemented with the material in the early talks. The audience will have an opportunity to learn more advanced topics as the tutorial proceeds.

Machine Learning at Netflix

5 AM PST (1 PM UTC)

Speakers: Yves Raimond, Sui Huang

Netflix is the world's leading streaming entertainment service with over 195 million paid memberships in over 190 countries enjoying TV series, documentaries and feature films across a wide variety of genres and languages.

In this workshop we will offer a deep-dive into some recent Machine Learning research at Netflix spanning many different areas, including:

- Personalization: How can we help our members find the content they'd enjoy the most?
- Content: How can we help shape our catalog of movies & TV shows? How can we help best showcase this catalog?
- Systems: How can Machine Learning help optimize our Netflix infrastructure & systems?

We'll conclude with a live 45 minutes Q&A session including all the speakers.

Mining and Learning with Graphs at Scale

10 AM PST (6 PM UTC)

Speakers: Bryan Perozzi, Vahab Mirrokni, Jonathan Halcrow, Jakub Lacki

This workshop focuses on methods for operating on massive information networks. We begin by highlighting applications of graph-based learning and graph algorithms for a wide range of areas such as detecting fraud and abuse, query clustering and duplication detection, image and multi-modal data analysis, privacy-respecting data mining and recommendation, and experimental design under interference.

The main body of the presentation is divided into three sections:

In our first segment, we cover graph learning and graph building algorithms which we apply to graphs with billions of nodes, and trillions of potential edges. We also discuss similarity ranking over graphs, and the clustering and community detection methods which power numerous industrial applications. This section concludes with a discussion of graph-based semi-supervised learning techniques.

Our second segment covers the application of neural networks to graph structured data through both positional graph embeddings and graph neural networks (GNNs). We present challenges, and recent results from our team on scalable inference algorithms for GNNs, methods for dealing with bias in graph data, and ensemble approaches to representing nodes which allow more modeling flexibility.

Our final segment discusses different techniques for working with massive graphs. We focus on how to take advantage of both single- and multi-machine parallelism to run algorithms on graphs of up to trillions of edges.

Real World RL with Vowpal Wabbit: Beyond Contextual Bandits

10 AM PST (6 PM UTC)

Speaker: Jacob Alber

In recent years, breakthroughs in sample-efficient RL algorithms like Contextual Bandits enabled new solutions to personalization and optimization scenarios. Unbiased off-policy evaluation gave Data Scientists superpowers on real-world data volumes, giving them confidence in putting machine learning into production. Vowpal Wabbit (https://vowpalwabbit.org) is an open source machine learning toolkit and research platform, used extensively across the industry, providing fast, scalable machine learning.

Dive beyond Contextual Bandits in the Real World: * Solve multi-slot scenarios with Conditional Contextual Bandits and Slates, and optimize systems with Continuous Action-Space CB * Learn about advanced off-policy evaluation and introspection options with new estimators and visualizations * Discover how to warm-start your learning using Apprentice Mode and put RL into production with confidence

Perspectives on Neurosymbolic Artificial Intelligence Research

10 AM PST (6 PM UTC)

Leslie Pack Kaelbling, Jiajun Wu, David Cox, Ronald Fagin, Achille Fokoue, Chuang Gan, Alexander Gray, Pavan Kapanipathi, Tim Klinger, Ryan Riegel, Salim Roukos, Akash Srivastava

Neuro-symbolic AI approaches have recently begun to generate significant interest, as urgency in the field appears to be growing around various ideas for somehow extending the strengths and success of neural networks (or machine learning, more broadly) with capabilities typically found in symbolic, or classical AI (such as knowledge representation and reasoning). A general aim of this research is to create a new class of far more powerful than the sum of its parts and leverage the best of both worlds while simultaneously addressing the shortcomings of each. Typical advantages sought include the ability to:

- Perform reasoning to solve more difficult problems
- Leverage explicit domain knowledge where available
- Learn with fewer examples
- Provide understandable or verifiable decisions.

These abilities are particularly relevant to the adoption of AI in a broader array of industrial and societal problems where data is scarce, the stakes are higher, and where the scrutability of systems is important. This research direction is at once an old pursuit and nascent, and several perspectives are expected to be needed in order to solve this grand challenge. In this workshop we will explore several points of view, both from industry and academia, and highlight strong recent and emerging results that we believe are providing new fundamental insights for the area and also beginning to demonstrate state-of-the-art results on both the theoretical side and the applied side.

DAQA – Domain Adaptation for Question Answering:

3 PM PST (11 PM UTC)

Speakers: Eneko Agirre, Thomas Wolf, Salim Roukos

Question Answering (QA) in its various flavors has made notable strides in recent years thanks in part to the availability of public datasets and leaderboards. Large datasets are not representative of many real world scenarios of interest; this is especially true for industry data and specialized field data. Small datasets cannot be used to train QA systems from scratch: domain adaptation techniques are required. In this proposal, we use the term domain adaptation broadly, to cover techniques that leverage out-of-domain data, or in-domain data that does not match the task at hand. The workshop is intended to highlight innovative approaches that have the potential to yield significant improvement in QA scenarios where limited labeled data is available and to promote the development and use of real-world datasets for domain adaptation. Topics of interest include established and emerging approaches that have notable potential to substantially impact domain adaptation for QA. Notable examples are: adversarial training, automatic augmentation of a training set, unsupervised transfer learning, joint learning of QA and question generation, multi-task learning, domain-specific knowledge graphs, and using large models with few-shot learning. The invited talks represent both the industry and the academic perspective: industry has pressing needs for techniques that address the small amount of labeled data that one can expect from customers; academia is leading the path towards innovative breakthroughs that can quickly advance the field. In addition to the invited talks we will a case study, for which we will distribute material (including one or more Jupyter notebooks and corresponding data) prior to the workshop date.

New Challenges in User-Generated Content

8 PM PST (4 AM UTC Monday)

Speakers: Yaliang Li, Bolin Ding, Jinyang Gao

Understanding the user-generated content (UGC) is one of the key components for various applications of machine learning such as social networks, online image/video-sharing platforms, and ecommerce. Today, UGC has evolved from plain texts to more enriched formats including live comments , images, and short videos, which brings new challenges to understand the unique characteristics of these different formats of UGC. Meanwhile, with the development of deep learning techniques, nowadays, we have powerful tools to learn representations of various UGC and extract useful information from them, which enables a wide range of new applications that would otherwise be infeasible due to difficulty in processing such UGC. In particular, this workshop would like to promote research and discussions on the following challenges in the new generation of UGC. • Heterogeneity (multi-modality) of contents. • Interconnected entities. • Large volumes. • Privacy and ethics issues. • Data management and analytical systems for UGC.

This workshop will provide a forum for both academic and industrial researchers to review the recent progress of new-generation user-generated content understanding and utilization, with an emphasis on novel approaches and systems that tackle the above challenges.

Please find a more detailed workshop description at https://files.alicdn.com/tpsservice/47fe0d23af04b3bd2b60c770bbe892ef.pdf

Machine Learning for All-Inclusive Finance

8 PM PST (4 AM UTC Monday)

Speaker: Hui Tian

The goal of all-inclusive finance is to bring reliable and high quality financial service to everyone everywhere, rich or poor. Managing large scale multi-agent interactions lies in the very nature of the all-inclusive finance, and many emerging problems in this new space involve sophisticated cooperation and competition between a diverse range of entities, such as people, small businesses, online platforms, financial units and even fraudsters. For instance,

- Good customer service requires cooperative problem solving by customer and automated systems;
- Fraudulent transaction detection is a game between attackers and platform defenders;
- Mutual insurance needs to incentivize millions of people in order to be effective and low cost;
- A healthy online lending product requires good policies for cash flow management;
- A good investment recommendation needs to understand the intricate relations between companies, financial assets and economic environment.

Machine learning techniques, such as multi-agent reinforcement learning, algorithmic game theory, generative adversarial learning, imitation learning, graph neural networks, construction and reasoning over knowledge graph, building interpretable and fair models, are playing increasingly important roles in addressing these problems in all-inclusive finance. In this workshop, we will invite technical leaders from both all-inclusive finance platforms to talk about these emerging problems and their solutions, as well as experts from such as Matei Zaharia, Raluca Ada Popa, Virginia Smith, and John Duchi. We will also invite Fintech tech leaders to talk about current status and their views for all-inclusive finance.

Using Sparse Quantization for Efficient Inference on Deep Neural Networks

12 PM PST (8 PM UTC)

Speakers: Mark J Kurtz, Dan Alistarh

Today's state of deep neural network inference can be summed up with two words: complex and inefficient. The quest for accuracy has led to overparameterized deep neural networks that require heavy compute resources to solve tasks at hand, and as such we are "rapidly approaching outrageous computational, economic, and environmental costs to gain incrementally smaller improvements in model performance (State of AI Report 2020)." Furthermore, there is no lack of research on achieving high levels of unstructured sparsity, but putting that research into practice remains a challenge. As a result, data scientists and machine learning engineers are often forced to make tradeoffs between model performance, accuracy, and inference costs.

There is a better way.

After years of research at MIT, the team at Neural Magic concluded that throwing teraflops at dense models is not sustainable. So we've taken the best of known research on model compression (unstructured pruning and quantization, in particular) and efficient sparse execution to build a software solution that delivers efficient deep neural network inference on everyday CPUs, without the need for specialized hardware.

Join Neural Magic ML experts to learn how we successfully applied published research on model compression and efficient sparse execution to built software that compresses and optimize deep learning models for efficient inference with ease.

You'll walk away with an overview of:

- SOTA model compression techniques;
- A demo of the first-ever general-purpose inference engine that translates high sparsity levels into significant speedup, and
- Next steps on using the Neural Magic Inference engine and ML tools to make your inference efficient, with less complexity.

Beyond AutoML: AI Automation & Scaling

3 PM PST (11 PM UTC)

Speakers: Lisa Amini, Nitin Gupta, Parikshit Ram, Kiran Kate, Bhanu Vinzamuri, Nathalie Baracaldo, Martin Korytak, Daniel Karl

There has been a recent burst in "AutoML" techniques as a means to automate the creation of ML models without necessary domain expertise. This demonstration looks well beyond AutoML's current narrow focus on automated model building, to tackling automation across the full end-to-end AI/ML lifecycle. In industry settings, the AI/ML lifecycle typically includes a series of labor-intensive tasks such as preparing data, training models, deploying the selected model in cloud, monitoring performance, identifying faults, and taking corrective actions when failures or new business requirements occur. Enormous opportunities exist for scaling, automating, and accelerating this AI/ML lifecycle.

In this session, we demonstrate tools and research results in driving automation across the entire AI/ML lifecycle: from assessing data readiness and recommending mitigations, to semantically-driven automation based on concept discovery and knowledge augmentation, to advanced ML model building with business and fairness constraints, to novel pipelines for industry-critical modalities, to automation for monitoring models in deployment, recognizing deficiencies and recommending corrective actions. We will also demonstrate practical methods for scaling in multi-cloud environments with federated learning, and accelerated cloud-based inference of widely-popular classical ML algorithms such as XGBoost and LightGBM.

Medical Transcription Analysis

4 PM PST (12 AM Monday UTC)

Speakers: Ujjwal Ratan, Erika A Pelaez Coyotl, Vidya Sagar Ravipati

Medical transcription involves the use of automatic speech recognition (ASR) technology that is able to understand key medical terminology from audio formats and is able to convert it into text. Moreover, to understand what's clinically relevant in these transcripts, it's important to integrate the ASR engine with a natural language processing engine that is able to extract medical terminology. In this demo, we will introduce the audience to two services from AWS specifically designed for the medical domain:

1. Amazon Transcribe Medical: A dedicate Speech to Text service that understands medical terminology in audio files and converts them into text.
2. Amazon Comprehend Medical: A natural language processing service that extracts key medical entities and ontologies from free form clinical text.

The demonstrated solution will integrate these two services to create a medical transcription analysis pipeline. The pipeline is able to process medical transcriptions and extract key medical entities from them. The solution will show how these entities can then be summarized into a report for sharing or storage.

AWS Computer Vision Science

5 PM PST (1 AM UTC Monday)

Speaker: Yuting Zhang

AWS has a fast-growing team of scientists to support its computer vision services with cutting-edge science. In this demonstration, Stefano Soatto will first deliver an opening talk on AWS AI Applications. We will then show a selected set of the latest computer vision services and product-driven scientific developments at AWS. In particular, we will demonstrate Amazon Rekognition, the AWS services to automate your image and video analysis with machine learning. The demos will cover Amazon Rekognition Labels, Content Moderation, Personal Protection Equipment Detection, and Custom Labels, as well as our work on video analysis. We will also demonstrate Amazon Textract services, which can automatically extract printed text, handwriting, and structured data from documents. The demos will cover DetectText (OCR), Text in Image, and AnalyzeDoc Forms and Tables. Meanwhile, AWS takes the utmost care about the fairness of the algorithms and models. We will have a panel discussion on Fairness in AI delivered by Michael Kearns, Michelle Lee, Pietro Perona, and Nashlie Sephus. In the final 15-min QA session, you will have the opportunity to chat with AWS scientists. Please join us to learn more about AWS computer vision science.

Whale: Accelerate EasyTransfer training workloads within one unified distributed training framework

Speaker: Wei Lin

6 PM PST (2 AM UTC Monday)

Recent advances in deep learning have led to substantial gains in various fields. With the increase of datasets and model size, it is a common practice to speed up the training workload by using data parallel (DP). However, in our practice with EasyTransfer (a modeling framework designed for NLP developers to implement algorithms with usability and flexibility), DP loses its magic for giant models that cannot fit into single GPU memory. Moreover, for different model architectures, it is nontrivial to find an efficient parallel strategy that can make full use of the resources.

To address the above challenges, we present Whale, a unified distributed training framework that can boost AI training tasks with usability and efficiency. It provides comprehensive parallel strategies including data parallel, model parallel, operator splitting, pipeline, hybrid strategy, and automatic parallel strategy. As far as we know, this is the first work that supports various distributed strategies within one framework. To effectively express different training strategies, a new intermediate representation of models is designed for distributed paradigms and execution cost. The automatic parallel strategy is generated upon using the cost model. The parallelization of execution is built by editing the computational graph, which can be applied to different training frameworks. Whale can easily distribute training tasks by adding a few code lines without changing user model code.

In our experiment of BertLarge model that built with EasyTransfer, Whale pipeline strategy attains 57% speedup when compared to Horovod data parallel strategy (HDP) on 64 GPUs. In the large-scale image classification task (100,000 classes), Whale hybrid strategy, which consists of operator splitting and DP, is 14.8 times faster than HDP on 64 GPUs. For models that cannot fit into the GPU device memory, Whale enables the training of T5 and image classification tasks with 100 billions of classes.

Accelerating Deep Learning for Entertainment with Sony's Neural Network Libraries and Console

7 PM PST (3 AM UTC Monday)

Speakers: Takuya Narihira, Akio Hayakawa, Andrew Shin, Yoshiyuki Kobayashi, Kazumi Aoyama, Akira Nakamura

Sony has open soured its own deep learning framework as Neural Network Libraries and has also developed a unique GUI-based integrated deep learning development environment as Neural Network Console. Its application has spread to a wide range from electronics products to entertainment service areas such as video games and movies. We introduce our unique deep learning framework & tools, and demonstrate the latest results from AI-based audio and video processing.

The Intelligent Vision Sensor

8 PM PST (4 AM UTC Monday)

Speakers: Seigo Hirikawa, Hareesh Gowtham, Seiya Nishimura

Sony Corporation has developed two models of intelligent vision sensors, IMX500 and IMX501, which are equipped with AI processing functions. The new sensors feature a stacked configuration consisting of a pixel chip and logic chip. They are the world's first image sensor to be equipped with AI image analysis and processing functions on the logic chip. We demonstrate an example which shows capability of edge AI with the intelligent vision sensor. The demo shows that AI processing is possible even with a small system consists of an intelligent vision sensor and Wi-Fi module. It also demonstrates the programmable potential of the intelligent vision sensor to selectively perform multiple inference tasks.

Discovering genetic medicines using the Deep Genomics AI Drug Discovery Platform

9 PM PST (5 AM UTC Monday)

Speakers: Shreshth Gandhi, Amit G Deshwar

Our AI Workbench enables us to efficiently find therapeutic targets and drug candidates with desirable properties. We are focusing on the discovery of oligonucleotide therapies for genetic disorders. These genetic diseases are mediated by altered molecular phenotypes, such as transcription, splicing, translation or protein binding. Our predictors leverage deep learning to model these molecular phenotypes with high accuracy. This enables us to, in-silico, pin-point the disease-causing genetic mutation(s), identify their molecular consequences, and design oligonucleotides that directly restore the molecular phenotype.

GAN Applications in Fashion Article Design and Outfit Rendering

12 AM PST (8 AM UTC Monday)

Speakers: Gökhan Yildirim, Nikolay Jetchev

Advances in deep learning enabled sampling realistic images via generative modeling. This leads to new avenues in visual design and content creation, e.g. in fashion, where visualization is a key component. GANs can be used to create personalized visual content - e.g. rendering an outfit on a human body and creating unique designs - which can enrich shopping experience on e-commerce platforms. We will demo two projects, where we used GANs to create fashion images and enable novel applications:

Fashion Outfit Renderer

We work on generating high-resolution images of fashion models wearing desired outfits and standing in different poses. At Zalando, we provide quality photographs of fashion models wearing the articles in our online selection. These photographs help customers visualise the garments they browse and enhance the shopping experience. But what if our customers wish to visualise an individually created outfit? Zalando has a large and evolving assortment of garments, which makes it infeasible to photograph all outfit combinations. To solve this challenge, we work on a "Fashion Renderer", which creates a computer-generated image of a fashion model wearing an input outfit for an input body pose.

Generative fashion design and search

Fashion customers often have a visual idea of what they would like to buy. However, finding the right article can be a time-consuming process, as people need to convert their visual ideas into accurate linguistic search terms, and search engines should correctly interpret customers' search queries and retrieve relevant results. We enable search in a visual-only space by allowing customers to generate and breed different dress designs with using a style-based GAN. Created designs are used as a visual query to retrieve existing dresses in real time. This approach attempts to eliminate representation and interpretation problems in the word-based search and provides a novel way for searching fashion items.