ARIA

APPLIED RESEARCH IN ACTION

MASTER OF SCIENCE IN APPLIED COMPUTING (MScAC) PROGRAM

2017–2018 Projects

Computer Science
UNIVERSITY OF TORONTO

Ravin Balakrishnan

MESSAGE FROM THE CHAIR

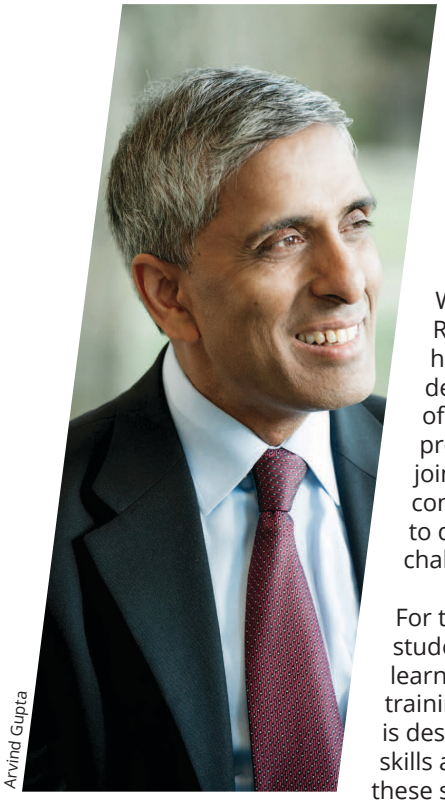
Today, all aspects of our lives are impacted by technological innovations, so much so that technology weaves through nearly every action we take. It is little wonder that the demand for highly-skilled and highly-trained computer scientists continues to accelerate. Toronto has emerged as a world leader in incubating and growing new technology firms. The Department of Computer Science at the University of Toronto is committed to being an integral player in this phenomenon, ensuring that our teaching and learning environment leverages the many technological developments occurring in broader society, while building bridges to industry to ensure our research has the greatest possible impact.

The Master of Science in Applied Computing (MScAC) program has emerged as a critical component of our outreach effort. The program is premised on a partnership model that builds industry partnerships, and ensures these students get the best of both academic and industry research training. We are proud of the contributions these students are already making, in tackling local and global challenges, as we see them contributing to building healthy, sustainable communities.

The MScAC program is part of our broader efforts to create engagement opportunities with industry partners across all of our graduate and research programs. For more information on research partnerships or to learn more about our research programs, please contact our Associate Chair for Research, Yashar Ganjali at **acrir@cs.toronto.edu**.

I wish all of our guests at ARIA a stimulating and enjoyable time learning about this year's internship projects.

Ravin Balakrishnan
Professor & Chair
Department of Computer Science



Arvind Gupta

MESSAGE FROM THE PROGRAM DIRECTORS

Welcome to ARIA 2018, our Applied Research in Action showcase, which highlights the exciting innovations being developed by the students in our Master of Science in Applied Computing (MScAC) program. We are delighted that you are joining us in celebrating the remarkable contributions of these incredible students, to developing world-class solutions to the challenges posed by our industry partners.

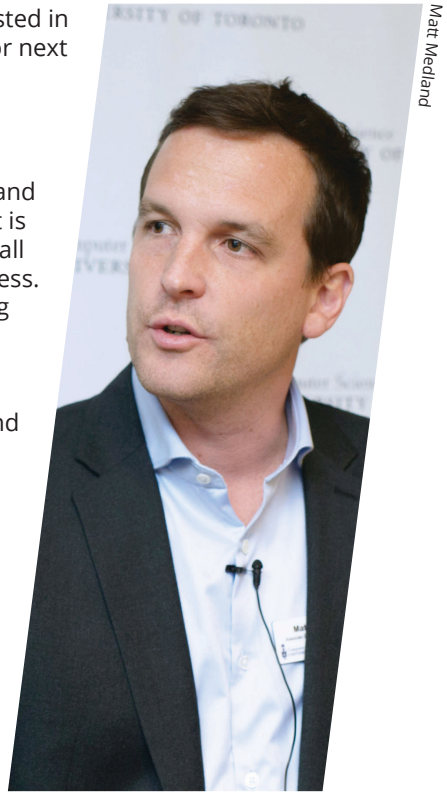
For the past 15 months, these MScAC students have been immersed in a unique learning environment that fuses academic training with industry research. The program is designed to give them advanced technical skills and to create a platform for applying these skills in creative ways. Today we are seeing first-hand the end-result of this process. ARIA is a milestone event in their degree progression - a time to honour their hard work, commitment and dedication.

We also want to recognize the extraordinary group of companies who have contributed so much to our students' career successes. We are delighted to have ROSS Intelligence as the presenting sponsor of this year's event. We also thank our platinum sponsors; Deloitte, Ethoca, Pelmorex, Surgical Safety Technologies, and gold sponsors; Autodesk Research, CaseWare International Inc., ecobee, Kindred, Inc., Layer 6 AI, and Uber ATG. A special thanks to Mitacs for their generous support to the MScAC program since its inception.

Our Applied Research Internship Expo (ARIE), where our industry partners meet and start a collaborative rapport with our 2018/19 cohort, will take place in the New Year. If you are interested in hosting a research internship project for next year's students, please contact us at **mscac@cs.toronto.edu**.

Finally, we thank the many people who continue to make the MScAC program and ARIA such a success. First and foremost is Claire Mosses, who so adeptly ensures all the pieces are in place for student success. Ryan Perez did much of the heavy lifting in putting together ARIA. And a thanks to the entire team in the Department of Computer Science for all their contributions to the MScAC program and ARIA.

Once again, congratulations to our 2017/18 applied computing cohort!



Matt Medland

Arvind Gupta
Professor & Academic Director, Professional Programs, MScAC

Matt Medland
Managing Director, MScAC
Assistant Professor, Teaching Stream

PRESENTING SPONSOR

ROSS

PLATINUM SPONSORS

Deloitte.

ethoca™

Pelmorex Corp

SURGICAL SAFETY TECHNOLOGIES

GOLD SPONSORS

AUTODESK RESEARCH

CASEWARE

ecobee

KINDRED

LAYER 6 AI

UBER

FUNDING PARTNER

Mitacs

Special thanks to all of our sponsors who made ARIA 2018 possible.

3D SEGMENTATION

Phenomic AI

High-content imaging of live tissues is an important phenotypic assay to describe healthy and diseased states of living human and animal tissues. 3D techniques allow for the study of the complex spatial arrangements and interactions of cells in their native environment. Analysis of large quantities of such image data poses unprecedented challenges to scientists and often still involves laborious manual annotation of the content. For example, it is often of interest to distinguish between and quantify different types of cells, based on their morphology. At the tissue level, one often finds other complex structures of interest, such as blood vessels. Automatically labeling and quantifying the morphologies (shapes) of all components, as well as their interactions (physical connections, relative distances, etc.), can dramatically increase the experimental throughput and accelerate scientific research. Fluorescent labels are often used to distinguish different cellular structures via different color emissions. Furthermore, living tissues evolve over time, and quite often, time series of cellular processes are captured. This provides an additional dimension to image analysis where in addition to static features, dynamic features can be extracted as well, such as the speed of migrating cells, or the appearance or disappearance of certain structures.

The goal of this internship project is to develop a deep-learning based 3D segmentation package that can be used to segment different cell types and vascular structures in the dataset and quantify features (i.e. length, volume, protrusion number, marker intensity) of these objects. These features will be used to evaluate the effectiveness of therapeutic treatments. The dataset for this project is provided by a major pharmaceutical company.

Kshitij Gupta
Industry Supervisor (IS): Oren Kraus
Academic Supervisor (AS): Sanja Fidler

A REINFORCEMENT LEARNING APPROACH FOR A DERMATOLOGICAL QUESTION ANSWERING SYMPTOM CHECKER USING CNNS

Triage

The project's objective is the improvement of the overall skin condition classification methodology by introducing a Question-Answering (QA) model and component.This is based on two different approaches: (1) maximizing the long-term reward of simulated patients via the Reinforcement Learning (RL) framework, and (2) maximizing the information gain over a multitude of symptoms via the implementation of traditional decision tree structures. These two methods enable us to not only increase the classification confidence and accuracy of the deployed Convolutional Neural Network (CNN) system, but also enables the emulation of the conventional approach of doctors asking the relevant questions in refining the ultimate diagnosis and differential. The sequential pipeline of a pretrained CNN followed by a QA model look to use the CNN output, in the form of classification probabilities, as a "prior" to the QA model. This "prior", which includes the image's associated textual description, allows for the subsequent determination of the most relevant symptoms to ask and identify. We demonstrate that combining the QA model with the CNN increases the skin classification accuracy by up to 10%, as compared to a standalone CNN, and increases accuracy by more than 30% when compared to a standalone QA model.

Mohamed Akrout
IS: Latif Abid
AS: Amir-massoud Farahmand and Richard Zemel

LATENT SEMANTIC REPRESENTATION FOR RANKING

ROSS Intelligence

ROSS helps the world's biggest and smallest firms increase their human-power with artificial intelligence. To achieve its mission, ROSS leverages state-of-the-art technology in natural language processing, general machine learning, and information retrieval combined with legal signals, to show an ordered list of retrieved results, relevant to the user query. My research at ROSS revolves around improving the current ranking algorithm employed at ROSS. Since users perform their queries in natural language, the semantic understanding of text is more crucial than keyword matching.

The aim of the project is to create a model that generates a latent semantic representation for query and retrieved passages. The results are then ranked by the similarity scores of the passage representations to the query representation. The experimental architectures were built using Recurrent Neural Networks (Bi-LSTMs), attention, and legal embeddings. They were then trained with a custom loss function, which pushes the most semantically similar passages as the top-K results. These architectures and loss function can be generalised for Learning to Rank objective in any Information Retrieval (IR) domain.

Manasa Bharadwaj
IS: Jimoh Ovbiagele
AS: Frank Rudzicz

USING SIMULATION TO TEST & OPTIMIZE HIGH-VOLUME AD BIDDING SOFTWARE

Pelmorex Corp (formerly Addictive Mobility)

Pelmorex Audience is a company specialized in mobile advertisement. The business relies on their real-time bidding platform, and as it grows more complex, they need more novel ways to test the different subsystems, and to run experiments without disturbing their critical production operations.

The bidding platform exhibits two characteristics that make regular unit and integration tests insufficient. First is the accumulated effect over time; their operations are time sensitive, adaptive, and change behaviour according to many time-related issues. Second is the distributed nature of the environment; they have hundreds of replicas of the bidding logic running at the same time to handle the amount of concurrent traffic that they see. This amplifies the effect of a decision made by any given system because other instances are probably taking the same decision, and unaware of each other.

These dynamics are hard to predict and test, and has forced Pelmorex Audience to test their logic changes in production and waste a test budget in the process. The objective of this project is to use a simulation environment in place of the test logic.

Vincent B. Tembo
IS: Lubna Khader
AS: Cristiana Amza and Matt Medland

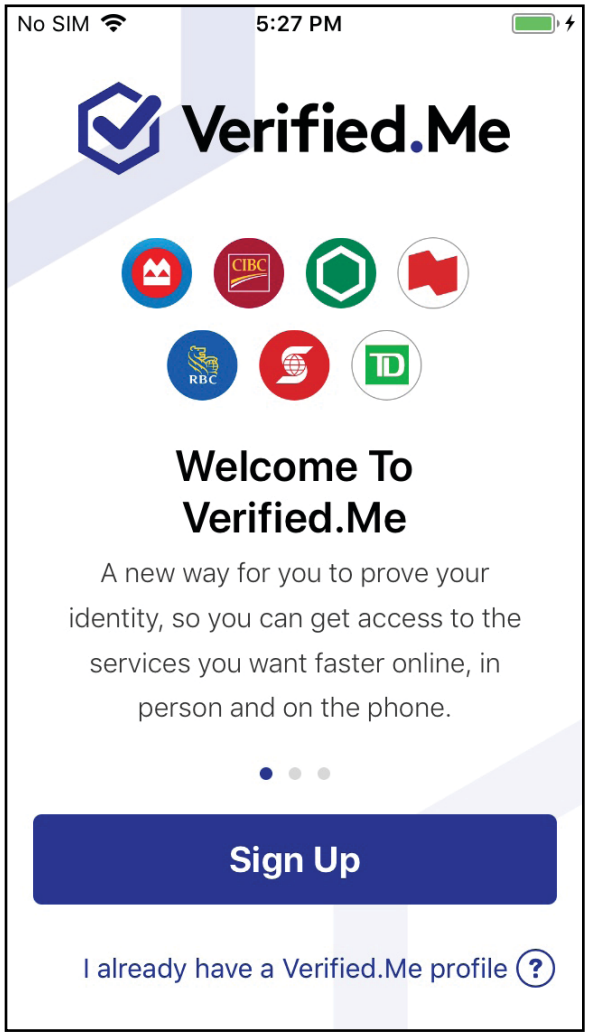
A MOBILE APPLICATION FOR USER-CENTRIC IDENTITY SHARING AND MANAGEMENT SYSTEM

SecureKey Technologies Inc.

Today most identity verification still requires a face-to-face process. With online services becoming more prevalent, and many of them requiring users to provide their identity information, people would like to access their sensitive information and identify themselves in a more secure and convenient manner. Verified.Me is a system proposed by SecureKey that aims to provide a user-centric, privacy-enhanced and secure identity sharing and verification solution. In this system, there is no central point of failure or central point of trust: trusted organizations run distributed network consensus protocols that collectively determine the state of the network, participants, digital assets, and users.

The goal of this project is to build a mobile application as the user agent of the Verified.Me system. The application manages user experience and workflow screens; guiding users through registration, collecting digital assets, executing transactions, and managing the contents of digital lockboxes. As a user agent, the mobile application needs to be intuitive enough to be integrated into the online service transaction flow, and correctly reflect the concept of the user-centric identity management system. In the application, all identity data transactions require user consent. No personally identifiable user data is ever stored in the application and the access to all application functionality is guarded by the state-of-the-art user authentication.

Lijun Pan
IS: Velibor Mandic
AS: Marsha Chechik



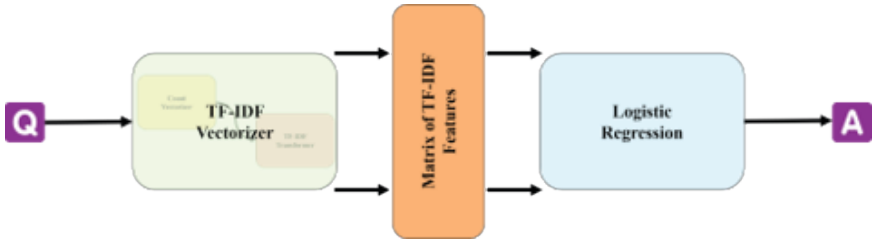
A SEQ2SEQ-BASED CONVERSATIONAL AGENT FOR AD HOC QUERIES IN ANALYTICS AND PRODUCT SUPPORT

CaseWare International Inc.

The main aim of this project is to develop a pure natural language conversational agent (PNLCA), i.e. a chatbot, that provides its users (CaseWare's clients) with answers to their questions regarding product support. Current conversational agents typically do not make use of natural language when talking to its users. This means that in order to interact with a chatbot, one needs to provide command-like phrases with the necessary keywords instead of human-like conversation. Not only does this pose a problem for users who are unfamiliar with the technical jargon necessary, but the development of a PNLCA would be a major breakthrough in the field.

Simply speaking, this is an information retrieval problem, where the chatbot takes the user's questions and returns appropriate answers. Technically speaking, the chatbot is built upon a Seq2Seq (LSTMs & GRUs) model that passes a user question through an encoder that returns an embedding of the question as a hidden state. This is later used as input to the decoder. The embedding is a sequence of numbers, where each number represents a particular word in the training vocabulary. The decoder then tries to generate an appropriate sequence of numbers, which are later converted to words against the chosen vocabulary, that matches the input question. In addition, the Seq2Seq model makes use of transfer learning by being trained against a corpus of different datasets, including a fraction of the CaseWare documentation, SQuAD, and StackOverflow, to finetune the hyperparameters of the model.

Yomna Omar
IS: Brett Kennedy
AS: Frank Rudzicz



ABNORMAL TRADING ACTIVITY DETECTION

TMX

In the current era of high-frequency trading (HFT), institutional investors are disadvantaged because of their relative high-latency compared to co-located traders. To level the playing field, we seek to understand what features are present in the market when there is normal versus abnormal trading activity. To do so, we use various cluster analysis methods like Gaussian Mixture Models (GMM) to determine what conditions are better for slower traders. Due to the high dimensionality of the data, we also explore dimensional reduction techniques before clustering (Principal Component Analysis + GMM), and those that simultaneously perform dimensional reduction and clustering (DEC).

Franco Ho Ting Lin
IS: Charlie Frantowski
AS: Sebastian Jaimungal

ACCURATE IDENTIFICATION OF POI VISITS FOR REFINED TRADE ANALYSIS & AUDIENCE SEGMENTATION

The Weather Network

Pelmorex (owner of the brands *The Weather Network* and *MeteoMedia*) operates weather information services throughout Canada for various media channels. In particular, The Weather Network mobile application for smartphones/tablets is one of the most downloaded and top-rated in the country in both Apple and Google ecosystems. The mobile application is designed with a "follow-me" feature, which enables collection of location data for the purpose of receiving locally relevant weather information and advertisements. Our data collection policies adhere to the guidelines and standards set out by the Canadian privacy commissioner, PIPEDA and GDPR. These strict rules ensure that no personally identifiable information is collected.

The goal of the research project is to improve the accuracy of the identification of a Point of Interest (POI) visited by users and to increase the total number of unique visits that are identified. The visits can be identified by training the machine learning model with features such as time spent, point of interest attributes, day of week, hour of visit etc. The methodology proposed disentangles and determines accurately the POI visited given the case of overlapping POIs. These accurately identified visits, on their turn, can be utilized to produce valuable services for consumers and advertisers to engage the audience with personalised advertisements based on location history. In addition, this data is further leveraged to determine POI visits to enable the creation of audience segments and aggregated visitor insights, such as popular visit times.This approach helps recognizing the pattern observed based on anonymous ID and allows accurate profiling for weather related content and advertising targeting for an enhanced user experience.

Shilpa Rajagopal
IS: Wilson A Higashino
AS: Matt Medland and Nathan Taback

ADAPTIVE ARTIFICIAL AGENT FOR SMART HOMES

ecobee

Ecobee is a Canadian home automation company that makes smart home devices for residential and commercial use. The devices connect to the internet and help automate homes, while saving energy and keeping users comfortable. Ecobee builds superior products for smart homes, in terms of design and function. By integrating a robust AI capability, ecobee tends to make people's lives simpler by providing comfort, safety, and convenience.

The adaptive artificial agent for smart homes is a smart hub, which can receive a variety of signals through sensors placed in ecobee thermostats, light switches and other smart devices. It utilizes machine learning solutions to better understand the devices' environment and help it to make better decisions. Various sensors in the home create a sensor network that can be used collectively to infer an accurate detection of occupancy, location and person identification. These inferences give the artificial agents the ability to make better decisions to control smart home devices, such as thermostats, light switches and home monitoring systems.

Qi He
IS: Sina Shahandeh
AS: Roger Grosse

AI SOLUTIONS FOR SMART HEALTHCARE

Deloitte

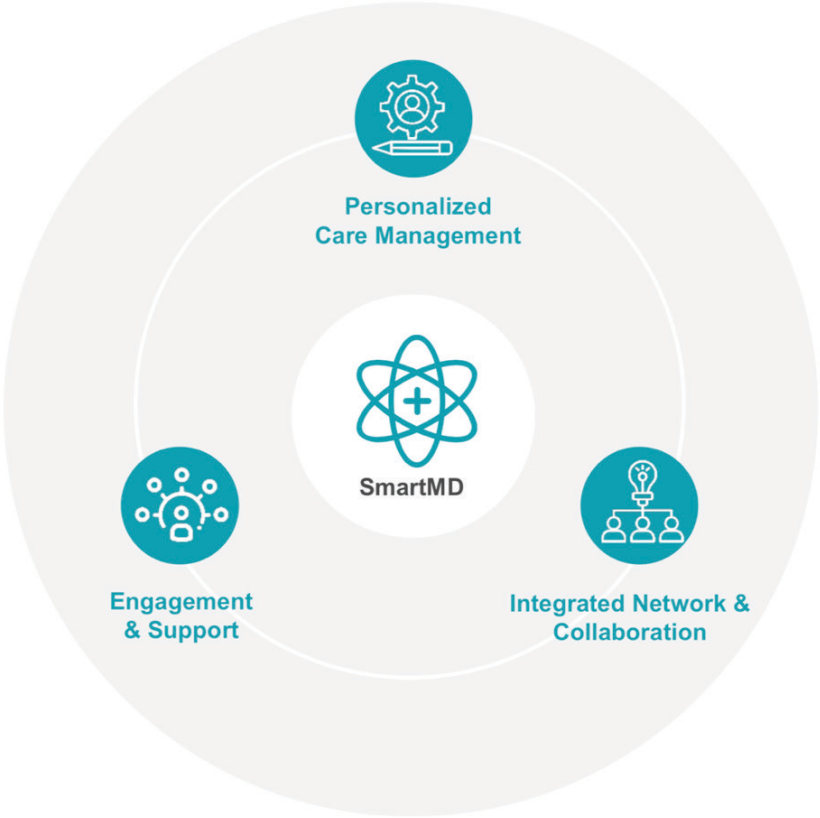
SmartMD is a patient-physician engagement AI solution that is designed to optimize the patient-physician interaction and overall patient journey through an automated voice-to-insight algorithmic system.

The solution is developed through a customized cloud and AI infrastructure that consists of five components. The first is the 'interaction data capture' which records physician-patient interaction using an interactive recording tool. Audio dictations and conversations are then converted into freeform text using a customized software API.

This transcript is then passed into a natural language processing module, where AI and natural language processing algorithms analyze text to understand what and how words are used, and organizes them into a structured database. This is the key deliverable in the project and it involves exploring supervised and semi-supervised knowledge extraction & language representation approaches, such as named entity recognition, relation extraction, sentence encodings and intents classification.

The next component is the predictive modelling for the insights. The goal here is to build predictive models and machine learning algorithms on the structured data to extract insights on the patient, using medical literature and all the pertinent information collected. The final component is insight deployment where insights are re-integrated back into a clinical or hospital electronic medical record (EMR), enabling physicians to make more informed, data-driven and faster decisions for patients through a visual dashboard.

Mali Sankaranarayanan
IS: Alik Sokolov
AS: Suzanne Stevenson



BUILDING A DATA PLATFORM FOR REVENUE REPORTING AND BI-EXPLORING FACTORS IMPACTING REVENUE

Boat Rocker Media

Multi Channel Networks (MCNs) are media companies with large libraries of media assets. In partnership with YouTube, MCNs provide global services such as content monetization, asset management, and right protection to their clients. These range from media studios to companies dealing with several YouTube channels. This results in big data from many sources for each asset, including revenue, ads performance, claims, demographics, traffic sources, and device types.

The size of the media library is comprised of hundreds of thousands of assets. Data is collected in individual CSV files, which makes reporting revenue a challenge. In addition, it is desirable to add additional factors that impact client's revenue on YouTube.

The objectives of this project are to create an automated data platform in the cloud to ingest, process, and store data, and make it available for reporting, business intelligence, and real-time predictions. Additionally, we wanted to develop an automated revenue reporting solution and a business intelligence framework for revenue and asset performance analysis. This would allow us to identify factors impacting revenue on YouTube and deploy a model for revenue prediction.

The data platform, revenue reporting, and BI were built using Python, and Google Cloud Products including BigQuery, AppEngine, Cloud Storage, Cloud SQL, Developer API's, and Data Studio.

Multiple Linear Regression and MLP-Regressor models, consisting of 260 features, were fitted to the data with an adjusted R2 of ~0.9. Predictive revenue modelling and deployment using Google Dataflow and Cloud ML is scheduled for December 2018.

Sayed Hassan Naqawi
IS: TJ Alston
AS: Nathan Taback

CANNIBALIZATION AND HALO EFFECT IDENTIFICATION AND AUTOMATED MODEL TUNING FOR RETAIL

Rubikloud Technologies Inc.

Promotions in retail, such as price discounts, always has a significant impact on the sales of the product. However, the sales promotion of one product may, in addition, have significant secondary effects on the sales of other products not included in the promotion. For example, someone who buys butter may not buy margarine, or someone who buys diapers will buy baby wipes.

Rubikloud's machine learning pipeline is the backbone of their retail data science products. The machine learning pipeline needs to be configured and customized for each new client. This process involves choosing the right models from the available set of models, using the appropriate features per model, and also other configuration parameters. This is currently tedious manual work, and adds several weeks to the deployment time per client. As the company scales, Rubikloud would like to have a short and predictable deployment time per client.

Therefore, there are two main objectives of this project. First, to build a machine learning model to identify products that cannibalize other products and products that have a halo effect on other products. Second, to design and develop a system that can automatically come up with a set of models with corresponding features for a specific problem (e.g., promotional forecasting) that provides excellent or acceptable performance for a given client. The automated process would help to build the machine learning pipeline. This involves a smart (non-brute force) method for speeding up some of the choices required (e.g., model/feature combinations). Adding these features to the machine learning pipeline would sharply improve the accuracy of promotion forecast, as well as the efficiency and scalability of the existing machine learning infrastructure.

Lan Yao
IS: Kanchana Padmanabhan
AS: Anthony Bonner

COMPREHENSIVE VAR MODEL TO MEASURE THE MARKET RISK FOR TRADING BOOK ASSETS

BMO Financial Group

BMO has a large number of financial products and instruments and the complexity and number of products are growing. The market risk valuation system BMO currently uses sees challenges in evaluating different financial products in a timely and efficient manner. With these challenges, BMO is developing its Market Risk Next Generation (MRNG) system. The MRNG system uses full revaluation historical simulation techniques. It is designed and developed to replace the current system, which uses the Monte Carlo method. It also comes with a more streamlined and integrated technology platform for the end-to-end process.

Banks are also required to calculate their stressed VaR taking into account a one-year observation period relating to significant losses. One of the challenges is the incompleteness of data in the previous historical period for relevant market factors. The project focuses on 1) implementing the new MRNG system with the historical VaR model 2) measuring its impact on the bank's trading book portfolio as well as the capital requirements 3) exploring new methods to improve the stressed period selection process. This project benefits the bank by producing timely and reliable market risk metrics and providing accurate global reporting, to facilitate decision making by senior management and compliance with regulatory requirements.

Siyu Ji
IS: Roy Gunawan
AS: Sebastian Jaimungal

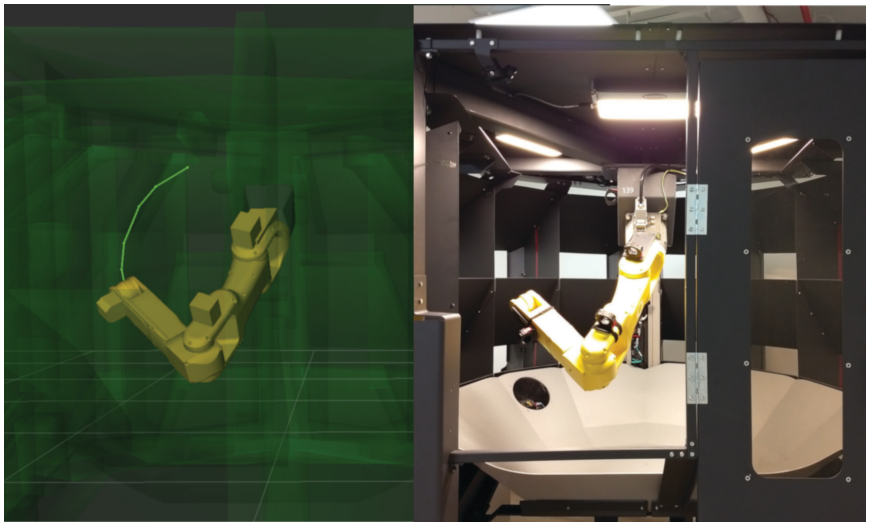
CONTINUOUS VIDEO CLASSIFICATION FOR ROBOTIC GRASP VERIFICATION

Kindred, Inc.

Kindred's SORT system consists of a robotic arm capable of grasping items from a cluttered bin and sorting them into putwall cubbies. Grasp verification refers to the crucial step during the sortation loop of verifying the number of items held in the gripper prior to attempting to stow in a cubby. The goal of this step is to quickly determine whether there are zero (empty pick), one (single pick), or multiple (multi-pick) items in the gripper. In the case of an empty pick or a multi-pick, the grasp must be aborted and re-attempted.

Historically, grasp verification has been formulated as an image classification problem with only two images as input (front and back images), making it difficult to distinguish between single picks and multi-picks. In an attempt to improve the accuracy of the grasp verification classifier, the proposed approach aims to take advantage of the additional information contained in video clips while minimizing the increase in the time-to-decision. The proposed classifier continuously captures images from the front and back cameras to improve the quality of its predictions, and employs a reinforcement learning agent to dynamically trade-off between prediction speed (number of images processed) and accuracy, in alignment with a business-driven reward function. The resultant model improved the true positive classification rate on the challenging double pick vs. single pick classification task from 49.9% to 63.4%, while maintaining a false positive rate of under 10.0%.

Ryan Dick
IS: James Bergstra
AS: Sven Dickinson



AUTOMATED END-TO-END TESTING OF ROBOTICS SOFTWARE

Kindred, Inc.

Robots depend on numerous hardware devices and often human interaction in order to operate. At Kindred, their robot SORT uses an industrial arm, custom gripper, barcode scanners, and machine-vision cameras to autonomously sort items from an input bin into cubbies. When autonomous sorting fails, the robot will refer to a remote human operator to complete the task. These dependencies on hardware devices and human interaction, in addition to the distributed nature of robotics software, makes automated end-to-end testing a challenge. This prevents the robot's software from being verified as part of continuous integration.

The focus of this project is to develop a set of application program interfaces (APIs) that will enable automated end-to-end testing of robotics software. The benefits include full system tests on incremental changes and improved software reliability. To accomplish this, approaches including full simulation and robot-in-the-loop simulation were investigated. However, in the end, an approach called state induction was devised. This method relies on inducing sensor state from pre-recorded data and observing the behavior of simulated actuators in order to test different scenarios. In comparison to the other approaches, this method allows use of real sensor data to create realistic scenarios, is easier to implement and can enable testing with or without the physical hardware or human interaction. In terms of test coverage, it allows for testing business logic software, sensor and actuator abstractions above the driver level, integration testing of cloud services, and testing of other various parts of the system.

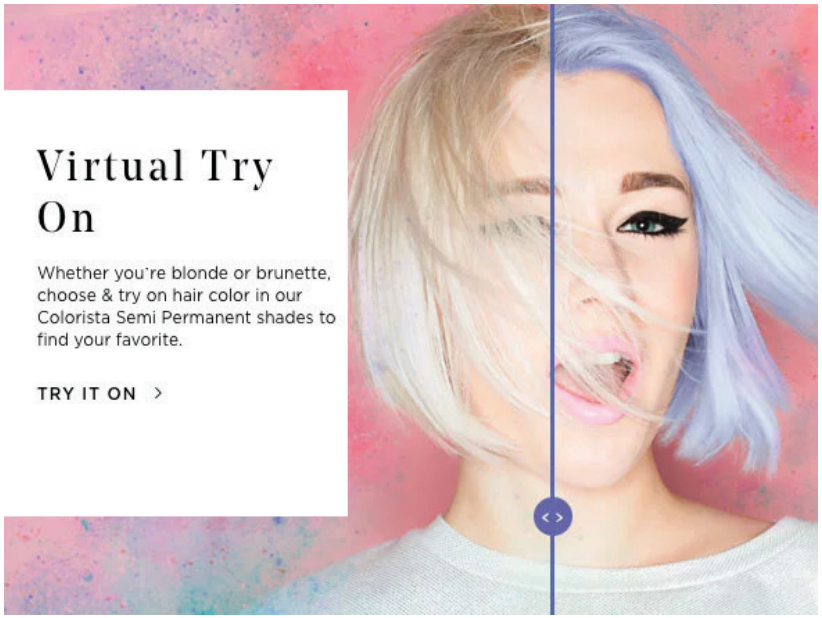
Sachit Ramjee
IS: Neil Isaac
AS: Ashvin Goel

DISRUPTOR IDENTIFICATION

CPPIB

Disruptors are forward-thinking, innovative and ambitious companies that are disrupting marketplaces and industries. The perfect example of these companies are the FAANGs, they are easy to recognize ex-post. The value lies in predicting which companies will be disruptors in the future. The information we use to identify disruptors include traditional financial statement data, stock market data, as well as features extracted by NLP from relevant texts and web scraped data. Due to the high dimensional feature space, we use Machine Learning techniques such as Logistic Regression as baseline and Neural Networks to dynamically forecast disruptor companies within each industry in the the US economy. We then build a trading strategy on top of the identified disruptors.

Franco Ho Ting Lin
IS: Jonathan Briggs
AS: Sebastian Jaimungal



DNN COMPRESSION FOR REAL-TIME HAIR SEGMENTATION

ModiFace

ModiFace is a leading AR experience provider in the beauty industry. Our main product offerings provide the ability to perform beauty try-on simulations via live video on mobile and web, and to track the face, facial features and hair in precise detail. Regarding hair color augmentation, the main challenge is to segment the hair in real time. At the same time we need an accurate and fast model to achieve this. We also need to respect model size constraints on our platforms.

Our hair segmentation network is built with a MobileNetV2 encoder and our customized decoder. The current smallest model of such architecture with acceptable accuracy has a size of 7.4MB. However, a desired model size is 2~3MB. The goal of this work is to compress a bigger but more accurate network to a smaller model while preserving an acceptable accuracy level. In this work, we compare methods, such as knowledge distillation and data distillation, on compressing the model sized 7.4MB to 2.1MB. We have found that the data distillation method and knowledge distillation improves the IoU of the model with a MobileNetV2 encoder (alpha=0.5) by 1.5% and 1.3% respectively, which is quite significant.

Jeremy Ma and Ruowei (Irene) Jiang
IS: Alex Levinshtein and Irina Kezele
AS: Sanja Fidler

DETECTING ACCIDENTS AND IDENTIFYING CAUSALITY WITHIN FLEETS

Geotab Inc.

Geotab is advancing security, connecting commercial vehicles to the internet and providing web-based analytics to help customers better manage their fleets. Processing billions of data points per day, Geotab leverages data analytics and machine learning to help customers improve productivity, optimize fleets through the reduction of fuel consumption, enhance driver safety, and achieve strong compliance to regulatory changes. Geotab is continuously working on new projects to solve customers' fleet management problems.

Car accidents are common in life, but they are expensive to Geotab's customers because of missing scheduled work, losing customers, increased insurance, vehicle repair cost, and vehicle replacement cost. To better help Geotab's 30,000+ customers understand the real risks that exist in their fleet, Geotab started a data science project to detect accidents and identify causality within fleets. This project consists of two phases: the first phase consists of building a machine learning model that is used to detect accidents and to identify the instant that the accident occurred with a precision level down to the millisecond. The second phase is a statistical analysis that is applied to the vehicle's data around the time of the accident, in order to discover the causality of the incident to see if the accidents were avoidable.

The results of this project could increase road safety, prevent accidents, reduce the cost incurred by crashes and even save lives. The next steps would involve incorporating this analysis into our real-time streaming analytics in an attempt to create alerts before collisions occurred.

Meng Zhang
IS: Daniel Dodgson
AS: Roger Grosse

DETECTING ANOMALIES IN VOLUME OF ALERT MESSAGES

Ethoca

Ethoca is the leading, global provider of collaboration-based technology, that enables card issuers, e-commerce merchants and online businesses to increase transaction acceptance, stop more fraud, recover lost revenue and eliminate chargebacks from both fraud and customer service disputes. Through its global collaboration network, Ethoca processes a large volume of alert messages between issuers and merchants on a daily basis. The objective of the project is to explore different machine learning techniques and to investigate how they can be used to detect possible anomalies in volumes of alert messages. The approach adopted is a combination of a supervised machine-learning algorithm to predict volumes and a statistical test to detect when the actual volumes and the predicted volumes may differ significantly. The project also involves working on how the adopted approach could be integrated into existing volume processing environment and tools.

Hervé Dukuze
IS: Neel Punna
AS: Anthony Bonner and Matt Medland

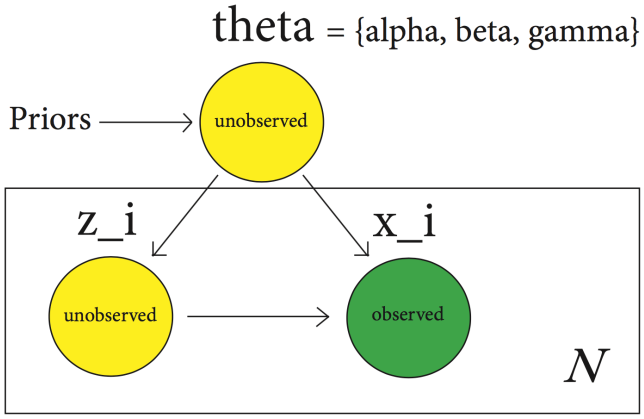
DISCOVERING TRUE SIZES FROM TRANSACTION DATA

Georgian Partners

Georgian Partners is a venture capital firm based out of Toronto. In addition to investing, Georgian also conducts applied AI research via an in-house team. For this project, we worked with one of Georgian's portfolio companies in the e-commerce fashion space to help them better predict what size a piece of clothing a user is going to purchase.

Accurately predicting garment size preference can greatly reduce returns from online purchases and thereby cut costs. Using the fact that most users purchase and keep items that fit them well, we were able to learn a mapping of each garment size to a universal "fit" space, from where we could make size recommendations that work better than just looking at size charts. Specifically, we applied a latent variable model on a large amount of user transaction data, and had the pleasure of working with some modern Bayesian inference methods. Some challenges include coming up with a reasonable model based on expert knowledge, dealing with sparse data, and tweaking gradient estimators to have low variance.

Shu Jian (Eddie) Du
IS: Madalin Mihailescu
AS: Murat Erdogdu and Sanja Fidler



DYNAMIC DATA MODEL MANAGEMENT

Goldman Sachs

To facilitate electronic trading, it is vital to store and manage complex configurations for each client. The settings to be stored depend on the type of client, their trading preferences, the products they use, regions, and regulations. Client Settings is an application that stores and manages these settings. The application stores the settings as structureless documents in a NoSQL database. The various consumers of the data insert and query data through a web interface and a REST API as part of the trade execution workflow.

The data model, which describes the structure of the data that is stored and served, is embedded in the domain logic component and is treated as part of the code. As changes to the data model became more frequent than changes to the code, embedding the model resulted in a slow and inflexible software development lifecycle (SDLC) for Client Settings. The project aims to automate much of the manual process required in a model change, and allow the users of Client Settings to facilitate model changes for their needs without intervention from the development team. The data model is migrated into SecDB, an internal system, for widespread access and ease of release, and separated into contextual components to facilitate ease of changes from each user group.

Robert Kwang Bok Lee
IS: Amrita Rajagopal
AS: Eyal de Lara

KAPPA ARCHITECTURE SERVING LAYER SUITABILITY

Ethoca

Ethoca's Alerts is responsible for fighting fraud and stopping chargebacks through collaboration between merchants, banks and credit card issuers. To support an increasing number of clients and multi-region presence, architectural and technology changes are necessary, moving towards distributed and currently available horizontally scalable solutions. The standard kappa architecture pattern, used to explore how the product could be improved, is a data-oriented software architecture that relies on stream processing. It depends heavily on a suitable serving layer, the software responsible for serving the current application state, largely in the form of one or more data stores (as distributed databases). The complexity of such a system is increased significantly by business constraints, requiring regional data residency (selective geo-replication) and high performance using large datasets. Without a suitable solution, customers concerned with the residency of their sensitive data will refuse to use (or continue using) provided business services. Also, regulatory enforcement agencies may restrict such business in their regions. The objective is to design a correct replication model that ensures data consistency and availability, while complying with regional data restrictions and meeting business performance targets. Every option evaluated should follow high standards regarding fault tolerance, availability, horizontal scalability and query flexibility.

Alexandre Luiz Brisighello Filho
IS: Ronak Patel
AS: Ashvin Goel

KNOWLEDGE REPRESENTATION IN CORPORATE TAX

PwC

With the growing relevance of data, PwC has embraced data & analytics as part of its culture. The Tax practice, specifically, strives to bring value to market by empowering domain experts with automation and data-driven services. Among their initiatives, the Tax Technology team is investigating methods to evaluate the quality of corporate tax returns (T2). Given the complex nature of these documents, this problem requires exploring distinct issues:

1) A T2 can contain hundreds of sub-forms supported by large professional teams within distinct specializations. Scalable methods are needed to identify finalized tax returns among drafts and the large volume of work product produced by these teams.

2) A mapping between each field in a tax return and its corresponding sub-form does not exist.

3) Tax forms contain slips—copies of a tax question which must be answered for any number of relevant cases—yielding forms of vastly different structures.

With the University of Toronto's support, PwC hopes to assess the integrity of its data and develop proof-of-concepts to optimize service delivery, along with identifying practices and procedures that will differentiate them in the marketplace.

Zain Nasrullah
IS: Michael Charette
AS: Nathan Taback

STOCHASTIC SEQUENCE MODELING, AND APPLICATIONS TO NLP AND HEALTHCARE

Layer 6 AI

How can we best model patterns among causal sequences of events? Recurrent Neural Networks (RNNs) are general deep learning frameworks for sequence modeling that build one hidden state for each sequence element, where weights are learned via backpropagating the error. LSTMs help combat the vanishing gradient problem with vanilla RNNs. The above models are usually trained in a deterministic fashion. Deep Kalman Filters (DKFs) propose to learn from a sequence in an unsupervised manner. They are a generative models representing each state by a multivariate Gaussian distribution.

In this project, we study how to improve sequence modeling via introducing stochasticity. We choose natural language processing (NLP) and healthcare as two main domains of applications. The former consists of numerous benchmarked tasks, of which performance keeps being pushed up over years (surpassing humans on question answering for instance). The latter presents tremendous opportunity for improvement with AI. Medical diagnosis often takes the form of sequence codes (ICD-9 or ICD-10), and the history of a patient heavily impacts on a new diagnosis. Thus, we can draw resemblances between modeling population health data and NLP.

In NLP, language modeling consists of going through words one by one, and predicting a probability distribution over the next word. We introduce LatentShift, a stochastic add-on to RNNs that enables us to push the state-of-the-art perplexity on the WikiText-2 and Penn TreeBank datasets.

During this internship, Layer 6 AI partnered with the Institute for Clinical Evaluative Science (ICES). ICES has an enormous population health database, covering every aspect of healthcare for Ontario residents since 1991. We are interested in people with diabetes, the most common chronic disease in Canada. Our goal is to perform unsupervised clustering to identify subtypes of diabetes. Indeed, a recent study suggests the existence of five types of diabetes, questioning the traditional type I/type II classification.

Mathieu Ravaut
IS: Tomi Poutanen
AS: David Duvenaud

ENTITY RESOLUTION ON FINANCIAL DOCUMENTS

Scotiabank

The network ecosystem contains rich entity information. This entity information is complex, has a geographical scope and contains proprietary data, and data from third parties (i.e. – government, customs, and legal entity). Currently, Scotiabank has developed and built an attribute-based entity resolution system, which extracts and links various forms of entity references to canonical entity representations. Then, Scotiabank would like to extend the algorithm to be purely contextual-based and be able to detect and disambiguate entities from noisy text with high precision and decent recall.

For this project, we will leverage the data in the network ecosystem, retrieve useful information, and create a consolidated view of entities. We will extend the classic and state-of-the-art entity resolution framework to new data sources by domain adaptation and model optimization. More specifically, we will focus on named entity linking and coreference resolution on financial documents.

Jindong Liu
IS: Junjie Zhu
AS: Graeme Hirst

EXTRACT INSIGHTS FROM NEWS MEDIA

CaseWare International

Advances in technology are bringing changes to the profession of auditing. Machine learning techniques can potentially bring enhanced insight into the audit process, thus identifying where auditors would better spend their time during a lengthy and complicated audit. During an audit, auditors often need to review a large amount of electronic documents, such as company reports, financial statements, press releases, and news articles. By analyzing information obtained from these sources, auditors can identify areas of high risk and address the issues with clients. This project aims to using natural language processing techniques to extract key information from electronic documents, in order to assist auditors through the document review process.

This project started with collecting and parsing financial news. After mining and analyzing the largely unlabelled financial news data, we developed topic modelling and sentiment analysis algorithms. The topic modelling algorithm learns topics discussed in the news and summarizes them in easy to understand illustrations. The sentiment analysis algorithms can then determine the general tone used by the media for each topic. This system can be applied for an individual company, its business sector, and also the general economy as a whole.

Ran Zhang
IS: Brett Kennedy
AS: Suzanne Stevenson

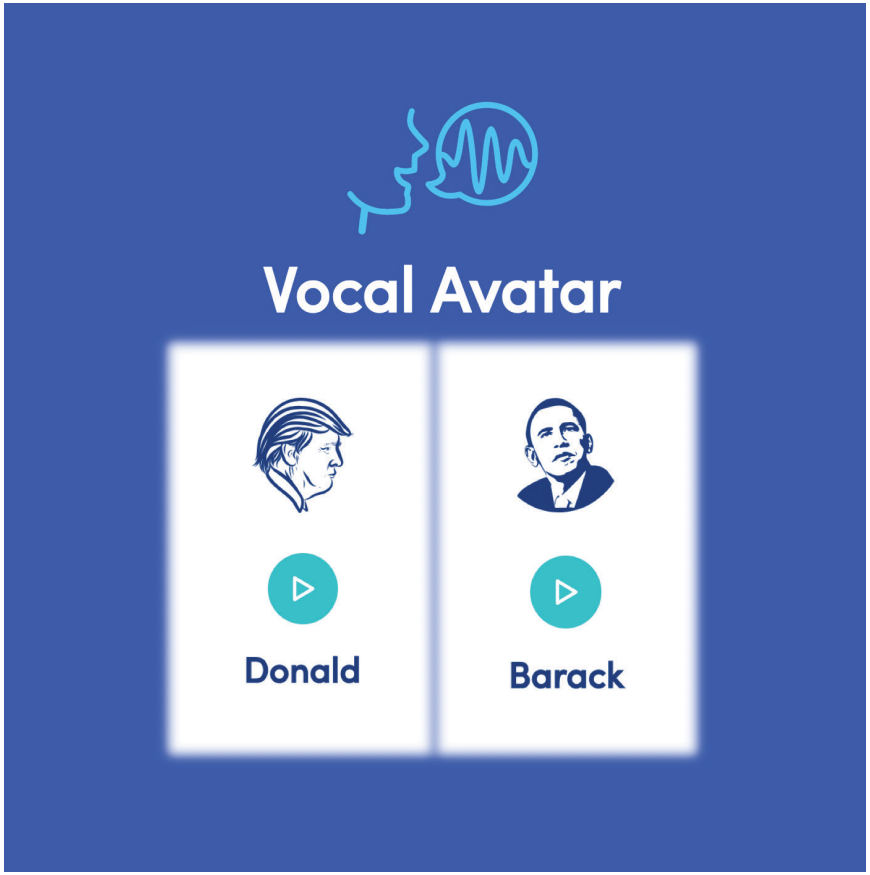
FLEXIBLE DATA READER ON DISTRIBUTED FILE SYSTEMS FOR TRAINING DEEP LEARNING MODELS

Uber ATG

When training machine learning models, and in particular computer vision and other sensor data applications, large datasets (millions of samples) comprised of small sized files are usually read into the model, and presented in a random order at each iteration of the model. This data access pattern is needed in order to avoid local minima and model overfitting. Therefore, the ability to access large datasets randomly is crucial. With the ever-growing size of datasets, it has become increasingly difficult to store such datasets in traditional file systems. However, storing datasets on large-scale distributed file systems pose several difficult challenges: 1) such systems are usually meant for storing large files (a few hundred megabytes each), and 2) such systems use caching and prefetching methods based on assumptions that do not fit the random access pattern. The differences in operating assumptions of data access cause performance degeneration and could considerably slow down model training, an already time-consuming process that can take hours or days.

This project focuses on finding proper solutions for storing large datasets in distributed file systems, while maintaining the crucial ability to access each small size sample in a random manner, and with high performance for training deep learning models. In this project, a monitoring system has been developed to determine the required data ingestion rate of machine learning models. Studies were conducted to collect the data reading performance of existing reading methods. Multiple possible solutions were explored and benchmarked to evaluate how well they perform. Finally, the best solution will be further analyzed and optimized to meet industrial standards, and put in use in the production system.

Hongbo Fan
IS: Inmar Givoni
AS: Yashar Ganjali



GENERATING EMOTIONAL SPEECH IN CLONED VOICES

Lyrebird

Lyrebird has developed a state-of-the-art Text-to-Speech (TTS) model for voice cloning, which is capable of learning from less than 300 seconds of a speaker's recordings to generate new speech in the speaker's voice. This learning process takes less than 90 seconds of GPU time to achieve high fidelity performance. In this project, we work towards a TTS model for cloned voices with controllable prosody. Our goal is to generate high-fidelity speech in cloned voices, with emotions not previously observed in the cloned speakers' recordings. We want to achieve this while maintaining a time-and-data efficient cloning pipeline.

We divide this task into two stages. At the first stage we develop a large multi-speaker TTS model with controllable prosody. We build on top of a backbone sequence-to-sequence neural network model with attention. This model learns to generate summarized audio representations, conditioned on phonetic information. These audio representations can be transformed into audio waveform using signal processing techniques. We augment the model with a style encoder layer, which extracts prosody information from the data as fixed length vectors. These vectors condition the decoding of phonetic information into audio representations. We train this base model using a large dataset with many speakers. We only recorded emotional data for two of those speakers. At the second stage, we fine-tune the trained model on less than 300 seconds of an unseen speaker's recordings. By varying the style input source in the fine-tuned model, we are able to vary the emotion of generated speech in the cloned voice.

Wei Zhen Teoh
IS: Thibault de Boissière
AS: David Duvenaud

NATURAL LANGUAGE PROCESSING TO AUTOMATE THE REQUEST FOR QUOTE WORKFLOW

Goldman Sachs

Clients of large broker-dealer firms regularly communicate with salespeople (through multiple channels, including voice, email, and chat) in order to discover the price of over-the-counter equity derivatives. Salespeople must then coordinate with traders to provide a response (i.e. the "Request for Quote" or "RFQ" process). Often these client requests involve complicated structures comprised of multiple constituent securities, making them tedious to specify manually. Existing systems for parsing such requests rely heavily on rigid syntaxes that fail to accommodate the fluid textual representations of quotable securities often encountered. Automating the RFQ process would therefore lead to a considerable improvement in the speed and volume of quotes flowing through sales and trading desks.

The objective of this project is to build a parser capable of translating free-form text representations of options trading strategies into data formats capable of being fed into existing pricing systems. We implemented a pipeline that combines supervised learning with combinatorial optimization to classify tokens in a given input string before situating them within predefined frameworks that describe the possible structures a trading strategy might have. We also implemented a diagnostic interface to allow user feedback on incorrectly parsed strings and automatically produce new training samples.

Andrew Nelles
IS: Stephen Kekicheff
AS: Gerald Penn and Ken Jackson

OPENSEQ2SEQ

NVIDIA

OpenSeq2Seq is an open source toolkit that implements a variety of state of the art neural network models for natural language processing and speech. The toolkit enables distributed training across multiple NVIDIA GPUs and supports mixed precision arithmetic for acceleration using tensor cores. This enables anyone with a recent NVIDIA GPU to achieve state of the art performance. OpenSeq2Seq has models for translation, speech recognition, speech synthesis, language modeling, and image classification. All models come with an example configuration that can be used to reproduce neural network training and performance. In addition, all models have trained checkpoints for developers who are eager to test these models without having to redo the entire training process.

This internship project involves implementing the speech synthesis part of OpenSeq2Seq, integrating the Tacotron 2 model into OpenSeq2Seq, and improving the performance of the speech recognition models. The Tacotron 2 model translates English text to speech spectrograms which are converted to English speech. Once the speech synthesis portion of OpenSeq2Seq was completed, it was relatively easy to generate a large amount of high-quality synthetic speech. By using this synthetic speech and combining it with natural speech datasets, the speech recognition models for OpenSeq2Seq improved from 6.58% and 19.61% WERs to 4.32% and 14.08% WERs on the LibriSpeech test datasets.

Jing Yao (Jason) Li
IS: Boris Ginsburg
AS: Jimmy Ba

INTELLIGENT REVIEWER AND CONTRACT MANAGEMENT SYSTEM

Deloitte

Contract analysis and management is a laborious process and often proves to be a rate-determining step in business operations. Furthermore, manual annotations by legal subject matter experts can lead to inconsistent (and costly) reviews. In an effort to help in-house legal teams serve their businesses with increased efficiency and accuracy and streamline the contract process, Deloitte is developing a customizable, AI-enabled contract management suite—dTrax. Specifically, the NLP stream for dTrax is focused on development of additional modules that can provide intelligent extraction and serve as an automatic reviewing system.

The primary objective of this research project is automating contract analysis, including clause extraction, text classification, and risk identification. To achieve this goal, we experimented with a mix of unsupervised, semi-supervised, and supervised learning during different stages of development. As a supervised learning task, identifying relevant clauses requires clean and annotated data from legal subject matter experts. To deal with few labelled positive examples, we first used unsupervised clustering and topic modelling. We eventually shifted to a semi-supervised approach to learn from positive and unlabelled data. Finally, we converged on a two-step supervised solution involving a tf-idf and SVM, followed by a universal sentence encoder and XGBoost. We then created a pipeline around this model to extract relevant clauses. Continuing development of this tool aims to allow users to label their own clauses and train their own models.

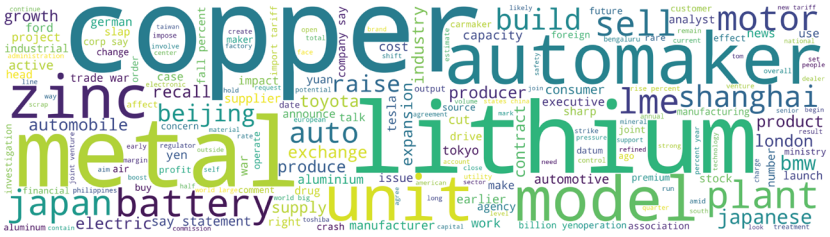
Christine Yuen
IS: Ofer Shai
AS: Qiang Sun

PERCEPTION AND TRACKING USING LIDAR AND RADAR

Uber ATG

For an Autonomous Vehicle (AV) to make decisions and drive independently, it needs to perceive all of the objects in its surroundings, and track them across time to know the state of all objects at any given point in time. In this work, we exploit both LIDAR and RADAR sensors to detect and track objects across time. These two sensors are complementary: RADAR returns sparse point clouds with doppler velocity, while LIDAR generates 3D point clouds. Our initial experiments show the benefits of using both sensors.

Satya Krishna Gorti
IS: Raquel Urtasun
AS: Sanja Fidler



MULTI-INTERACTION FRAMEWORK AND SCHEDULING

Google

Google's mission is to make the world's information more accessible and useful. The role of the Display Ads Infrastructure team at Google is to implement infrastructure that enables emerging use cases for Display Ads. This project contributed to the central request routing server in Google's Display Ads system, which sits behind the display ads front end and is responsible for collecting ad candidates from multiple sources, deciding the winner, and performing a variety of post-ad-selection processing tasks (e.g., fetching ad-rending data, logging, customizing and returning the winner to the front end).

This project was divided into two parts. The first part involves designing and implementing a framework through which internal developers could make callouts with inter-dependencies to multiple backends within a single request. The framework is now used in both the ad-rendering data fetch ad-selection portions of the ad serving stack. The main contributions to this part of the project were around proposing multiple potential frameworks while considering simplicity, flexibility, and performance at scale.

The second part of the projectias to analyze a component in the framework responsible for scheduling remote tasks. Scheduling could be done arbitrarily, but this project proposed an optimal ordering, as well as evaluated the gains in performance that could be achieved as the the number of multi-interactions scaled.

Michelle Arkhangorodsky
IS: Kimi Chung
AS: Matt Medland

METHODS FOR EFFICIENT AND EFFECTIVE DOCUMENT READING COMPREHENSION IN THE LEGAL DOMAIN

ROSS Intelligence

Machine reading comprehension (MRC) is the task of having a computer automatically answer a question, given a passage of text. Usually, but not always, the answer can be found in the passage. MRC has proven to be a very challenging task, given that it requires some combination of reading, processing, comprehension, and in some cases, summarization.

Deep Learning has been successfully applied to MRC, reaching and sometimes even beating human performance in popular benchmarks. In this project, ROSS is leveraging state of the art deep models to perform MRC in the legal domain. With this, ROSS can answer questions in the context of legal documents, powered by its proprietary dataset. This dataset was transformed and enhanced by this project to better support the task at hand. This project is also taking advantage of two key recent developments that have further improved the performance of deep learning models: attention mechanisms and word embeddings.

Attention mechanisms allow machine comprehension models to learn the correlation between words in sentences automatically. In this way, the models achieve an increased understanding of the structure and meaning of the text they are parsing.

Embedding map words and characters into vector representations better capture the higher-level conceptual relationships between them. GloVe and Word2Vec are two examples of popular pre-trained embeddings typically applied to Machine Reading Comprehension. This project utilizes these and other embeddings to enhance ROSS' capabilities in this task.

Simon Rojas and Antonia Mouawad
IS: Jimoh Ovbiagele
AS: Frank Rudzicz

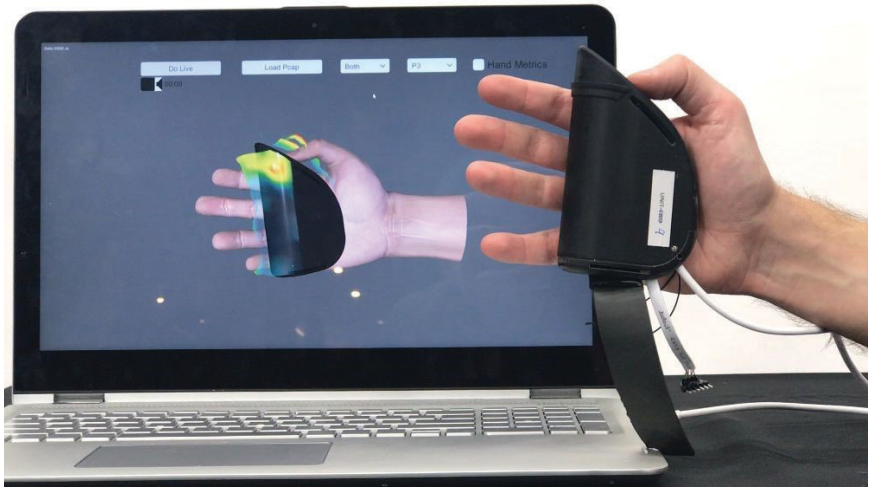
REAL-TIME HAND POSE RECONSTRUCTION THROUGH FAST MULTI-TOUCH SENSORS

Tactual Labs

Hand pose estimation is a challenging but active direction of research. Its success is significant to an improved seamless and immersive user experience in applications such as virtual reality (VR) and augmented reality (AR). To date, the research on hand pose estimation has predominantly relied on data captured by depth camera. In contrast, the sensing technology at Tactual Labs captures the distance information between the sensor and user's hand through capacitive sensing. This could serve as another data source for addressing the problem. In particular, the capability of the sensor in surrounding free-form surface and conducting real-time sensing gives it the potential to provide a better user experience over current hand controllers.

This project aims to accomplish hand pose reconstruction, based on the fast multi-touch sensors provided by Tactual Labs. A hand model containing 22 key point positions has been developed to explain the hand kinematics (e.g. flexion/extension, abduction/adduction), which are used for both computing training datasets and reconstructing the virtual hand. Machine learning algorithms including convolutional neural networks (CNNs), support vector machines (SVMs), k-nearest neighbors (KNN), and random forest (RF) are experimented with to regress the hand kinematic parameters. The resulting system can produce real-time visualization of the hand pose reconstruction when users interact with Tactual Labs' hand controller.

Yanjun Jiang
IS: Bruno De Araujo
AS: Karan Singh



PREDICTING VENTURE SUCCESS AND EQUITY VALUE CREATION (EVC) USING MACHINE LEARNING

Creative Destruction Lab

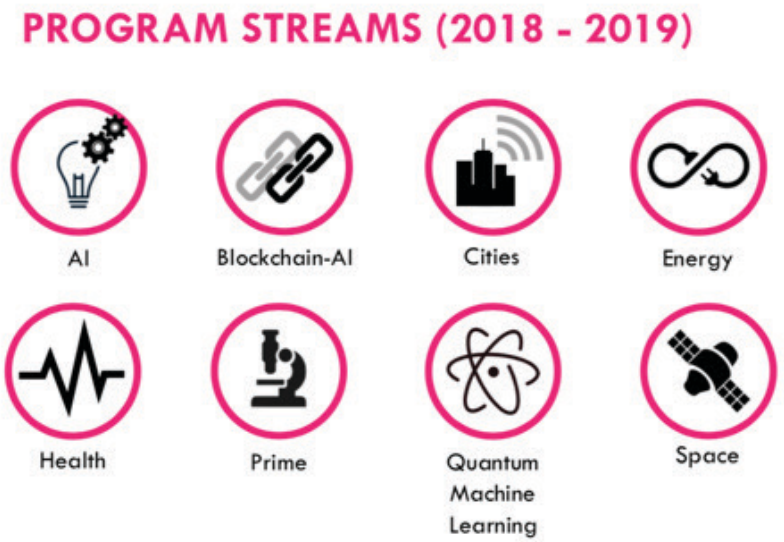
Creative Destruction Lab (CDL) is an entrepreneurship program for highly-scalable science-based start-ups. Currently, it spans across six North American locations and has partnered with five other universities across Canada and the United States in 2018, with an overall cohort size of 375 companies.

The CDL-AI Project (CAP) aims to develop an AI infused predictive engine for venture success and Equity Value Creation (EVC), using the extensive and detailed data generated from the evolution of early-stage ventures that participated in the CDL's 9-month program. The significance of this project is motivated by two primary imperatives: one scientific and the other economic.

The internship entails building a prediction engine from the ground-up, applying state-of-the-art research techniques in natural language processing (NLP) and machine learning. The work includes streamlining various raw sources of CDL's proprietary multi-modal venture data (including imbalanced categorical samples, unstructured textual content, and meeting dialogues), and building a production pipeline. A research grade textual data corpus was compiled with a modular context-aware word embeddings (ELMo) generator.

The promising preliminary results and work during the internship will help future scaling of the project and contribute to the scientific exploration of 'venture success and causalities' – a topic of great interest both from an academic research and economic development perspective.

Raeid Saqur
IS: Amir Sariri
AS: Frank Rudzicz



OPERATING ROOM TRAFFIC ASSESSMENT-A VIDEO ANALYSIS APPROACH

Surgical Safety Technologies

Operating room (OR) traffic, typically surgical staff movement, has a close relation with surgical safety. Much displacement in the OR not only distracts the surgical team but also pollutes the sterile environment and increases the risk of infection. To make a quantitive assessment of OR traffic, detecting, tracking and counting people in the OR became an important research topic in the field of surgical safety. These tasks are aimed at different groups of staff in the OR, including anaesthetists, surgeons, nurses, technical support, observers, students, researchers, etc. Performing analysis on the number of people present, as well as their location and trajectory in the OR, can help to find the correlation between OR traffic and events/errors during the operation. Furthermore, the result can provide information to assist in training surgical staff and building OR guidelines.

Recently, with the boom in deep learning, computer vision problems have become much easier to solve with the features captured from neural networks. The algorithms of object detection, tracking and counting in other fields like sports, surveillance, and self-driving inspire the improvement in OR traffic assessment. Starting from convolution neural networks (CNNs) and long short-term memory (LSTM), the surgical staff detection and counting achieve promising results, and outperform common baselines in crowded, high-occlusion and similar-dressing scenarios. The outcome of our algorithms will assist surgical analysts and are to be deployed in the OR safety monitoring system in the future. Together with high-quality labeled data, the project shows the way of future research of OR traffic and surgical safety.

Tianbao Li
IS: Teodor Grantcharov
AS: Sanja Fidler

INTRAOPERATIVE PERFORMANCE MEASUREMENT OF SURGICAL OPERATORS USING DEEP LEARNING

Surgical Safety Technologies

Performance evaluation of surgical operators is currently done by experienced doctors who either rely on real-time live observations, or the review of surgical videos, to assess the ability of surgeons based on set criterion. This task is tiresome and prone to subjectivity. Due to the large number of surgical procedures performed daily, it is unfeasible to have doctors review every procedure. Thus, there is a loss of invaluable performance data that would be invaluable to surgeons as they hone their craft to become more effective at their jobs. The aim of this project is to leverage computer vision and deep learning techniques to create pipelines that can provide a quantitative analysis of surgical performance. The ability to classify or grade surgical performance has the added benefit of potentially improving surgical outcomes, as studies have shown there exists a correlation between performance and outcomes.

Shuja Khalid
IS: Teodor Grantcharov
AS: Sanja Fidler

QUESTION ANSWERING-QUERY ON DEMAND WITH NATURAL LANGUAGE

Scotiabank

Scotiabank has developed and built an ecosystem that contains Scotiabank proprietary data and data from third parties (such as customs, government contracts and legal documents). This ecosystem contains data that has been transformed into network data structures and entity reference data. In many cases the data is collected daily and the size of the data grows substantially. Having such a rich and vast pool of data, various tools have been developed to help access data and retrieve information. However, we want to make it even faster and more convenient for our partners to gain instant access to all the information they need. The ultimate goal of this project is to build a tool that will search several data sources to extract relevant information and display the results, all in one centralized application.

Our Question Answering (QA) system allows end users to query data by posing questions in natural language. The QA system then answers these questions by pulling data from structured and unstructured data sources. This makes information accessible to both technical and non-technical users since questions can be asked using natural language rather than being queried using a programming language.

Navid Kaihanirad
IS: David King and Menaka Kiriwattuduwa
AS: Graeme Hirst

RECOMMENDATION SYSTEM FOR RETAIL SHOPPING

Loblaw Digital

Loblaw Digital is responsible for building and operating the digital channels for Canada's largest food retailers. At Loblaw Digital, we create experiences that span the physical and the digital, including online grocery offerings, traditional e-commerce, loyalty, and pharmacy products. This project involves building a cloud-native recommendation system, with machine learning and data mining models, on the Google Cloud Platform. The system understands the customers base and products, and also improves shopping experience by offering personalized recommendations. The retail recommendation is different from traditional e-commerce as the basket is substantially larger and customers tend to buy the same product over and over again. The models use implicit feedback, like purchases history and click streams for ten million users on 200,000 products. The interaction data is so sparse that users only interact with a few popular items. The models also use the product metadata, such as nutrition, brand, and name. This project uses item-to-item collaborative filtering and matrix factorization models, such as weighted regularized matrix factorization and Bayesian Personalized Ranking. The matrix factorization methods learn latent factors for the users and item, which are used to calculate user preferences on items and similarity between items or users. Additionally, this project also explores the frequent pattern approach to find items that are frequently bought together.

Wei Zheng
IS: Richard W Downe
AS: Nick Koudas

REDUCING UNNECESSARY RETURNS TO THE EMERGENCY DEPARTMENT AT THE HOSPITAL FOR SICK CHILDREN

SickKids Research

Every year, approximately 50,000 children and their families visit the Emergency Department (ED) at the Hospital for Sick Children, and while an overwhelming majority get sent home, a small but significant number return within 72 hours. Of these patients that return, most do so and receive no further tests, treatments, or diagnoses; instead, they simply receive reassurance on the predetermined course of treatment and are once again, sent home. By returning to the ED unnecessarily, these children and their parents are not only subjected to long wait-times for what proves to be a fruitless visit, but they may spread or pick up new infections while waiting. In addition, the unnecessary returns add to the burden on the ED, increasing wait times for all and placing more pressure on the physicians and nurses.

Using ED data from 2008 – 2018 comprising of 655,944 unique visits made by 303,431 unique patients, this project seeks to predict the risk of an unnecessary return for each individual patient in the ED. Comprising of administrative, demographic, physical exam and treatment data captured in 762 numerical, categorical and natural language inputs, we seek to build machine learning classifiers to first, predict the risk of a return to the ED within 72 hours, and second, predict the likelihood of the return being unnecessary. Once these high-risk patients have been identified, they can be targeted for interventions to help prevent their unnecessary returns.

Lebo Radebe
IS: Anna Goldenberg
AS: Lei Sun

MACHINE PREDICTION OF PATENT SUCCESS

Legalicity

Patents allow companies and individuals to protect their inventions. In many cases, patents can become valuable business assets. However, the process of obtaining a patent is very costly and time-consuming. Therefore, it is important to evaluate whether an idea is worth pursuing before spending considerable resources trying to patent it.

Legalicity's flagship product, NLPatent, helps lawyers and inventors evaluate the patentability of new ideas by identifying and retrieving the most relevant existing published patents. This is accomplished using the latest insights from Natural Language Processing (NLP) research to capture the semantic content of patent documents.

The objective of this internship project is to predict the success of any patent application based on its textual content. This involves automatically estimating the likelihood of rejection of the application on several grounds, as well as suggesting possible improvements to the invention description (e.g., by detecting problematic language).

We experimented with a variety of Machine Learning (ML) and NLP techniques using a large, publicly-available dataset from the U.S. Patent Office (USPTO). We developed prediction models for a variety of technology areas, covering over two million U.S. patent applications from 2008 to 2017. Our approach not only provided a novel application of ML and NLP for patentability prediction, but we achieved prediction accuracies up to 89%. This is a significant improvement over existing research, which typically focuses on small numbers of patents within a single technology area.

Chenfei Wang
IS: Yaroslav Riabinin
AS: Yang Xu

MEASURING SENTENCE SEMANTIC SIMILARITY FOR QUESTION ANSWERING SYSTEMS

RSVP.ai

Sentence semantic similarity measurement has been a very popular Natural Language Processing (NLP) research topic. It is also an important task as it helps to solve a number of general NLP problems, e.g. sentence classification, question answering (QA), semantic parsings. In QA systems, query questions are matched with the most similar candidate question from a question database by measuring their semantic similarities. The answer to the best matching, i.e. most semantically similar candidate is returned as the answer to the query question. Constructing a question-question (QQ) semantic similarity measurement model that works on various domains is very challenging, as data from different domains have various characteristics. Also, measuring sentence similarity is a high-dimensional problem — a change of one word would make the sentence mean a very different thing. In this project, we propose a Word Mover's Distance to add to an unsupervised baseline model. This model works on data from various domains. We also propose a transfer learning model that can easily adapt to new vertical domains with less training data. We show that our models outperform the state-of-the-art QQ similarity ranking algorithms.

Tianyi Chen
IS: Kun Xiong
AS: Frank Rudzicz

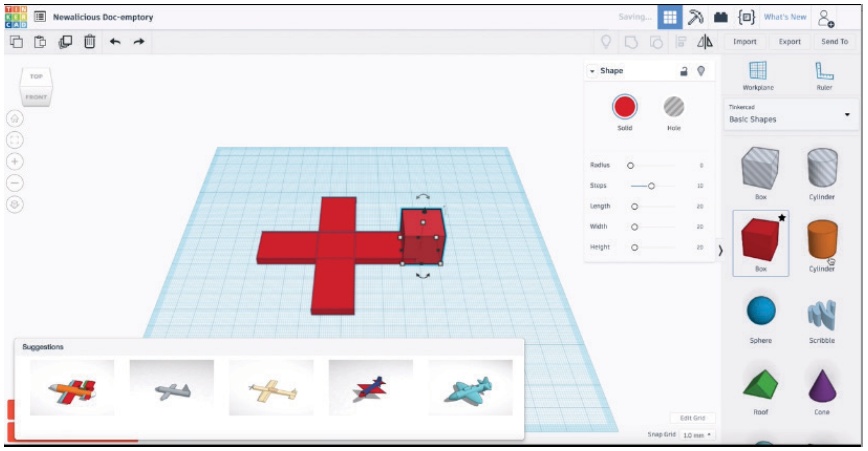
UNDERSTANDING THE SEMANTICS AND CREATION PROCESS OF 3D OBJECTS

Autodesk Research

Understanding the 3D world around us is a key goal in computer science. If we are endowed with 3D understanding, we can solve problems in several areas in computer science such as robotics, fast 3D modelling and physical reasoning. One such task involves modelling 3D objects in a scene, which is very tedious, and automating certain parts can significantly simplify this process. The task usually involves adding and changing the internal parameters as well as the affine matrix of primitives in a scene. So, there is a complex design process and trail of history in creation of a complete 3D object. We can simplify this process by using machine learning models on this design process and on discerning the geometric properties of the object. The goal of my internship involves researching this process and the semantics of 3D objects.

The geometric understanding of the complete 3D object can help in recommending users similar or more interesting 3D objects. Also, with this knowledge, 3D objects can be classified. This can help in internal tagging for a real-world software. Instead, we can also use machine learning on the creation process. Understanding the creation of a 3D object allows a more systematic approach to do these tasks but also gives us the additional ability to autocomplete/suggest complete objects based on partial shapes. The project involved investigating these interesting ideas and developing real world applications based on them.

Aditya Sanghi
IS: Ara Danielyan
AS: Alec Jacobson and David Duvenaud



EARLY PREDICTION OF ALZHEIMER'S DISEASE FROM SPONTANEOUS SPEECH

Winterlight Labs

Winterlight Labs has developed an AI diagnostic platform to detect cognitive impairment, associated with dementia and mental illness, by analyzing short snippets of speech and extracting hundreds of markers in speech and language. One of the most important challenges for applications of the technology for clinical trial screening or use in primary care, has been the ability to detect the symptoms of cognitive impairment early, in order to increase the effectiveness of potential treatments and therapies. The pathophysiological process of Alzheimer's Disease (AD) has been shown to begin years before clinical diagnosis, and the study of the preclinical stages of the disease has been highlighted as a critical opportunity for disease-modifying therapy and intervention. Modelling of longitudinal speech data using complex machine learning models could prove helpful in providing a better understanding of the non-linear trends of cognitive decline.

Winterlight has collected a dataset of spontaneous connected speech recordings from individuals with a confirmed clinical diagnosis of AD, and from individuals in healthy aging specifically curated for the purpose of evaluating the feasibility of early detection of AD. Through exploration of the trajectories of pre-clinical cognitive decline and significant features across subjects in the dataset, we observed that cues of AD can be detected up to six years before diagnosis. We present a novel dual feature decomposition model to correctly classify healthy people from those with Alzheimer's disease years before clinical diagnosis.

Aparna Balagopalan
IS: Jekaterina Novikova
AS: Yang Xu

SLINKY: QUERYING LARGE-SCALE BEHAVIOURAL ANALYTICS DATA AT INTERACTIVE SPEEDS

Shopify

Shopify provides a platform for merchants to run and grow their business. We collect a massive amount of data on behalf of our merchants, and utilize it to help them make successful decisions. The data we collect, and the insights drawn from it, provide us (and our merchants) with a competitive advantage.

This research project tackles the challenge of querying large-scale analytics data in order to draw insights in (as close as possible to) real time. Specifically, we've built an in-memory behavioural query engine that can be used to answer questions about how and why our users perform actions at interactive speeds (under 30 seconds). Our project mainly focuses on providing answers to funnel queries, which are specified by a sequence of events, in order to understand conversion rates between each step of the funnel.

Performance of behavioural analytics queries on top of traditional data stores are highly dependent on having a full view of users events at the same time, limiting parallelism and requiring expensive join operations. Our system is built on top of an open-source data storage and processing layer - TrailDB - that stores discrete event data organized by a unique user key. With this user-centric model, we were able to effectively parallelize workers to run on disjoint groups of users. A global aggregation can then performed by simply summing the results from individual workers. In this way, our model is similar to MapReduce, but avoids the expensive shuffle step. In addition, we are able to capitalize on TrailDB's impressive compression capabilities to improve performance by storing transformed data in-memory. Further optimizations were researched and implemented to further improve performance, such as low-overhead user and event filters. Time-sharding and stitching was also implemented in order to extend the amount of data the system could support in real-time from days to months.

Michelle Arkhangorodsky
IS: Zeeshan Qureshi
AS: Bogdan Simion

SUPPORTING CONTEXTUAL INFORMATION IN OUR SEARCHES

ROSS Intelligence

Currently, for most relevance matching tasks valuable information, which reflects on the context of the documents, are purposely discarded by ad hoc retrieval models and thus impact the ranking negatively for the gain of performance. The goal of this project is to use the latest research in information retrieval & natural language processing (NLP) to improve on the relevance matching task. Concepts in NLP such as focusing on contextualization or semantic inference at a character, word, or entence level found in semantic matching models, were shown to provide significant improvements over those which do not.

Semantic matching models use similarity matching signals, compositional meanings, and global matching requirements. Ad hoc relevance matching uses exact matching signals, query term importance, and a diverse matching principle (Verbosity Hypothesis). Representation-focused models attempt to match a single text between compositional and abstract text representations. Interaction-focused models build local interactions between two pieces of text and learn hierarchical interaction patterns for matching. “Interaction-focused deep matching models and representation-focused deep matching models address the ranking task problem from different perspectives, and can be combined”, as stated by Professor Jiafeng Guo. It is also not popular to have a model which supports both semantic and ad-hoc relevance matching properties while remaining performant. This idea serves as the main precursor for this research. Such a model would tackle different perspectives that aligns with industrial needs.

Phileas Hocquard
IS: Jimoh Ovbiagele
AS: Scott Sanner

TWO-STREAM NETWORK FOR VIDEO-BASED THERMAL INJURY DETECTION

Surgical Safety Technologies

Surgical Safety Technologies specializes in enhancing patient safety through analysis on perioperative factors, which includes intra-operative analysis on adverse events such as bleeding and thermal injury. Traditional thermal injury detection includes post-procedure reviews from clinical analytics, or using electrodiagnostic medicine techniques during procedures.

In this project, we propose a machine learning model that can detect thermal injury from internal surgical videos. Object-level detection tasks can be generally tackled by a classifier using features from still images. Thermal injury, however, has a variety of appearances regarding its size, shape and color. Meanwhile, thermal injury due to the unintended operation of a surgeon only appears for less than one second in a video and most of them are unrecognizable in a still image without video context. Therefore the project uses temporal features to detect thermal injury in the video.

The two-stream architecture uses an Inception V3 and an LSTM with linear interpolation to extract features from still images and learn temporal correlations among features in the clip. Another fold of the network obtains optical flow from FlowNet 2.0, classifying thermal injury regarding the motion of surgical tools in view. The result demonstrates a good performance of the proposed network on thermal injury detection tasks, even for cases where thermal injury is subtle and short in view and difficult to capture by a frame-based classifier.

The network is trained and evaluated on clinically validated video data provided by Surgical Safety Technologies and St. Michael’s Hospital.

Yichen Zhang
IS: Teodor Grantcharov
AS: Sanja Fidler

FRAUD PREVENTION IN REAL-TIME B2B PAYMENTS USING STREAMING ALGORITHMS

Pungle

Pungle, is a Toronto based fintech company providing real-time payment solutions that deliver enhanced cash flow for businesses and increase operational efficiencies. Pungle provides a simple solution to a complex web of financial technologies to deliver business payments in as little as three seconds. Pungle is the fintech for fintechs. The product-Pungle Payment Platform-delivers low cost, real-time, friction-free business disbursements, B2B supplier and B2C payments.

The problem that arises with digitization of business payments is a higher risk for fraud due to its electronic nature. Therefore, there is a need to be absolutely certain that both the sender and recipient of payments are the intended parties and that there are no anomalies in payment volume and frequency. This project is to build a streaming data pipeline, including data warehousing, that allowed us to log and persist transaction data for both audit trails and as a data set with which Pungle developed the real-time fraud prevention algorithms. This includes determining details about the origination of the transaction; for example, who made the business disbursement, where in the world the transaction originated, whether the device used to make the payment has been used before or flagged for fraud in the past, etc. The solution allows us to recognize patterns in real-time, therefore reducing account takeover fraud, friendly fraud and chargebacks associated with online payment systems.

Xu Sun
IS: Braulio Lam
AS: Richard Zemel

SOUND CLASSIFICATION

ecobee

Ecobee is a thermostat company, which intends to bring energy-saving, comfortability and safety features to internal spaces by integrating artificial intelligence. The company is a part of the smart-home industry and currently in second place in the smart-thermostat field. Its products currently include smart thermostats, light switches, and room sensors. According to internal analysis, our products have already helped ecobee customers in the U.S. save ~23% of their heating and cooling costs. Ecobee is making huge efforts to build the next generation of smart homes by providing more valuable services, including: schedule prediction for energy saving, occupancy detection for comfortability and safety, basement leaking prevention, and whole-house invisible guard.

The research project is about using sound to help occupancy and safety detection. Sound can provide valuable information about the inhabitants in an internal space. Deep learning model designs and implementations were involved. Specifically, the sound classification model contained two parts. The first part was preprocessing the sound data to get its transformed format. Then, it is sent to the second part, a deep learning model on cloud side, to classify what kind of sound it is. After that, the classification result is sent to a higher-level artificial agent with other features (e.g. PIR, Wi-Fi, Bluetooth, geofence) to let our devices understand the current occupancy and safety state.

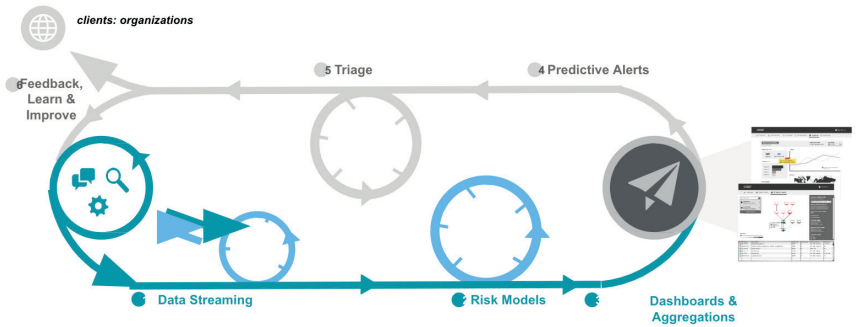
Han Meng
IS: Sina Shahandeh
AS: Roger Grosse

RELEVANCE RANKING OF LEGAL PASSAGES

ROSS Intelligence

A key factor within legal research is the matching of passages to a lawyer's legal issue. With recent advancements of deep learning in information retrieval, this project aims to explore a state-of-the-art model for application within the legal domain. By using a deep learning approach to determine the relevancy of a passage, this can be used to improve a user's ranked search results. This state-of-the-art architecture implements an approach based on the human judgement process to determine the relevance ranking. The local relevancies are detected within the document, and then are measured utilizing a convolutional neural network. The tensors are then aggregated with a recurrent neural network using a gating function, based on query term importance to output a relevancy label.

Nicole Langballe
IS: Jimoh Ovbiagele
AS: Frank Rudzicz



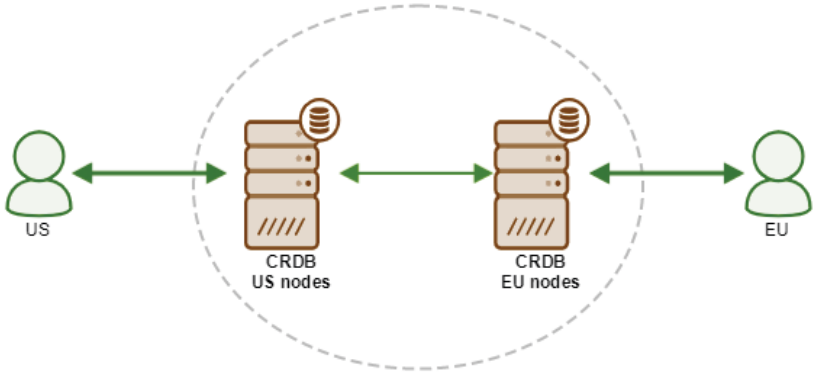
THE COGNITIVE RISK INTELLIGENCE AND SENSING PLATFORM

Deloitte

The Cognitive Risk Intelligence and Sensing Platform (CRiSP) project is Deloitte's social media risk sensing platform that uses deep learning to identify risks in clients' products and services. The CRiSP platform analyzes social media sentiments towards a target product or service and pinpoints risks that may endanger the client's business. With this platform, Deloitte clients can accurately detect problems within their product that are causing customer dissatisfaction, before they materialize into formal complaints or reputation-damaging events. The speed at which CRiSP can yield trustworthy results gives clients a significant advantage compared to traditional methods of analyzing customer feedback.

In the last seven months, our research and development efforts have guided the CRiSP project from early development to the production stage. Many deep learning models were developed and tested in an attempt to find a model that best accomplishes the project's risk classification problem. The major challenges include selecting an appropriate set of evaluation metrics by which candidate models are measured, and designing deep learning models that optimize these metrics. Models that were tested include variants of Convolutional Neural Networks and LSTMs. In addition, a significant effort was made to improve the existing software framework to support transfer learning and online machine learning. So far, the CRiSP platform has attracted the attention of one major player in the retail sector, who is interested in leveraging CRiSP for proactive risk mitigation.

Scarlett Guo
IS: Alik Sokolov
AS: Frank Rudzicz



EVOLVING DATA STORAGE DESIGN FOR A HIGHLY REGULATED FINANCIAL SECTOR

Ethoca

Evolving a standard 3-tier architecture to support external customer requirements, to restrict the movement of their confidential data (financial transaction information, possibly including PCI-DSS-regulated data and/or personally identifying information (PII)) into and out of regions (selective geo-replication) is hindered by the complexity of synchronization problems, especially relational integrity collisions arising from omni-directional writes. While a distributed log-based streaming architecture supports those requirements more directly, adopting one may not meet time constraints, in terms of performance, imposed by the business (data needs to be live for 12-18 months). Given the current database layer does not support selective geo-replication, the objective is to determine if there exists a suitable replacement for (or augmentation to) existing database technology that does support selective geo-replication and which can be deployed within reasonable time limits (6-12 months): retaining the relational model is desirable, but not a strict requirement (there is a mix of highly structured and semi-structured data). Specific targets for consistency latency (upper bound of 120s) and parity performance for read and write rates must be met (existing processing rates of 25-100 msg/s). Parity performance is measured via msg/s and is achieved if the modified/replacement system has at least the same performance profile (average, median, and 99% percentile) as the existing system.

Dana Alpysbayeva
IS: Ronak Patel
AS: Matt Medland

USER RECOGNITION USING GAIT RECOGNITION

Nymi

Authentication and identification systems aim to protect and secure an individual's data on websites, networks and various other platforms. Such systems only allow personnel with the necessary access rights to access specific and associated data. However, systems that are based on knowledge factors (something the user knows) have been susceptible to various failures and attacks over the years. It is also conceivable to see the vulnerability associated with systems using ownership factors (something the user has). For example, RFID fobs, ID/access cards, and NFC wristbands (e.g., Nymi band) provide a convenient way to authenticate an individual. However, they can still provide unauthorized individuals access to restricted resources since these devices are easily transferable through collusion or by an impostor stealing them.

The main problem with the current state of authentication technology is that it only identifies an individual when they begin to use the technology (T0 authentication). Examples include unlocking a phone with fingerprint or face scan, and entering a username and password to log into a computer. The aim of this research project is to explore the possibility of making this technology more robust and secure by ensuring the band continuously authenticates the user, TN authentication, using gait recognition.

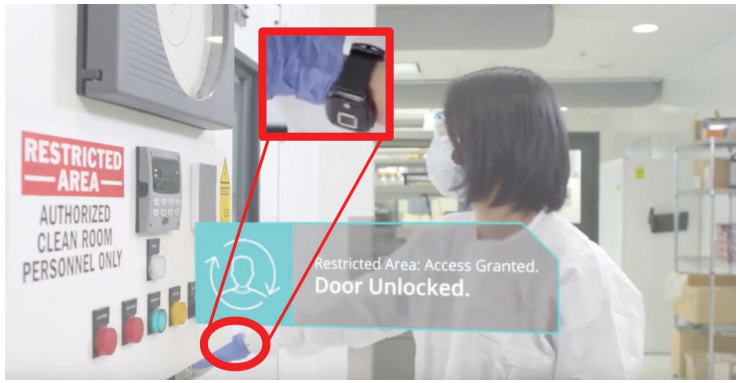
Rohit Rathi
IS: Abhishek Ranjan
AS: Khai Truong

LEGAL DECISION ENTAILMENT USING STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDINGS

ROSS Intelligence

Using existing information retrieval systems for a legal research tool has its limitations. These systems don't consider the meaning of query sentences, and do not retrieve documents that entail the sentences exactly. Using tasks like COLIEE and MultiNLI can train a sentence embedding model to pay attention to, and match, specific parts of the query text and candidate answers such that their meanings are matched. This project is an effort to solve COLIEE's Task 2 and Task 4 using a modification of structured self-attention sentence embeddings built for SNLI and MultiNLI. As a part of this, we are building a LSTM-attention model that first trains for MultiNLI. Then, we are attempting to learn to find legal passages that entail a given decision passage.

Siddhartha Thota
IS: Jimoh Ovbiagele
AS: Frank Rudzicz



UV MAPPING ASSISTANCE THROUGH DEEP LEARNING

Autodesk Research

UV mapping involves the projection of 3D surfaces to 2D representations. This involves unwrapping the 3D mesh from the surface for texture and color assignment. Unfortunately, this process of UV unwrapping can be tedious for complicated 3D meshes. The goal of the research project is to explore the unsolved UV mapping/unwrapping problem for 3D geometries with Machine Learning (ML). This involved investigating several domains of deep learning, involving point-clouds, meshes, and image based data. Of these approaches, the team have chosen to take multiple 2D camera projections from over 184 positions around a unit sphere. The views are then processed independently using state-of-the-art segmentation algorithms to predict where to cut and unfold the mesh. Attempts to merge different views to maintain information correlating different views of the model have been used using state of the art techniques, such as view pooling and LSTMs. In addition to solving the UV problem, the project's broader objective is to create a dedicated learning algorithm that can be used in the product directly, in order to adapt and learn from the user's workflows in a guided way. This will provide a more generic solution that can be extended to other application in 3D design via active learning.

Salvatore Vivona
IS: Herve Lange
AS: David Duvenaud

WHY I LIKE IT: MULTI-TASK LEARNING FOR RECOMMENDATION AND EXPLANATION

Layer 6 AI

Recommender systems have become an ever-present part of online user experiences. By analyzing our consumption habits, recommender systems can learn about our preferences and predict our needs with a high degree of accuracy. They routinely help apps, sites and services to present the right information (items, products, etc.) to the right users at the right time and in the right way. Modern collaborative filtering algorithms attempt to exploit latent features to represent users and items, which can lead to a lack of transparency. Such transparency issues can become severe when it comes to the case of e-commerce websites since latent features cannot be easily labelled. To build trust between a recommender system and its users, it has become important to complement recommendations with explanations so that users can understand why a particular item has been suggested.

This project attempts to extend recent progress in collaborative filtering and natural language processing, for review mining in a recommendation setting, by combining them into a multi-task learning framework. Briefly speaking, we employ a matrix factorization model for rating prediction, and a sequence-to-sequence learning model for explanation generation by generating personalized reviews for a given recommendation, user pair. We exploit the natural overlap between the latent factors learned by matrix factorization and the textual features learned by sequence autoencoders, allowing individual models to regularize each other. The jointly trained model achieves strong results on both the task of rating prediction, beating the state-of-the-art, and on the task of recommendation explanation.

Yichao Lu
IS: Tomi Poutanen
AS: Richard Zemel

WEAKLY SUPERVISED EMBEDDINGS OF FIBROBLAST CELLS

Phenomic AI

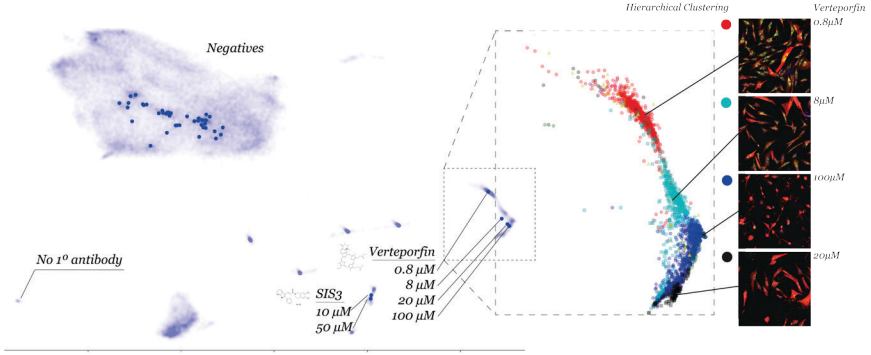
Phenomic AI seeks to discover new treatments for idiopathic pulmonary fibrosis. Using microscopy image screens, neural network models can be trained upon healthy/diseased control images of cells, and then used to predict drug "hits" based upon its capacity to recognize the presence of disease.

In the competitive pharmaceutical industry, merely identifying drug candidates in this manner is not enough. Considering that a newly-discovered drug candidate's underlying mechanism-of-action (MOA) might be very similar to another patented drug's MOA, it is also important to build a profile for drug candidates to compare and contrast them. In this way, we can identify first-in-class drugs by seeing if they cluster separately from known drugs/MOAs.

Weakly supervised single-cell image embeddings are one such profiling strategy invented by researchers at the Broad Institute. By defining an auxiliary task where the model recognizes a cell's experimentally-defined condition (expanding our dataset beyond just control images), a neural network's trained layers can then act as an embedding space for our dataset in which MOA clusters can emerge.

We have applied this methodology to our own internal data, and are also exploring our own variations of this method, such as the extension to full images from single-cell crops, or using more structured prediction in the auxiliary task (multi-headed neural networks). Once well validated, we eventually plan to use this methodology to discover drugs.

Grant Watson
IS: Oren Kraus
AS: Jimmy Ba



WEB-BASED SLAM FOR AUGMENTED REALITY

Rakuten Japan

SLAM (Simultaneous Localization and Mapping) is one effective approach that simultaneously localizes sensors with respect to their surroundings, while at the same time mapping the structure of the environment. As accurately positioning, orienting, and tracking a user's device are always desired in the Augmented Reality (AR) world, SLAM is becoming an increasingly important topic for AR applications. To help bring AR to as many users as possible, we plan to build a web-based AR framework which can enable state-of-the-art AR features on the web browsers. The objective of this project is to achieve real-time tracking of the user device's location and position through web-browser, webcam and the inertial measurement unit (IMU) by making use of SLAM technology. Subsequently, visual components such as 3D object and animation can also be added to the canvas based on the location information. We hope that this project can build a foundation for future web-based AR applications in Rakuten.

Chenfei Wang
IS: Tomoyuki Mukasa
AS: Karan Singh

Join the conversation online

#DCSARIA2018

 @UofTCompSci

 UofTCompSci

 U of T – Department of Computer Science

Master of Science in Applied Computing (MScAC)
Department of Computer Science
University of Toronto
40 St. George Street
Toronto ON M5S 2E4

mscac@cs.toronto.edu
www.cs.toronto.edu/mscac
+1-416-978-5180