✉ *johnn.chen@mail.utoronto.ca*
johntiger1.github.io
johnn-chen
uoftml
johntiger1

John Chen

Education

2019–2021 **Master of Science in Computer Science**, *University of Toronto*, Toronto, Canada, 4.00/4.00 cumulative GPA.
Research area in Deep Learning and Natural Language Processing. Advised by Professor Frank Rudzicz. Thesis-based research degree, eligible for direct entry to PhD.

2014-2019 **Bachelor of Science in Computer Science**, *University of Toronto*, Toronto, Canada, 3.98/4.00 cumulative GPA.
Graduated with high distinction. Focus in Artificial Intelligence. Completed one year of co-op work experience.

Research Grants

2020 **Mitacs Accelerate Grant IT17182: Machine learning in the operating room: focus, performance, and the medical record** *supervised by Frank Rudzicz et. al.*

Awards and Scholarships

2019 **Dorothy Helen McRobb Scholarship**
University of Toronto

2019 **Undergraduate Student Research Award**
Natural Science and Engineering Research Council of Canada (Declined)

2018 **Pierre Karch and Mariel O'Neill-Karch Scholarship**
University of Toronto

2018 **Vector Institute Research Grant / University of Toronto Computer Science Award**
Vector Institute and University of Toronto

2018 **Galleria Supermarket Scholarship**
Galleria Supermarket Charitable Foundation

2018 **Woodsworth College Student's Association Scholarship**
University of Toronto

2018 **Overall Winner**
ETHUofT Blockchain Hackathon
Created DataX, a data exchange marketplace on Ethereum. Responsible for entire web stack (Node, Express, Bootstrap and Solidity)

2017 **Overseas Outstanding Youth Award**
Federation of Overseas Chinese Association Taiwan
Awarded to one of two youth worldwide.

2017 **Samuel Beatty In-Course Scholarship**
University of Toronto

2017 **Canada Chinese Computer Association Scholarship In Computer Science**
Canada Chinese Computer Association

2017 **Undergraduate Student Research Award**
SickKids Hospital and Natural Science and Engineering Research Council of Canada
Award held at SickKids Hospital in the lab of Dr. Michael Brudno, the Centre for Computational Medicine

2017 **Goodwin O'Grady Academic Achievement Award**
University of Toronto
Awarded for most improved academic performance

2017 **Most Useful Tool for Students Prize**
Computer Science Student Union and Connected Lab
Built StudyTrackr, a productivity app for logging study time.

2016 **1st Place**
Cisco DevNet Toronto Hackathon
Came 1st place in a hackathon with over 150 participants.
Created an IoT app for real-time monitoring of plants, using Cisco's Tropo VoIP service and Relayr API

2015 **Junior Coach of the Year Award**
Toronto Blue Jays Rookie League
1 award is given per geographical region of camp locations. Award conferred for downtown-central region, which consisted of 8 campsites and over 30 staff.

2014 **Community Champion Scholarship**
Rotary Club of Toronto and Toronto Argonauts
Full-tuition scholarship to University of Toronto. Selected as sole recipient from open compeititon across both Toronto area public school boards

2014 **Randy Padmore Scholarship / University of Toronto Scholarship**
CHFT and University of Toronto
Renewable four-year scholarship

2014 **President's Entrance Scholarship**
University of Toronto

2014 **TD Bank Higher Education Award**
University of Toronto

2014 **Burger King Scholarship**
Burger King Scholars

2014 **Lu Zhen Chan Education Scholarship**
Chinese Canadian Intercultural Association

2014 **Platform 9 $\frac{3}{4}$ Award**
Northern Secondary School
Selected as sole recipient from the graduating class of over 300 students

Research Experience

June 2018–Present **Deep Learning Research Intern**, *SPOCLab*, Vector Institute and University of Toronto.

Basic research and implementation of several end-to-end deep learning solutions as part of coursework and Master's thesis. Projects include:

- framework for batched sequence-to-sequence modelling. Implemented features and experiments include word/character level analysis, teacher forcing, and network sizing.
- Active learning via reinforcement learning (policy gradient and deep Q-learning). **Technical report in preparation for submission to PKDD 2020, Interactive Adaptive Learning Workshop**
- medical image captioning using convolutional neural network encoder and recurrent neural network decoder. **Doubled F1 performance** by reformulating problem as multilabel classification problem.
- paragraph-level style transfer in text between doctor and patient corpora. **Extended abstract submitted to ACL 2020 Workshop on Neural Generation and Translation**.
- Massively multilabel, highly imbalanced cell image classification. **Improved F1 performance by 156**% over baseline pretrained network via **image augmentation and finetuning**.

May 2017–August 2017 **NSERC USRA Summer Research Student**, *Centre for Computational Medicine*, Peter Gilgan Centre for Research and Learning, SickKids Hospital.

My summer research project focused on adapting Bayesian Ontology Query Algorithm (BOQA) into an online setting. BOQA is a Bayesian network-based algorithm for automatic disease diagnosis of rare diseases. A Bayesian network is constructed from a set of phenotypes and a medical ontology linking phenotypes to genotypic disease markers. Given a set of observed phenotypes, simplified and approximate Bayesian inference is run to predict the rare disease that best explains the phenotypes. The online version incrementally provides partial candidate diagnoses with each additional phenotype input. Sped up naive algorithm by 500x via dynamic programming and multithreading. Work was accepted as an **oral presentation** at the 2017 Canadian Undergraduate Computer Science Conference.

Publications

2020 **"Standardizing the Style Transfer in Text Task: an Opinionated Taxonomy"**, *John Chen*, in submission.

2020 **"Active Learning via Reinforcement Learning"**, *John Chen and Eric Lin*, in preparation.

Research Projects

Winter 2020 **"Fairness in multimodal clinical machine learning"**

Ongoing master's thesis work

My master's thesis centers around multimodal machine learning. As data capture processes and techniques improve, our machine learning models will increasingly receive access to more and different types of data, like doctor-patient conversation recordings, video, knowledge bases, echocardiograms and high-resolution radiographs. Data fusion is an exciting and promising line or research, as demonstrated by the incredible recent empirical performance of visual-linguistic models like VisualBERT, LXMERT, ViLBERT, and VL-BERT. Additionally, it provides a potential avenue to unite traditional symbolic AI with recently popular deep learning methods.

In the clinical setting, fairness and electronic medical record modalities of text and tabular medical data are special concerns. Therefore, I am investigating methods to improve fairness on medical datasets that include both patient narratives and vital sign lab measurement data, such as in MIMIC-III. Current preliminary results show that it may be possible to achieve both greater fairness *and* performance when combining modalities.

Winter 2020 **"Extending local superlinear convergence results of some Quasi Newton methods under noise and random switching"**

Course project undertaken for CSC2532: Statistical Learning Theory

In this course project, we attempted to prove convergence results for Quasi Newton optimization methods under some stochastic conditions. We were able to prove that the addition of bounded random noise does not affect convergence results, and that randomly switching between Quasi-Newton update rules also maintains the overall convergence properties. We also provide experimental results that support the validity of our proof. Experimentally, we also discovered that introducing random switching of the QN-update rule (between Greenstadt update and BFGS update) can improve convergence and stability. We propose to study these empirical findings in a future work.

Winter 2020 **"Introducing Free-form Visual Question Answering"**

Course project undertaken for CSC2547: Inverse Graphics

In this course project, we investigate applying text generation methods towards Visual Question Answering. The VQA task as introduced by Agarwal's 2015 paper conventionally leverages a fixed vocabulary answer set. That is, the goal of the VQA model is to simply select the correct answer from the given possible answers. There are several large assumptions in this formulation, which may or may not be appropriate. Namely, we assume the correct answer exists, and it must be given as a possible candidate. This has implications on creativity, and different but correct answers. We therefore propose a different task formulation: generate the text answer instead of simply selecting it. Preliminary results show that this is a much more challenging task; when we apply state of the art VQA models like LXMERT, we find the performance drops considerably. At the same time, we find that LXMERT embeddings are still more performant than simply a seq2seq model trained on the text only. This is ongoing work, which we hope to extend to a conference paper; namely, we wish to argue the importance of answer generation and argue for a new type of free-form text generation VQA task.

Winter 2020 **"Exploring novel nutritional predictors of diabetes across geoeconomic status : a fairness perspective"**

In collaboration with Population Health Research Institute and Hamilton Health Sciences

In this project, we explore new techniques for discovering novel nutritional predictors of diabetes, using numeric data from the Prospective Urban Rural Epidemiology study, which encompasses data from over 225 000 patients across the geoeconomic spectrum. Current results indicate that there is a fairness gap on poor and rural patients when performing binary diabetes prediction; we recreate baseline results that show that classifiers trained on class-conditional data can provide better performance for some groups. Ongoing work seeks to validate envy-freeness between groups and develop an ensemble model that improves performance overall.

Fall 2019 **"Active Learning via Reinforcement Learning"**

Course project undertaken for CSC2547: Learning to Search

In this course project, we explored the application of reinforcement learning to learn the querying heuristic employed by active learning. In pool-based active learning, the goal is to select a minimal number of examples from a pool of data to label by an oracle in order to achieve a given performance on a given task. Normally, an active learning algorithm will have a querying algorithm that determines which data points from the point are most impactful and useful to label. This querying algorithm is often heuristic, such as random, uncertainty sampling or diversity sampling. In our course, we investigated the use of reinforcement learning to learn the query algorithm. We experimented with two main RL algorithms, namely policy gradient and Deep Q-Networks. For the policy gradient approach, we define a state and action space using clustering, and feed in signals related to the current performance of a model trained on the current labelled data. For the deep Q network, we directly estimate the Q-value of a datapoint from its features based on prior samples. We experimented with a custom toy dataset, and MNIST, and found small performance improvements. We are currently preparing a paper for conference review.

Fall 2019 **"Exploring Proxy Fairness in Text"**

Course project undertaken for CSC2541: AI and Ethics

In this project, we reproduce and then critically analyze results from the 2019 NAACL Best Thematic Paper Award winner, "What's in a Name? Reducing Bias in Bios Without Access to Protected Attributes". In this paper, they propose novel methods for improving fairness in situations where access to a protected attribute (such as race or gender) is not available, but access to a proxy attribute (such as a person's first name and last name) is available. The crux of their method involves adding a fairness regularization term to the loss that is computed in the word-embedding space of a person's name, in order to reduce downstream fairness gaps between groups defined by protected attributes. We investigate how correlated these proxy attributes are to the real protected attributes, finding that there is only weak correlation between the names and some protected attributes such as race. Furthermore, we identify a mismatch between the training objective and the evaluation metric which can explain their somewhat non-decisive results. Finally, as a novel experiment we investigate the recently popular phenomenon of overtraining, or training even after zero training loss has been achieved, on the fairness of the resulting model. As part of this work, we created and presented a poster for the final class presentation day held at Vector Institute.

Work Experience

Summer 2019 — **Data Scientist Intern**, *Air Miles (a LoyaltyOne company)*, Toronto.

Data scientist on CRM team. Provide program-wide analysis and modelling on 10 million+ loyalty members. Mentor: Somayyeh Aghababei, PhD.

- Developed flight redemption propensity model that was used in a marketing campaign targeting **500 000 collectors** for a major Canadian airline
  - **xgboost** model achieved **0.97 AUROC** and **9x lift**
  - Improved model via **seasonality adjustments**, **model-free feature selection algorithm** and **model-based genetic feature selection algorithm**, as well as hyperparameter tuning via **Bayesian Optimization** using scikit-optimize.
- Performed research and development in recommendation engine prototype for leading Canadian grocery chain. Achieved **111% basket coverage improvement** over baseline by using market basket analysis and frequency-based model.

March 2018–December 2018 (part-time) — **Machine Learning Consultant**, *DataX*, Toronto.

Designed, architected and implemented machine learning models end-to-end for DataX data marketplace platform.

- Built full machine learning pipeline for unsupervised equation parser of PubMed abstracts
- Developed novel uncertainty-based **unsupervised character-level recurrent neural network language model** for noisy detection of equations. Achieved good qualitative performance.
- Developed BootStrap website for reporting results

September 2017–May 2018 — **Software Developer Intern**, *Ontario Teachers' Pension Plan*, Toronto.

Worked two co-op terms at OTPP within the Investment IT divison, first on the DevOps team and then Risk Analytics product team.

- Managed the entire deployment lifecycle within Solutions Engineering including JIRA, Github, Jenkins, Nexus, SonarQube, UrbanCode Deploy and UrbanCode Release.
- Improved deployment set-up time by 30% by developing full-stack RESTful C# Web API + AngularJS + Oracle database web app for managing UrbanCode agent and server registry information.
- Contributed to performance gains of 50x in financial software through developing hybrid Powershell/C# solution to automatically generate, format and send e-mail reports to senior management daily.

Summer 2014, Summer 2015 — **Junior Program Leader**, *Toronto Blue Jays Rookie League*, Toronto.

Camp counsellor for baseball day camp organized by Toronto Blue Jays and Toronto Community Housing.

- Developed, organized and implemented various community-specific activities, such as inter-camp barbeques and trips to the 2015 Pan Am Games events to enrich the summer experience of kids aged 5-12
- Founded first-ever summer tutoring program under own initiative, providing valuable summer review for those struggling in school
- Recognized in 2015 with **Junior Coach of the Year Award**

Teaching Experience

I have had the pleasure to serve as a teaching assistant in many courses at both University of Toronto and Vector Institute with many great professors. Teaching and nurturing the next generation of academics and industry professionals is one of my core values.

Winter 2020 Mathematical Foundations for AI with Professor Graeme Taylor, at Vector Institute (Professional Development Course)

Winter 2020 JSC370H1S (Data Science II) with Professor Nathan Taback, at University of Toronto, St. George Campus

Winter 2020 CSC413H1S (Neural Networks and Deep Learning) with Professor Jimmy Ba, at University of Toronto, St. George Campus

Fall 2019 CSC420H1F (Introduction to Image Understanding) with Professor Babak Taati, at University of Toronto, St. George Campus

Winter 2019 CSC324H1S (Principles of Programming Languages), with Professor Albert Lai, at University of Toronto, St. George Campus

Winter 2018 CSC165H1S (Mathematical Expression and Reasoning for Computer Science), with Professors Tom Fairgrieve, David Liu and Toniann Pitassi at University of Toronto, St. George Campus

Winter 2017 CSC263H1S (Data Structures and Analysis), with Professors Sam Toueg, Francois Pitt and Sasa Milic, at University of Toronto, St. George Campus

Fall 2016 CSC209H1F (Software Tools and Systems Programming), with Professor Marina Barsky, at University of Toronto, St. George Campus

Invited Talks and Presentations

2020 **Fairness in multimodal clinical machine learning**
University of Toronto Graduate Visit Day
University of Toronto, Toronto, Canada
Presented my ongoing masters' work to prospective new matriculants to University of Toronto Graduet School of Computer Science

2018 **CongoToken: blockchain based enterprise transaction verification**
19th Annual Canadian Undergraduate Technology Conference
University of Waterloo, Waterloo, Canada
One of 8 teams invited to pitch a blockchain based startup research idea.

2017 **Refined Online Bayesian Ontology Query Algorithm**
Canadian Undergraduate Computer Science Conference
University of Toronto, Toronto, Canada
Presented my summer research project with Professor Michael Brudno at SickKids Hospital on Refined Bayesian Ontology Query Algorithm

Service

2020 **Research mentor**
Review of Undergraduate Computer Science (RUCS) Research-a-thon, Toronto, Canada
Presented my work on fairness in multimodal clinical machine learning and mentored research projects during the RUCS research-a-thon

2019 **Student Panelist**
2nd AI-Squared Conference, Toronto, Canada
Invited as a student researcher and industry panelist

2018– Present **Director of Communications**
Review of Undergraduate Computer Science (RUCS), University of Toronto
RUCS is a peer-reviewed journal and club for undergraduate computer science research.

Responsible for maintaining and growing RUCS social media presence. Coordinated with RUCS leadership to organize campaigns to advertise events and special dates.

- Created social media campaigns and hosted several events, which each attracted 30+ students on average:
  - March 2020: Research-a-thon
  - Jan 2020: Undergrad CS Research Social
  - Jan. 2019: Undergrad CS Research Social
  - Nov. 2018: Undergrad research info session, attended by 50+ first and second year students
  - Sept. 2018: **First ever Undergraduate Computer Science Research Conference at U of T** (co-hosted with TURCS).
- Conducted interviews for undergraduate research guide and contributed three featured interviews to the final publication.

2017-2018 **Editor-in-Chief**
3rd Annual Review of Undergraduate Computer Science (RUCS), University of Toronto

As editor-in-chief of RUCS, I was responsible for the strategic planning that transformed RUCS from primarily a peer journal to a more inclusive community-oriented club. On top of regular editing duties, I was also responsible for all daily logistics involved in running the publication and releasing the 3rd annual edition. These included hiring for new positions, sourcing funding conducting social media outreach, and later, event planning.

- Oversaw publication including hiring for new positions, sourcing funding and conducting social media outreach
- Expanded publication abstract sourcing to UBC and MIT
- Hosted successful full-day undergraduate research conference jointly with Toronto Undergraduate Research in Computer Science (TURCS) group. **Attracted 100+ participants**, with keynote lectures by Roger Grosse and Ishtiaque Ahmed.
- Oversaw website redesign and overhaul

Conferences

2019 **Pan-Canadian Self-Organizing Conference on Machine Learning**
MILA, Montreal, Canada. Attended with funding from Vector Travel Grant.

2019 **Analytics By Design**
TIFF Bell Lightbox, Toronto, Canada Attended with funding from LoyaltyOne.

Volunteer and Community Involvement

Feb. 2020 – Present **Hospice Volunteer**, *Kensington Hospice*, Toronto, Canada.
Completed 30 hours of hospice palliative care training. Provided direct patient care, including bed transfer and feeding.

Sep 2019 – Present **Committee Member**, *Alexandra Park Revitalization Working Group*, Toronto, Canada.

The working group meets monthly to discuss and shape the future of Alexandra Park's revitalization.

- Judged solicited RFP bid for passive housing architect. Voted on winning RFP design.
- Member of Alexandra Park Delegation to 2020 Economic Club of Canada's Future of Work & New Economy Series (A Conversation with Barack Obama)
- Advocated for increased youth involvement and participation in the community

Sep 2014 – Present **Scholarship Manager**, *Chinese Canadian Intercultural Association (CCIA)*, Toronto, Canada.

I have served in various volunteer roles with the CCIA, including as a web developer, tutor, social media manager, and scholarship manager.

- Created and maintained website and social media accounts for CCIA
- Tutored students from Grade 9 to Grade 12 in Mathematics and Biology, weekly at Saturday School
- Ran in-house advertising and digital media outreach for tutoring and scholarship programs
  - In 2017, reached 10,000 first-year university students organically on Facebook regarding the Lu Zhen Chan scholarship resulting in year-over-year increase in applications by 460%
  - In 2018, refined and optimized digital strategy, resulting in 40% year-over-year increase in Facebook post reach and a 45% increase in applications. Redesigned and re-implemented website from scratch. Coded up new website in Gatsby frontend static site generator framework, moving to Netlify hosting platform and adding Google Analytics to website. Tracked 450 monthly users during peak time.