YEMING WEN

Email: ywen@utexas.edu

RESEARCH INTEREST

My research focuses on building a machine learning framework to generate code with human-like efficiency. In the meantime, I'm also interested in enhancing the efficient adaptation framework for Large Language Models (LLMs), specifically for code generation tasks. Before my PhD study, I worked on the development of efficient learning algorithms for deep neural networks, with a focus on large batch training, ensemble methods and uncertainty modelling.

EDUCATION

University of Texas, Ausin, Ph.D in Computer Science — Jan. 2021 - Jan. 2025
Advisor: Swarat Chaudhuri

University of Toronto, M.Sc. in Computer Science — Sept. 2018 - Jan. 2020
Machine Learning Group and Vector Institute
Advisor: Jimmy Ba and Roger Grosse

University of Toronto, B.Sc. in Mathematics and Computer Science — Sept. 2013 - June 2017
Cumulative GPA: 3.98/4.0, Average: 92.7/100

PREPRINTS

Yeming Wen, Pengcheng Yin, Kensen Shi, Henryk Michalewski, Swarat Chaudhuri, Alex Polozov. *Grounding Data Science Code Generation with Input-Output Specifications.* Instruction Tuning and Instruction Following Workshop at NeurIPS, 2023.

Amitayush Thakur, George Tsoukalas, **Yeming Wen**, Jimmy Xin, Swarat Chaudhuri. *COPRA: An In-Context Learning Agent for Formal Theorem-Proving.* Math-AI Workshop at NeurIPS, 2023

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, **Yeming Wen**, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, Jimmy Ba. *Benchmarking Model-Based Reinforcement Learning.* arXiv preprint arXiv:1907.02057, 2019.

PUBLICATIONS

Yeming Wen, Swarat Chaudhuri. *Batched Low-Rank Adaptation of Foundation Models.* International Conference on Learning Representations (**ICLR**), 2024 (Oral, 1.2%).

Pengcheng Yin, Wen-Ding Li, Kefan Xiao, Abhishek Rao, **Yeming Wen**, Kensen Shi, Joshua Howland, Paige Bailey, Michele Catasta, Henryk Michalewski, Alex Polozov, Charles Sutton. *Natural language to code generation in interactive data science notebooks.* Association for Computational Linguistics (**ACL**), 2023.

Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, **Yeming Wen**, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, Balaji Lakshminarayanan. *A simple approach to improve single-model deep uncertainty via distance-awareness.* Journal of Machine Learning Research (**JMLR**), 2023.

Rohan Mukherjee, **Yeming Wen**, Dipak Chaudhari, Thomas Reps, Swarat Chaudhuri, Chris Jermaine. *Neural Program Generation Modulo Static Analysis.* Advances in Neural Information Processing Systems (**NeurIPS**) , 2021 (Spotlight).

Yeming Wen*, Ghassen Jerfel*, Rafael Muller, Mike Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, Dustin Tran. *Combining Ensembles and Data Augmentation Can Harm Your Calibration.* International Conference on Learning Representations (**ICLR**), 2021.

Mike Dusenberry, Ghassen Jerfel, **Yeming Wen**, Yi-an Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, Dustin Tran. *Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors.* International Conference on Machine Learning (**ICML**), 2020.

Yeming Wen, Dustin Tran, Jimmy Ba. *BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning.* International Conference on Learning Representations (**ICLR**), 2020.

Yeming Wen*, Kevin Luk*, Maxime Gazeau*, Guodong Zhang, Harris Chan, Jimmy Ba. *Interplay Between Optimization and Generalization of Stochastic Gradient Descent with Covariance Noise.* International Conference on Artificial Intelligence and Statistics (**AISTATS**), 2020.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, Roger Grosse. *Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches.* International Conference on Learning Representations (**ICLR**), 2018.

RESEARCH EXPERIENCE

Research Internship at Google May 2022 - May 2023
Advisor: Alex Polozov *Mountain View, CA*

- Developing algorithms on language models to generate code with better fidelity
  - Build a static analyzer in Python to feed the semantic information in the code to language models.
  - Train a language model with the additional information from static analyzer along with the code.
  - Evaluate the model on the notebook dataset (ARCADE). Accepted to ACL 2023 (https://arxiv.org/pdf/2212.09248.pdf).

Graduate Research Assistant at UT Austin Jan 2021 - Now
Advisor: Swarat Chaudhuri *Austin, TX*

- Develop algorithms on automatic code generation with large scale language models
  - Use automata to generate equivalent programs to increase the size of training data, leading to improved performance of large language models on code generation.
  - Applied static analysis to generate JAVA method automatically, published at NeurIPS 2021 https://openreview.net/pdf?id=yaksQCYcRs.

Research Internship at Google Feb 2020 - Sept 2020
Advisor: Dustin Tran *Mountain View, CA*

- Combining Ensembles and Data Augmentation Can Harm Your Calibration
  - Made a large scale empirical study on the combination of BatchEnsemble/MC-dropout/Deep Ensembles and various data augmentation methods (including AugMix and Mixup).
  - Implemented BatchEnsembles related codebase in the open-source uncertainty baselines, https://github.com/google/uncertainty-baselines.
  - Built a data augmentation pipeline which is extensively reused in other research projects, https://github.com/google/edward2/tree/master/experimental/marginalization_mixup.

Research Internship at Google August 2019 - Dec 2019
Advisor: Dustin Tran *Toronto, Canada*

- Rank-1 Net: An Alternative Approach to Efficient Ensembles and Lifelong Learning

- Extended BatchEnsemble (Rank-1 net) to more complicated lifelong learning set-up, including a new benchmark dataset SPLIT-ImageNet.
- Demonstrated that Rank-1 Net is capable of learning a large number of lifelong learning tasks (up to 100) without forgetting, which no previous methods can achieve.
- Experiments in uncertainty modelling showed that Rank-1 Net is orthogonal to existing ensemble methods. Combining Rank-1 net with existing ensemble methods such as MC-dropout leads to better uncertainty predictions.

M.Sc. Research Project March 2019 - Dec 2019
Advisor: Prof. Jimmy Ba *Toronto, Canada*

· BatchEnsemble: Ensembles of Neural Networks in a Mini-Batch Friendly Way

- Proposed an efficient ensemble method which is mini-batch friendly. It incurs negligible computational and memory costs.
- Demonstrated its effectiveness in image classification and machine translation. BatchEnsemble also captures model uncertainty in contextual bandits task and achieves compelling calibrated predictions on CIFAR-10 corrupted dataset.
- Demonstrated BatchEnsemble can be used in large-batch training and continual learning.

Research Intern at Borealis AI Sept 2018 - Feb 2019
Advisor: Prof. Jimmy Ba *Toronto, Canada*

· Large-Batch Stochastic Optimization with Curvature Noise

- Explored different intrinsic noise structures in SGD optimization.
- Analytically showed that the convergence rate of noisy SGD optimization not only depends on the marginal variance of the noise but also the Frobenius norm of the noise matrix.
- Empirically verified the above conclusion and showed that adding diagonal Fisher noise to large batch gradient leads to better generalization without increasing the number of training iterations.

University of Toronto Excellence Awards May 2017 - Sept 2017
Research Assistant, Advisor: Prof. Roger Grosse *Toronto, Canada*

· Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches

- Analytically showed that Flipout is unbiased and gives lower gradient variance than naive stochastic neural networks.
- Implemented the Flipout upon multiplicative perturbation algorithm with various neural network architectures, such as MLP, LeNet, VGG. Empirically evaluated that Flipout achieves an ideal variance reduction effect.
- Extended the algorithm to Bayesian neural networks (trained with Bayes by Backprop) and evolution strategies in both supervised learning and reinforcement learning. Evaluated by MNIST data set and Mujoco environment.

OTHERS

Reviewer	ICML 2022, NeurIPS 2021, ICML 2021, ICLR 2021, NeurIPS 2020, ICML 2020, NeurIPS 2019
Programming Languages	Python, Matlab, R
Frameworks & Tools	Tensorflow, MXNet, PyTorch
Teaching	TAed Calculus, Theory of Computation, Probability and Statistics