Brendon Reperttang

(619) 895-0022
brendonrt@gmail.com
https://www.linkedin.com/in/brendon-rt

EDUCATION

Massachusetts Institute of Technology December 2023
Cambridge, MA
Bachelor of Science in Computer Science and Engineering

SKILLS

Computer Languages & Software: Python, Mojo, Type/JavaScript, Java, Lua, C++, C, C#, LaTeX, Julia, RISC-V Assembly, Jupyter, Gurobi, Linux/Unix/Windows CLI, GitHub, Docker, Kubernetes, CSS, HTML, React/Angular, PostgreSQL, Prisma, n8n, Windmill

Specializations: LLMs and Generative AI, Pipelining and Architectural Patterns, Category Theory, Supervised and Unsupervised Learning, Ensemble Learning, Recurrent Neural Networks and Anomaly Detection, Reinforcement and Multi-Agent Learning, Computer Vision, Data Synthesis and Feature Engineering, MLOps, Research Paper Implementation, Unit and Integration Testing, Ablation Studies

Relevant Coursework: Deep Learning, Computer Vision, Machine Learning, Multi-Agent Learning, Software Construction, Fundamentals of Programming, Linear Algebra, Multi-variable Calculus, Math for CS, Algorithm Design, Computational Structures, Embedded Systems, Computer Systems Engineering, Optimization, Modeling for ML, Software Studio

Languages: English (Fluent), Mandarin (Fluent), Cantonese (Semi-fluent)

WORK EXPERIENCE

AI Research Associate June 2024 - September 2024
Society for AGI, Pasadena, CA
- Discussed and proposed LLM research and evaluation methods from a Category and Control Theory perspective
- Experimented with compute efficiency focused research papers, resulting in iso-FLOP to accuracy improvements of over 40 percent for the best ablated component combination, while also holding parameter count constant
- Enhanced depth of ML architecture expertise using Flax

Zephyr AI Researcher August 2023 - December 2023
DAI Lab, Cambridge, MA
- Studied the effects of including static covariates on time series data forecasting, increasing accuracy by 16 percent over the baseline XGBoost without covariates
- Deployed Temporal Fusion Transformer models as a pipeline with MLBlocks
- Theoretical propositions and experiments for other potential predictive maintenance improvements on error code predictions

Machine Learning Engineer and Data Analyst May 2020 - August 2021
Bamboo Mortgage, Portland, OR
- Created a random forest estimator model to find high ROI housing investments
- Informed acquirements of 6 properties with a hybrid model/heuristic system
- Scraped data and compiled expenses to allow for investment analysis and cash-flow analysis

Web Development Intern and Financial Assistant June 2016 - July 2020
Bamboo Mortgage, Portland, OR
- Solely reformatted the company website using HTML and JavaScript
- Organized and kept track of hundreds of financial documents for many clients
- Assisted in organizing over a dozen open house events with Xiao Realty

OTHER EXPERIENCE

Efficient Transformer Project February 2024 - Present

Developing a SOTA language model in Flax (A flexible end-to-end ML framework using JAX), utilizing the most recent research and novel techniques to optimize performance and efficiency. With Transformer as a base, feature engineering, pre-training, and fine-tuning stages also were iterated on. For pre-training, a BERT-inspired approach with dynamically scheduled masked language modeling, and custom bilevel optimization algorithms for training. This approach, combined with various sampling techniques, significantly boosted training efficiency. Select model features include Grapheme Pair Encoding (GPE) for tokenization, hypersphere normalization (nGPT), multi-head attention mechanisms, and RELU^2 activations with dropout. The project will also feature mixed-precision training, novel fine-tuning processes, and advanced attention mechanisms such as SageAttention combined with Differential Transformer. Focusing on a modular approach facilitates experimentation with emerging techniques and enables comprehensive ablation studies. This work demonstrates expertise in advanced NLP, efficient model design, rigorous empirical evaluation, and best practices for large-scale DevOps.

Multi-Agent LLM Truth Elicitation September 2023 - December 2023

In 6.S890, a special subject in Multi-Agent Learning, collaborating with PhD students Charlotte Siegmann and Stewart Slocum, the aim of the project is to verify theoretical claims about applying a Bayesian Truth Serum (BTS) to a multi-agent LLM system. Summarizing the groundwork for this BTS setup, provided by Ray Weaver and Drazen Prelec, given questions $q \in Q$ with a finite number of unique answer choices, a query is made to each agent for their answer as well as guessed probability mass for every answer option, the predicted chance that others would output that. Leveraging this unique approach to truth elicitation (with theoretical guarantees of a Pareto-Dominant Nash Equilibrium), many setups contained statistically significant results, with p-score values at most $7 * 10^{-19}$ for outlier BTS setups compared to the control of querying the question by itself. For context, on the Massive Multitask Language Understanding dataset, an important benchmark for STEM knowledge, the Brier score went from .074 to .035 on GPT-4, a significant improvement on the proper scoring metric. This result is only from prompting a small number of agents, and there is feasibility of applying this setup to pre-training tasks such as next-word prediction, which provides larger potential for model performance. This experience provided knowledge about the nuances of modularity and automation, advanced game theory theorems for extensive form games, and statistical analysis techniques.

Recurrent Neural Network for Online Market Futures February 2023 - May 2023

In 6.S052, a special subject ML project class, a long short-term memory recurrent neural network was created to perform market analysis in multiple decentralized online markets. There exists a major opportunity for larger profit margins in these markets, due to the saturation of algorithmic trading on the stock market. A simulated a 60-day trading run with price, demand, and volume data from the Steam Community Market, provided an average profit of over 7,000 USD, starting with only 2,500 USD! To get these results, state-of-the-art RNN techniques were used, along with lesser-known techniques, and even some novel ones. An uncommon technique of encoding important conditional and static features directly into the time-series data is done by affinely transforming the input vector at $t = 0$. This, paired with a novel static feature vector formulation, which provides a unique encoding for the set of commodities focused on (Counter Strike skins), average profit improved by over 2,700 USD over the baseline RNN. For selection, ROI "slopes" are generated from the outputted price predictions, top-k ROI "peaks" are selected, and at times corresponding to peaks in this list, the points are regenerated. If any updated slopes from this current point are found to be higher than the prediciton, the asset is held, otherwise, it is listed. This project provided valuable experience regarding creating a custom pipeline and solution for nearly any ML architecture.