Yilin Wang

+1-4129967954 yilin_wang@fas.harvard.edu yilin-wang-838916310 TonyW42

EDUCATION

Harvard University, GPA: 3.94/4.00 **Cambridge, MA**
Master of Science in Data Science *09/2023 – 05/2025 (Expected)*

Carnegie Mellon University, GPA: 3.98/4.00 **Pittsburgh, PA**
Bachelor of Science in Economics and Mathematical Science *09/2019 – 05/2023*
Bachelor of Science in Statistics and Machine Learning *09/2019 – 05/2023*
Dietrich College Dean's List, with high honors *Fall 2019 – Spring 2022*

RESEARCH & INDUSTRY EXPERIENCE

NLP Research Assistant for Professor Matthew Gormley at CMU **08/2022 – Present**

- Responsible for Designing, training, and evaluating novel algorithms for sequence labeling and representation learning. Experimented with pre-training BERT-sized language model on multiple GPUs. Conducted various baseline experiments on machine translation.
- We have an ongoing project on improving LLM reasoning, using LLama-3-8B models.

(Capstone) Software Developer for Wearable Research at DtAK Lab, Harvard **09/2024 – 12/2024**

- Developed a Web app to visualize user data from wearable devices. The website interacts with users to gather input and provide preliminary statistical analyses for research purpose. Managed backend database using AWS's RDS. Worked with other researchers, translating research ideas to features in the app.
- The website is expected to be deployed in behavioral research with 30-50 participants in March 2025

SELECTED PROJECTS

Thesis in CMU, advised by Prof. David Childers

- Finetuned a GPT-2 model to mimic Federal Reserve (Fed) statements, archives a BLEU score of 0.66. Built forecaster with a language model as the backbone to predict changes in economic indicators from Fed statements, evaluated the model on historical data, and deployed the model for online prediction

Self-supervised Denoising via Speech-Noise Reconstruction

- Proposed a novel training paradigm for enhancing the de-noising ability of large speech models. Trained a HuBERT model, deployed it on speech recognition, reducing WER by 3 points on LibriSpeech's noisy split.

Course Projects on Computer Vision, NLP, and Multimodal ML

- Improving compositional reasoning in vision-language models by incorporating syntactic information.
- Built a VGGNet for image classification. Got 1% improvement by applying ML techniques for robustness.
- Proposed Label-Aware Attention network for fine-grained emotion recognition, raising macro F1 by 5 points

PUBLICATIONS

Yilin Wang, Xinyi Hu, Matthew R. Gormley. 2024. Learning Mutually Informed Representations for Characters and Subwords. *In Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Jianing Yang, Harshine Visvanathan, Yilin Wang, Xinyi Hu, Matthew R. Gormley, Improving Autoregressive Training with Dynamic Oracles. (Arxiv Preprint)

TECHNICAL SKILLS

- **Programming Languages:** Python, R, C, SQL (MySQL, PostgreSQL, SQLite3), Latex, Shell
- **Machine Learning Toolkits:** Pytorch, TensorFlow, Scikit-Learn, NLTK, Pandas, Numpy, Transformers
- **Technologies/Frameworks:** Linux, Git, AWS, GCP, HTML, CSS, CI/CD, Docker, W&B, Flask, Streamlit

SELECTED COURSEWORK

- **Machine Learning**: Natural Language Processing, Multimodal ML, TinyML/Efficient Deep Learning, Multiagent systems, Speech Processing, Reinforcement Learning,
- **Computer Science**: Software Development, Algorithms & Data Structures, MLOps
- **Math:** Analysis, Discrete Math, Linear Algebra, Calculus, Stochastic Process, Financial Engineering
- **Statistics:** Probability Theory, Statistical Inference, Time Series Forecasting, Econometrics