Corey Hu

contact@coreyhu.com | linkedin.com/in/coreyhu | github.com/coreyhu

PROJECTS

TruEra: April 2022 – Present

Machine Learning Engineer

- Pioneering research and productionization on actionable & interpretable strategies for LLM app evaluation at scale
- Successfully launched TruEra's NLP diagnostics platform for explaining and debugging Transformer models, with support for Tensorflow, HuggingFace, and PyTorch
- Took neural network drift analysis and explainability projects from exploration to production
- Started and led TruEra's first-ever mentorship and career development program

NVIDIA: June 2020 – April 2022

Deep Learning Engineer

- Oversaw end-to-end development and deployment of deep learning models for AI-assisted chip design
- Trained a transformer model to identify root crashes and errors from log files with 94% line classification accuracy
- Designed a recommendation system for clustering bugs and recommending improvements with historical data
- Co-authored 2 papers on gate sizing using Transformers with 98% accuracy and exponential runtime improvements against traditional EDA tools
- Founded and co-chairing NVIDIA's first-ever Asian & Pacific Islander Community, organizing company-wide events and campaigns around cultural education, advocacy, and professional development for over 200 members

Qualcomm Research: May 2019 – August 2019

Computer Vision Intern

- Developed GLANCE, a low-power computer vision sensor for object detection with ensemble cascading classifiers in a low-resolution and low-framerate environment
- Improved detection accuracy by 8% by designing a post-processing step involving dual IIR filters and stratification
- Developed an optimizer for tuning filtering parameters using analytical solvers on convex optimization problems that could be deployed and run offline

Berkeley Artificial Intelligence Research (BAIR) Lab: January 2018 – May 2019

Undergraduate Researcher

- Worked on the AIKA project for automated data modeling via optimizing machine learning/deep learning pipelines and architectures to generalize to a breadth of datasets and applications
- Researched lifted neural network (LNN) frameworks and adaptive activation functions with Professor Laurent El Ghaoui

Tencent AI Lab: May 2018 – August 2018

Machine Learning Research Intern

- Developed a two-tower Mask-RCNN and ensemble U-Net model designed to be robust towards small datasets and different cell types for nuclei instance segmentation (30 training images with ~22,000 nuclear boundary annotations)
- Co-authored a manuscript (Generalized Nuclear Segmentation using a Deep Convolutional Neural Network Method) detailing our model architecture, training schedule, and results to be published in a scientific journal
- Model performance ranked 9th and 14th in the MICCAI MoNuSeg and Digital Pathology Challenges respectively

EDUCATION

University of California, Berkeley

Bachelor's, Computer Science (2016 – 2020)

- Dean's Honor List, Jacobs Institute of Design Innovation Certificate

PROJECTS

Concierge

- Built a LLM information retrieval web app for recommending activities and points of interest near you, powered by GPT4 and Google Places API. Frontend created using React, NextJS

Malta

- Created a Python library for online training and dataset streaming with support for live dynamic dataset tuning
- Scaled the streaming backend service with Kubernetes and support for AWS EKS

SKILLS

Python, TensorFlow, PyTorch, HuggingFace, AWS, React, Java, C++, HTML, CSS, SQL, JavaScript, git, Linux/Unix