SANGWU LEE

+1(585) 537-9227 ⋄ New York City NY

sangwu@quizard.ai ⋄ linkedin ⋄ google scholar ⋄ github

EDUCATION

BS Computer Science, University of Rochester, GPA 3.98/4.0, Dean's List May 2024
Relevant Coursework: Artificial Intelligence, Computer Vision, and Deep Learning

BS Honors Mathematics, University of Rochester, GPA 3.98/4.0, Dean's List May 2024
Relevant Coursework : Linear Algebra, Analysis, Differential Equations, Computational Statistics

WORK EXPERIENCE

Founding ML Engineer Dec 2023 - Present
Quizard AI *New York City, NY*

- Optimized topic modeling and visualization pipeline, achieving 25× speedup on 40M+ student questions using NVIDIA cuML and Google BigQuery on A100 GPUs.
- Enhanced QnA RAG pipeline performance, reducing latency by 2× and costs by 10× through strategic migration of document retrieval engine using LaunchDarkly feature flags.

Machine Learning Researcher Aug 2019 - Present
Language Technology Institute, Carnegie Melon University *Pittsburgh, PA*

- Enabled large-scale finetuning of Huggingface models with Weights and Biases, totaling 5000+ GPU hours.
- Attained 85.7% accuracy setting state-of-the-art performance on 2 multimodal datasets using BERT and XLNet.
- Developed a novel vision-language transformer model (VLM) by integrating InstructBLIP and ViT achitecture, reducing model parameters by 10× and cutting inference costs by 15×, while retaining model performance.
- Published 3 papers in top AI conferences at AAAI, ACL, and ICLR.

PROJECTS

Generative Model Training [report]

- Trained latest image generation models on TPUv3 cluster for 100+ hours. Implemented various GAN, diffusion, autoregressive models including ViT-VQGAN, MUSE, Stylegan, EDM, and DDPM in JAX.
- Released 3.6M+ illustration dataset containing RGB image, text caption, sketches, and depth map.

Neural Celluar Automata [demo]

- Implemented neural cellular automata using JAX inside Google Colab environment.
- Deployed a working public demo on Netlify using tensorflow.js and SvelteKit.

PUBLICATIONS

Integrating Multimodal Information in Large Pretrained Transformers [pdf] [code]
ACL 2020 - 450+ citations

Humor Knowledge Enriched Transformer for Understanding Multimodal Humor [pdf] [code]
AAAI 2021 - 60+ citations

Self-Imagine: Effective Unimodal Reasoning with Multimodal Models using Self-Imagination [pdf]
ICLR 2024 Workshop on LLM Agents

SKILLS

Technical Skills	Python (6 years), HTML/CSS/JS (7 years), React (6 years), Next.js (2 years), Svelte (1 year)
ML Skills	PyTorch (6 years), PyTorch lightning (2 years), JAX (2 years), accelerate (2 year), TPU (2 years)