Siddharth Verma

https://siddharthverma314.github.io | siddharthverma314@gmail.com

About me

I am an ML Engineer with deep expertise in NLP and Reinforcement Learning. I have experience training SoTA 100B parameter LLMs. I have performed ML research in both academic and industrial settings, resulting in multiple papers published in venues like NeurIPS and ACL.

Languages: Python, Haskell, Rust, Java, C, Go, PureScript, RISC-V, SQL
Skills: Language Modeling, Distributed Training, Reinforcement Learning, Neural Networks, Statistics
Technologies: Pytorch, Pytorch Lightning, JAX, Docker, NixOS, Unix/Bash, Git, Google Cloud, AWS

Experience

Research Engineer

Character.ai | New York NY | Dec 2023-Current

- Contributed to all aspects of LLM pretraining from fundamental research to performant implementations
- Implemented multiple MoE variants for our flagship model trained across our entire cluster of GPUs
- Discovered the exact relation between attention and intelligence in LLMs
- Designed scaling laws to predict a 5 order-of-magnitude extrapolation in validation loss with tight error bounds

Senior Machine Learning Engineer

Square | Boston MA | Sep 2022-Dec 2023

- Finetuned open-source LLMs on merchant-buyer conversations to suggest replies to incoming messages
- Conducted an online A/B test and demonstrated a 5% increase in suggestion acceptance rate
- Designed and implemented a multi-task training system to incorporate classification tasks into an LLM
- Instruction finetuned FLAN-T5 on internal data and evaluated performance against individual classifiers

AI Resident

Meta (Facebook) | Seattle WA | Aug 2021-Sep 2022

- Wrote code to process 1TB of multimodal data using Rust and Parquet for a 20x speedup against Python
- Automated the training LLMs of up to 13B parameters on large multi-node clusters with up to 64GPUs
- Evaluated whether training on explanations improve reasoning capabilities of LLMs, and found that explanations mostly benefit mathematical reasoning
- Analyzed effect of masking rates and masking strategies in multimodal learning, showing that increasing masking rate nullifies effects of different masking strategies

Machine Learning Intern

Apple | Seattle WA | Jun 2021-Aug 2021

- Implemented Transformer architecture from primitive operations for an in-house deep learning framework
- Demonstrated correctness by replicating English-German translation results from 'Attention Is All You Need'
- Optimized self-attention for Apple Neural Engine by rewriting computation with supported operations

Undergraduate Researcher at Robotic AI and Learning Lab

Berkeley Artificial Intelligence Research Lab | Berkeley CA | Jan 2019-May 2021

- Worked with Prof. Sergey Levine and Prof. Chelsea Finn on RL and NLP in domains of robotics and chatbots
- Designed and implemented a multi-agent RL algorithm to learn composable locomotive skills without manual environment resets, subsequently using them to solve a maze. Published at NeurIPS
- Used Offline RL to finetune LLMs to bargain on craigslist items, beating supervised learning in human evals across all metrics. Accepted as oral presentation at NAACL

Education

BA Computer Science & Music

UC Berkeley | GPA: 3.965, Major: 4.0 | Aug 2017-May 2021

- **Courses**: Machine Learning, Artificial Intelligence, Probability and Random Processes, Theoretical Statistics*, Information Theory and Coding*, Security, Operating Systems, Data Structures, Computer Architecture, Algorithms, Real Analysis * *indicates graduate level*