Arpan Jain

Website: aj-prime.github.io/aj-prime/ | Email: arpan.jain1405@gmail.com | LinkedIn: aj-prime | GitHub: github.com/aj-prime

Summary

Experienced Machine Learning and Systems Engineer with a Ph.D. in High-Performance Deep Learning. Expertise in large-scale distributed training, LLM inference optimization, and ML systems. Proven track record in both academic and industry environments, including Microsoft and national labs. Passionate about building robust ML infrastructure and enabling fast iteration and reliable deployment of ML models.

Technical Skills

ML Frameworks: PyTorch, TensorFlow, DeepSpeed, LBANN, SGLang
Infrastructure: Distributed Training, Model Parallelism, Elastic Fault Tolerance, GPU/CPU/DPU Optimization, ML Systems Benchmarking
Languages: Python (Expert), MPI, C
Tooling: NVTX, NVIDIA Triton, Horovod, MVAPICH2, SLURM, Docker
Concepts: Parallel Computing, LLM Inference, DNN Optimization, HPC

Experience

- **Senior Researcher – AI Frameworks, Microsoft (CoreAI) | Redmond, WA | Aug 2024 – Present**
  - Developed and optimized internal inference stack for LLMs using SGLang.
  - Improved latency and throughput for online deployments of Microsoft-internal LLMs.
  - Benchmarked and characterized leading LLMs for production readiness.
  - Collaborated across research and infra teams to deliver robust inference systems.
- **Applied Scientist II – Bing Ads, Microsoft | Redmond, WA | Jan 2023 – Aug 2024**
  - Led infrastructure efforts to optimize GPU inference workloads, saving $1.5M+ annually.
  - Implemented dynamic batching, multi-GPU utilization, and model-specific acceleration.
  - Deployed and evaluated ANN search methods and AMD GPU integration.

- **Graduate Research Assistant – Network-Based Computing Lab, OSU | Columbus, OH | Dec 2018 – Dec 2022**
  - Designed novel distributed DNN training algorithms (Hy-Fi, SUPER, GEMS).
  - Integrated training optimization strategies into TensorFlow and PyTorch using MVAPICH2.
  - Led conversational AI for HPC initiative.
- **Research Intern – DeepSpeed Team, Microsoft | Bellevue, WA | Summer 2022**
  - Developed elastic fault-tolerant training strategies for large-scale Transformers.
- **Intern – LBANN Group, Lawrence Livermore National Lab | Livermore, CA | Summers 2020 & 2021**
  - Designed sub-graph parallelism and optimized Transformer inference/training.
  - Improved performance over traditional data parallelism for large models.

Education

The Ohio State University – Ph.D., High Performance Deep Learning
GPA: 4.0 | Advisor: Prof. D. K. Panda | 2018 – 2022
Thesis: Novel Parallelization Strategies for High-Performance DNN Training on HPC Systems

ABV-IIITM Gwalior – Integrated Post Graduate (B.Tech + M.Tech)
GPA: 8.83/10 | 2013 – 2018

Key Publications

- Hy-Fi: Hybrid 5D Parallel Training on GPU Clusters (ISC '22)
- SUPER: Sub-Graph Parallelism for Transformers (IPDPS '21)
- GEMS: GPU Memory-Aware Model Parallelism (SC '20)
- Optimizing Distributed DNN Training using CPUs & BlueField-2 DPUs (IEEE Micro '21)

Selected Open-Source and Talks

• Contributor to DeepSpeed, SGLang, MVAPICH2
• NVIDIA GTC '24 Speaker – ANN on GPU
• Invited tutorials at ISCA, SC, PPoPP, ASPLOS, HotI