BINFENG XU

+1 (336) 744-6809 ‖ billxbf@gmail.com ‖ 🏠 billxbf.github.io

EXPERIENCE

Senior Research Engineer @ Samsung Research America — Oct 2023 – Present

Large Foundation Models; On-device AI.

Responsibilities and Contributions:

I. **Post-Training.** Distributed training of 3 ∼72B LLMs on 1∼64×A100 GPUs for various tasks.

- Supervised Fine-tuning (SFT) LLMs for task-specific instruction following.
- Preference Alignment from user feedback data, with RL-derived algorithms such as DPO.
- Knowledge Distillation from server-side LLMs to on-device light-weight LLMs.
- Parameter Efficient Fine-tuning with LoRA and QLoRA for on-device deployment.

II. **RAG and Agents.** End-to-end development of two agentic systems for conversational chat assistant.

- Designed and implemented a modularized agent architecture for LLM inference, prompts, vector database, embedding models, retrieval, ranking, and function calling.
- Fine-tuned LLMs for enhanced function calling, context-aware conversation, and safety / rejection.
- Built and benchmarked various light-weight NLP sub-modules such as Query-to-SPARQL generation, Named Entity Recognition (NER), and Intent Classification wrapped for agent function calls.

Selected Projects:

I. **On-device Chat Assistant over Personal Knowledge Graph.** (Galaxy AI)

- Led a group of 6 researchers and engineers to develop an efficient agent system over Personal Knowledge Graph (PKG) for Samsung Gallery, where an on-device 3B LLM learns to effectively locate and retrieve information from users' private knowledge database (RDFox) and converse with users.

II. **Customer Service Troubleshooting Chat Assistant.** (Samsung Online Shop)

- Independently developed a RAG pipeline to assist users on electronics troubleshooting. Trained a series (7B, 27B, 70B) of LLMs to assist users with retrieved context from PDF user manuals, past troubleshooting chat log, etc. Hosted server instances for inference and provided RESTful APIs.

III. **Research and Side Projects.**

- LLM Reasoning: Mentored a PhD intern to augment LLM Mathematical reasonings (on AIME) with step-wise data synthesis from Monte Carlo Tree Search (MCTS) and dynamic step prediction.
- LLM Pretraining: Dumped Yahoo Finance database, tokenized Nasdaq stocks candle history, and trained (from scratch) a 7B transformer on 5 million tokens to predict the market.
- Efficient Inference: Benchmarked Speculative Decoding to accelerate Query-to-SPARQL.

Applied Researcher @ eBay — Jul 2022 – Oct 2023

Recommender Systems; Natural Language Processing; MLOps.

I. **Table-view Product Comparison Module.**

- End-to-end developed a recommendation algorithm for comparable / alternative items, popping up predicted key aspects and features users intend to compare.
- Trained and served (1) A two-tower User-Product BERT-based embedding model for product candidate recalls. (2) An Aspect Importance model to rank popping-up features.
- Collaborated with a front-end engineer to serve through production, which significantly improved page-level Click-Through-Rate (CTR) by 20%+ through multiple A/B tests.

II. **Semantic Book Recommendation.**

- Recreated Book Recommendation service on eBay with semantic embeddings from ISBN databases.
- Built and served a batch-updating ANN Index service for ISBN embeddings. Engineered downstream recall & ranking pipelines to merge ISBN top neighbors into recommendation candidates.

Research Intern @ eBay — May 2021 – Aug 2021

Fine-tuned an RoBERTa-based fraud detection model using seller-buyer chat log, and served with ONNX.

Research Intern @ Baidu AI Lab — May 2018 – Aug 2018

Benchmarked YOLO v3 for an internal face recognition model based on Siamese Network.

EDUCATION

NEW YORK UNIVERSITY *M.S. in Data Science* SEP 2020 – MAY 2022

Relevant Coursework: Deep Learning, Natural Language Processing & Understanding.

WAKE FOREST UNIVERSITY *B.S. doubling Computer Science and Statistics* SEP 2016 – MAY 2020

Relevant Coursework: Machine Learning, Computer Vision, Data Structures & Algorithms, Parallel Programming, Numerical Computation, Probability & Statistics, Markov Processes, Convex Optimization.

Research @ New York University DEC 2020 – MAR 2021

- *Prediction and Policy-Learning under Uncertainty* Advisor: *Yann LeCun, Alfredo Canziani*
  Involved in a policy learning project for self-driving cars, where I experimented with introducing penalization term to the loss of environmental cars in dense traffic to avoid colliding into the studied agent.

Research @ Wake Forest University SEP 2018 – DEC 2019

- *Tucker Decomposition with f-MRI Neural Activity Tensor* Advisor: *Grey Ballard*
  Developed a novel algorithm to compress high dimensional f-MRI tensors with Tucker Decomposition for data efficiency, benchmarked against various statistical models.
- *Object Recognition in Peru Forest* Advisor: *Paul Pauca*
  Trained YOLOv3 to detect illegal mining activities of drone-taken images above Peru forests.

WRITINGS

ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models 2023

Binfeng Xu*, *Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, Dongkuan Xu*

Gentopia: A Collaborative Platform for Tool-Augmented LLMs 2023

Binfeng Xu*, *Xukun Liu, Hua Shen, Zeyu H, Yuhan L, Murong Y, Zhiyuan P, Yuchen L, Ziyu Y, Dongkuan X*

Dynamic Noise Preference Optimization for LLM Self-Improvement via Synthetic Data 2024

Haoyan Yang, Ting Hua, Shangqian Gao,* ***Binfeng Xu****, Zheng Tang, Jie Xu, Hongxia Jin, Vijay Srinivasan*

HONORS

Kaggle: Competitions Master - Global Ranking Top 1% MAY 2018 – PRESENT

- Santander Customer Transaction Prediction (Banking, Classification): Rank #24 of 8,802 (Gold);
- Santa 2020 Contest (Competitive Reinforcement Learning): Rank #17 of 792 (Silver);
- BirdCLEF 2021 Birdcall Identification (Signal Processing, Time-Series): Rank #15 of 816 (Silver);
- Predicting Molecular Properties (Biochemical Modeling): Rank #182 of 2737 (Bronze);
- LMSYS - Chatbot Arena Human Preference Predictions (Reward Modeling): Rank #143 of 1849 (Bronze);

ACM ICPC: Regional 4th Place in North Carolina MAR 2019 – MAR 2019

Udacity Nanodegree in Deep Learning SEP 2018 – OCT 2018

TECH STACK

Area of Expertise: Large Foundation Models, NLP, MLOps, Search & Rec System, Data Science.

Programming: Python3*, Pytorch*, Scala*, Java, SQL, C#, C++; Spark, Spring, React, Unity3D (simulation).