Dr. ZHENHUAN YANG

E-mail: zhenhuan.yang@hotmail.com | Tel: (201) 551-9851 | Website: https://zhenhuan-yang.github.io/

Work and Intern Experience

Etsy Inc—Senior Applied Scientist 2024—now

- Pioneered the integration of Azure OpenAI, Google Cloud Vertex AI and other models with Langchain runnable interface for building LLM applications, driving innovative solutions within the company.
- Speed up LLM batch inference with the asyncio and the multiprocessing library. Complete company's 130 million inventory inference in 2 weeks by reaching the maximal available token per minute (TPM) through vendor.
- Lead the query and listing entity recognition project with LLM. Index human defined entities with FAISS library and implement retrieval augmented generation (RAG) during extraction prompting. The first use case in search filter leads to site-wise filter usage improvement by 2% and search click improvement by 0.5%.

Etsy Inc—Applied Scientist II 2022—2024

- Collaboratively prototype and productionize the company's first deep neural network (DNN) search re-ranking model for last-pass business logic enhancement in the cascade search ranking system.
- Develop purchase price based learning-to-rank loss function and evaluation metrics. Improve site-wise average order value (AOV) by over 3% in online A/B test, leading to a yearly 16M incremental gross merchandise sales (GMS), placing me in the top 10% of performance by headcount.
- Feature processing with Tensorflow transform (TFT) and Dataflow, DNN modeling and training with Keras and Vertex AI, prototyping workflow orchestration with YAML and Kubeflow, model deployment with Docker and Tensorflow serving (TFX).

Intern Experiences

- Research scientist intern at Etsy@2021. Improve click NDCG of cold-start users by 10% through integer programming via user group fairness constraints. Work is accepted by WWW'2022 industrial track.
- Research and development intern at Tencent@2019. Improve pix2pixHD-based 2D human broadcaster synthesis model training speed by 25% through stochastic gradient algorithmic optimization. Work is patented.

Selected Publications

Zhenhuan Yang, Y. L. Ko, K. Varshney, Y. Ying. *Minimax AUC Fairness: Efficient Algorithm with Provable Convergence.* **AAAI'2023**. #4 ranked AI publication by Google Scholar

Zhenhuan Yang*, Y. Lei*, T. Yang, Y. Ying. *Simple Stochastic and Online Gradient Descent Algorithms for Pairwise Learning.* **NeurIPS'2021**. #1 ranked AI publication by Google Scholar

Y. Lei*, **Zhenhuan Yang***, T. Yang, Y. Ying. *Stability and Generalization of Stochastic Gradient Methods for Minimax Problems.* **ICML'2021**. #3 ranked AI publication by Google Scholar

Education

State University of New York at Albany 2016—2018 / 2018—2022

M.A. in Mathematics / Ph.D. in Mathematics

Sun Yat-sen University 2012—2016

B.A. in Mathematics & Applied Mathematics

Skills and Services

- **Programming & Frameworks:** Python, Tensorflow, Pytorch, Scala, SQL, Git, Docker, GCP, Apache Beam, Kubeflow, Airflow, Kubenetes, LangChain, HuggingFace
- **AI / ML:** recommendation system, information retrieval, natural language processing, deep learning, prompt engineering, retrieval augmented generation (RAG), LoRA fine-tuning
- **Academic Services:** reviewer for NeurIPS, ICLR, ICML, CVPR, committee member for AAAI, IJCAL