Tian Zhang

0tianzhang0@gmail.com | 213-373-0955 | San Jose, CA | https://www.linkedin.com/in/tian-zhang-87b4a0165

WORK EXPERIENCE

Senior Machine Learning Engineer @ Walmart Global Tech | Sunnyvale, CA Oct 2021 – now

Improve Search Ranking for Walmart eCommerce

- Developed and deployed **Learning-to-Rank** models, enhancing Walmart's eCommerce search relevance and improving user engagement.
- Engineered production-grade features and signals in **Java** for real-time model inference.
- Scaled large datasets and trained state-of-the-art models **(XGBoost, BERT, LLMs)**, leveraging distributed computing for optimal performance.
- Led end-to-end model lifecycle, including feature engineering, training, hyper-parameter tuning, deployment, and A/B testing, accumulated more than 6% improvement in Walmart search relevancy (NDCG@5).

Knowledge Distillation from LLM to Scalable Deep Learning Ranking Model to improve Search Relevancy

- Fine-tuned a Mistral 7B LLM model to generate 170M high-quality pseudo-labels, leveraging its robust language understanding for enhanced training data.
- Implemented margin MSE loss to effectively distill knowledge from the teacher model, optimizing relevance metrics in search ranking.
- Developed a knowledge distillation pipeline to train a compact BERT student model, significantly reducing model size while maintaining high accuracy.
- Applied the new student model in Walmart search system, achieved 4% relevance gain.

BERT-Based Single Tower Cross Encoder Feature for Tail Query Ranking

- Proposed and developed one tower cross-encoder model to predict search query-item relevancy, enhancing ranking for tail queries and boosting NDCG by 10% and ATC by 0.37%.
- Optimized model deployment using TensorRT and TorchServe, achieving 40-70 QPS on A100 GPUs.
- Applied post-training quantization (PTQ) and quantization-aware training (QAT) to minimize latency while maintaining model performance.

Neural-Net-Based Ranking Model

- Replaced Xgboost ranking model with MLP models, improve the relevance of Walmart search by 1.23%
- Try different training data and loss functions to optimize both relevance and engagement metrics

Machine Learning Engineer @ JM Eagle | Los Angeles, CA Dec 2019 – Oct 2021

Sale Quotation Prediction

- Built 4 separated ***XGBoost*** models to predict quotation price and suitable shipping plant, order by date, and estimated shipping date based on 210,000 quotation data, the price has an accuracy of 96%(***mape***), the plant has an accuracy of 98.5%, the date has an ***RMSE*** of 3.74 days
- Rolled up models into ***APIs*** (get data, train, predict) by ***Flask***, ***Dockerized*** files that could serve local or cloud
- Deployed ***Docker Container*** into ***AWS*** by ***Serverless*** framework with ***AWS API Gateway, Lambda*** to receive and manage web requests, set ***SageMaker*** to train and update model daily
- Tuned and updated model by applying new features and shrinking training period to face the dramatic price change caused by COVID, the prediction accepted rate returned to 82% from a huge drop to 21%

Monthly Sales Forecast

- Collected 10 years of product sales data to build a monthly demand forecast model by using the ***DeepAR*** model, accuracy(***MAPE***) has improved by 10% compared with the previous manually forecast(65%)
- Deployed the forecast pipeline to AWS with usage of ***Lambda***, ***AWS Forecast***, ***Step Functions***
- Design ***REST APIs*** for training forecast model and sending recent forecast results to subscribers by email (***AWS SNS***) and set monthly forecast report training and sending to subscribers at the beginning of each month

SKILLS

Machine Learning & Deep Learning: Transformer, LLM, XGBoost, Reinforcement Learning, NLP, Recommendation Systems, Information Retrieval

MLOps & Cloud Deployment: Docker, Kubernetes, CI/CD, AWS, GCP

Programming Languages: Python, Java, SQL, C/C++

Frameworks & Tools: TensorFlow, PyTorch, TensorRT, TorchServe, Serverless, Spark, Git, Jupyter Notebook

EDUCATION

University of Southern California, Los Angeles, US **Master's: Electrical Engineering :** 3.9/4.0 2017 – 2019

Beihang University, Beijing, China **Bachelor's: Electrical Engineering:** 3.5/4.0 2013 – 2017