Hao-xuan (Horace) Wang

+86 18622468042

billweasley20092@gmail.com

Github: https://github.com/billweasley

Linkedin: https://www.linkedin.com/in/horace-haoxuan-wang

Personal Website: http://shellcottage.me

Work Experience

2025.03 - Now • **Senior Data Scientist, Integrity and customer experience** @ Grab

- Initial project is about producing an enhanced on-device keyword spotting model and combining it with service-side multimodality LLM to detect safety issue during taxi hailing.

- 2022.03 - 2025.02 **Machine Learning Engineer, ASR and Language Tech** @ Zoom

  - Independently experimented with LLM for ASR in multimodality setting, aiming to improve the consistency of ASR decoding results. The orthographic (fully-format) WER (word error rate), and the rare word WER achieved a better result compared with the production model.
  - LLM post-processing for transcription: used a closed-source LLM for ASR error correction post-processing, extracting the N-best list from the Zipformer-Transducer model, and wrote prompts combined with a biasing word list to send to Claude 3.5 Sonnet for named entity correction. Offline experiments on a medical dataset showed that the **Rare word WER decreased from 37.8% to 17.5%** .
  - Trained a speech recognition and text punctuation model, building a LAS-S2S Danish model from scratch. The initial WER on the test dataset was about **8%**, further reduced by data augmentation, **outperforming MS Teams' results**. The initial overall F1 score for case and punctuation was about **70%**.
  - Independently implemented **Whisper** inference support, optimization, and performance evaluation (WER, RTF/latency/throughput), using an in-house VAD (voice activity detection) model and open-source WhisperX. Achieved higher throughput compared to OpenAI's implementation and lower WER on most test sets.
  - Implemented multi-head attention (MHA) based time alignment on LAS/seq2seq models to provide good word-level timestamps, meeting the business needs for multilingual transcription. [Filed a US patent for this]

- 2019.08 - 2021.12 **Data Scientist and Software Engineer** @ Barclays

  - Company address matching and entity matching without internal GPU and labeled data available. Solve using an active learning method. Start from constructing some small datasets only with external data and training an XGBoost tree, then label samples in the boundary and fine-turn BERT models in an iterative way. Finish the inference on 6 million internal pair-wised samples with this model on a CPU cluster, using a DJL based pipeline built from scratch on my own. It achieved a very satisfying result of **94% F1 score on a noisy testing dataset from 89% where we started**. The model does inference offline on our Spark cluster in a distributed way. For 6 million pair-wised samples, the running time is under 1 hour (on a cluster with 80 CPUs).
  - Predict the aggregated user's transaction activity (volume and value) using the historical mean and Informer model, a variant of Transformer for time-series modeling. Following that, a counterfactual was constructed to provide an evaluation of how much finance loss that the bank suffers from system downtime and to find out the critical period for the system reliability.

Education

2018 - 2019	MSc Web Science and Big Data Analytics @ **University College London,** Distinction
2016 - 2018	BSc Internet Computing @ **University of Liverpool *,** First class
2014 - 2016	BSc Information and Computing Science @ **Xi'an Jiaotong-Liverpool University ***

***Note:** 2+2 pathway routine (first 2 years in Suzhou, China and final 2 years in Liverpool, UK), dual degree.

Personal Project

- 2024.06 - **Fine-tuning and evaluation of medical record data on Large Language Models (LLMs)**

  Using hundreds of thousands of Chinese medical case data, fine-tune different LLM foundation models (Llama3-instruct, Llama3 Chinese-chat, Qwen2) on tasks such as department classification, medical record summarization, and discharge certification. On the domain-specific test dataset it achieved significant improvements in scenarios like consultation summary/discharge summary (BLEU 0% ~ 30% -> 49% ~ 55%, ROUGE-L 20% ~ 30% -> 60% ~ 64%), and scenario like the department classification for multi-class accuracy (Accuracy 0% ~ 36% -> 69% ~ 71%). We plan to open source the data in the future.

Technical Article

- **"Accelerating Deep Learning on the JVM with Apache Spark and NVIDIA GPUs"**

  Author: Haoxuan Wang, Qin Lan [AWS], Carol McDonald [Nvidia]; Link: https://www.infoq.com/articles/deep-learning-apache-spark-nvidia-gpu/?itm_source=articles_about_ai-ml-data-eng&itm_medium=link&itm_campaign=ai-ml-data-eng

Project

- 2025.02 - 2025.02 **Work trail** @ Finalround.ai

  In a one-week project focused on intent detection from intermediate ASR results, I independently implemented a detection pipeline and achieved an F1 score of 87% on a validation meeting. Notably, half of the intents were detected ahead of the ASR final utterance. This work enables me to receiving a job offer from them. The complete pipeline included: 1) Rule-based handling of greeting utterances.2) Evaluating sentence completeness using segment-any-text, syntactic parsing, and perplexity scoring. 3)Detecting confirmation-type questions (e.g., "Can you hear me?") using Sentence-BERT embeddings. 4) Classifying final question intents with a small language model. Also, I developed prompts for extracting resume information, which improved the personalization and quality of LLM responses.

Technical Skills

- Programming Language: Python, Java, C
- Deep learning, ASR and NLP: PyTorch, HuggingFace, Whisper, K2, Transformers (and its variants), LLM fine-tuning (LoRA, Full-paramter)
- Data processing: Spark, Pandas
- Training and System Infra: DeepSpeed, AsyncMQ/Kafka, Docker, Kubernetes (Istio, Knative, ...)
- General: Jenkins, git, JIRA