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

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

  Tech stack: Torch/Pytorch, ONNX, Kubernates, Huggingface

  - Used a closed-source LLM for ASR error correction post-processing, extracting the N-best list from the Conformer-Transducer model, and wrote prompts combined with a biasing word list to send to GPT-4 for named entity correction. Offline experiments on a medical dataset showed that the **Rare word WER decreased from 37.8 to 17.5**.
  - Wrote meta-prompts, using open-source LLM models (e.g., Mistral MoE 8x7B) to generate hundreds of dialogue scene prompts, and combined them with different reading formats of numerical sequences generated by the LLM model. These were used to create dialogue scenarios with numerical texts, which were handed over to colleagues to generate audio using internal/Microsoft TTS services, resulting in a test set and part of the training data. Fine-tuned the production model with the training data resulted in a slight improvement on another internal digits dataset (**Absolute digit WER decreased by about 0.4%**).
  - (Ongoing) Use SLAM-LLM to experiment with ASR/LLM integration for speech + text modality SFT, aiming to improve the consistency of ASR decoding results.
  - Trained a speech recognition and text punctuation model, building a LAS-S2S Danish model from scratch. The initial word error rate (WER) on the test dataset was about **8%**, further reduced by phoneme 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, tested, maintained, and evaluated key functions of the inference system, including **decoding**, **timestamp generation**, and **post-processing** code.
  - 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]
  - Collaborated with downstream web and infrastructure teams to streamline the offline transcription system architecture and deployment, supporting **35+** single-language models across **10+** different regions for **5+** different downstream services.
  - Evaluated the performance of the speech recognition system, seeking the most suitable parameters to reduce real-time factors or improve system throughput.

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

  Tech stack: Spark / PySpark, Amazon Deep Java Library (DJL), Tensorflow / Keras, Pandas, Jupyter, Pretrained Transformers / Likelihood Ratio

  - 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.
  - Maintain the Spark cluster for the team, and build up pipelines for distributed inference by combining DJL / PySpark UDF with models.Collaborated with one of my colleagues, we created a team-wised package to start a Spark session within 4 lines of codes, which significantly reduces the overhead of using Spark for colleagues who are not with a distributed computing background.

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)**

  （Ongoing）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

Early Stage Project

- 2019.06 - 2019.09 **Project Internship (Master Degree Thesis)**@ Astroscreen

  Social media posting language source identification (tweets and gabs) project. Finished a crawler for collecting language (posts) data from Gab.com, pre-processed data using Regular Expression, built models for classifying the source of these data by fine-turning BERT and XLNet, visualized results using t-SNE, did "leave-one-hashtag-out" cross-validation, and evaluated models using some common matrics (Accuracy, F1 score, Confusion Matrix, Matthews Correlation Coefficient). After fine-turning, XLNet shows a 86% F1 score on hashtag-balanced test dataset, reducing from 97% on random-balanced test dataset. The results show the protential for doing source checking using a model and also indicate the importance for avoiding data leakage.