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
  - Led experiments on integrating **LLMs with ASR models** in multimodal settings, significantly improving consistency in ASR decoding. Achieved better **orthographic WER** and **rare word WER** compared to the production model.
  - Developed **LLM-based transcription post-processing pipelines**, leveraging N-best lists from **Zipformer-Transducer models** and customized prompts with biasing word lists sent to Claude 3.5 Sonnet. Offline experiments on a medical dataset reduced **Rare Word WER from 37.8% to 17.5%**.
  - Designed **LLM-driven data augmentation workflows**, utilizing **Mistral MoE 8x7B** to generate diverse dialogue scenarios and numerical reading formats. Resulting datasets improved ASR digit recognition performance (**Absolute digit WER reduced by ~0.4%**).
  - Built a **LAS-S2S Danish ASR model** from scratch, achieving an initial **WER of ~8%** and **punctuation/case F1 score of ~70%**, outperforming **MS Teams' benchmarks** after data augmentation.
  - Independently optimized **Whisper inference pipelines**, integrating in-house **VAD models** and **WhisperX** to deliver superior **WER and throughput** performance compared to OpenAI's implementation.
  - Implemented **multi-head attention-based time alignment** in LAS/Seq2Seq models to deliver precise **word-level timestamps** for multilingual transcription. (**Patent filed**)
  - Maintained and optimized **ASR inference pipelines**, resolving production-level issues and ensuring smooth operations.
- 2019.08 - 2021.12 **Data Scientist and Software Engineer** @ Barclays
  - Developed an **entity-matching pipeline** using **active learning techniques**. Constructed small, externally sourced datasets with fine-tuned **BERT models**, achieving a **94% F1 score** on noisy test datasets. Deployed inference on a distributed **DJL-based CPU cluster**, processing **6 million pairwise samples in under 1 hour**.
  - Applied **Informer models** for **time-series transaction forecasting**, enabling accurate predictions of transaction volume and counterfactual financial loss assessments during system downtimes.

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 program (first 2 years in Suzhou, China, 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)**

  Fine-tuning **LLaMA3-instruct**, **LLaMA3 Chinese-chat**, and **Qwen2** models on large-scale **Chinese medical datasets** for tasks such as **department classification**, **medical record summarization**, and **discharge report generation**. It was planned to **open-sourcing datasets**. Achieved notable improvements:

  - **Consultation/Discharge Summarization:** BLEU (**0%-30% → 49%-55%**), ROUGE-L (**20%-30% → 60%-64%**)
  - **Department Classification:** Accuracy (**0%-36% → 69%-71%**)

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

  Worked on **social media language source identification** (e.g., tweets and gabs).

  - Implemented a **crawler for Gab.com** to collect linguistic data.
  - Processed data using **Regular Expressions** and fine-tuned **BERT** and **XLNet** models for classification tasks.
  - Applied **t-SNE visualization** and **"leave-one-hashtag-out" cross-validation** to prevent data leakage.
  - Achieved **86% F1 score** on a **hashtag-balanced test dataset**, demonstrating the importance of avoiding biased splits during training.
- 2019.02 - 2019.03 **Integrated BERT and Embeddings in CommonsenseQA Challenge**

  Fine-tuned **Google BERT** for **CommonsenseQA Challenge 1.0**, integrating **ConceptNet Numberbatch and ELMo embeddings**. Achieved **68.79% accuracy** on validation datasets (BERT only: 67.47%).