ZIYUE WANG

07703773993 ◇ Wangziyue@hotmail.co.uk

SKILLS

Programming languages	Python 3, C++ 14, C#
Tools	AWS (Airflow, S3, Athena), Docker, Postman, LocalStack, Git, Jira
Libraries	Tensorflow, Keras, PyTorch, ONNX, OpenCV, Numpy, Pandas, OpenSSL, Curl Scikit-learn, LightGBM, Pydantic, DAGs
Languages	English (native), Mandarin (native)

EXPERIENCE

Senior Machine Learning Engineer, Edited **11/2023 - Present**

- Headed the implementation of DS research code into production-ready code for use in Airflow, including development of data input/output (IO) with Docker containers (using MongoDB, etc.). Hosted on AWS and serves to thousands of users.
- Expanded existing translation models to support additional languages (e.g., Portuguese), reducing translation costs by 40% annually.
- Refined pre-existing classification model using GPT-3.5 to correct errors, developing a pipeline that decreased manual labeling by 90%, significantly reducing the need for external labelers. Accuracy increased by 5% overall, resulting in an additional 3.5 million correctly classified products over a 6-month period.
- Mentoring junior DS and MLEs on coding and testing best practices, as well as CI/CD. Laid the groundwork for a testing best practice schema for junior members to follow

Senior Machine Learning Engineer, Evalueserve **9/2021 - 10/2023**

- Led team of 3 to build a fruit and vegetable recognition software, now deployed in over 10,000 stores
  - Trained CNN model to achieve 95% accuracy using tensorflow, optimized (openvino/NCNN/ONNX) to 200ms inference time and deployed as a C++ DLL and python service
  - Implemented integration links for C# (Windows) and Java (Android) frontend using C++/CLI and JNI respectively
  - Streamlined and automated library unit testing, packaging and encryption (model) to reduce human errors and deployment time by up to 60%
- Developed PDF parsing package to extract structural information from B2B PDFs (financial and annual reports) Used in NLP pipelines to reduce human workload by up to 40%
- Reduced several LLM models GPU usage through quantization. Improved inference performance on CPU by up to 2.5x through conversion to Openvino and Onnx
- Led the development of Human Action recognition using AlphaPose, a software aimed at identifying when a person had fallen through CCTV footage. Optimized the DL model so that it could run on low-cost hardware, improving CPU inference time by 40%

Machine Learning Engineer, Evalueserve **2/2020 - 9/2021**

- Build a personnel tracking software using YoloV5 and DeepSort to track behaviour of shoppers through retail store, achieving overall accuracy of 74%. Deployed on AWS through Django
- Implemented pipeline for MOT to be mapped to a 2D bird's eye view of the environment, through the use of OpenCV's image processing and homography, which allowed reconciliation of MOT through multiple camera feeds
- Optimized existing Safety Helmet detection deep learning model to run on low cost hardware (edge), increasing overall speed by 150% and saving hardware costs by 50%

EDUCATION

The University of Birmingham *September 2015 - July 2019*
MSci Physics Upper 2:1 Honours