Jaden Yifan He

3333 Forbes Ave, Pittsburgh, PA 15213

📞 814-880-8736   ✉ yifanhe@andrew.cmu.edu   in LinkedIn   ○ Github

Education

Carnegie Mellon University   *Aug 2021 – Dec 2023*

Master of Science - MS, Intelligent Information Systems; MS, Music & Technology | QPA: 4.07/4.0   *Pittsburgh, PA*

Coursework: Multi-modal Machine Learning, Natural Language Processing, Digital Signal Processing, Cloud Computing

University of North Carolina at Chapel Hill   *Aug 2019 – May 2021*

Bachelor of Science - BS, Information Science with Honors; Minor, Music | GPA: 3.90/4.0   *Chapel Hill, NC*

Coursework: Information Retrieval, Search Engines, Deep Learning, Object-Oriented/Web Development, Database Systems

Technical Skills

Languages: Python, Java, JavaScript/TypeScript, HTML/CSS, C/C++, Golang, MATLAB, SQL

Tools: Linux/Unix, Git, Docker, AWS, GCP, Kubernetes, Kubeflow, MySQL, Hadoop, Spark, Tableau, Gradle/Maven

Frameworks: PyTorch, TensorFlow, scikit-learn, pandas, Librosa, OpenCV, NLTK, Django, React, AntDesign, SpringBoot

Work Experience

Carnegie Mellon University   *May 2022 – Present*

Machine Learning Research Assistant   *Pittsburgh, PA*

- Proposed a novel **audio-language pre-training** architecture based on **Seq2Seq** structure, which can optimize multiple downstream tasks in a unified **multitask** paradigm, and achieved a few SOTA performances in the MIR domain.
- Proposed a novel **audio timbre transfer** method based on **autoencoder** structure, which realized the first method in **many-to-many** transfer results comparable to SOTA one-to-one GAN-based methods [ICASSP 2023].
- Proposed an automatic **noisy label detection** method for **automatic speaker verification** task; applied inconsistency ranking of label noise to detect and conducted a noise level estimator based on **beta mixture** model.

Tencent   *May 2019 – Nov 2020*

Machine Learning Engineer   *Shenzhen, China*

- Developed a standardized **label tagging** and **feature engineering pipeline** for ML datasets, applied **filter method** to select features based on **Pearson correlation**, and improved team's **audio fingerprinting and recognition algorithm** for song identification task for 7% in accuracy and 50% in time.
- Lead **sentiment analysis** for over 5000 user reviews about team's products on social media platforms and take **LDA** to perform topic modeling; reported to team manager weekly for **A/B testing**.
- Managed song copyright **data warehouse** and **ETL pipelines** utilizing **MySQL** and **PySpark**, and made over 3000 qualified files available on platform; planned and built team website using **Django** for visualizing research results.

Selected Projects

End-to-end NLP System for Scientific Entity Recognition [Github]   *Oct 2022 – Nov 2022*

- Built own dataset for model training/testing: scraped papers from conferences, parsed pdf into sentences and tokenized using **spaCy**, manually annotated with **Label Studio**, and annotated surrogate label with **knowledge distillation**.
- Trained a teacher system of pretrained **BERT** and **Linear-Chain CRF** on the dataset, and a small student system of **BiLSTM-CRF** on the augmented dataset, reached F1-score of over 93% and 62% respectively on private testset.

Mytorch Deep Leaning Library [Github]   *Jan 2022 – Dec 2022*

- Created a **PyTorch-like Deep Learning Library** from scratch that supports **MLP**, **CNN**, **RNN**, **GRU**, and **BERT** using only **Numpy**; implemented functions including activation functions (e.g., Sigmoid, ReLU), loss function (e.g., Binary CrossEntropy, MSE), backpropagation, autograd, SGD with momentum, Adam, batch norm, dropout, etc.
- Built and trained neural networks on **cloud services** (AWS/GCP) for Kaggle competitions tasks including phoneme state classification, face verification, and speech-to-text transcription on large datasets.

Audiovisual-based Multimedia Event Detection System [Github]   *Jan 2022 – May 2022*

- Created a large-scale scalable multimedia processing pipeline on **cloud services** (AWS/GCP), which extract MFCC (with **FFmpeg & Librosa**) and SIFT (with **OpenCV**) features from videos then applying **Bag-of-Words** representation through **K-means** clustering and classified by SVM and MLP; employed neural-based features including SoundNet, PANNs, PaSST for audio, and VGG, ResNet for vision to further extract high-level embeddings.
- Raised 14% accuracy on test set (reached 100% accuracy on validation set) by generating more effective embeddings by performing features fusion schemes including **early/late/double fusion**, progressed 72% speed by achieving multiple levels of abstract to execute parallel tasks in **Pyturbo**.

Collaborative Filtering Recommendation System for Music [Thesis]   *Apr 2020 – May 2021*

- Collected a new dataset from the streaming platform (NetEase) for music recommendation task utilizing **Selenium**.
- Conducted experiments comparing Memory-based (**k-nearest neighbors**) and Model-based (**Slope One, SVD, SVD++**) collaborative filtering algorithms on the dataset; exploited SOTA deep learning-based recommendation algorithms, including **LightGCN, DeepFM** to compare with non-DL methods.