Yiming Xu

A master student with AI and CS backgrounds, have a strong interest in data science and data engineering and looking to solve real-world business problems.
Datacamp certified Data Scientist and Udacity certified AWS ML Foundational Nanodegree

Amsterdam, the Netherlands | ✉ yiming.xu96@outlook.com | ☎ +31 (0)685488473 | Homepage | Github | LinkedIn

EDUCATION

Master of Science in Computer Science - Big Data Engineering *SEP. 2021 – **Present**.*

Vrije Universiteit Amsterdam & Universiteit van Amsterdam, Amsterdam, Netherlands.

Core Curriculum: Large Scale Data Engineering(8/10), High Performance Computing and Big Data(8.5/10), The Social Web(8/10), Web Data Processing systems(7.5/10), Data Mining Technology(8.5/10)

M1 - Data and Artificial Intelligence *Sep. 2020 – Jun. 2021*

Institut Polytechnique de Paris, Paris, France.

Core Curriculum: Navigation for autonomous systems, Learning for Robots, Machine and Deep Learning, Reinforcement Learning

Bachelor of Engineering in Computer Science *Sep. 2015 – Jun. 2019*

Henan University, Henan, China. **Core GPA: 89.21%**

Core Curriculum: Advanced Mathematics(94%), Probability & Mathematical Statistics(87%) , Mathematical Modeling(97%), Discrete Mathematics(93%), Basic Circuit and Electronics(88%), Operating System(86%), C++ Programming(92%)

- **Thesis:** *Cross-modal Information Retrieval Model Based on Hard Examples Fine-grained Label Learning* (*Outstanding Graduate Thesis Award*)

SKILLS & Awards

- Programming: Python, R, C++, SQL, JavaScript, HTML, D3.js (visualization), SQL
- Big Data & AI: Pyspark, Scala, Scikit-learn, NLTK, TensorFlow, Pytorch, Tensorflow, Keras
- Data Science: Data science pipeline (cleansing, wrangling,visualization, modeling, interpretation), Statistics, Time series
- Chinese Academy of Sciences College Students Innovative Practice Training Program Scholarship.
- Kaggle U.S. Patent Phrase to Phrase Matching competition Silver Award (public leaderboard, Top 0.02%).

PUBLICATIONS

Yiming Xu, Jing Yu, Yue Hu, Jingjing Guo, Jianlong Tan, "*Fine-Grained Label Learning via Siamese Network for Cross-modal Information Retrieval*," International Conference of Computational Science. Springer, Cham, 2019: 304-317.

WORK EXPERIENCE

Research Engineer Intern *Aug. 2018 – Feb. 2020*

Institute of Information Engineering, Chinese Academy of Sciences *Beijing, China*

- **Cross-modal Retrieval:** Trained a siamese network to achieve mutual retrieval between text and images, the retrieval accuracy of the model was higher than the best model results at the time and was published at the International Conference of Computational Science.
- **Video Events Search:** Designed a retrieval system for querying events that occur in the video.

Data Engineer Intern *Dec. 2017 – Jan. 2018*

Data Analysis Technology Lab, Henan University *Henan, China*

- **Automatic Check-in System based on Face Recognition:** Combined Tensorflow with ROS robot operating system to implement a lightweight face recognition-based check-in system.
- **Breast Tumor Classification:** Extracted features from CT images of breast tumors, use deep learning for classification and compare with classic machine learning methods.

PROJECT EXPERIENCE

Generate Images from Speech Descriptions [Code] *Jul. 2022 – Sep. 2022s*

GAN, ViT, ESResNeXt, PyTorch Lightning

- Build a Multimodal Attention model to train a image and speech embedding network for feature extraction. Then use a Densely-stacked Generator to generate images from speech features.
- Applying the model on the CUB, Oxford-102 and CelebAMask-HQ datasets produces images of higher quality than the results of the current state-of-the-art model.

U.S. Patent Phrase to Phrase Matching [Code] *May. 2022 – Jun. 2022s*

Bert, Pytorch, WandB

- Build a model ensemble Deberta, Roberta, and bert-for-patents is designed to predict the degree of similarity between two patent phrases in specific application scenarios.
- Position 47 out of 1,975 teams, Silver Award on the *public leaderboard*.

European Passenger and Cargo Aircraft Analysis [Display] *Sep. 2021 – Oct. 2021*

Pyspark, Scala, Scikit-learn, JavaScript, HTML, D3.js

- Use scala to build a pipeline on Databricks, extract the flight trajectories of passenger and cargo planes from 800G OpenSky data.
- Analyze and visualize the differences between passenger and cargo planes in speed, route, altitude, and time by using random forest.

Video Events Search System **[Code and Display]** *Oct. 2019 – Jan. 2020*

Tensorflow, Whoosh, HTML

- Sampled frames from each video, every few seconds and generates natural language captions for each frame using DenseCap.
- Indexed these captions as documents along with the corresponding video URL and timestamp.
- Retrieved the caption that best matches the user's search query, along with the video and the precise timestamp, within the video associated with the caption.