Ke Wu

12 Coledale Rd, Markham, Canada
kwu2@wpi.edu
karenwuke.github.io
774-303-6097

Engineer for **Data Science** and **Natural Language Processing**.

Education

Worcester Polytechnic Institute, Data Science — Massachusetts, USA
Master of Science in Data Science; GPS: 3.7/4.0 — *Aug 2017 - Dec 2019*

Relevant Coursework: Statistical Learning, Big Data Analysis, Machine Learning

Yunnan University, School of Management — Yunnan, China
Bachelor of Arts, Cultural Industry Management — *Aug 2011 - Aug 2015*

Experience

Population and Quantitative Health Sciences, UMass — Worcester, Massachusetts, USA
Research Engineer, Supervisor: Feifan Liu — *Dec 2019 - Now*

- **Readmission Prediction**: Write **automated scripts (shell/python/R)** to clean and maintain unorganized clinical notes from **multiple source**, adopted **sentence encoding** (clinical-BERT) on **SVM**, **Bi-LSTM**, and **BERT** to predict readmission rate for stroke patients.
- **Language Feature Engineering**: Introduce **attention model** on domain knowledge with a pretrained clinical Bi-LSTM model to enrich language features for small datasets.
- **Handling Imbalanced Data**: Rich experience on handling **extremely unbalanced dataset** with **hyper-parameter tuning**, **bootstrap batch**, **class weighting** and employed **focal tversky loss**.

Unison Global — Sterling, Virginia, USA
*Data Science Graduate Capstone, Supervisor: **Fatemeh Emdad*** — *Sep 2019 - Dec 2019*

- **Image Processing**: Use **OpenCV** to preprocess the ill-scanned documents, including **homography rectification**, **dilation/erosion**, **Hough-transformation**, line and **edge extraction**. This work expedited our **labelling workflow** by **10 times**.
- **Optical Character Recognition**: **Fine-tune** the **detection** module with **pytorch CTPN** and **recognition** module with **CRNN** on low-quality, ill-scanned documents, achieved **92.8**% **character recognition accuracy** on validation dataset.

Wayfair — Boston, Massachusetts, USA
*Data Engineer Intern, Mentor: **Sneha Jain*** — *July 2019 - Dec 2019*

- **Optimize regression and classification**: Deployed a **logistic regression model** to forecast advertisement click-through rate for web analysis team and employed **feature crossing** with **gradient boosting** decision tree to **speed up training** by **50**%.
- **Automated A/B test**: Revolutionized team A/B test pipelines by creating automated scripts, increasing **25**% team capacity.
- **Ad-hoc analysis**: Wrote a **real-time tracking** and **analysis framework** in python to process workstream log data from **multiple resources**, automatically generate **Tableau dashboard** for stakeholders to **monitor productivity** and **track workflow**.

Bank of China — Yunnan, China
*Business Data Analyst, Supervisor: **Kang Li*** — *Jan 2016 - Jul 2017*

- **Requirement Engineering**: Perform business and user **requirement analysis** for various banking systems and document them to requirement definition reports and act as an interface between the business users and the technical developers.
- **Data Visualization**: Collect, clean and visualize local customer seasonal consuming report with **matplotlib** and **Tableau**.
- **Anomaly and Fraud Detection**: Built fraud detection classifiers using gaussian **naive bayes** and **decision tress** to identify POIs (persons of interests) and applied machine learning techniques such as **features selection**, **precision and recall**, and **stochastic gradient descent** for model optimization in Python.

Projects

- **Ancient Chinese Translator**: Re-implement the **Transformer** model from "*Attention Is All You Need*". Self crawled and cleaned **ancient Chinese dataset**. Trained and deployed the **docker image** of translation model on **AWS SageMaker** for web-based access.
- **Deep Check-in**: Transfer pretrained **DeepFace** (github) with our colleague facial image data and automatic generate check-in csv report for our team manager.
- **Bilibili Bullet Generator**: Implemented a simple bullet screen comment **crawler** with Python and use the collected comments to train a neural network to **generate captions** for an image using **CNN** and **RNN** with **BEAM Search**.
- **Tiny Object Detection**: Self-collected and labelled datasets using OpenCV **object tracking** library. Applied **random warping**, **resizing**, **cropping** and **transformation** to **augment** the dataset. Fine-tuned a pretrained **faster RCNN** with **ResNet-18** backbone, employed **attention map** for **Region Proposal Network** to focus on tiny object. Achieved around **20**% accurarcy boost.
- **Predict Bitcoin Price from Twitter**: Crawled twitter data using Twitter API, collected and cleaned bitcoin related data. Applied **sentimental analysis** using **Textblob** and **NLTK** to predict the coin price trending from Twitter.

Skills

- **Programming Language**: **Python** (pandas, scikit-learn, **PyTorch**, **TensorFlow**), R, SQL, Java, Matlab
- **Data Management**: **MongoDB**, QCP, HP Vertica, **Hadoop**, **Spark**, Excel, Data Studio, **AWS**
- **Data Analysis**: A/B test, regular expression, **Machine learning**, **Deep Learning**, **Data Mining**, **NLP**, Computer Vision
- **Data Visualization**: Matplotlib, GGPlot, **D3.js**, **Tableau**, PowerBI