Xiang Liu

Email: xliu29@gmu.edu

Mobile: +86-15617865777

Education

- **The University of Hong Kong** — HK, CHINA
  Master of Science(MSc) - Computer Science — *Sep 2022 - Dec 2023*
- **George Mason University** — VA, USA
  Bachelor of Science - Computer Science; GPA: 3.71/4.0 — *Aug 2018 - Aug 2022*
  Honors/Awards: *Dean's List (2018-2020)*
  Courses: *Natural Language Processing, Visual Computing, Data Structure, Database Concepts*
- **Henan University** — Henan, CHINA
  Bachelor of Engineering - Networking Engineering; GPA: 88/100; Rank: 4/292 — *Sep 2017 - Jun 2022*

Experience

- **Baidu Research Cognitive Computing Lab** — Beijing,CHINA
  Intern (Full-time) — *Dec 2021 - June 2022*
  - **Research Papers**: Research OIE, Logical Reasoning papers and reproduce the code of the papers.
  - **Project Deployment**: Complete the deployment of OIA (*A Predicate-Function-Argument Annotation of Natural Language for Open-Domain Information eXpression*) project on Chinese corpus, and develop corresponding annotation rules, processing rules, conversion rules. The OIA on Chinese dataset can also play close to the effect of the English dataset.
  - **Most Representative Selection**: The problem of co-referencing based on WikiCoref data The model needs to select the most representative mention from the cluster which contain all mentions of a entity. Used graph-based topological sorting, confidence-based cumulative sorting, and LightGBM's list wise sorting.
  - **Personal gain**: Reading papers more efficiently, learn the paper method and think how to use the paper method to solve the current project. The ability to work on large projects more quickly, and to develop engineering skills. Enhanced the communication skills with colleagues and mentors.

Competitions

- **Kaggle — Feedback Prize - Predicting Effective Arguments Competition, Sliver Medal Top 2%** :
  Jun 2022 - Aug 2022
  Task: The goal of this competition is to classify argumentative elements in student writing as 'effective', 'adequate', or 'ineffective'.
  Main Solution: Using the token classification method instead of the baseline's sequence classification method. It can give the special token more context information, and provide position information on the relationship between the special tokens and the essay. In terms of efficiency, this method is around 3 times faster than sequence classification. Using focal loss to handle the label imbalance situation.
- **Kaggle — U.S. Patent Phrase to Phrase Matching, Bronze Medal Top 9%** :
  Mar 2022 - Jun 2022
  Task: The goal of this competition is to determine two key phrases are semantic similarity under patent documents.
  Main Solution: The mean pooling, multi dropout and layer normalization model structures are mainly used, and multiple pre-trained language models are used for ensemble. The Deberta-v3-large and electra-large-discriminator model have powerful performance.
- **CCF Collegiate Computer System & Programming Contest Top 5%**:
  Sep 2021 - Nov 2021
  Task: The goal of this competition is to predict default or non-default using user personal information and consumption information.
  Main Solution: Co-modeling using two internally discrepant data. Used lightGBM and XGBoost to build two single models, and eventually used model fusion stacking and weight fusion techniques to improve performance.

Projects

- **Rumor Prediction Model for Microblog Epidemic**:
  A Covid-19 rumor prediction model was built based on the pre-trained model BERT, which completes the deep learning model complete process.
  In this project, I went through the whole process of idea generation, data search, reference papers, code practice, model building and article writing, and gained a very deep understanding of the NLP project process.
- **Emotional analysis case**:
  Sentiment analysis is performed on the movie review text dataset, and the bag-of-words model, N-gram model, loss function, and gradient descent function are implemented manually.
  Through this project, I have gained a deeper understanding of the principles of deep learning, the real usage of gradient descent, and a deeper understanding of statistical-based language models.

Personal Summary

- **Languages**: Python, SQL, JAVA,C++
- **Frameworks**: PyTorch, Scikit, NLTK, SpaC