Yijun Yang

yyang5@ed.ac.uk https://thomasyyj.github.io/yangyijun/

Education

University of Edinburgh **UK**

MSc in Data Science *Sep. 2022 - Aug. 2023*

- **Relevant Courses:** Natural Language Understanding, Machine Learning Practical
- **GPA:** 4.0/4.0, Distinction
- **Research Interest:** Question Answering, Knowledge Extraction
- **Thesis:** *A Unified Framework on Unbiased and Robust Factual Knowledge Extraction*, advised by Prof. Jeff Z. Pan

University of Nottingham **UK**

BSc in Mathematics with Applied Mathematics *Sep. 2020 - Aug. 2022*

- **Relevant Courses:** Statistical Inference, Machine Learning, Optimization
- **GPA:** 4.0/4.0, First Class Honours

Publications

UniArk: A Holistic Approach to Unbiased and Consistent Factual Knowledge Extraction

Yijun Yang, Jie he, Pinzhen Chen, Gutiérrez Basulto V, Jeff Z. Pan

Submitted to The Conference of European Chapter of the Association for Computational Linguistics (EACL2024)

Exploring Effective and Efficient Question-Answer Representations

Zhanghao Hu, **Yijun Yang**, Junjie Xu, Yifu Qiu, Pinzhen Chen

Co-first author, Submitted to LREC-COLING 2024

Deep Interaction Enhanced Graph Reasoning Language Model

Yijun Yang, Junjie XU, Zhanghao HU

MLP Course Project Shortlisted as the 2023 IBM Prize at the University of Edinburgh (Top 5/88)

Research Experiences

Research Assistent

Institute for Language, Cognition and Computation, University of Edinburgh *Sep. 2023 - Dec. 2023*

- Developing systems for large language model based information Extraction.
- Exploring methods for precise scientific document informational retrieval and trustworthy ethical assessments.
- Constructing knowledge graph on scientific literature for epidemiology.

Research Assistant

School of Computer Science, University of Nottingham Ningbo China *Sep. 2020 - Dec. 2020*

- Designing scripts of the toolbox to detect and characterize hydropeaking of 32 reservoirs in Ningbo from 2017 to 2020.
- Using the K-means clustering method to cluster the data based on Euclidean distance and investigate common characters of reservoirs with hydropeaking. Writing a corresponding technical report.

Competition Experiences

Kaggle: H&M Personalized Fashion Recommendations (Rank: 45/3006, 2%) *Mar. 2022 - May. 2022*

- Background: A recommendation system competition targeting product recommendations based on data from previous transactions, and metadata including images and tabular data.
- Data processing: Augmenting data with sliding windows; Encoding image information by embeddings from pretrained ResNet. Encoding product names through Word2Vec embeddings.
- Feature engineering: We construct and combine 300+ features based on vector aspects such as similarity and business aspects such as popularity, repurchase rate, etc.
- Retrival Strategy: Items are retrieved based on hand-crafted rules, item collaborative filtering, and matrix factorization. Both Bayesian Personalized Ranking and Alternating Least Square are applied in the matrix factorization method. An accuracy-first strategy was chosen instead of recall-first to improve the ranking results.
- Ranking Strategy: Using down-sampling on a mixture of training data. Ensembling LightGBM and a two-layer neural network to make the final prediction.
- Our solutions are open sourced here

Kaggle: Google AI4Code (Rank: 25/1135, 3%) *May. 2022 - Aug. 2022*

- Background: NLP competition targeting understanding the relationship between code and comments in Python notebooks. Participants are challenged to reconstruct the order of markdown cells in a given notebook based on the order of the code cells, demonstrating comprehension of which natural language references which code.
- Text pre-processing: Removing stopwords and invalid words using NLTK. Using a special token to encode the title.
- Model Selection: Choosing codebert, deberta-base and deberta-large. Replacing the pooling head by max pooling and attention pooling to highlight important information within the sentence.
- Pre-training: Further pretraining on code cells for in-domain information adaption.
- Downstream tasks: We formulate our downstream tasks as pairwise and pointwise tasks respectively for code-reference pair prediction. A confidence rate was hand-crafted as a guidance for model ensembling.

Technical Skills

Languages: Chinese, English
Programming Languages: Python, R, LaTEX, Slurm, Shell
NLP Toolkits: Huggingface, NLTK, SpaCy, BeautifulSoup
Deep Learning Toolkits: PyTorch
Machine Toolkits: ONNX, Numpy, Pandas, Polars, Scikit-learn