XUANSHENG WU

Email: wuxsmail@163.com **Phone**: (706)4612812 **Github:** github.com/jacksonwuxs

EDUCATION

University of Georgia 08/2021 – Present

- *Doctor of Philosophy in* ***Computer Science***

Shanghai University of International Business and Economics 09/2016 – 06/2020

- *Bachelor of Science in* ***Applied Statistic***

HONORS & AWARDS

Earned **550 Stars** on Github for my open source data analysis framework: DaPy	Present
2021 Tencent Advertising Algorithm Competition - **Top 1.7%**	07/2021
Baidu Python Good Coder - **1/22**	03/2021
Shanghai Aijian Scholarship - **Top 1%**	01/2020
Research Pioneer Award of Shanghai University of International Business and Economics - **Top 1%**	08/2019
2nd Prize of China Programming Contest for College Students - **Top 4.5%**	08/2019
Third-class Scholarship of Shanghai University of International Business and Economics - **Top 15%**	04/2019
Honorable Winner of MCM/ICM - **Top 20%**	04/2018
Single-subject Scholarship of Shanghai University of International Business and Economics - **1/125**	11/2017
1st Asian University Archery Championship - **3rd** in Men's Group	04/2017

INTERNSHIP

Baidu - Department of Natural Language Process 09/2020 – 04/2021

- Put forth the Long Text Semantic Segmentation task to reduce the negative impact caused by wrong paragraphs on downstream tasks (e.g. Machine Reading Comprehension, Query Generation, and ES Selection); Introduced several methods for automatic data construction including "Subtitle Data Augment", "Mixing Same Page Paragraph" and "Mixing Same Topic Paragraph"; Redesigned the loss function of ERNIE; The Model has been applied as a service to support the Baidu Top-1 Search and the FAQ Mining System with **F1=72.31%** on general documents.
- Investigated the cause of low GPU utilization of the old FAQ Mining System; Rebuilt the system into three independent-process components, the Proxy, the Scheduler, and the Operator, and as a result, the new system accepts a large number of requests and schedule the GPU-based services dynamically according to the payload of different tasks; Came up with a low computation schedule algorithm to assign services for multiple GPU automatically; Improved utilization ratio of GPU by **2.9X** and QPS by **4.1X**.
- Worked with coworkers to improve experience of the Baidu Top-1 Search under metrics **Good:Same:Bad=28:6:1** by optimizing services of Truncation Detection, Enumerate Answers, and Multi-Resources Recall.
- Deployed a query filter service for queries that require videos as answers (VQA) based on GBDT with **F1=93.43%**.

Pingan OneConnect Co., Ltd - Institute of Big Data 01/2019 – 06/2019

- Proposed to use Feature Context-Free Grammar for Seq2SQL task, new solution supports 34 external patterns.
- Completed intention recognition task with **Kappa=77.21%** through TF-IDF + BoW + MLP in one day.
- Developed an automatic machine learning module including correlation, clustering and time series analysis.

Shanghai Ronghao Investment and Management Co., Ltd 01/2018 – 02/2018

- Designed and tested a short-term trading strategy by analyzing over 40 millions transaction records.

Researches

FastDCE: Non-Parametric Self-Attention for Dynamic Contextual Sentence Embedding 09/2021 – 11/2021

- Proposed a novel sentence encoder by integrating the non-parametric self-attention into the bag-of-words model.
- Conducted the conjunction matrix between each two positional word embedding to capture word-word relationships, then applied a non-linear kernel between them to construct feature embedding to describe this relationship.

- Evaluated FastDCE on eight different tasks and exceeded the bag-of-words-based baseline by **Acc=2.89%** on average.
- Observed from visualization that FastDCE could detect contextual topics, common phrases, and word causalities.

2021 Tencent Advertising Algorithm Competition: Multimodal Video Ads Tagging 05/2021 – 07/2021

- Investigated that the baseline InceptionV3+Vggish+BERT+Resnet50-->NextVLAD-->ContextGate-->MLP has two issues: overfitting and outdated feature extractor, used the following solutions to improve it from 76.10% to **GAP=82.10%**.
- Removed the center frame modal as a redundant modal; Fine-tuned dropout rate; Applied data augment strategies.
- Replaced the original InceptionV3 embedding with the embedding from the last two blocks of the EfficientNet to provide additional high and low level semantic information; Enhanced time information to the image flow by simply shifting parts of the embedding referring the idea of the Temporal Shift Model; Concatenated Word2Vec of ASR tokens with EfficientNet embedding to enhance relevant and suppress unnecessary information to the image flow.

2020 Tencent Advertising Algorithm Competition: Users Demographic Attributes Prediction 05/2020 – 06/2020

- Estimated embedding of advertisement resource (AdvEmb) using Word2Vec Algorithm on 1 million user clicking log.
- Predicted age and gender of users by deploying an AdvEmb+BiLSTM+Attention model with **Accuracy=72.43%**.

Learning Sentence Embedding as Humans Do: Syntactic Attention-based Random Walk Model 09/2021 – 11/2021

- Introduced the syntactic attention into the Random Walk Model to improve the level of modeling word semantic.
- Pretrained on 4.2 million corpus and improved **6.36%** Pearson correlation than the baseline on 5 different datasets.

Lifestyle-Based Cervical Cancer Screening 03/2018 –03/2019

- Applied Lasso and Z-Test for feature selection and solved non-random missing value issue with proxy variables.
- Trained a GBDT classifier with the Bayesian Optimization algorithm to reach **AUC-ROC=65.6%** and beat SOTA over 5%.

Optimization of Value Average Strategy in China Stock Market 03/2017 – 03/2018

- Proposed a new investment strategy and back-test it on over 0.5 millions records of China Stock data.

PROJECTS

DaPy: An open-source and easy-to-use data analysis framework for humans 09/2017 – Present

- Achieved comparable efficiency with Pandas by using MemoryView, Cache-Friendly Operations, and Binary Search Index.
- Implemented machine learning models (e.g., Decision Tree, Linear Regression, Language Model, Hidden Markov Model).

Distributed Archery Events Supporting System based on B/S+C/S Hybrid Architecture 12/2017 –12/2018

- Proposed “B/S + C/S Hybrid Layout Architecture” to support high efficiency and high stability concurrently.
- Designed a series of strategies for data transmission security using Caesar encryption algorithm and MD5 algorithm.
- Implement data synchronization across firewalls between system nodes in LAN environment with HTTP protocol.
- Used Sqlite3 and wxPython to implement local client and built the online web server with Flask.
- Served 2018 Chinese University Archery Championships (peak of the system was **1430 visits per second**)

Online Financial Info Query Robot based on RASA Framework 07/2018 – 08/2018

- Implemented intention recognition and named entity recognition with Rasa library.
- Designed a micro-services architecture to support multiple-round and multi-language conversation.
- Extended the above service to Wechat platform with Itchat library.

PUBLICATION

- ***FastDCE: Non-Parametric Self-Attention for Dynamic Contextual Sentence Embedding***, reviewing by 2022 Meeting of the Association for Computing Linguistics (ACL).
- ***Rethinking the Impacts of Overfitting and Feature Quality on Small-scale Video Classification***, published in 2021 ACM Multimedia Conference (ACMMM).
- ***Syntactic Attention-based Random Walk Model***, reported in B.S. Dissertation Defense.
- ***Lifestyle-based Approach for Cervical Cancer Screening***, reported in 2018 Int. Conference on Data Science (ICDATA).
- ***Optimization of Value Average Strategy in China Stock Market***, published in China Collective Economy, 2018, 69(4).

Skills & Hobbies

- Python (50,000+ lines code), C, Linux, SQL, SPSS, MATLAB.
- Archery (part-time coach 2+ years), Basketball (core member in school team), Skiing.