GE ZHANG

5 Yiheyuan Road, Haidian District, Beijing, China 100871
(+86)15957610055 ⋄ zhangge9194@pku.edu.com

EDUCATION

Peking University, Beijing, China *September 2016 - Present*
B.S., Department of Intelligence Science, School of Electronics Engineering and Computer Science
GPA: 3.2/4.0

PUBLICATIONS/MANUSCRIPTS

1. Anonymous Authors
   Mining Collective Data Science Knowledge from Code on the Web to Suggest Alternative Data Analysis Approaches
   In submission to WWW 2021
2. **Ge Zhang***, Mike Merrill*, Yang Liu, Jeffrey Heer, Tim Althoff (*: equal contribution)
   CORAL: COde RepresentAtion Learning with Weakly-Supervised Transformers for Analyzing Data Analysis
   arxiv 2020 [pdf][code]
3. Zijun Sun*, **Ge Zhang***, Junxu Lu, and Jiwei Li (*: equal contribution)
   LOP-OCR: A Language-Oriented Pipeline for Large-chunk Text OCR
   2019 [pdf]

EXPERIENCE

Social Futures Lab, University of Washington, Seattle August 2020-Present
Research Intern (remote)

- Supervisor: Prof. **Amy X. Zhang**, Assistant Professor, University of Washington, Seattle
- Designed a DAG structure computational notebook to support reproducibility and collaboration.

BData Lab, University of Washington, Seattle March 2020-Present
Research Intern (remote)

- Supervisor: Prof. **Tim Althoff**, Assistant Professor, University of Washington, Seattle
- Collaborators: Mike Merrill
- Progress towards a paper on Notebook Analysis, intended for publication at **WWW 2021**.
- Worked on multiverse analysis to support robust data science through auto decision making and alternative decisions suggestion.
- Proposed a span-alternative-prediction task for code representation learning.
- Pretrained a multitask variant of BART with a new token type prediction task to help model learn more about syntax.

BData Lab, University of Washington, Seattle June 2019-Present
Research Intern (visiting student)

- Supervisor: Dr. **Tim Althoff**, Assistant Professor, University of Washington, Seattle
- Collaborators: Mike Merrill, Yang Liu, Jeffrey Heer
- Created a dataset of decision points and alternatives by clustering based on similarity of functions(including signals of return values, arguments, cooccurrence and sommon subtokens).
- Proposed a novel weakly supervised transformer architecture for computing joint representations of data science code from both abstract syntax trees and natural language annotations.
- Presented a new classification task for labeling computational notebook cells as stages in the data analysis process (i.e., data import, wrangling, exploration, modeling, and evaluation).
- Annotated cells with one of the above stages for 100 data science Jupyter notebooks and produced a standardized rubric for qualitative coding.

Shannon.ai Nov 2018-May 2019

Full-time research intern

- Supervisor: Dr. **Jiwei Li**, Chief Executive Officer, Shannon.ai
- Collaborators: Junxu Lu, Zijun Sun
- Designed a model based on seq2seq with auxiliary image information to help scene text recognition tasks.
- Extended the model to a system to improve the performance of OCR, increasing the accuracy significantly from 77.9 to 88.9.
- (As a product of the company) Extracted structured data from prospectus(PDFs) and helped design a data structure to format long text used for information extraction. Reached 0.95 accuracy on the whole corpus.

Language Computing & Web Mining Group, Peking University March 2018-Jan 2019

Research intern

- Supervisor: Dr. **Xiaojun Wan**, Professor, Peking University
- Crawled data to construct a fiction-story summarization dataset.
- Used SVM with designed features(length, sentence representation) to summarize public online fiction-stories.

AWARDS

2nd Prize of ACM Competition, Peking University, **2017**

People's Choice Prize, Google Girls' Hackathon (as team leader), **2019**

SELECTED PROJECTS

Multiverse Analyses on Jupyter Notebook

With the neural generative model backing up, the demo allowed people to upload jupyter notebook, highlighted decision points and suggested alternatives. Allowed people to dynamically add comments and feedback.

Helped data scientists better understand data analyses and provided a platform to better define decision points in data science process.

Github: `https://github.com/behavioral-data/CORAL`

Splendor AI (Google Girls' Hackathon)

An AI bot of Splendor, which could beat most human players and won People's Choice Prize in the hackathon.This AI bot computes scores based on weighted custom features, which also had a try on reinforcement learning.

Github: `https://github.com/AshleyZG/splendor`

LOP-OCR: A Language-Oriented Pipeline for Large-chunk Text OCR

Together with Jiwei Li, Zijun Sun, Junxu Lu.

LOP-OCR, a language- oriented pipeline to help scene text recognition tasks based on seq2seq models with auxiliary image information.

Improve the performance of the CRNN-based OCR models, increasing sentence-level accuracy from 77.9 to 88.9 and position-level accuracy from 91.8 to 96.5.

SKILLS

Chinese(native), English (fluent)
Python, c/c++, Pytorch, Javascript(D3 library)
Familiar with Linux system.