Zhenghui Wang

251 10th Street, NW Apt. F318B – Atlanta, GA 30318

+1 (404) 863-3240 • zhwang@gatech.edu • zhenghuiwang.net
zhwang • *Seeking Full-time Software Engineer Position*

Education

Georgia Institute of Technology **Atlanta, GA**
Aug. 2019 - Now

- M.S. in Computer Science - GPA: 4.0/4.0

Shanghai Jiao Tong University **Shanghai, China**
Sep. 2015 - Jun. 2019

- B.Eng. (Hons) in Computer Science and Technology - GPA: 3.92/4.3
- Advisors: Prof. Weinan Zhang and Prof. Yong Yu

Internship

Pinterest **Atlanta, GA (remote)**
Pinterest Labs Research Intern - Mentor: Dr. Jinfeng Zhuang and Vijai Mohan *May. 2021 - Jul. 2021*

- Worked on automatically extract shopping metadata without human labeling.
- Built a workflow that uses the merchants provided feeds data for shopping products to build training data and train Wrapper Induction models automatically in hundreds of different domains at scale.

ByteDance AI Lab **Shanghai, China**
Machine Learning Engineer Intern - Mentor: Dr. Hao Zhou *Feb. 2019 - Jun. 2019*

- Constructed a huge Chinese reading comprehension dataset by cleaning and linking the data of Wikipedia and Wikidata. The new dataset contains more than 800K articles and 3M QA pairs.
- Implemented a novel QA4IE (Question Answering for Information Extraction) system with neural NER, QA and Entity Linking modules. Deployed the system with GPU using Flask and Bootstrap. The system achieved F1=43.64% on title triplet extraction and F1=22.73% on full triplet extraction.
- Work accepted as a demo paper at **SIGIR 2020**

Synyi AI **Shanghai, China**
Machine Learning Engineer Intern - Mentor: Dr. Ken Chen *Nov. 2017 - Jun. 2018*

- Proposed and implemented a label-aware double transfer learning framework for named entity recognition and applied in the medical domain with new state-of-the-art performance.
- Introduced a label-aware assumption which is critical in real-world named entity recognition systems. Proved the equivalence of the L2 distance in parameter space and the KL-divergence in model output distributions.
- Work accepted as a long oral paper at **NAACL HLT 2018**.

Research Experience

NLP research at SALT Lab, Georgia Institute of Technology **Atlanta, GA**
Advisor: Prof. Diyi Yang, GaTech *Jan. 2020 - Dec. 2020*

- **Project 1**: Proposed and implemented LADA, a local additivity based data augmentation for semi-supervised NER, which performed interpolations in hidden space among closed examples to generate augmented data and improved NER performances with limited training data. (work accepted at **EMNLP 2020**)
- **Project 2**: Worked on personalized response generation. Proposed a new evaluation metric and a model which incorporated tensor factorization for better personalized response generation. (work accepted at **GEM 2020**)
- **Project 3**: Focused on neural networks' interpretability. Proposed a subspace probing method to better understand pre-trained language models like BERT. (work in submission)

ML research at Ma Lab, Carnegie Mellon University **Pittsburgh, PA**
Advisor: Prof. Jian Ma, CMU *Jul. 2018 - Jan. 2019*

- Worked as a research assistant focusing on interpretable prediction of the gene sequence. Incorporated advanced NLP techniques (e.g., deep contextualized word representations) into the modeling of the gene sequence. Improved *learning to explain* framework with attention mechanism to do interpretable prediction of the gene sequence.

ML research at APEX Lab, Shanghai Jiao Tong University **Shanghai, China**
Advisors: Prof. Weinan Zhang and Prof. Yong Yu, SJTU *Dec. 2016 - Jun. 2019*

- **Project 1**: Worked on multi-task learning. Discovered and analyzed the *negative task gain* problem of gradient boosting decision tree (GBDT) in a multi-task learning scenario. Proposed a multi-task GBDT model and improved the performance of diabetes prediction in diabetes data collected from 21 different centers in China. (work accepted at **KDD 2021**)
- **Project 2**: Focused on graph embedding. Proposed a new explanation of Random Walk from the perspective of neighborhood joint probability. Proposed an efficient sampling policy which reduces more than 99.9% training pairs compared with Random Walk. Proposed an inductive graph embedding model to make full use of textual information on graphs. (work accepted at **WWW 2019**)
- **Project 3**: Surveyed the automated ICD coding problem. Cleaned massive raw medical records data from hospital. Implemented a state-of-the-art multi-label document classification model, hierarchical attention networks, for automated ICD coding.

Publications and Preprints

Not All Dimensions Are Created Equal: Subspace Probing for Pre-trained Language Models
- **Z. Wang**, L. Luo, and D. Yang
- In submission

Personalized Response Generation with Tensor Factorization
- **Z. Wang**, L. Luo, and D. Yang
- In *Proceedings of the First Workshop on Generation Evaluation and Metrics at ACL 2021*. **GEM 2021**.

Task-wise Split Gradient Boosting Trees for Multi-center Diabetes Prediction
- **Z. Wang***, M. Chen* et al.
- In *Proceedings of* **KDD 2021**.

Local Additivity Based Data Augmentation for Semi-supervised NER
- J. Chen*, **Z. Wang***, R. Tian, Z. Yang, D. Yang
- In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. **EMNLP 2020**.

QuAChIE: Question Answering based Chinese Information Extraction System
- D. Ru, **Z. Wang**, L. Qiu, H. Zhou, L. Li, W. Zhang, and Y. Yu
- In *Proceedings of* ***SIGIR 2020*** *Demo track*.

Sampled in Pairs and Driven by Text: A New Graph Embedding Framework
- L. Chen, Y. Qu, **Z. Wang**, L. Qiu, W. Zhang, K. Chen, S. Zhang, and Y. Yu
- In *Proceedings of the 30th Web Conference*. **WWW 2019**.

Label-aware Double Transfer Learning for Cross-specialty Medical Named Entity Recognition
- **Z. Wang**, Y. Qu, L. Chen, J. Shen, W. Zhang, S. Zhang, Y. Gao, G. Gu, K. Chen, and Y. Yu
- In *Proceedings of* **NAACL HLT 2018** (oral, 6.73%).

Honors and Awards

- China National Scholarship, Lenovo Scholarship, Wish Inc Scholarship, Yitu Tech Scholarship, Zhiyuan Scholarship, Rong Chang Innovation Scholarship, EIC Education Scholarship *2016-2019*
- Finalist Winner Prize, Mathematical Contest in Modeling (**F** Prize, < 1% worldwide) *2017*

Skills

Python, C++, MATLAB, Java, R, Linux, MySQL, Git, LaTeX, Jupyter, PyTorch, TensorFlow, Hadoop, Flask, Bootstrap, Airflow.