Statement of Purpose

Ph.D. in Computer Science

Tianxin Wei (rouseau@mail.ustc.edu.cn)

My research interest lies in data mining and natural language processing (NLP). As an undergraduate, I was lucky to lead several research projects at the University of Science and Technology of China (USTC), University of California Los Angeles (UCLA), University of Texas at Austin (UT-Austin), and Amazon, resulting in one first-author publication at the top-tier data mining conference and four first-author submissions, one of which has been presented at a top-tier workshop. To further pursue my research interest, I am determined to apply for the Ph.D. program in Computer Science.

Modern deep learning systems have achieved great success. However, their performance largely hinges on large amounts of high-quality, human-labeled data. The availability of such high-quality training data can often be limited in novel real-world applications and can easily become the bottleneck of a deep learning system. I found three particularly important and intriguing questions when resolving such a bottleneck: 1) When annotated data is difficult to obtain, how to design and train an effective model from unlabeled data? 2) Existing deep learning models are mainly likelihood-based, leading to spurious correlations and bias, which hurt the generalization ability. How to eliminate such bias in training data to train robust and fair models? 3) In practical applications, how to learn from heterogeneous types of information? My undergraduate research is carried out from these perspectives and mainly focuses on exploring solutions to low-resource conditions and data bottlenecks.

My first venture was about tackling the cold start problem of collaborative filtering (CF) recommender models, supervised by **Professors Wei Wang** and **Yizhou Sun** at UCLA. Current CF models suffer from three critical problems for new (cold) users: 1) The gap between the number of user interactions of training users and testing cold-start users; 2) Sparse user interactions of fresh users; 3) Training instability of the model. To address these issues, I proposed a meta-learning paradigm, named MetaCF, in which I treated the adaption on each new user as a task and learned the CF model with similar tasks on existing users to obtain the ability to adapt in the future. Firstly, I constructed representative training tasks to avoid over-fitting, in which I dynamically sampled subgraph centered at each user in the training phase to account for the effect of limited interactions of fresh users. Secondly, I expanded the historical interactions, by incorporating potential Interactions to mitigate the data sparsity problem. Thirdly, I optimized the fine-grained learning-rate to perform fine-tuning according to the performance of cold-start user adaption. Our method achieved up to 38.23% gain in hit ratio over the state-of-the-art baselines on three real-world datasets. Our work was published in **ICDM 2020**, in which I was the **first author**. I am also now doing a new project with the same professors about automated meta-path discovery on heterogeneous graphs via reinforcement learning (RL) to ease the difficulty of manual selection. I combined instance and schema knowledge graphs to design the reward for better performance.

Another long-standing data bottleneck of recommender systems is the popularity bias of user behavior data – popular items are recommended much more frequently than niche ones while ignoring the matching between user and item preferences. In collaboration with **Professor Xiangnan He** of USTC and **Dr. Jinfeng Yi** of JD AI Research, I proposed a counterfactual reason method to eliminate popularity bias in recommender systems. I confronted several critical challenges in this work: the representation, capture and removal of popularity bias. Firstly, I presented a causal view of the popularity bias in recommender systems and formulated a causal graph for the recommen-

dation. Secondly, I proposed a model-agnostic counterfactual reasoning (MACR) framework that trained the recommender model according to the causal graph to capture bias. Thirdly, I performed counterfactual inference to eliminate popularity bias in the inference stage of recommendation. We achieved an average 197.56% improvement over two representative models, MF and LightGCN, on five large-scale datasets. This work was submitted to the **SIGIR 2021** Conference, of which I was the **first author**. Meanwhile, I began to investigate how to take advantage of temporal popularity drift in online recommendation systems to better correct for popularity biases.

Besides recommender systems, I have also made contributions to resolving data bottleneck challenges in various other domains by learning from unlabeled data. In the financial domain, I built a novel graph neural network to accurately predict stock trends by leveraging rich information in the stock knowledge graph with **Professor Zhangyang Wang** of UT-Austin. In particular, I designed a geometric augmentation approach to discover long-range hidden dependencies between stocks. Also, I leveraged self-supervised learning to facilitate GCN training and to enforce global and local graph structure awareness. Our method achieved the improvements of 30.48% and 65.77% on NYSE and NASDAQ datasets over state-of-the-art models. The preliminary work has been accepted by **AAAI 2021 Workshop** on Knowledge Discovery from Unstructured Data as oral presentation, of which I was the **first author**.

For the problem of speaker identification, we proposed a self-supervised task with a reconstruction objective on the frame and spectrogram information in the audio. Together with adversarial training techniques, the self-supervised task enabled the model to improve itself on unlabeled data for speaker identification. Our approach achieved a 11% improvement over the baseline models. I submitted this project to **NAACL 2021** as the **first author** with **Dr. Ruirui Li** and **Dr. Oguz Elibol** from Amazon.com.

After learning from unlabeled data through self-supervision, I was excited to explore building neural machine translation (NMT) systems in a completely unsupervised manner. To tackle this problem, multimodal content was introduced to help build an NMT system without parallel corpora. In this project, I proposed a RL method to build an NMT system by introducing a sequence-level supervision signal as the reward. Based on the fact that visual information can be a universal representation to ground different languages, I designed two different rewards to guide the learning process: (1) the likelihood of a generated sentence given source image and (2) the distance of attention weights generated by image caption models. Our model Improved 1.0 – 3.0 BLEU over baselines on the Multi30K, IAPR-TC12, and IKEA datasets. The paper has been accepted to **DASFAA 2021** as a full research paper with **Professors Qi Liu** and **Enhong Chen**, of which I was the **co-first author**.

After more than one year's immersion in data mining and natural language processing, I am determined to pursue an academic career and focus on the essence and application in this field. I am convinced that with the help of the extraordinary minds and the top research environment, I can definitely continue to contribute to the community.