Statement of Purpose

Yifan Hou (yfhou@cse.cuhk.edu.hk)

I want to pursue a PhD in artificial intelligence, and plan to continue my career in research after obtaining my PhD. My main research interests include machine learning, data mining and theory. I am currently working with Professors James Cheng and Ming-Chang Yang at The Chinese University of Hong Kong (CUHK) as an MPhil, with a focus on graph neural networks. Previously, I worked with Professors Wei Wang and Pan Zhou on reinforcement learning at Huazhong University of Science and Technology (HUST). I have published my research in conferences three times, twice as first author [1, 2, 3], and presently one conference paper is under review.

My interest in AI research comes from the experience working with Professors Wei Wang and Pan Zhou on the course recommendation when I was undergraduate at HUST. My task was to design a personalized recommendation algorithm for the massive open online course platform to offer courses for users. I learned about reinforcement learning, specifically the multi-armed bandits method, for recommendation. One key challenge was that the algorithm needed to recommend prerequisite before high-level courses for primary user. To address this problem, I constructed a graph, with course connected by prerequisite relationships. Simply, high-level courses would not be recommended for primary users until finishing related prerequisites. I published the algorithm as a first-author paper at IEEE International Conference on Computer Communications Workshops: Knowledge Centric Networking [1]. The complex prerequisite relationships among courses inspired me that data are not independent with each other.

To further understand the relationship, I took the network analysis course taught by Professor Jana Diesner in University of Illinois Urbana-Champaign (UIUC) during my summer exchange. This experience first introduced me to the intriguing graph world. In that course I learned about indicators of centrality such as betweenness and their meanings in social networks. The course project was to visualize and analyze several real-world social networks in UCI Network Data Repository by Gephi. The analysis method was fairly limited since it was mainly completed by human based on experience and hand-crafted indicators, which spurred me that I could utilize machine learning algorithms to extract features of the graphs (networks) for following analysis, or even tasks.

After graduating from HUST, fortunately, I began to pursue MPhil on graph data mining supervised by Professors James Cheng and Ming-Chang Yang at CUHK. I was firstly involved in a project about graph query system. My task was to clean data and evaluate several popular graph databases such as OrientDB. I crawled three large-scale graphs derived from Wiki, Twitter and Amazon (up to 500GB) and transformed them into property graph format under distributed implement. I also tested the latency and throughput of these graph databases on those graphs. The work was published in International Conference on Management of Data [2]. A general picture of graphs including graph mining, graph analytics and graph database was built in my mind.

With the experience analyzing large-scale graphs, I was encouraged to explore graph-based machine learning. I started from graph embedding, also called graph representation learning. The purpose was to learn a low-dimensional vector for each node, which could be used in following tasks such as node classification. During my survey on related fields, I found that existing works e.g., DeepWalk-based methods, graph neural networks-based methods, only considered graph structure and node attributes, without differentiating the importance of neighborhood and considering edge types. Thus I designed a strategy to measure the importance of neighbors and introduced several independent weight matrices to handle edges with properties. The improvement of performance on node classification was significant. I published the research as a first-author paper at International Conference on Knowledge Discovery and Data Mining [3]. After finishing this project, my focus did not stay on the application-level algorithms. My curiosity about the mechanism behind machine learning algorithms on graphs had been growing.

To figure out the reason why graph neural networks achieved significant performance in representation learning, I worked with Professor Richard T. B. Ma in National University of Singapore (NUS) as a summer intern. In that summer, I tried to analyzed graph neural networks from the perspective of information theory. Specifically, I generalized existing representative graph neural networks and its variants into one framework, which was composed of two parts: aggregation and combination. Two metrics to measure the quantity and quality of information gain in that framework were then defined. Based on the analysis conclusions, a new model that could utilize these two metrics to improve following task performance was proposed. The work was submitted to International Conference on Learning Representations (open review and double blind) which is currently under review. After exploring graph data and algorithms, it was the first time that I touched theory in my academic career.

Though I am open to a wide variety of research within AI, my experience in relational data and graph representation learning has inspired an interest in graph-based machine learning. More concretely, I am interested in exploring deep learning methods on graph data and understanding the mechanism hidden behind them. I want to continue my study in ETH Zurich, because of the strong AI group within the Computer Science Department. I find some professors whose projects are especially attractive and relevant to my research: Professors Andreas Krause, Martin Vechev, David Steurer, Rasmus Kyng, Valentina Boeva and Julia Vogt. After reading several papers of these groups, I am confident that it is a great place for me to pursue a PhD.

References

[1] **Yifan Hou**, Pan Zhou, Jie Xu, and Dapeng Oliver Wu. Course recommendation of MOOC with big data support: A contextual online learning approach. In *IEEE INFOCOM Workshops*, pages 106–111, 2018.

[2] Hongzhi Chen, Xiaoxi Wang, Chenghuan Huang, Juncheng Fang, **Yifan Hou**, Changji Li, and James Cheng. Large scale graph mining with g-miner. In *SIGMOD*, pages 1881–1884, 2019.

[3] **Yifan Hou**, Hongzhi Chen, Changji Li, James Cheng, and Ming-Chang Yang. A representation learning framework for property graphs. In *KDD*, pages 65–73, 2019.