

Haofei Yu

haofei@cs.cmu.edu | +1 (412)-537-0110 | www.haofeiyu.me

EDUCATION

Carnegie Mellon University - School of Computer Science
Master of Science in Intelligence Information Systems

Pittsburgh, PA
May. 2024

Zhejiang University

Bachelor of Engineering in Computer Science and Technology

Hangzhou, China
June 2022

- GPA: **3.96**/4.00 | Rank: 7/134
- Awards: Provincial Scholarship (**top 5%**)
- Selected Coursework: Discrete Mathematics and Application(**98**), Ordinary Differential Equations(**95**), Introduction to Applied Operations Research(**94**), Computer Network(**96**), Object-Oriented Programming(**96**)

RESEARCH EXPERIENCE

Westlake University

Supervisor: Prof. Yue Zhang

Hangzhou, China
Sep. 2021 – Feb. 2022

An Encoder as Partial Re-ranker for Open-Domain Question Answering

- Considered the encoder part in the Fusion-in-Decoder architecture to be partially responsible for re-ranking and enhanced its re-ranking ability by modifying the training objective.
- Conceptually modeled a new latent variable to describe the correlation between retrieved candidate passages and answers and used labeled golden passages as supervision signals for training.
- Achieved more than 1.5 points EM improvements on both NaturalQuestions and TriviaQA datasets.

Georgia Institute of Technology

Supervisor: Asst. Prof. Diyi Yang

Atlanta, GA(online)
Jul. 2021 – Sep. 2021

Empirical Study on Personalized Dialog Generation

- Reviewed personalized response generation papers from speaker model (using persona embedding) to P-square Bot (using RL methods to model mutual persona perception).
- Utilized tensor factorization to explicitly learn the latent representation of users' embedding in order to tackle the data sparsity problem of explicit and high-quality persona or demographic information.
- Proposed to add user-user-generation signals to the existing Transformer-based Dialog Generation framework to achieve mutual persona perception in the dialog.

Westlake University

Supervisor: Asst. Prof. Zhenzhong Lan

Hangzhou, China
Mar. 2021 – Jul. 2021

Uni-Encoder: A Fast and Accurate Response Selection Paradigm for Generation-Based Dialogue Systems

- Developed a new paradigm called Uni-Encoder, that keeps the full attention over each pair as in Cross-Encoder while only encoding the context once, as in Poly-Encoder.
- Designed in-batch negatives mechanism that makes other responses in one batch as negative samples for one history-response pair.
- Reached state-of-the-art performance on 4 benchmark datasets (Ubuntu v1, Ubuntu v2, ConvAI, and Douban).

PROFESSIONAL EXPERIENCE

Tencent AI Lab

Research Intern

Shenzhen, China
Feb. 2022 – Jul.2022

- Responsible for a research project related to Long-range Language Modeling.
- Proposed a unified framework for Long-range Language Modeling which considers Transformer-XL, Routing Transformer, ExpireSpan, and other long-range transformer variants to be special cases.
- Made modifications based on K-means clustering and improved Transformer-XL to become a general form.
- Achieved 0.3 Perplexity drop on Wikitext-103 dataset with a constant extra time cost.

SKILLS

- **Programming Languages:** Python, C/C++, Verilog, SQL.
- **Technologies /Frameworks:** Linux, Git, PyTorch, Transformers, React.