Qian Chen

qianchen901005@gmail.com | github.com/qq31415926

Education

East China Normal University
Master of Computer Science and Technology *Sep. 2022 – Present*

Hefei University of Technology
Bachelor of Computer Science and Technology *Aug. 2018 – June 2022*

Research Experience

My research interests include Natural Language Processing, Interpretability and Large Language Models.

CIDR: A Cooperative Integrated Dynamic Refining Method for Minimal Feature Removal Problem

AAAI2024 First Author

- Minimal feature removal problem aims to find the minimum feature set. Previous works rely on monotonic assumptions, which cannot be satisfied in general scenarios.
- Prove that using Integrated Gradients we can transform the original problem into a knapsack problem and propose a plug-and-play method for generating minimum feature candidate sets.
- Extensive evaluation on Eraser benchmarks demonstrate the effectiveness of our method.

PE: A Poincare Explanation Method for Text Hierarchy Generation

EMNLP2024 under review First Author

- To model non-contagious feature interactions, prior studies build the hierarchical attribution tree by enumerating all combinations, neglecting underlying syntax and semantics.
- Propose a hyperbolic probing method and introduce a fast algorithm for generating hierarchical attribution trees.
- Experimental results show the effectiveness of our approach in building high-quality hierarchical explanations.

A Concept Decomposition Perspective for Interpreting Continuous Prompts

EMNLP2024 under review First Author

- Word-based explanation methods lacks comprehensive semantic understanding and cannot generalize to BPE encoding based models.
- Propose a framework to decompose continuous prompts to human-readable concepts.
- Experimental results show our framework can achieve the same results as the original P-tuning and word-based methods with a few concepts while providing more plausible results.

Internship

Xiaohong shu inc. Sep. 2023 – Dec. 2023

Awards

The First Prize Scholarship (**top 10%**)	2021.11
The Second Prize Scholarship	2020.11

Skills

Languages: Mandarin (Native), English (CET6:540)

Programming Languages: Python, LaTeX

Tools: Git/Github, Linux shell, VS Code, PyCharm, Markdown

Relevant Course: Natural Language Understanding(93/100), Machine Learning Basic(93/100), Linear Algebra(99/100), Probability Theory and Mathematical Statistics(98/100), Advanced Mathematics A(93/100), Advanced Mathematics B(95/100)