2024 年度年次報告書
次世代 AI を築く数理・情報科学の革新
2023 年度採択研究代表者

ファン インゾウ

京都大学 大学院情報学研究科
特定助教

Incorporating Meta-information in Machine Unlearning for Large Language Models.（メタ情報による大規模言語モデルの機械アンラーニング）

研究成果の概要

We explore the task of machine unlearning in the context of large language models (LLMs) in this research project. In this fiscal year, we specifically focused on the following parts of the research project:

(1) Memorization mechanism: Continuing the work from last year, we utilized representation engineering approach to analyze how specific knowledge/concepts are stored in LLMs. Specifically, we analyzed hidden activations evoked by some concept-related stimuli and extracted the common patterns. Further, we use Sparse Autoencoders to identify the concept-relevant layers. Our experimental results showed that we can locate concept-relevant layers and further manipulate model output and behaviors. This provides a foundation for identifying and targeting memorized content.

(2) Machine unlearning evaluation: We proposed an evaluation framework that verifies whether a concept has been effectively unlearned by observing its absence in the activation space. Specifically, we showed that if cosine similarity between hidden states and a control vector stays below a threshold across inputs, the model no longer expresses the concept. Based on this finding, we track cosine similarity between activations and hidden state vectors that represent some specific concept and define unlearning success as keeping this similarity below a threshold across all input ranges.

Next, we will further explore the method based on the above findings, combining representation engineering with certified perturbation analysis to develop verifiable machine unlearning techniques.

【代表的な原著論文情報】

1) Yin Jou Huang, Prakhar Saxena, Zi Cheng Zhao. Shaping Personality of Large Language Models: An Approach Based on Representation Engineering. 言語処理学会 第 31 回年次大会 発表論文集, 943-947 (2025)

2) Yin Jou Huang, Rafik Hadfi. How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models. Findings of the Association for Computational Linguistics: EMNLP 2024, 10336-10351. (2024)