2023 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 machine unlearning in the context of large language models (LLMs) in this research project, which aims to erase a specific piece of knowledge or concept from a trained LLM. In this fiscal year, we specifically focused the following the following parts of the research project: (1) Machine unlearning evaluation, and (2) Memorization mechanism.

For machine unlearning evaluation, we explored a cryptology-inspired machine unlearning evaluation method, which applies the concept of zero-knowledge protocol (ZKP) to design a repeated trial for assessing the knowledge level of a language model. Experiments based on password-guessing scenario and price negotiation scenario is conducted to showcase the idea. We observe the difference in performance (password-guessing accuracy or utility score of negotiation) among language models with different knowledge levels. This illustrates the possibilities of using this kind of trial to assess the knowledge level of a language model.

Further, we explored the memorization mechanism of LLMs. Among the bottom-up and top-down approaches for model interpretability, we choose the top-down representation engineering approach. Inspired by the emergent ability of LLM of scaling across multiple languages, we obtain the model control vectors that is derived from a specific concept (target of unlearning). This control vector could be used to control LLM behavior to align with the given concept. As a next step, we plan to explore LLM unlearning methods based on Fisher information between the model parameter and the control vector.