Natural Language Processing

Midterm Proposal

Date: May 15, 2024

Jiaao Wu
Fudan University

Yidong Wu
Fudan University

Lingjie Chen
Fudan University

Directions

This research explores **Machine Unlearning**, aiming to erase copyrighted information and reduce harmful speech in LLMs. Various approaches will be tested to determine which method achieves the best results.
Large Language Models have revolutionized natural language processing but face challenges in data privacy and content moderation. Without the ability to forget certain content, LLMs can encounter copyright issues due to their auto-regressive nature. Additionally, if harmful or contaminated data is found in a dataset, retraining the model to remove this data is costly and can provide a method for adversarial attacks.

Data

We use Harry Potter original series as the main data which is renowned and integrated into most LLM's training data.

Methods

Our study attempts to try different methods mentioned in past research on copy-right content unlearning problems.

- **Task Vector:** This approach allows for a targeted and efficient unlearning process, focusing on the parameters linked to specific data points that need to be forgotten.
- **Gradient Ascent:** Adjusts the model's parameters to maximize the loss for the data points we want the model to forget, effectively increasing the model's error on these specific points to ensure they are unlearned.
- **Reinforcement Learning - Negative Sampling:** Trains the model to avoid generating harmful or undesirable outputs by providing negative samples during training, minimizing the likelihood of producing such outputs and reinforcing the unlearning process.

Evaluation

Empirical testing will assess the effectiveness of our unlearning techniques by comparing post-unlearning model performance to pre-unlearning benchmarks. We expect to see significant reductions in copyright infringement without substantially affecting the LLMs' general utility.

Challenges

Technical Challenges

- **Engineering:** Setting up a remote server to support our experiments could be challenging, especially since we will be using the school's server, which lacks the pre-configured environment found in commercial servers. Deploying and fine-tuning the LLM will require careful setup and troubleshooting.

Conceptual Challenges

- **Research Landscape:** As LLM unlearning is a nascent field, there are many scenario-specific techniques and a lack of unified verification methods. We need to identify the most suitable approach for our goal of copyright unlearning.
- **Method Comparison:** The comparison dimensions for different unlearning methods are vast. Analyzing the final results will require thorough reasoning and justification to ensure accurate conclusions.

The uniqueness of this work lies in its exploration of a rapidly developing field with many unresolved questions. Through this project, we aim to gain a deeper understanding of what machine unlearning can achieve in LLMs and its broader implications. This project will serve as a pioneering effort, guiding us into this fascinating and valuable field. And I believe we can gain valuable on-hand experience from this project.

References

[1] A. Blanco-Justicia, N. Jebreel, B. Manzanares, D. Sánchez, J. Domingo-Ferrer, G. Collell, and K. E. Tan. Digital forgetting in large language models: A survey of unlearning methods, 2024.

[2] J. Davis. latex-homework-template. https://github.com/jdavis/latex-homework-template, 2020.

[3] R. Eldan and M. Russinovich. Who's harry potter? approximate unlearning in llms, 2023.

[4] H. Hu, S. Wang, T. Dong, and M. Xue. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning, 2024.

[5] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic, 2023.

[6] Z. Liu, G. Dou, Z. Tan, Y. Tian, and M. Jiang. Towards safer large language models through machine unlearning, 2024.

[7] N. Si, H. Zhang, H. Chang, W. Zhang, D. Qu, and W. Zhang. Knowledge unlearning for llms: Tasks, methods, and challenges, 2023.

[8] D. Trippa, C. Campagnano, M. S. Bucarelli, G. Tolomei, and F. Silvestri. $\nabla\tau$: Gradient-based and task-agnostic machine unlearning, 2024.

[9] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. Machine unlearning: A survey, 2023.

[10] Y. Xu. Machine unlearning for traditional models and large language models: A short survey, 2024.

[11] Y. Yao, X. Xu, and Y. Liu. Large language model unlearning, 2024.