

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

CS2205 - PHƯƠNG PHÁP LUẬN NGHIÊN CỨU KHOA HỌC

Lớp học

CS2205.CH181

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

04/01/2024 – 07/03/2024

----- *Trang này cố tình để trống* -----

THÔNG TIN CHUNG CỦA BÁO CÁO


- Link YouTube video của báo cáo (tối đa 5 phút):

<https://www.youtube.com/playlist?list=PLUapSYP99LaOdPVNtdE4snwnNAwuNX86n>

- Link slides (dạng .pdf đặt trên Github):

<https://github.com/hannh18/PPLNCKH.CS2205.CH181/blob/f4fa6b1c494e0a35822b54f0c79a97ea988af8eb/CS2205.CH181.DeCuong.FinalReport.Slide.AIO.230202025.pdf>

- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

<ul style="list-style-type: none">● Họ và Tên: Nguyễn Hữu Hân● MSSV: 230202025 	<ul style="list-style-type: none">● Lớp: CS2205.CH181● Tự đánh giá (điểm tổng kết môn): 8/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân:● Link Github: https://github.com/hannh18/PPLNCKH.CS2205.CH181
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI: SỬ DỤNG KỸ THUẬT KNOWLEDGE UNLEARNING ĐỂ GIẢM THIỂU RỦI RO VI PHẠM QUYỀN RIÊNG TƯ TRONG MÔ HÌNH NGÔN NGỮ

TÊN ĐỀ TÀI TIẾNG ANH: KNOWLEDGE UNLEARNING FOR MITIGATING PRIVACY RISKS IN LANGUAGE MODELS

TÓM TẮT

Các mô hình ngôn ngữ được huấn luyện trước – Pretrained Language Model (PLM) ghi nhớ một lượng lớn kiến thức trong quá trình huấn luyện ban đầu, bao gồm cả thông tin có thể vi phạm quyền riêng tư của đời sống cá nhân và danh tính. Các nghiên cứu trước đây giải quyết các vấn đề riêng tư cho PLM chủ yếu tập trung vào tiền xử lý dữ liệu và các phương pháp bảo vệ riêng tư bằng cách làm nhiễu dữ liệu, cả hai đều yêu cầu huấn luyện lại mô hình ngôn ngữ. Chúng tôi thực nghiệm phương pháp Knowledge Unlearning như một phương pháp thay thế để giảm thiểu rủi ro riêng tư cho ngôn ngữ sau khi huấn luyện. Sau khi thực nghiệm cho thấy rằng việc thực hiện gradient ascent (điều chỉnh tham số theo hướng ngược lại với gradient thông thường) trên các chuỗi token mục tiêu có hiệu quả trong việc làm cho PLM quên chúng mà hầu như không làm giảm hiệu suất của mô hình ngôn ngữ tổng quát đối với các PLM có kích thước lớn hơn.

GIỚI THIỆU (Tối đa 1 trang A4)

Sự phát triển nhanh chóng của mô hình máy học và kéo theo sự phát triển của việc xây dựng các tập dữ liệu huấn luyện, nguồn gốc của dữ liệu này rất đa dạng từ sách, báo, các cuộc hội thoại,... bao gồm dữ liệu riêng tư và dữ liệu công khai. Thời gian gần đây xuất hiện các vấn đề như có thể trích xuất dữ liệu huấn luyện từ các Mô hình Ngôn ngữ Tiên huấn (LMs), bao gồm Thông tin Nhận dạng Cá nhân (PII) như tên, số điện thoại và địa chỉ email, cũng như các thông tin khác như mã được cấp phép, ghi chú lâm sàng

riêng tư, vv¹. Vào năm 2021, một chatbot AI mang tên Iruda đã trở thành hệ thống AI đầu tiên bị kiện về vi phạm Luật Bảo vệ Thông tin Cá nhân sau khi tạo ra các địa chỉ nhà và số tài khoản ngân hàng chính xác của các cá nhân một cách không cố ý². Heikkilä cũng đã chỉ ra rằng GPT-3³, một trong những Mô hình Ngôn ngữ được biết đến nhất hiện đang được sử dụng thương mại, cung cấp thông tin cá nhân chi tiết về Tổng biên tập của MIT Technology Review bao gồm các thành viên trong gia đình, địa chỉ làm việc và số điện thoại của anh ấy. Mỗi cá nhân đều có quyền được quên, xóa bỏ thông tin cá nhân của mình (Right to be forgotten - RTBF) để có thể hạn chế việc sử dụng trực tiếp và gián tiếp mục đích thương mại với thông tin cá nhân của họ. Các phương pháp trước đây giải quyết các rủi ro về quyền riêng tư cho các mô hình ngôn ngữ là cố gắng loại bỏ tất cả thông tin riêng tư từ dữ liệu huấn luyện (tiền xử lý dữ liệu)⁴⁵ hoặc cố gắng thiết kế các thuật toán đảm bảo quyền riêng tư Differential Privacy (DP). Cả hai phương pháp đều yêu cầu huấn luyện lại mô hình ngôn ngữ cơ bản mỗi khi cá nhân muốn thực hiện quyền RTBF của mình, điều này làm cho chúng không phù hợp cho các mô hình ngôn ngữ lớn, vì việc huấn luyện lại cực kỳ tốn kém. Kỹ thuật Knowledge Unlearning như một giải pháp hiệu quả có thể được áp dụng chỉ với một vài cập nhật tham số thay vì huấn luyện lại mô hình ngôn ngữ cơ bản.

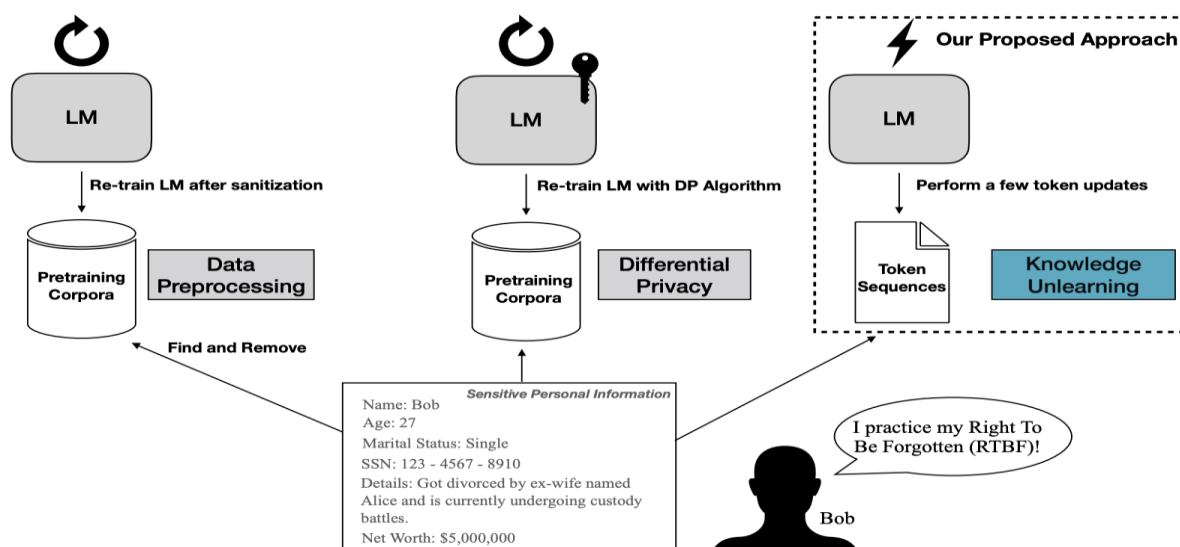
¹ Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21), pages 2633–2650

² Jasmine Park. 2021. South Korea: The first case where the personal information protection act was applied to an AI system

³ Melissa Heikkilä. 2022. What does gpt-3 "know" about me?

⁴ Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In Proceedings of the 59th Annual

⁵ Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. arXiv preprint arXiv:2202.06539



Hình 1: Mô hình thực nghiệm - <https://github.com/joeljang/knowledge-unlearning>

Input: Sử dụng mô hình ngôn ngữ dữ liệu lớn GPT-NEO (125M, 1.3B, 2.7B) và OPT (125M, 1.3B, 2.7B) LMs trong việc thử nghiệm.

Output: Kết quả thực nghiệm khi áp dụng các phương pháp giảm thiểu rủi ro rò rỉ dữ liệu riêng tư của các phương pháp: Data Pre/Post-Processing, Differential privacy và Knowledge Unlearning.

MỤC TIÊU

- Thử nghiệm kỹ thuật Knowledge Unlearning trong việc giảm thiểu rủi ro về vi phạm quyền riêng tư trong các mô hình ngôn ngữ (PLM).
- Đánh giá mức độ hiệu quả và khả năng ứng dụng khi áp dụng kỹ thuật Knowledge Unlearning trong việc giảm thiểu rủi ro về vi phạm quyền riêng tư trong mô hình ngôn ngữ so với các mô hình hiện có như: Data Pre/Post-Processing và Differential Privacy, vv.

PHẠM VI

- Thực nghiệm các kỹ thuật giảm thiểu rủi ro về vi phạm quyền riêng tư trong mô hình ngôn ngữ lớn: GPT-NEO (125M, 1.3B, 2.7B) và OPT (125M, 1.3B, 2.7B).

NỘI DUNG

- Nghiên cứu các kỹ thuật xóa bỏ dữ liệu riêng tư hiện có

- Nghiên cứu kỹ thuật Knowledge Unlearning
- Đánh giá hiệu quả và cải tiến kỹ thuật bảo vệ quyền riêng tư trong mô hình ngôn ngữ

PHƯƠNG PHÁP

➤ Nghiên cứu các kỹ thuật xóa bỏ dữ liệu riêng tư hiện có

- Khảo sát, phân tích và so sánh các kỹ thuật xóa bỏ dữ liệu riêng tư phổ biến, bao gồm Data Pre/Post-Processing, Differential Privacy, k-anonymity, v.v.
- Thực nghiệm đánh giá mức độ hiệu quả của các kỹ thuật xóa bỏ dữ liệu riêng tư trong việc bảo vệ quyền riêng tư và duy trì hiệu suất mô hình ngôn ngữ.

➤ Nghiên cứu kỹ thuật Knowledge Unlearning

- Khảo sát và tìm hiểu kỹ thuật Knowledge unlearning.
- Thực hiện áp dụng các kỹ thuật Knowledge unlearning cho mô hình ngôn ngữ để giảm thiểu rủi ro vi phạm quyền riêng tư.

➤ Đánh giá hiệu quả và cải tiến

- Thiết kế các chỉ số đánh giá phù hợp để đo lường hiệu quả của các kỹ thuật bảo vệ quyền riêng tư trong mô hình ngôn ngữ.
- Thực nghiệm và đánh giá mức độ hiệu quả của Knowledge unlearning trong việc giảm thiểu rủi ro dữ liệu riêng tư và duy trì hiệu suất mô hình ngôn ngữ so với các kỹ thuật bảo vệ quyền riêng tư khác.
- Phân tích kết quả đánh giá để xác định kỹ thuật bảo vệ quyền riêng tư hiệu quả nhất cho các trường hợp cụ thể.
- Nghiên cứu cách thức và phương pháp cải tiến kỹ thuật bảo vệ quyền riêng tư gồm cải tiến độ chính xác cũng như đảm bảo hiệu suất của mô hình sau khi áp dụng.

KẾT QUẢ MONG ĐỢI

- Đưa ra kết quả thực nghiệm của các kỹ thuật bảo vệ quyền riêng tư trong mô hình ngôn ngữ.

- Đánh giá mức độ hiệu quả và chi phí áp dụng của kỹ thuật Knowledge Unlearning so với các kỹ thuật như: Data Pre/Post-Processing và Differential Privacy, vv.
- Thực nghiệm nhiều lần để cố gắng xác định yếu tố góp phần vào độ khó của việc áp dụng Knowledge unlearning.

TÀI LIỆU THAM KHẢO

- [1] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? arXiv preprint arXiv:2202.05520
- [2] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. arXiv preprint arXiv:2202.06539.
- [3] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In International Conference on Learning Representations.
- [4] Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models. arXiv preprint arXiv:2205.13621.
- [5] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In International Conference on Learning Representations