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

Tác giả: Nguyễn Hữu Hân

Trường Đại học Công nghệ Thông tin-Đại học Quốc gia

What ?

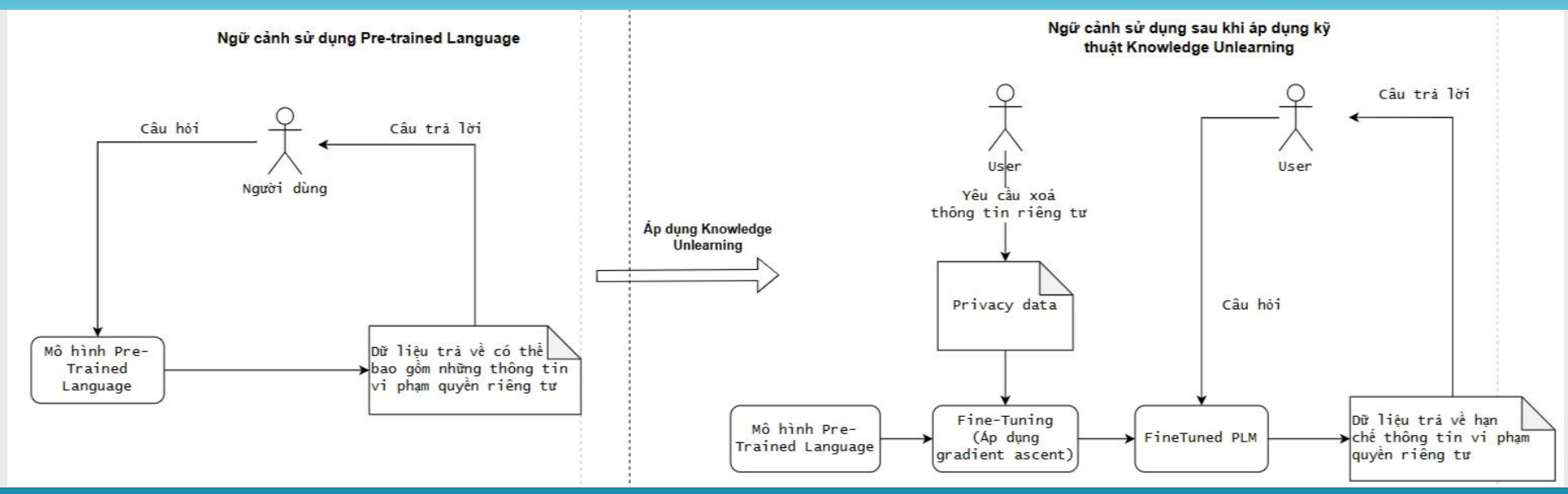
Chúng tôi thực nghiệm kỹ thuật Knowledge Unlearning:

- Nghiên cứu kỹ thuật Knowledge Unlearning trong việc giảm thiểu các rủi ro về vi phạm quyền riêng tư trong mô hình ngôn ngữ.
- Thực nghiệm và so sánh hiệu quả, chi phí khi áp dụng kỹ thuật Knowledge Unlearning, Data Pre/Post processing, Differential Privacy

Why ?

- Pre-Trained Language Model (PLM) ghi nhớ một lượng lớn kiến thức trong quá trình huấn luyện, bao gồm các thông tin có thể vi phạm quyền riêng tư.
- Các kỹ thuật trước đây đều yêu cầu phải huấn luyện lại mô hình khi có các yêu vầu về xử lý dữ liệu vi phạm quyền riêng tư, gây mất thời gian và lãng phí tài nguyên.
- Chúng tôi thử nghiệm các mô hình để xác tìm ra cách thức đảm vệ dữ liệu riêng tư nhưng vẫn giữ được hiệu suất của mô hình và giảm thiểu chi phí tái huấn luyện

Overview



Description

1. Pre-trained Language Model

- Trong đề tài ngày chúng tôi chử nghiệm các kỹ thuật giảm thiểu rủi ro vi phạm quyền riêng tư trên các mô hình ngôn ngữ sau:
  - GPT-NEO (125M, 1.3B, 2.7B)
  - OPT (125M, 1.3B, 2.7B) LMs
- Trong đó, 125M, 1.3B, 2.7B là số lượng tham số dùng trong việc huấn luyện mô hình. Số lượng tham số càng lớn thì độ chính xác của mô hình càng cao nhưng sẽ tiêu tốn nhiêu tài nguyên trong quá trình huấn luyện.

2. Triển khai Knowledge Unlearning

- Fine-tuning là một kỹ thuật quan trọng trong học sâu, cho phép chúng ta điều chỉnh mô hình đã được huấn luyện trước (pre-trained model) trên một tập dữ liệu mới
- Với phương pháp Knowledge Unlearning, chúng tôi áp dụng Gradient ascent đối với những dữ liệu cần loại bỏ. Với phương pháp này, mô hình sau khi được tạo ra sẽ không còn nhớ về dữ liệu bị yêu cầu xoá nữa, từ đó gây ra khó khắn khi cố gắng trích xuất các thông tin đã bị yêu cầu xoá.

3. Một số kỹ thuật khác

Ngoài việc triển khai áp dụng kỹ thuật Knowledge Unlearning, chúng tôi còn áp dụng thêm các kỹ thuật khác như:

- Data pre-processing
- Differential privacy

4. Kết quả dự kiến

- 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.

NII

Nguyễn Hữu Hân – Trường Đại học Công nghệ Thông tin-Đại học Quốc gia

Email : hannh.18@grad.uit.edu.vn