TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

PHƯƠNG PHÁP XÁC ĐỊNH VÀ XÓA NƠ-RON TRI THỨC TRONG MÔ HÌNH TRANSFORMER

(Determining and erasing knowledge neurons in Transformer model)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– PGS.TS. Lê Hoàng Thái (Khoa Công nghệ Thông tin)

Nhóm sinh viên thực hiện:

1. Nguyễn Thiên Phúc (MSSV: 20127681)
2. Nguyễn Trương Hoàng Thái (MSSV: 20127625)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *01/2024* đến *08/2024*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Ngày nay, những hệ thống máy tính nắm giữ lượng lớn dữ liệu cá nhân của người dùng. Bên cạnh lợi ích về mặt cải tiến những mô hình học máy còn tiềm ẩn nguy cơ làm rò rỉ dữ liệu của người dùng [1]. Quá trình xoá bỏ dữ liệu từ phía cơ sở dữ liệu không đảm bảo rằng những mô hình học máy sẽ loại bỏ hoàn toàn lượng dữ liệu dẫn đến tình trạng mô hình vẫn còn "ghi nhớ" lượng thông tin đó [2]. Các cuộc tấn công đối kháng (adversarial attack) nhắm đến việc tấn công vào dữ liệu huấn luyện của mô hình nhằm trích xuất thông tin đặc trưng của dữ liệu [3].

Tình trạng các mô hình học máy "ghi nhớ" thông tin cá nhân của người dùng tiềm ẩn những cuộc tấn công trong tương lai xa. Do đó, việc cho phép người dùng loại bỏ hoàn toàn dữ liệu riêng tư của mình khỏi mô hình học máy được gọi là lĩnh vực Machine Unlearning. Lĩnh vực này cho phép loại bỏ hoàn toàn dữ liệu khỏi mô hình học máy mà không yêu cầu quá trình huấn luyện lại mô hình [4].

Đề tài này nghiên cứu về phương pháp xác định và xóa các nơ-ron tri thức trong mô hình Transformer dựa trên cơ chế tự chú ý (Self-Attention) và mạng nơ-ron truyền thẳng (Feed Forward Network). Cụ thể, chúng tôi thực hiện tìm hiểu phương pháp giúp nhận diện các nơ-ron tri thức, từ đó có thể thực hiện thao tác xoá đi các nơ-ron tri thức này mà không cần tái huấn luyện lại mô hình. Cuối cùng, chúng tôi tiến hành thực nghiệm nhằm đánh giá và so sánh hiệu suất đối với bài toán điền từ vào ô trống. Bài toán điền từ vào ô trống (Cloze Task) [5] là một tác vụ trong Xử lý ngôn ngữ tự nhiên tập trung vào việc dự đoán từ phù hợp nhất với chỗ trống trong một câu văn dựa trên ngữ cảnh của các từ xung quanh.

Đề tài có ý nghĩa đối với các mô hình học máy và sự riêng tư của người dùng, các ý nghĩa nổi bật có thể được đề cập như:

- Việc áp dụng các phương pháp Machine Unlearning giúp tiết kiệm tài nguyên vì các phương pháp đảm bảo được khả năng của mô hình mà không yêu cầu

quá trình tái huấn luyện.

- Tăng cường tính bảo mật cho hệ thống: giảm nguy cơ bị tấn công bởi các cuộc tấn công đối kháng nhắm vào dữ liệu huấn luyện [3, 6].
- Tăng cường niềm tin của người dùng vào các hệ thống trí tuệ nhân tạo: khi người dùng có khả năng kiểm soát dữ liệu cá nhân, sự tin tưởng của người dùng dành cho các hệ thống AI sẽ được cải thiện đáng kể.
- Thúc đẩy sự phát triển của các hệ thống AI có đạo đức: việc bảo vệ quyền riêng tư của người dùng là một yếu tố quan trọng trong việc phát triển các hệ thống AI có đạo đức [1].

2.2 Mục tiêu đề tài

Đề tài này hướng đến nghiên cứu phương pháp xác định và xóa các nơ-ron tri thức trong mô hình Transformer. Đây là một phương pháp trong lĩnh vực Machine Unlearning, giúp loại bỏ dữ liệu khỏi mô hình học máy mà không yêu cầu huấn luyện lại mô hình. Phương pháp ứng dụng tích hợp tích phân ở bước chọn lọc nơ-ron tri thức giúp lọc ra được lượng nơ-ron tri thức chất lượng, làm tiền đề cho một quá trình tinh chỉnh tri thức hiệu quả [4]. Cụ thể, mục tiêu của đề tài bao gồm:

- Xác định các nơ-ron tri thức trong mô hình Transformer có thể hiện thực tế quan hệ (relational facts).
- Áp dụng phương pháp xóa tri thức ở các nơ-ron đã được xác định.
- Đánh giá hiệu quả của phương pháp được đề xuất đối với bài toán điền vào ô trống.

2.3 Phạm vi của đề tài

Trong đề tài này, phương pháp xác định nơ-ron tri thức dựa trên điểm phân bổ được giới hạn ở hai phương pháp là phương pháp cơ sở và phương pháp tích hợp

tích phân được đề xuất ở bài báo của D. Dai [4]. Phương pháp cơ sở sử dụng các nơ-ron của mô hình tiền huấn luyện để phục vụ cho việc tính toán. Như đã được đề cập, phương pháp được đề xuất sử dụng kết hợp các nơ-ron của mô hình tiền huấn luyện với tích phân và được tính toán với xấp xỉ Riemann.

Có rất nhiều mô hình Transformer tiền huấn luyện có cấu trúc gồm hai thành phần chính là tầng tự chú ý và mạng nơ-ron truyền thẳng, nổi bật như Generative Pre-trained Transformer (OpenAI's GPT-1, GPT-3) [7, 8], RoBERTa [9]. Với tiêu chí lựa chọn là các mô hình có cấu trúc phù hợp với phương pháp được đề xuất đồng thời có kích thước phù hợp cho quá trình thực nghiệm, mô hình được sử dụng trong đề tài này là mô hình Transformer tiền huấn luyện BERT-base-cased [10].

Tập dữ liệu được sử dụng trong đề tài này là PARAREL [11]. Tập dữ liệu PARAREL ban đầu là một tập dữ liệu dành cho bài toán điền vào ô trống được xây dựng bởi các chuyên gia, bao gồm nhiều mẫu câu cho 38 quan hệ khác nhau. Tuy nhiên với mục đích đảm bảo tính đa dạng cho tập dữ liệu, nhóm tác giả đã phân tích và chỉ giữ lại 34 quan hệ từ tập dữ liệu ban đầu.

2.4 Cách tiếp cận dự kiến

Tìm hiểu mô hình

Ở giai đoạn này, chúng tôi tiến hành nghiên cứu về kiến trúc của mô hình Transformer [12] và cơ chế key-value trong mạng nơ-ron truyền thẳng [13] giúp nhận diện nơ-ron tri thức.

Tiếp theo, chúng tôi tiến hành nghiên cứu mô hình BERT-base-cased [10] được xây dựng dựa trên nền tảng của Transformer. Sau khi có kiến thức về mô hình, chúng tôi tiếp tục nghiên cứu về phương pháp xác định nơ-ron tri thức và tiến hành xóa tri thức với phương pháp được đề xuất [4].

Chuẩn bị dữ liệu

Ở đề tài này, chúng tôi sử dụng tập dữ liệu PARAREL đã được chuẩn bị và xử lý bởi nhóm tác giả [4]. Tập dữ liệu PARAREL [11] ban đầu là một tập dữ liệu cho bài toán điền vào ô trống được xây dựng bởi các chuyên gia, bao gồm nhiều mẫu câu cho 38 quan hệ khác nhau. Tuy nhiên với mục đích đảm bảo tính đa dạng cho tập dữ liệu, nhóm tác giả đã phân tích và chỉ giữ lại 34 quan hệ từ tập dữ liệu ban đầu. Các quan hệ thực tế (relational facts) được cấu tạo dưới dạng bộ ba $\langle h, r, t \rangle$, trong đó h là phần đầu, t là phần đuôi và r thể hiện mối quan hệ giữa h và t.

Áp dụng phương pháp được đề xuất

Tiến hành thực nghiệm với phương pháp xác định nơ-ron tri thức dựa trên bài toán điền từ vào ô trống. Sau khi xác định được các nơ-ron tri thức, chúng tôi tiến hành xóa tri thức đối với các nơ-ron đã xác định ở mô hình BERT-base-cased.

Kiểm định và đánh giá

Thực hiện thống kê số lượng nơ-ron tri thức đã xác định được và tiến hành phương pháp xóa tri thức. Chúng tôi thực hiện tính toán các độ đo bao gồm:

- Độ chính xác của mô hình: tỉ lệ giữa số lần dự đoán chính xác và tổng số lần dự đoán của mô hình.
- Độ đo Perplexity (PPL): là một phương pháp phổ biến để đánh giá hiệu quả của mô hình ngôn ngữ.

Cuối cùng, chúng tôi áp dụng mô hình vào bài toán điền vào ô trống và so sánh hiệu suất của mô hình sau khi thực hiện xóa nhằm đánh giá tác động của quá trình.

2.5 Kết quả dự kiến của đề tài

Không chỉ dừng lại ở việc đánh giá kết quả dự đoán của mô hình tiền huấn luyện, đề tài nghiên cứu một phương pháp phân tích sâu hơn về mạng nơ-ron truyền thẳng trong mô hình Transformer và phương pháp xóa tri thức được đề xuất, hay còn gọi là Machine Unlearning. Kết quả kỳ vọng của đề tài sẽ bao gồm:

- Số liệu thống kê về số nơ-ron tri thức ở mô hình BERT-base-cased.
- Độ đo đánh giá của phương pháp xóa tri thức, bao gồm độ chính xác của mô hình và số liệu Perplexity (PPL).

2.6 Kế hoạch thực hiện

Mỗi thành viên cần nắm được nội dung của đề tài, công việc cụ thể được phân bố phụ thuộc vào tình hình thực hiện đề tài thực tế. Nội dung thực hiện tổng quát sẽ bao gồm 3 phần chính:

1. Tìm hiểu lý thuyết về đề tài.
2. Thực nghiệm với phương pháp được đề xuất.
3. Viết báo cáo và chuẩn bị bảo vệ đề tài.

Mốc thời gian	**Nội dung công việc**
Tháng 1 - ngày 10/4	Tìm hiểu các kiến thức liên quan đến đề tài như Machine Unlearning, cơ chế chú ý, mô hình Transformer, cấu trúc key-value, phương pháp đề xuất của bài báo.
Ngày 11/4 - 14/6	Thực nghiệm phương pháp được đề xuất bao gồm xác định nơ-ron tri thức, xóa tri thức cho mô hình tiền huấn luyện đã chọn.
Ngày 15/6 - 15/7	Viết báo cáo và chuẩn bị bảo vệ đề tài.

Tài liệu

[1] K. Siau and W. Wang, "Artificial intelligence (ai) ethics: Ethics of ai and ethical ai," *Journal of Database Management (JDM)*, vol. 31, pp. 74–87, 2020.

[2] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning." https://doi.org/10.48550/arXiv.2209.02299.

[3] Y. Chang, Z. Ren, T. T. Nguyen, W. Nejdl, and B. W. Schuller, "Example-based explanations with adversarial attacks for respiratory sound analysis." https://doi.org/10.48550/arXiv.2203.16141.

[4] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei, "Knowledge neurons in pretrained transformers." https://doi.org/10.48550/arXiv.2104.08696.

[5] Z. Hu, R. Chanumolu, X. Lin, N. Ayaz, and V. Chi, "Evaluating nlp systems on a novel cloze task: Judging the plausibility of possible fillers in instructional texts." https://doi.org/10.48550/arXiv.2112.01867.

[6] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015.

[7] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training." https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark,

C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners." https://doi.org/10.48550/arXiv.2005.14165.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach." https://doi.org/10.48550/arXiv.1907.11692.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding." https://doi.org/10.48550/arXiv.1810.04805.

[11] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg, "Measuring and improving consistency in pretrained language models." https://doi.org/10.48550/arXiv.2102.01017.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need." https://doi.org/10.48550/arXiv.1706.03762.

[13] M. Geva, R. Schuster, J. Berant, and O. Levy, "Transformer feed-forward layers are key-value memories." https://doi.org/10.48550/arXiv.2012.14913.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày ... tháng ... năm 2024
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)