KHOÁ LUẬN TỐT NGHIỆP

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

Giảng viên hướng dẫn
PGS.TS. Lê Hoàng Thái

Sinh viên thực hiện
Nguyễn Trương Hoàng Thái - 20127625
Nguyễn Thiên Phúc - 20127681

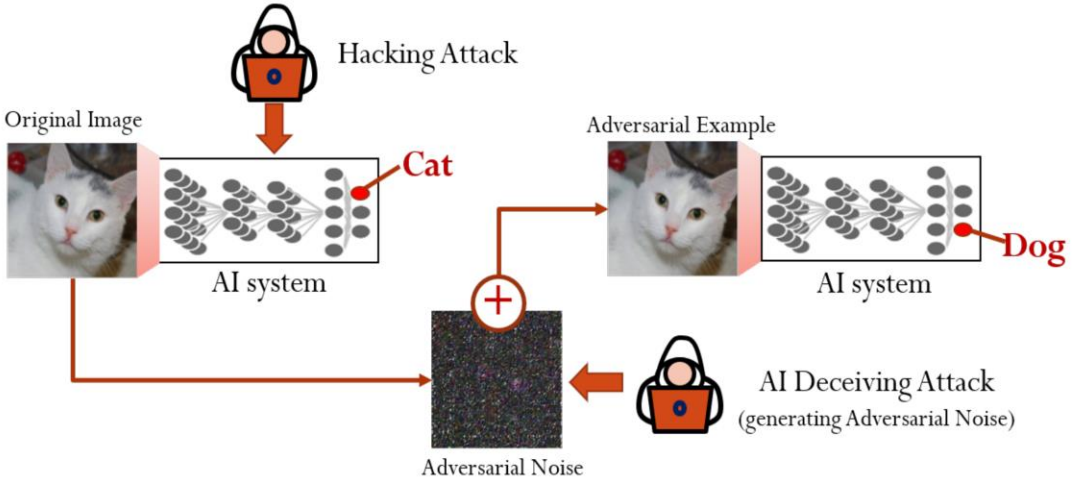
KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nội dung

- Giới thiệu đề tài
- Các công trình liên quan
- Phương pháp tìm hiểu
- Kết quả thực nghiệm
- Kết luận
- Tài liệu tham khảo

Giới thiệu đề tài

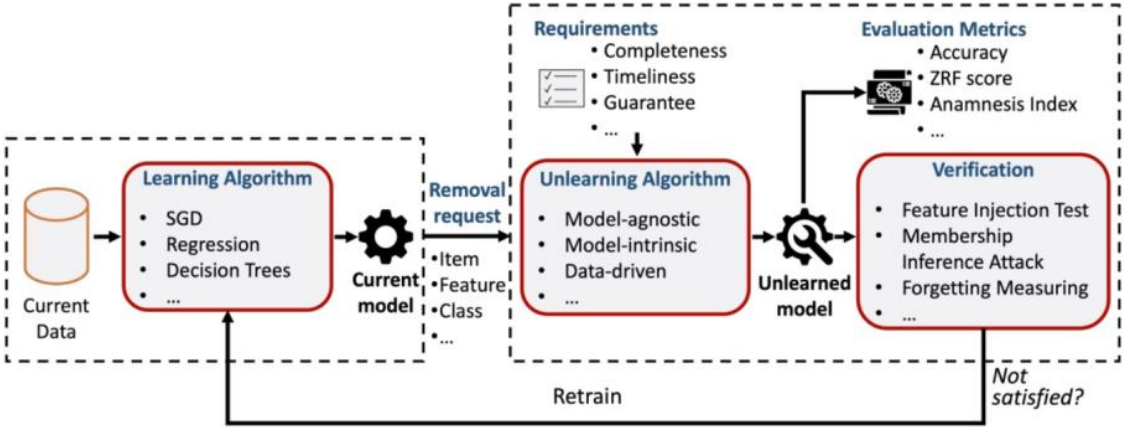
- Mô hình học máy nắm giữ lượng lớn dữ liệu cá nhân của người dùng.
- Các cuộc tấn công đối kháng hướng đến dữ liệu huấn luyện của mô hình nhằm trích xuất thông tin đặc trưng của dữ liệu.



Hình ảnh được tham khảo từ bài báo *Real-Time Adversarial Attack Detection with Deep Image Prior Initialized as a High-Level Representation Based Blurring Network*. Hình ảnh minh họa cho việc tấn công đối kháng (adversarial attack).

Giới thiệu đề tài

- Đề tài của chúng tôi nghiên cứu xác định và loại bỏ tri thức được chỉ định ở mô hình ngôn ngữ.
  - Đảm bảo quyền riêng tư dữ liệu.
  - Không yêu cầu tái huấn luyện mô hình học máy.



Hình ảnh được tham khảo từ bài báo *A Survey of Machine Unlearning*. Hình ảnh trực quan khung sườn tổng quát của phương pháp loại bỏ tri thức.

Giới thiệu đề tài

- Tính toán điểm phân bổ áp dụng độ dốc dựa trên giá trị kích hoạt của các nơ-ron.
- Phương pháp thể hiện sự hiệu quả, nổi bật là kết quả loại bỏ tri thức về dữ liệu nghề nghiệp P106.

Quan hệ xóa	Số lượng nơ-ron	Trên quan hệ xóa (PPL)		Trên các quan hệ khác (PPL)	
		Trước	Sau	Trước	Sau
P937 (work_location)	15 (baseline)	58.0	64.6 (+11.3%)	138.0	139.0 (+1.3%)
	15		129.9 (+123.8%)		153.3 (+11.1%)
	30		173.5 (+199.0%)		160.5 (+16.3%)
	40		184.2 (+217.3%)		160.2 (+16.1%)
	50		187.2 (+222.5%)		158.1 (+14.6%)
P19 (place_of_birth)	15 (baseline)	1,450	1,341 (-7.5%)	120.3	115.7 (-3.8%)
	15		2,873 (+98.1%)		120.7 (+0.3%)
	30		3,105 (+114.1%)		121.6 (+1.0%)
	40		4,025 (+177.6%)		128.5 (+6.8%)
	50		4,384 (+202.3%)		130.6 (+8.5%)
P27 (country_of_citizenship)	10 (baseline)	28.0	27.7 (-1.2%)	143.6	147.9 (+3.0%)
	10		34.8 (+24.3%)		152.3 (+6.1%)
	30		39.0 (+39.3%)		151.5 (+5.5%)
	40		40.2 (+43.6%)		152.4 (+6.1%)
	50		43.6 (+55.7%)		148.3 (+3.3%)
P106 (occupation)	14 (baseline)	2,279	1,956 (-14.2%)	120.1	118.9 (-1.0%)
	14		5,048 (+121.5%)		125.1 (+4.2%)
	27 (max)		5,536 (+142.9%)		127.1 (+5.8%)
P178 (developer)	9 (baseline)	204.5	53.3 (-74.0%)	133.6	142.5 (+6.6%)
	9		391.9 (+91.6%)		136.5 (+2.2%)
	17 (max)		618.8 (+202.5%)		142.2 (+6.4%)

Độ đo perplexity của một số quan hệ dữ liệu trước và sau quá trình loại bỏ tri thức.

Các công trình liên quan

Phương pháp tích hợp độ dốc

- Là một kỹ thuật giải thích dự đoán của mạng nơ-ron sâu.
- Phân tích dự đoán của mô hình học máy bằng cách xác định mức độ quan trọng của từng đặc trưng đầu vào đối với kết quả dự đoán của mô hình.
- Kết hợp ưu điểm của phương pháp Gradient và phương pháp LRP/DeepLIFT.
- Được xây dựng dựa trên hai tiên đề: độ nhạy cảm (Sensitivity) và tính bất biến khi thực thi (Implementation Invariance).

Các công trình liên quan

Công thức mô tả phương pháp tích hợp độ dốc

$$\text{IntegratedGrads}_i(x) := (x_i - x'_i) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

- $\text{IntegratedGrads}_i(x)$: độ phân bổ của đặc trưng thứ i đối với đầu vào x.
- $x_i,\ x'_i$: giá trị của đặc trưng thứ i trong đầu vào x, x'.
- F: hàm kích hoạt.
- Đạo hàm riêng $\frac{\partial F}{\partial x_i}$: đo lường độ nhạy cảm của kết quả dự đoán đối với các thay đổi trong đặc trưng thứ i.
- Hệ số α: tham số nội suy thay đổi từ 0 đến 1.

Các công trình liên quan

Điểm phân bổ ở cơ chế tự chú ý

- Điểm phân bổ này cho biết mức độ quan trọng của kết nối ở quá trình chú ý đối với dự đoán của mô hình.
- Điểm phân bổ được tính bằng tích phân độ dốc của hàm mục tiêu (xác suất dự đoán của mô hình) theo trọng số chú ý của kết nối đó trên một đường đi từ ma trận chú ý với giá trị 0 đến ma trận chú ý ban đầu với giá trị tương ứng là 1.

$$\mathrm{Attr}_h(A) = A_h \odot \int_{\alpha=0}^{1} \frac{\partial F(\alpha A)}{\partial A_h} d\alpha \in \mathbb{R}^{n \times n}$$

Các công trình liên quan

Điểm phân bổ ở cơ chế tự chú ý

$$\mathrm{Attr}_h(A) = A_h \odot \int_{\alpha=0}^{1} \frac{\partial F(\alpha A)}{\partial A_h} d\alpha \in \mathbb{R}^{n \times n}$$

- Attr_h(A): ma trận điểm phân bổ cho đầu chú ý thứ h.
- A_h: ma trận trọng số của đầu chú ý thứ h.
- $\odot$: phép nhân tương ứng từng phần tử (element-wise) giữa ma trận A_h và ma trận kết quả của tích phân.
- Hệ số α: tham số nội suy thay đổi từ 0 đến 1.

Các công trình liên quan

Kiến trúc Transformer

- Mô hình gồm tầng tự chú, mạng nơ-ron truyền thẳng với số lượng tham số phù hợp.

$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V, \tag{2.7}$$

$$\text{Self-Att}_h(X) = softmax(\frac{Q_h K_h^T}{\sqrt{d_{size}}})V_h, \tag{2.8}$$

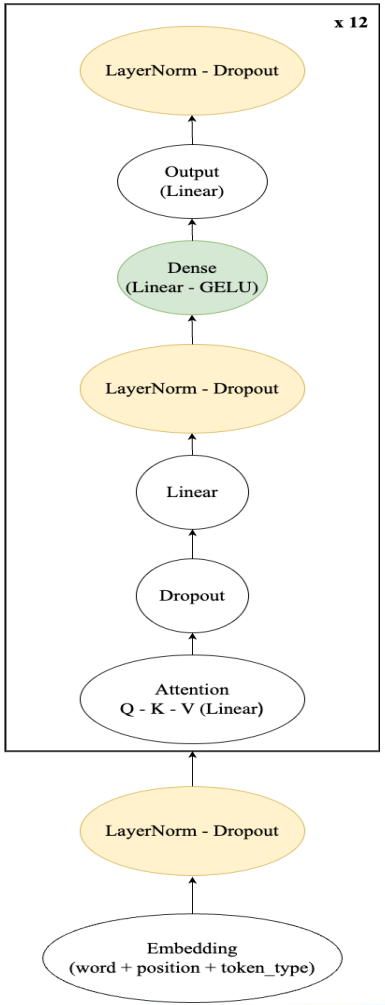
$$FFN(H) = gelu(HW_1)W_2 \tag{2.9}$$

Biểu thức biểu diễn tầng tự chú ý và mạng nơ-ron truyền thẳng.

Các công trình liên quan

Kiến trúc Transformer

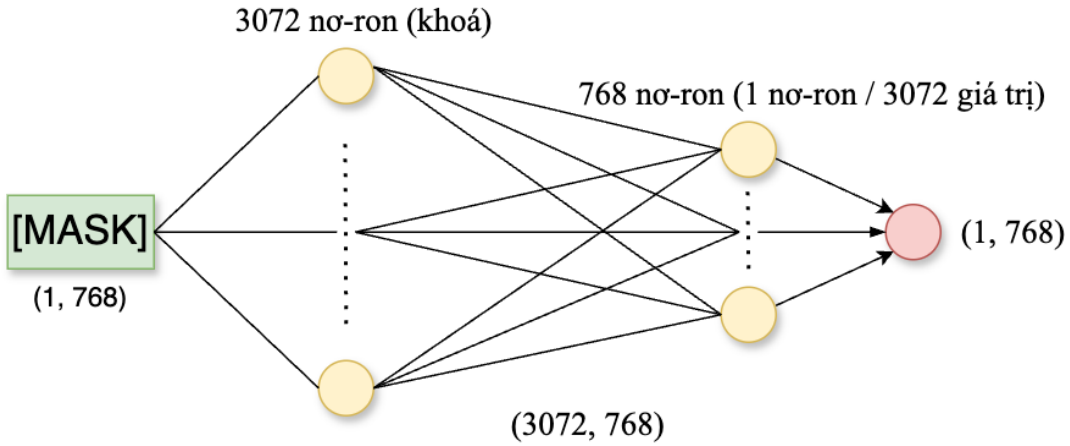
- Mô hình BERT cơ sở (bert-base-cased)
  - 110 triệu tham số.
  - Chiều dữ liệu d = 768.
  - Số lượng từ ngữ thuộc từ điển là 28996.



Các công trình liên quan

Cơ chế khóa - giá trị

- Mô hình Transformer liên hệ chặt chẽ với cấu trúc khóa - giá trị.

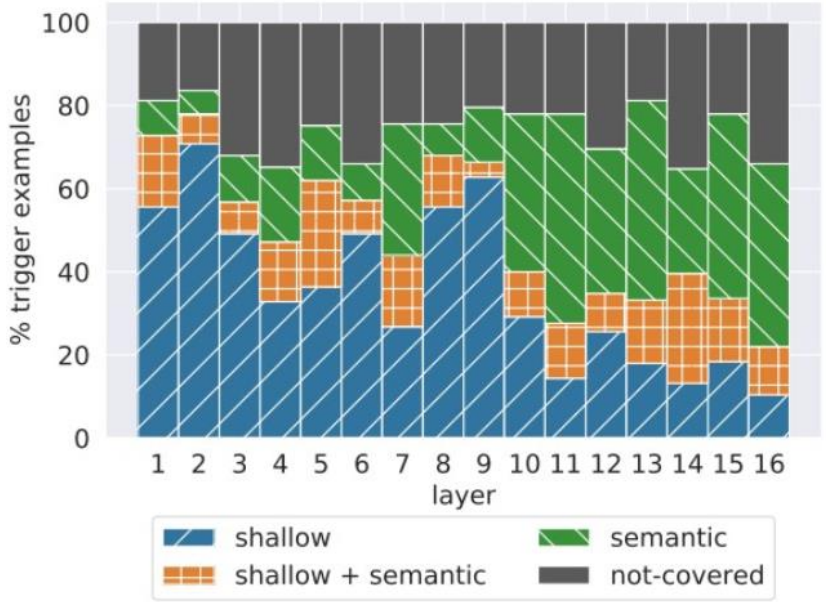


Hình ảnh minh họa cho cấu trúc khóa - giá trị của mạng nơ-ron truyền thẳng thuộc mô hình BERT cơ sở.

Các công trình liên quan

Cơ chế khóa - giá trị

- Các khóa K xác định mẫu câu cấu tạo nên văn bản.
- Các giá trị V thực hiện phân phối.



Hình ảnh được tham khảo từ bài báo *Transformer Feed-Forward Layers Are Key-Value Memories*. Hình ảnh trực quan kết quả chú thích các tiền tố ở mỗi tầng Transformer.

Phương pháp tìm hiểu

Dữ liệu

- Phương pháp sử dụng tập dữ liệu PARAREL.
- Tập dữ liệu về bài toán điền vào ô trống, gồm nhiều mẫu câu cho 34 quan hệ dữ liệu khác nhau được tham khảo từ tập dữ liệu T-REx.
- Mỗi quan hệ có trung bình 8.63 mẫu câu. Các mẫu câu này tạo ra 253,448 câu dữ liệu tri thức với 27,738 thực tế quan hệ.

Phương pháp tìm hiểu

Điểm phân bổ với phương pháp tích hợp độ dốc

$$\text{Attr}(w_i^{(l)}) = w_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial P_x(\alpha \overline{w_i}^{(l)})}{\partial w_i^{(l)}} d\alpha$$

Công thức tính toán điểm phân bổ cho một nơ-ron tri thức trong mô hình ngôn ngữ lớn.

- $w_i^{(l)}$: giá trị kích hoạt của nơ-ron thứ i thuộc tầng ẩn l trong mô hình.
- $P_X\left(\alpha \overline{w}_i^{(l)}\right)$: xác suất dự đoán nhãn đầu ra là nhãn đúng của mô hình.
- Hệ số α: tham số nội suy thay đổi từ 0 đến 1.

Phương pháp tìm hiểu

- Xấp xỉ Riemman cho độ phân bổ với phương pháp tích hợp độ dốc.

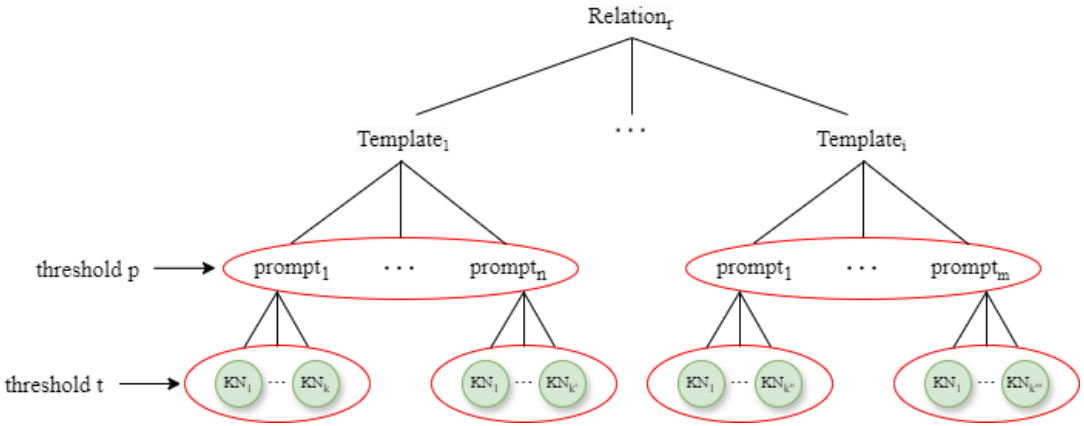
$$\tilde{\mathrm{Attr}}(w_i^{(l)}) = \frac{\overline{w_i}^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial P_x\left(\frac{k}{m}\overline{w_i}^{(l)}\right)}{\partial w_i^{(l)}}$$

Với giá trị tham số m được chọn trong quá trình thực nghiệm là 20.

Phương pháp tìm hiểu

Chắt lọc nơ-ron tri thức

- Tập dữ liệu đảm bảo sự đa dạng.
- Phương pháp gồm ba bước:
  1. Tính toán điểm phân bổ với mỗi câu truy vấn.
  2. Chắt lọc với ngưỡng t sử dụng điểm phân bổ.
  3. Chắt lọc với ngưỡng p.

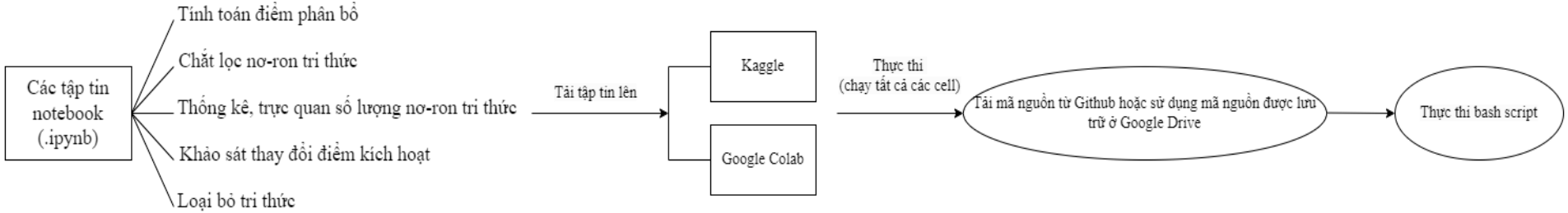


Hình ảnh trực quan phương pháp chắt lọc nơ-ron tri thức sử dụng cấu trúc cây.

Kết quả thực nghiệm

Cách thức tiến hành

- Sử dụng kết hợp tài nguyên của Kaggle và Google Colab.
  - Kaggle: GPU P100.
  - Google Colab: GPU L4, A100.



Hình ảnh trực quan cách thức tiến hành thực nghiệm.

Kết quả thực nghiệm

Thống kê số lượng nơ-ron tri thức

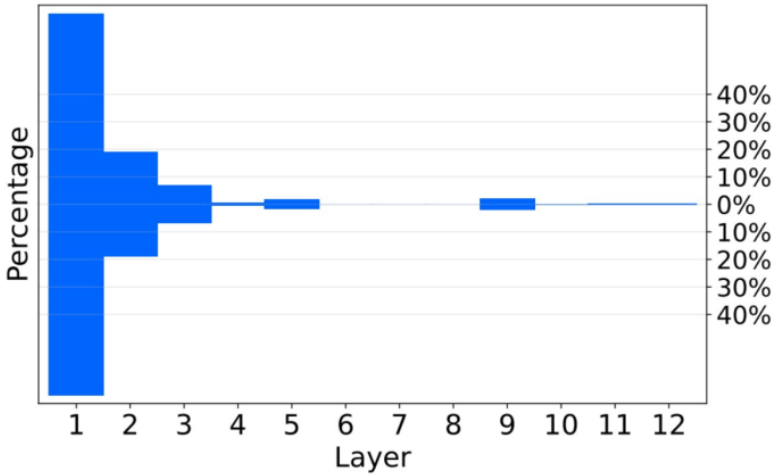
- Điểm tham chiếu là 0
- Bước xấp xỉ m = 20
- Giá trị ngưỡng t = 0.2
- Giá trị khởi tạo ngưỡng p = 0.7

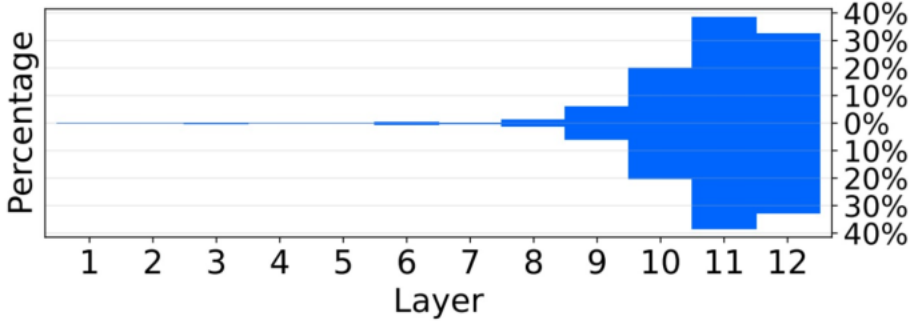
	Phương pháp tích hợp độ dốc	Phương pháp cơ sở
Số lượng nơ-ron tri thức trung bình của mỗi mẫu câu	4.1338	3.9632
Số lượng nơ-ron tri thức giao nhau giữa hai mẫu câu bất kỳ thuộc cùng một quan hệ	1.2279	2.8466
Số lượng nơ-ron tri thức giao nhau giữa hai mẫu câu bất kỳ thuộc hai quan hệ bất kỳ	0.0932	1.9235

Bảng thống kê số lượng nơ-ron tri thức của phương pháp tích hợp độ dốc và phương pháp cơ sở.

Kết quả thực nghiệm

Thống kê số lượng nơ-ron tri thức

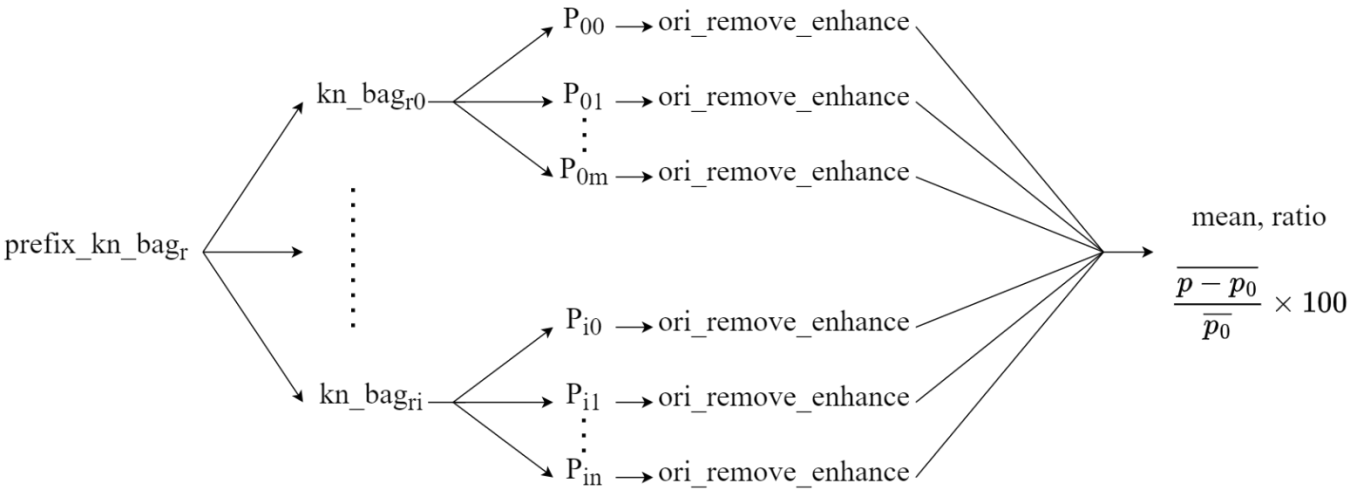




Tỉ lệ phân bổ các nơ-ron dựa trên điểm phân bổ cơ sở (hình trái) và điểm phân bổ tính toán với phương pháp tích hợp độ dốc (hình phải) theo các tầng con trong mô hình.

Kết quả thực nghiệm

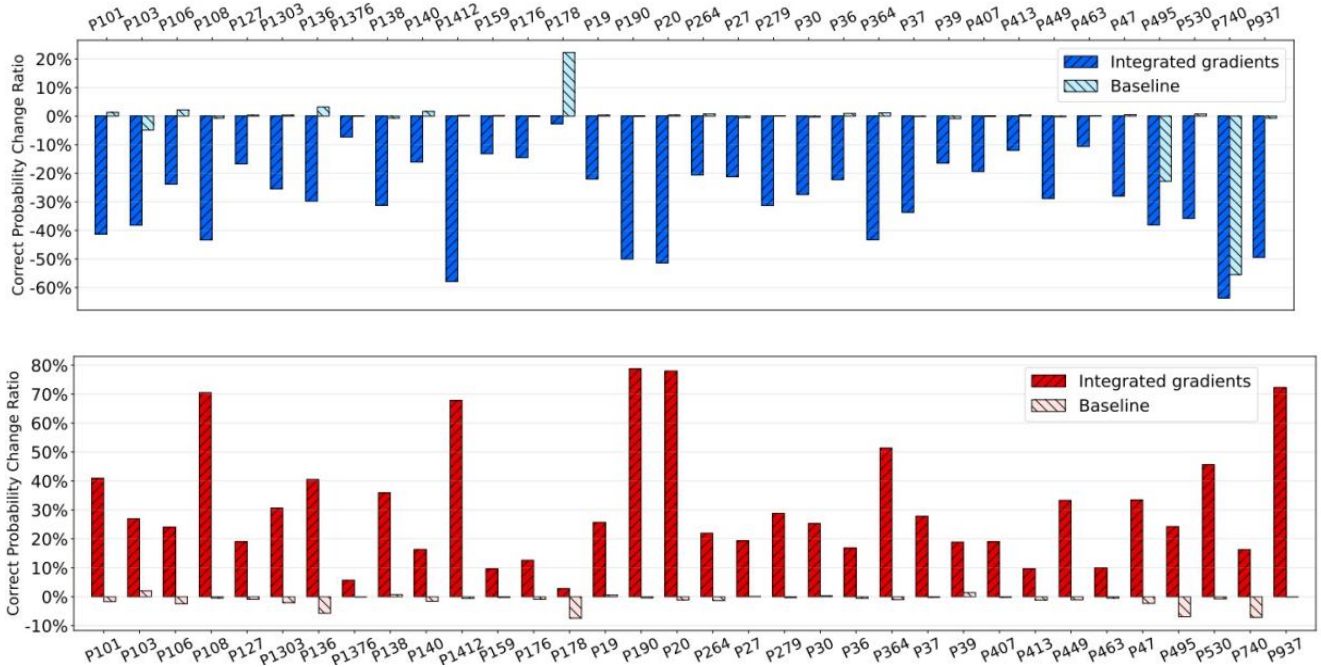
Khảo sát thay đổi giá trị kích hoạt



Hình ảnh trực quan quá trình thay đổi điểm kích hoạt với hai phương thức chính là gán giá trị 0 (remove) và tăng cường bằng cách nhân với hệ số dương (enhance).

Kết quả thực nghiệm

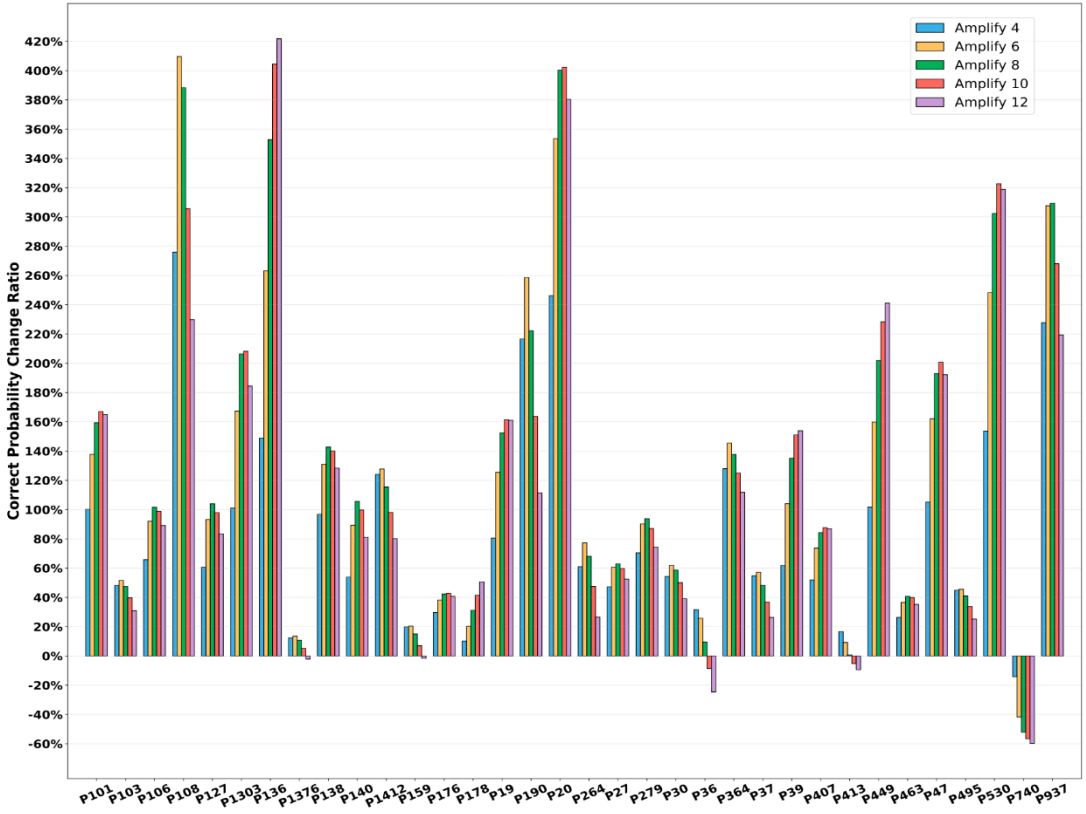
Khảo sát thay đổi giá trị kích hoạt



Tỉ lệ chênh lệch dự đoán của mô hình là nhãn đúng sau và trước khi gán giá trị kích hoạt bằng 0 (hình trên) và khi nhân với hệ số 2 (hình dưới) theo từng quan hệ dữ liệu.

Kết quả thực nghiệm

- Xảy ra ba trường hợp: (1) **tăng** (P136, P449); (2) **giảm** (P36, P413); (3) **tăng-giảm** (P101, P103).
- Phần lớn các quan hệ dữ liệu rơi vào trường hợp tăng-giảm.



Tỉ lệ chênh lệch dự đoán của mô hình là nhãn đúng sau và trước khi nhân giá trị kích hoạt với hệ số dương dựa trên phương pháp tích hợp độ dốc.

Kết quả thực nghiệm

Loại bỏ tri thức khỏi mô hình

- Độ đo perplexity tăng mạnh ở quan hệ được loại bỏ (11.3% đến 222.5%), trong khi ở các quan hệ khác tăng nhẹ (0.3% đến 16.3%).
- Kết quả dựa trên phương pháp cơ sở là bất thường vì độ giảm perplexity.

Quan hệ xóa	Số lượng nơ-ron	Trên quan hệ xóa (PPL)		Trên các quan hệ khác (PPL)	
		Trước	Sau	Trước	Sau
P937 (work_location)	15 (baseline)	58.0	64.6 (+11.3%)	138.0	139.0 (+1.3%)
	15		129.9 (+123.8%)		153.3 (+11.1%)
	30		173.5 (+199.0%)		160.5 (+16.3%)
	40		184.2 (+217.3%)		160.2 (+16.1%)
	50		187.2 (+222.5%)		158.1 (+14.6%)
P19 (place_of_birth)	15 (baseline)	1,450	1,341 (-7.5%)	120.3	115.7 (-3.8%)
	15		2,873 (+98.1%)		120.7 (+0.3%)
	30		3,105 (+114.1%)		121.6 (+1.0%)
	40		4,025 (+177.6%)		128.5 (+6.8%)
	50		4,384 (+202.3%)		130.6 (+8.5%)
P27 (country_of_citizenship)	10 (baseline)	28.0	27.7 (-1.2%)	143.6	147.9 (+3.0%)
	10		34.8 (+24.3%)		152.3 (+6.1%)
	30		39.0 (+39.3%)		151.5 (+5.5%)
	40		40.2 (+43.6%)		152.4 (+6.1%)
	50		43.6 (+55.7%)		148.3 (+3.3%)
P106 (occupation)	14 (baseline)	2,279	1,956 (-14.2%)	120.1	118.9 (-1.0%)
	14		5,048 (+121.5%)		125.1 (+4.2%)
	27 (max)		5,536 (+142.9%)		127.1 (+5.8%)
P178 (developer)	9 (baseline)	204.5	53.3 (-74.0%)	133.6	142.5 (+6.6%)
	9		391.9 (+91.6%)		136.5 (+2.2%)
	17 (max)		618.8 (+202.5%)		142.2 (+6.4%)

Độ đo perplexity của một số quan hệ dữ liệu trước và sau quá trình loại bỏ tri thức.

Kết quả thực nghiệm

Thời gian thực thi

1. Tính toán điểm phân bổ của các nơ-ron với mỗi câu truy vấn: GPU P100 ≈ 1.9515, GPU A100 ≈ 0.9176 (giây/câu truy vấn).
2. Chắt lọc nơ-ron tri thức: GPU A100 ≈ 16 phút.
3. Thống kê số lượng nơ-ron tri thức: GPU P100 ≈ 10 phút.
4. Khảo sát thay đổi giá trị phân bổ: GPU P100 ≈ 3.253 giờ.
5. Loại bỏ tri thức khỏi mô hình: GPU P100 ≈ 2.4 giờ.

Kết luận

- Phương pháp đem lại kết quả về quá trình loại bỏ tri thức.
- Phương pháp có thể được áp dụng cho các mô hình Transformer.
- Vì là nghiên cứu sơ bộ nên kết quả phụ thuộc vào tập dữ liệu cụ thể, không đảm bảo sự hiệu quả với các tập dữ liệu khác.

Tài liệu tham khảo

1. Nguyen, Thanh Tam et al. “A Survey of Machine Unlearning”. In: arXiv preprint arXiv:2209.02299 (2022).
2. Dai, Damai et al. “Knowledge Neurons in Pretrained Transformers”. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. 2022, pp. 8493-8502.
3. Sundararajan, Mukund, Taly, Ankur, and Yan, Qiqi. “Axiomatic Attribution for Deep Networks”. In: Proceedings of the 34th International Conference on Machine Learning. Ed. by Precup, Doina and Teh, Yee Whye. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 3319-3328. url: https://proceedings.mlr.press/v70/sundararajan17a.html.
4. Geva, Mor et al. “Transformer Feed-Forward Layers Are Key-Value Memories”. In: Empirical Methods in Natural Language Processing (EMNLP). 2021.
5. Hao, Yaru et al. “Self-Attention Attribution: Interpreting Information Interactions Inside Transformer”. In: Proceedings of the AAAI Conference on Artificial Intelligence 35.14 (2021), pp. 12963-12971. doi: 10.1609/aaai.v35i14.17533. url: https://ojs.aaai.org/index.php/AAAI/article/view/17533.
6. Vaswani, Ashish et al. “Attention is All you Need”. In: Advances in Neural Information Processing Systems. Ed. by Guyon, I. et al. Vol. 30. Curran Associates, Inc., 2017. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
7. Elsahar, Hady et al. “T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples”. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Ed. by Calzolari, Nicoletta et al. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. url: https://aclanthology.org/L18-1544.

XIN CẢM ƠN SỰ THEO DÕI CỦA QUÝ THẦY CÔ
CẢM ƠN SỰ THEO DÕI CỦA CÁC BẠN

THANK YOU!