(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0053723**

(51)[2019.01] **G09B 5/00**

(13) **B**

(21) 1-2023-08940

(22) 14/12/2023

(45) 25/11/2025 452

(43) 26/02/2024 431

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)
Lô D26 khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội.

(72) BÙI KHẮC HOÀI NAM (VN); TRẦN QUANG MINH (VN); NGUYỄN THÀNH ĐÔ (VN); PHẠM VIỆT HOÀNG (VN); BÙI CHÍ MINH (VN); NGUYỄN THẾ LÂM (VN); NGUYỄN QUANG VINH (VN); NGUYỄN NGỌC DŨNG (VN).

(74) Công ty TNHH NACILAW (NACILAW)

(54) PHƯƠNG PHÁP XÂY DỰNG MÔ HÌNH NGÔN NGỮ LỚN VÀ ỨNG DỤNG TRỢ LÝ ẢO DỰA TRÊN MÔ HÌNH NGÔN NGỮ LỚN CHO TIẾNG VIỆT

(21) 1-2023-08940

(57) Sáng chế đề xuất phương pháp xây dựng mô hình ngôn ngữ lớn và ứng dụng trợ lý ảo dựa trên mô hình ngôn ngữ lớn cho tiếng Việt. Sáng chế này đề xuất phương pháp xây dựng mô hình ngôn ngữ lớn ViettelLLM với 7 tỉ tham số được huấn luyện theo chiến thuật mô hình ngôn ngữ nhân quả (Causal language modeling), tập trung vào hỗ trợ cho tiếng Việt. Phương pháp theo sáng chế gồm các bước: thu thập và xử lý dữ liệu tiền huấn luyện hai ngôn ngữ tiếng Anh và tiếng Việt, xây dựng kiến trúc huấn luyện cho ViettelLLM, huấn luyện mô hình ngôn ngữ lớn ViettelLLM, xây dựng ứng dụng ViettelLLM-Chat sử dụng mô hình ngôn ngữ lớn ViettelLLM.

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề xuất phương pháp xây dựng mô hình ngôn ngữ lớn (large language model, LLM) và ứng dụng trợ lý ảo dựa trên mô hình ngôn ngữ lớn cho tiếng Việt. Cụ thể, phương pháp được đề cập trong sáng chế là mô hình được huấn luyện trên nhiều dữ liệu tiếng Việt nhất ở thời điểm hiện tại, mang đến hiệu quả đáng kể cho các ứng dụng trợ lý ảo nói riêng và các bài toán xử lý ngôn ngữ tự nhiên (natural languge processing, NLP) nói chung.

Tình trạng kỹ thuật của sáng chế

Các bài toán có đầu vào và đầu ra là chuỗi (sequence to sequence, seq2seq) là một lớp các bài toán nâng cao trong NLP. Hướng tiếp cận các bài toán seq2seq được nghiên cứu và phát triển mạnh mẽ nhất trong những năm gần đây là sử dụng LLM, với công nghệ cốt lõi là cơ chế chú ý (attention mechanism) để dự đoán từ tiếp theo của một ngữ cảnh đầu vào. Các bài toán sử dụng kiến trúc seq2seq rất đa dạng, bao gồm dịch máy (machine translation), hỏi đáp (question answering), tổng hợp văn bản (text summarization), và nhiều ứng dụng NLP khác. Trong đó, ứng dụng Trợ lý ảo (TLA), với bản chất là được phát triển từ bài toán hỏi đáp, là một trong những ứng dụng tiêu biểu nhất, được phát triển rộng rãi trong thời gian gần đây, sử dụng công nghệ mô hình ngôn ngữ lớn.

Hiện nay, các mô hình tiên tiến trên thế giới cho bài toán seq2seq tập trung vào tận dụng tri thức của các mô hình ngôn ngữ tạo sinh (Generative AI) như GPT, T5 hay Llama để sinh ra văn bản nối tiếp văn bản ngữ cảnh thông qua bộ giải mã (Decoder). Cụ thể, cách tiếp cận này đã được ứng dụng vào trong sản phẩm thực tế như ChatGPT hay Google Bard, đồng thời cũng cho thấy hiệu quả cao trên hầu hết các tập dữ liệu mở tiêu chuẩn (benchmark datasets) như MMLU, HellaSwag, và GSM8K.

Tuy nhiên, hầu hết các LLM mã nguồn mở hiện tại hỗ trợ tiếng Việt rất hạn chế. Các LLM hiện tại được huấn luyện với lượng dữ liệu rất ít tiếng Việt, dẫn đến việc tri thức của LLM trong tiếng Việt bị thiếu sót, nội dung tiếng Việt sinh ra không tự nhiên. Mặt khác, khả năng hiểu ngôn ngữ của LLM bị giới hạn bởi độ dài ngữ cảnh trong khi các bộ tách từ (tokenizer) hiện có không tối ưu dẫn đến việc LLM chỉ có thể xử lý văn bản tiếng Việt ngắn. Do đó, việc xây dựng mô hình ngôn ngữ lớn chuyên dùng cho tiếng Việt vẫn là nghiên cứu mở trong lĩnh vực này. Ngoài ra, hiện cũng có rất ít ứng dụng sử dụng LLM cho tiếng Việt trong khi tiềm năng của LLM cho những ứng dụng như TLA là rất lớn.

Do đó, dựa vào quan sát rằng LLM có hiệu quả rất tốt với dữ liệu tiếng Anh và có khả năng học đa ngôn ngữ, sáng chế này đề xuất xây dựng mô hình ngôn ngữ lớn, được huấn luyện trên hai ngôn ngữ chính là tiếng Anh và tiếng Việt, trong đó tập trung lượng dữ liệu lớn vào tiếng Việt. Thêm vào đó, sáng chế cũng đề xuất phương pháp xây dựng bộ dữ liệu chỉ dẫn để tinh chỉnh LLM (finetune instruction) để phù hợp với ứng dụng TLA.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp xây dựng mô hình ngôn ngữ lớn và đưa mô hình ngôn ngữ lớn vào các ứng dụng liên quan đến phát triển ứng dụng TLA cho tiếng Việt. Cụ thể, sáng chế tập trung vào hai vấn đề chính, bao gồm: i) tiền huấn luyện LLM với nguồn dữ liệu hai ngôn ngữ tiếng Anh và tiếng Việt, được thu thập và có chọn lọc từ internet. Mô hình ngôn ngữ lớn này được đặt tên là ViettelLLM; và ii) đưa ra phương pháp thiết kế bộ dữ liệu chỉ dẫn và phương pháp huấn luyện tinh chỉnh ViettelLLM với dữ liệu chỉ dẫn cho ứng dụng TLA ViettelLLM-Chat.

Sáng chế cung cấp những giải pháp để giải quyết từng vấn đề trên. Cụ thể, sáng chế đã thu thập dữ liệu thô từ Internet, xử lý bằng nhiều bước sàng lọc và kiểm định chất lượng để thu được bộ dữ liệu huấn luyện dưới dạng văn bản với hai ngôn ngữ tiếng Anh và tiếng Việt. Sau đó, sáng chế đề xuất kiến trúc mô hình ngôn ngữ, phương pháp huấn luyện bộ tách từ (tokenizer), và tiến hành huấn luyện mô hình ngôn ngữ với dữ liệu song ngữ. Nhờ việc huấn luyện đồng đều với dữ liệu hai ngôn ngữ, bộ tách từ có khả năng mã hóa văn bản

tiếng Anh và tiếng Việt đều hiệu quả, đồng thời ViettelLLM có khả năng hiểu tri thức trong cả hai ngôn ngữ. Sau khi hoàn thành tiền huấn luyện ViettelLLM, sáng chế đề xuất cách xây dựng bộ dữ liệu chỉ dẫn với ý tưởng chính là dịch dữ liệu tổng hợp từ nhiều nguồn với những điểm ưu việt khác nhau, làm đầu vào cho việc huấn luyện mô hình ViettelLLM-Chat. Để đạt được mục đích trên, sáng chế bao gồm các bước:

Bước 1: thu thập và xử lý dữ liệu tiền huấn luyện hai ngôn ngữ tiếng Anh và tiếng Việt;

Bước 2: xây dựng kiến trúc huấn luyện cho ViettelLLM;

Bước 3: huấn luyện mô hình ngôn ngữ lớn ViettelLLM;

Bước 4: xây dựng ứng dụng ViettelLLM-Chat sử dụng mô hình ngôn ngữ lớn ViettelLLM.

Mô tả vắn tắt các hình vẽ

Hình 1 là hình vẽ mô tả các bước thực hiện của sáng chế để xây dựng ứng dụng ViettelLLM-Chat sử dụng mô hình ngôn ngữ lớn ViettelLLM;

Hình 2 là hình vẽ mô tả kiến trúc của mô hình ngôn ngữ lớn ViettelLLM;

Mô tả chi tiết sáng chế

Sáng chế được mô tả chi tiết dưới đây có dựa vào hình vẽ kèm theo, các hình vẽ này nhằm mục đích minh họa các phương án của sáng chế mà không làm giới hạn phạm vi bảo hộ sáng chế.

Hình 1 mô tả các bước xử lý để xây xây dựng mô hình ngôn ngữ lớn ViettelLLM và sử dụng ViettelLLM để phát triển ứng dụng trợ lý ảo: i) thu thập và tiền xử lý dữ liệu; ii) huấn luyện mô hình theo chiến thuật học tự giám sát (self-supervised learning) để tạo thành mô hình ngôn ngữ lớn ViettelLLM, kiến trúc của mô hình ViettelLLM; iii) tạo bộ dữ liệu chỉ dẫn và tiến hành huấn luyện chỉ dẫn để tạo ứng dụng mô hình ViettelLLM-

Chat. Cụ thể hơn, phương pháp đề xuất cho việc xây dựng mô hình ngôn ngữ lớn và ứng dụng dựa trên mô hình ngôn ngữ cho tiếng Việt bao gồm các bước:

Bước 1: thu thập và xử lý dữ liệu tiền huấn luyện hai ngôn ngữ tiếng Anh và tiếng Việt;

Tại bước này, tập trung vào thu thập và xử lý dữ liệu để huấn luyện mô hình ngôn ngữ lớn có tổng kích thước là 500GB trong đó 400GB tiếng Việt và 100GB tiếng Anh. Cụ thể, *dữ liệu tiếng Việt* được thu thập từ những văn bản được tổng hợp từ internet như các trang báo điện tử uy tín, wiki, pháp luật Việt Nam với tổng số là 800GB. Sau khi xử lý, dữ liệu tiếng Việt có kích thước 500GB. Theo đó, bộ dữ liệu được xử lý theo các bước như sau: i) tách tập dữ liệu theo mức văn bản; ii) loại bỏ các câu ít hơn ba từ, nhằm loại bỏ các câu không có nhiều ý nghĩa; iii) loại bỏ các văn bản có ít hơn ba câu, nhằm loại bỏ các văn bản không có nhiều thông tin; iv) loại bỏ các văn bản có chứa các ký tự cảm xúc (emoji); v) loại bỏ các văn bản chứa các từ thô tục. *Dữ liệu tiếng Anh* được thu thập từ bộ dữ liệu C4 (The Colossal Clean Crawled Corpus) với tổng kích thước là 750GB. Bộ dữ liệu này sau khi tiến hành xử lý, thu được 100GB, theo các bước như sau: i) giữ lại các văn bản có kết thúc bằng các dấu kết thúc câu (dấu chấm, dấu chấm than, dấu chấm hỏi,...); ii) loại bỏ các văn bản ít hơn ba câu và chỉ giữ lại các câu có ít nhất ba từ; iii) loại bỏ các văn bản có chứa các từ thô tục, nhạy cảm; iv) loại bỏ dòng có chứa mã JavaScript; v) loại bỏ các mã (code) có xuất hiện trong các văn bản.

Bước 2: xây dựng kiến trúc huấn luyện cho ViettelLLM;

Việc thiết kế bộ từ vựng tác động đáng kể đến hiệu quả huấn luyện và khả năng của các nhiệm vụ tầng dưới (downstream tasks). Trong sáng chế này, bộ từ vựng (vocab) được huấn luyện từ đầu sử dụng thuật toán BPE trên dữ hai ngôn ngữ tiếng Anh và tiếng Việt. Cụ thể, bộ từ vựng được thiết kế có kích thước là 50K nhằm đảm bảo khả năng tách từ hiệu quả trên cả tiếng Anh và tiếng Việt.

Kiến trúc của mô hình ngôn ngữ trong sáng chế áp dụng kiến trúc của mô hình Transformer Decoder, với một số thay đổi nhỏ. Chi tiết kiến trúc của LLM trong sáng chế

được thể hiện trong Hình 1. Cụ thể, những đặc điểm nổi bật của kiến trúc mô hình ViettelLLM được mô tả như sau:

- Kích thước mô hình: 7 tỉ tham số với 32 mạch giải mã (Decoder)
- Chiều dài vectơ: 4096
- Mã hóa vị trí: sử dụng vị trí mã hóa xoay (Rotary Positional Embedding, RoPE) cho phép khả năng ngoại suy tốt với ngữ cảnh dài.
- Chuẩn hóa: sử dụng cơ chế tiền chuẩn hóa (pre-normalization) và lớp chuẩn hóa RMSNorm. Phương pháp chuẩn hóa này cho phép giảm khối lượng tính toán nhưng tăng độ chính xác so với hậu chuẩn hóa (post-normalization) với lớp chuẩn hóa LayerNorm thường thấy.
- Hàm kích hoạt: hàm SwiGLU được sử dụng thay vì hàm kích hoạt ReLU thường thấy vì có độ chính xác cao hơn.

Bước 3: huấn luyện mô hình ngôn ngữ lớn ViettelLLM;

Hạ tầng sử dụng để huấn luyện LLM của sáng chế bao gồm bốn cụm máy DGX, mỗi cụm máy bao gồm tám NVIDIA A100 Tensor Core GPU 40GB. Để huấn luyện LLM với kích thước 7 tỷ tham số, sáng chế đã áp dụng kỹ thuật song song hóa dữ liệu (data parallelism) tiên tiến nhằm giảm yêu cầu bộ nhớ để có hạ tầng có thể tải được kích thước mô hình. Cụ thể, sáng chế áp dụng kỹ thuật ZeRO 1 với ý tưởng chính là khai thác sự dư thừa bộ nhớ trong việc lưu trữ các biến trạng thái của bộ tối ưu (optimizer states) khi huấn luyện song song dữ liệu. Ngoài ra, sáng chế cũng áp dụng những cải tiến mới nhất về giao tiếp nhanh giữa các GPU để cải thiện thông lượng và tăng khối lượng giao tiếp (Infiniband intra-node, NVLink inter-node). Sau bước này, mô hình ngôn ngữ lớn dành cho tiếng Việt được tạo ra, được nhóm sáng chế đặt tên là mô hình ngôn ngữ ViettelLLM.

Bước 4: xây dựng ứng dụng ViettelLLM-Chat sử dụng mô hình ngôn ngữ lớn ViettelLLM;

Các mô hình trò chuyện (chat) được tinh chỉnh (finetune) dựa trên LLM có thể được sử dụng để giải quyết nhiều bài toán xử lý ngôn ngữ tự nhiên, có thể kể đến như dịch thuật,

hỏi đáp, nhận diện thực thể có tên, phân loại văn bản... So với các phương pháp truyền thống trước đây khi tiếp cận các bài toán này, cần phải chuẩn bị dữ liệu chuyên biệt cho từng bài toán cụ thể, LLM có khả năng linh hoạt giải quyết nhiều tác vụ khác nhau mà không yêu cầu dữ liệu huấn luyện riêng cho từng tác vụ đó. Tuy nhiên để LLM có thể tổng quát hóa hiệu quả như vậy, nó cần được tinh chỉnh với dữ liệu chỉ dẫn đa dạng để trở thành LLM-chat, gồm hai thành phần chính: phần đề dẫn (prompt) mô tả nhiệm vụ mà mô hình cần thực hiện và phần phản hồi là kết quả mong muốn. Vì vậy, quá trình xây dựng ViettelLLM-Chat bao gồm hai bước là *xây dựng bộ dữ liệu chỉ dẫn* và *huấn luyện mô hình chỉ dẫn*, được mô tả cụ thể như sau:

- *Xây dựng bộ dữ liệu chỉ dẫn*: do chưa có các bộ dữ liệu chỉ dẫn được xây dựng cho tiếng Việt, nhóm tác giả sáng chế xây dựng bộ dữ liệu bằng cách tổng hợp và dịch các bộ dữ liệu chỉ dẫn tiếng Anh. Thêm vào đó, để bộ dữ liệu chỉ dẫn tiếng Việt có được những đặc điểm như mong muốn, nguồn dữ liệu tiếng Anh cũng phải được chọn lọc với các tiêu chí sau:
    - Được xây dựng bởi con người. Ví dụ: Dolly (3k)
    - Chỉ dẫn đơn giản, có nhiệm vụ đa dạng. Ví dụ: Alpaca (27k)
    - Chỉ dẫn phức tạp, có nhiệm vụ đa dạng. Ví dụ: WizardLM (5k)
    - Các chỉ dẫn tập trung vào khả năng suy luận, tổng hợp thông tin từ một ngữ cảnh cho trước (hỏi đáp, tóm tắt). Ví dụ: Orca, Dolphin (15k).
- *Huấn luyện mô hình chỉ dẫn*: về bản chất, việc tinh chỉnh và tiền huấn luyện LLM đều là cùng giải quyết bài toán mô hình hóa ngôn ngữ nhân quả (Causal Language Modeling, CLM). Tuy nhiên, để việc tinh chỉnh hiệu quả mô hình cần tập trung vào việc học phản hồi thay vì học cả đề dẫn. Do đó nhóm tác giả khi tinh chỉnh LLM chỉ tính hàm mất mát với các nội dung thuộc phần phản hồi.

Sau bước này, thu được bộ dữ liệu chỉ dẫn bao gồm 100 nghìn mẫu dữ liệu chỉ dẫn, chia đều cho hai loại ngôn ngữ tiếng Anh và tiếng Việt, mỗi mẫu bao gồm một cặp *#chỉ_dẫn* và *#phản_hồi*. Bộ dữ liệu chỉ dẫn sau đó được huấn luyện theo cơ chế học có giám sát

(supervised learning) để tạo ra ứng dụng trợ lý ảo, ViettelLLM-Chat có khả năng trò chuyện, cung cấp thông tin hữu ích cho người dùng.

Ví dụ thực hiện sáng chế

Giải pháp đã được đưa vào thử nghiệm trên tập dữ liệu thử nghiệm (text dataset), là các văn bản trên Wikipedia, với khoảng 400 trang wiki được thống kê lượt xem nhiều nhất trong năm 2022 bao gồm 200 trang tiếng Anh và 200 trang tiếng Việt (nguồn tham khảo: https://pageviews.wmcloud.org/topviews) để đánh giá khả năng sinh từ tự nhiên của mô hình ngôn ngữ lớn. Ngoài ra, để đánh giá khả năng ứng dụng của mô hình LLM-Chat, bộ dữ liệu bộ dữ liệu thi tốt nghiệp THPT Việt Nam, VNHSGE (nguồn tham khảo: https://github.com/Xdao85/VNHSGE), được sử dụng để thực hiện việc đánh giá ứng dụng trợ lý ảo của sáng chế.

Về thông số đánh giá hiệu quả, sử dụng hai độ do: i) độ đo phức tạp (Perplexity) để đánh giá khả năng sinh nội dung của mô hình ngôn ngữ lớn (LLM); và ii) độ đo trung bình điểm (Average Score) để đánh giá khả năng đưa ra đáp án đúng của ứng dụng trợ lý ảo sử dụng mô hình ngôn ngữ lớn (LLM-Chat). Đây là những độ đo phổ biến được sử dụng rộng rãi. Cụ thể, các thông số đánh giá trên được mô tả cụ thể như sau:

- Độ đo phức tạp (Perplexity): là một trong những độ đo phổ biến nhất để đánh giá các mô hình ngôn ngữ, được định nghĩa là trung bình lũy thừa hàm đối log hợp lý (negative log-likelihood) của một chuỗi đầu vào. Cụ thể, với chuỗi cho trước:$X = \{x1, x2, .. xn\}$, Perplexity của X được tính toán như sau:

$$Perplexity(X) = exp\{-\frac{1}{t}\sum_{i}^{t} log\, p_\theta(x_i|x_{<i})\}$$

  Trong đó $log\ p_\theta(x_i|x_{<i})$ là hàm log hợp lý (log likehood) của từ thứ x_i so với các từ đứng trước nó. Có thể nhận thấy rằng, giá trị độ đo phức tạp (*perplexity*) càng bé thì mô hình ngôn ngữ lớn càng tốt.

- Độ đo điểm trung bình (Average Score): là độ đo để đánh giá tính chính xác của mô hình. Cụ thể, đối với thử nghiệm của sáng chế này, độ đo điểm trung bình được

tính bằng tổng số đáp án đúng trên tổng số tất cả các đáp án. Có thể nhận thấy rằng, giá trị độ đo điểm trung bình (*Average Score*) càng lớn thì mô hình trò chuyện (Chat) càng tốt.

Về phương pháp đánh giá so sánh, tiến hành thực nghiệm trên những mô hình ngôn ngữ lớn mã nguồn mở phổ biến nhất có hỗ trợ tiếng Việt ở thời điểm hiện tại là mô hình Bloom, mô hình Llama-2, mô hình Sealion dành riêng cho ngôn ngữ các nước Đông Nam Á, và mô hình PhoGPT dành riêng cho tiếng Việt. Trong đó, một số mô hình có hỗ trợ ứng dụng trò chuyện (Chat) bao gồm Bloomz-mt và Llama-2-Chat, để so sánh kết quả với các nghiên cứu đề xuất trong sáng chế này.

Kết quả đánh giá mô hình ngôn ngữ lớn và ứng dụng trợ lý ảo sử dụng mô hình ngôn ngữ lớn cho thấy sự hiệu quả của phương pháp đề xuất, cụ thể như sau:

Dữ liệu	Mô hình	Độ đo phức tạp (Perplexity)	Độ đo trung bình điểm (Average Score)
200 trang Wikipedia tiếng Anh	PhoGPT	-	-
	Sealion	10,96	-
	Llama-2	5,70	-
	Bloom	13,83	-
	ViettelLLM	16,98	-
200 trang Wikipedia tiếng Việt	PhoGPT	7,65	-
	Sealion	6,98	-
	Llama-2	-	-

	Bloom	8,39	-
	ViettelLLM	4,66	-
VNHSGE	Llama-2-Chat	-	23,44
	Bloomz-mt	-	23,50
	ViettelLLM-Chat	-	32,16

Hiệu quả đạt được của sáng chế

Ưu điểm đặc biệt liên quan đến sáng chế này đó là xây dựng được mô hình ngôn ngữ lớn từ được huấn luyện với số lượng lớn dữ liệu tiếng Việt, từ đó làm nền tảng cho ứng dụng trợ lý ảo nói riêng, và các tác vụ xử lý ngôn ngữ tự nhiên nói chung. Ngoài ra, nhờ được huấn luyện trên hai ngôn ngữ tiếng Anh và tiếng Việt, mô hình ngôn ngữ có khả năng hiểu tốt hai ngôn ngữ và bộ mã hóa tối ưu cho cả tiếng anh và tiếng việt.

Để triển khai ứng dụng trợ lý ảo sử dụng mô hình ngôn ngữ lớn, sáng chế đã đưa ra phương pháp xây dựng, thiết kế bộ dữ liệu chỉ dẫn và đưa mô hình ngôn ngữ đã được tinh chỉnh vào ứng dụng trò chuyện (Chat). Cụ thể với phương pháp đề xuất, có thể thực hiện bài toán hỏi đáp miền đóng với ngữ cảnh dài, nhiều nhiệm vụ đơn giản đến phức tạp với độ chính xác cao và có câu trả lời tự nhiên.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

Yêu cầu bảo hộ

1. Phương pháp xây dựng mô hình ngôn ngữ lớn và ứng dụng trợ lý ảo dựa trên mô hình ngôn ngữ lớn cho tiếng Việt bao gồm các bước:

bước 1: thu thập và xử lý dữ liệu tiền huấn luyện hai ngôn ngữ tiếng Anh và tiếng Việt;

tại bước này, tập trung vào thu thập và xử lý dữ liệu để huấn luyện mô hình ngôn ngữ lớn có tổng kích thước là 500GB trong đó 400GB tiếng Việt và 100GB tiếng Anh; cụ thể, dữ liệu tiếng Việt được thu thập từ những văn bản được tổng hợp từ internet như các trang báo điện tử uy tín, wiki, pháp luật Việt Nam với tổng số là 800GB; sau khi xử lý, dữ liệu tiếng Việt có kích thước 500GB; theo đó, bộ dữ liệu được xử lý theo các bước như sau: i) tách tập dữ liệu theo mức văn bản; ii) loại bỏ các câu ít hơn ba từ, nhằm loại bỏ các câu không có nhiều ý nghĩa; iii) loại bỏ các văn bản có ít hơn ba câu, nhằm loại bỏ các văn bản không có nhiều thông tin; iv) loại bỏ các văn bản có chứa các ký tự cảm xúc (emoji); v) loại bỏ các văn bản chứa các từ thô tục; dữ liệu tiếng Anh được thu thập từ bộ dữ liệu C4 (the colossal clean crawled corpus) với tổng kích thước là 750GB; bộ dữ liệu này sau khi tiến hành xử lý, thu được 100GB, theo các bước như sau: i) giữ lại các văn bản có kết thúc bằng các dấu kết thúc câu (dấu chấm, dấu chấm than, dấu chấm hỏi,…); ii) loại bỏ các văn bản ít hơn ba câu và chỉ giữ lại các câu có ít nhất ba từ; iii) loại bỏ các văn bản có chứa các từ thô tục, nhạy cảm; iv) loại bỏ dòng có chứa mã JavaScript; v) loại bỏ các mã (code) có xuất hiện trong các văn bản;

bước 2: xây dựng kiến trúc huấn luyện cho ViettelLLM;

việc thiết kế bộ từ vựng tác động đáng kể đến hiệu quả huấn luyện và khả năng của các nhiệm vụ tầng dưới (downstream tasks); trong sáng chế này, bộ từ vựng (vocab) được huấn luyện từ đầu sử dụng thuật toán BPE trên dữ hai ngôn ngữ tiếng Anh và tiếng Việt; cụ thể, bộ từ vựng được thiết kế có kích thước là 50K nhằm đảm bảo khả năng tách từ hiệu quả trên cả tiếng Anh và tiếng Việt; cụ thể, những đặc điểm nổi bật của kiến trúc mô hình ViettelLLM được mô tả như sau:

kích thước mô hình: 7 tỉ tham số với 32 mạch giải mã (decoder);

chiều dài vectơ: 4096;

mã hóa vị trí: sử dụng vị trí mã hóa xoay (rotary positional embedding, RoPE) cho phép khả năng ngoại suy tốt với ngữ cảnh dài;

chuẩn hóa: sử dụng cơ chế tiền chuẩn hóa (pre-normalization) và lớp chuẩn hóa RMSNorm; phương pháp chuẩn hóa này cho phép giảm khối lượng tính toán nhưng tăng độ chính xác so với hậu chuẩn hóa (post-normalization) với lớp chuẩn hóa layernorm thường thấy;

hàm kích hoạt: hàm SwiGLU được sử dụng thay vì hàm kích hoạt ReLU thường thấy vì có độ chính xác cao hơn;

bước 3: huấn luyện mô hình ngôn ngữ lớn ViettelLLM;

hạ tầng sử dụng để huấn luyện LLM của sáng chế bao gồm bốn cụm máy DGX, mỗi cụm máy bao gồm tám NVIDIA A100 tensor core GPU 40GB; để huấn luyện LLM với kích thước 7 tỷ tham số, sáng chế đã áp dụng kỹ thuật song song hóa dữ liệu (data parallelism) tiên tiến nhằm giảm yêu cầu bộ nhớ để có hạ tầng có thể tải được kích thước mô hình; cụ thể, sáng chế áp dụng kỹ thuật ZeRO 1 với ý tưởng chính là khai thác sự dư thừa bộ nhớ trong việc lưu trữ các biến trạng thái của bộ tối ưu (optimizer states) khi huấn luyện song sọng dữ liệu; ngoài ra, sáng chế cũng áp dụng những cải tiến mới nhất về giao tiếp nhanh giữa các GPU để cải thiện thông lượng và tăng khối lượng giao tiếp (infiniband intra-node, NVLink inter-node); sau bước này, mô hình ngôn ngữ lớn dành cho tiếng Việt được tạo ra, được nhóm sáng chế đặt tên là mô hình ngôn ngữ ViettelLLM;

bước 4: xây dựng ứng dụng ViettelLLM-Chat sử dụng mô hình ngôn ngữ lớn ViettelLLM;

các mô hình trò chuyện (chat) được tinh chỉnh (finetune) dựa trên LLM có thể được sử dụng để giải quyết nhiều bài toán xử lý ngôn ngữ tự nhiên, có thể kể đến như dịch thuật, hỏi đáp, nhận diện thực thể có tên, phân loại văn bản; để LLM có thể tổng quát hóa hiệu quả như vậy, nó cần được tinh chỉnh với dữ liệu chỉ dẫn đa dạng để trở thành LLM-chat,

gồm hai thành phần chính: phần đề dẫn (prompt) mô tả nhiệm vụ mà mô hình cần thực hiện và phần phản hồi là kết quả mong muốn, vì vậy, quá trình xây dựng ViettelLLM-Chat bao gồm hai bước là *xây dựng bộ dữ liệu chỉ dẫn* và *huấn luyện mô hình chỉ dẫn*, được mô tả cụ thể như sau:

xây dựng bộ dữ liệu chỉ dẫn: do chưa có các bộ dữ liệu chỉ dẫn được xây dựng cho tiếng Việt, nhóm tác giả sáng chế xây dựng bộ dữ liệu bằng cách tổng hợp và dịch các bộ dữ liệu chỉ dẫn tiếng Anh; thêm vào đó, để bộ dữ liệu chỉ dẫn tiếng Việt có được những đặc điểm như mong muốn, nguồn dữ liệu tiếng Anh cũng phải được chọn lọc với các tiêu chí sau:

được xây dựng bởi con người;

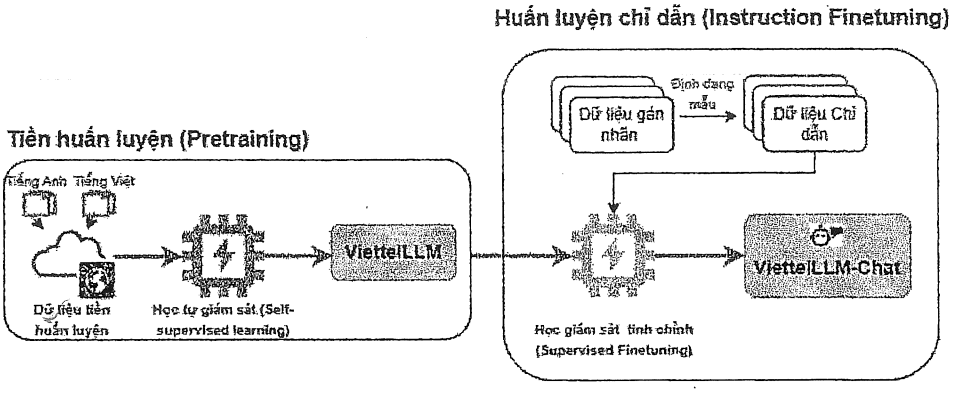
chỉ dẫn đơn giản, có nhiệm vụ đa dạng;

chỉ dẫn phức tạp, có nhiệm vụ đa dạng;

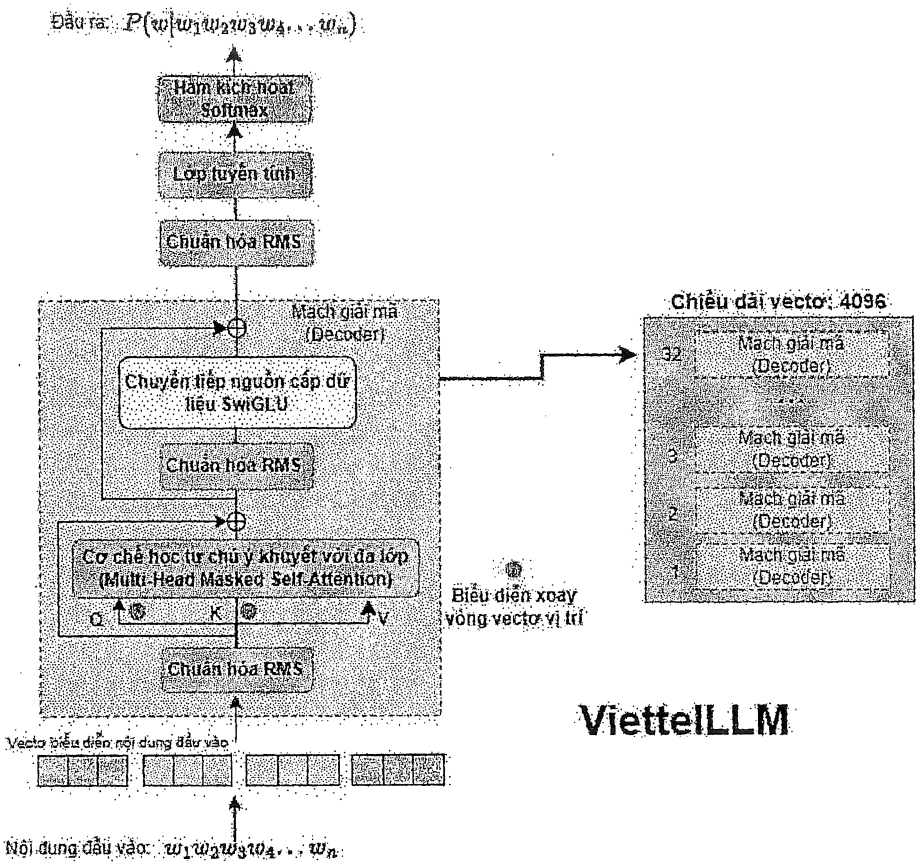
các chỉ dẫn tập trung vào khả năng suy luận, tổng hợp thông tin từ một ngữ cảnh cho trước (hỏi đáp, tóm tắt);

huấn luyện mô hình chỉ dẫn: về bản chất, việc tinh chỉnh và tiền huấn luyện LLM đều là cùng giải quyết bài toán mô hình hóa ngôn ngữ nhân quả (causal language modeling, CLM); tuy nhiên, để việc tinh chỉnh hiệu quả mô hình cần tập trung vào việc học phản hồi thay vì học cả đề dẫn; do đó nhóm tác giả khi tinh chỉnh LLM chỉ tính hàm mất mát với các nội dung thuộc phần phản hồi;

sau bước này, thu được bộ dữ liệu chỉ dẫn bao gồm 100 nghìn mẫu dữ liệu chỉ dẫn, chia đều cho hai loại ngôn ngữ tiếng Anh và tiếng Việt, mỗi mẫu bao gồm một cặp *#chỉ_dẫn* và *#phản_hồi*; bộ dữ liệu chỉ dẫn sau đó được huấn luyện theo cơ chế học có giám sát (supervised learning) để tạo ra ứng dụng trợ lý ảo, ViettelLLM-Chat có khả năng trò chuyện, cung cấp thông tin hữu ích cho người dùng.


Huấn luyện chỉ dẫn (Instruction Finetuning)
Tiền huấn luyện (Pretraining)
Tiếng Anh Tiếng Việt
Dữ liệu tiền huấn luyện
Học tự giám sát (Self-supervised learning)
ViettelLLM
Định dạng mẫu
Dữ liệu gán nhãn
Dữ liệu Chỉ dẫn
ViettelLLM-Chat
Học giám sát tinh chỉnh (Supervised Finetuning)

Hình 1


Đầu ra: $P(w|w_1w_2w_3w_4...w_n)$
Hàm kích hoạt Softmax
Lớp tuyến tính
Chuẩn hóa RMS
Mạch giải mã (Decoder)
Chuyển tiếp nguồn cấp dữ liệu SwiGLU
Chuẩn hóa RMS
Cơ chế học tự chú ý khuyết với đa lớp (Multi-Head Masked Self-Attention)
Q
K
V
Chuẩn hóa RMS
Vecto biểu diễn nội dung đầu vào
Nội dung đầu vào: $w_1w_2w_3w_4...w_n$
Biểu diễn xoay vòng vectơ vị trí
Chiều dài vectơ: 4096
32
3
2
1
Mạch giải mã (Decoder)
ViettelLLM

Hình 2