Finetuning BERT cho hai bài toán Phân tích Cảm xúc và Phân loại Chủ đề

Đồng Mạnh Hùng Đoàn Quang Huy Nguyễn Đình Khải

Viện Trí tuệ Nhân tạo, Trường Đại học Công nghệ, ĐHQGHN

Ngày 10 tháng 06 năm 2025

Tóm tắt: BERT (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ tiền huấn luyện mạnh mẽ, có khả năng hiểu ngữ cảnh hai chiều của từ trong câu. Trong nghiên cứu này, chúng tôi trình bày quy trình fine-tuning BERT cho hai tác vụ phân loại văn bản phổ biến: phân tích cảm xúc và phân loại chủ đề văn bản. Mô hình được tinh chỉnh bằng cách thêm một tầng phân loại trên token [CLS], sử dụng các siêu tham số phù hợp. Kết quả thử nghiệm cho thấy BERT đạt độ chính xác cao, trong đó đối với bài toán phân tích cảm xúc đạt độ chính xác 87,74% và bài toán phân loại chủ đề văn bản đạt .

Từ khóa: BERT, fine-tuning, phân tích cảm xúc, phân loại chủ đề, NLP, Transformers

1. Giới thiệu

Trong thời đại bùng nổ dữ liệu số và sự phát triển của trí tuệ nhân tạo, việc hiểu và xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) trở thành một trong những hướng nghiên cứu và ứng dụng quan trọng. Các bài toán như phân tích cảm xúc (Sentiment Analysis), phân loại chủ đề (Topic Classification) hay nhận dạng thực thể có tên (Named Entity Recognition – NER) đều đòi hỏi mô hình phải hiểu được ngữ nghĩa và ngữ cảnh phức tạp trong ngôn ngữ.

Sự ra đời của các mô hình ngôn ngữ tiền huấn luyện như ELMo (Peters et al., 2018a), Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018) và đặc biệt là BERT (Bidirectional Encoder Representations from Transformers) đã tạo ra bước ngoặt lớn trong nghiên cứu NLP. BERT, do Google đề xuất năm 2018, là mô hình học sâu dựa trên kiến trúc Transformer, được huấn luyện trên lượng lớn dữ liệu văn bản theo cách không có giám sát (unsupervised), với hai nhiệm vụ chính là Masked Language Modeling (MLM) và Next Sentence Prediction (NSP). Ưu điểm vượt trội của BERT là khả năng học biểu diễn ngữ nghĩa hai chiều (bidirectional) của từ trong câu, giúp mô hình hiểu rõ hơn mối quan hệ giữa các từ và cụm từ trong ngữ cảnh rộng hơn.

Một trong những ưu điểm chính của BERT là khả năng "fine-tuning"– tức là điều chỉnh mô hình đã được tiền huấn luyện để áp dụng cho các bài toán cụ thể, chỉ với một lượng dữ liệu nhỏ và thời gian huấn luyện ngắn.

Trong bài báo này, chúng tôi tập trung nghiên cứu việc fine-tuning mô hình BERT cho hai tác vụ : (i) phân tích cảm xúc (Sentiment Analysis), giúp xác định thái độ của người viết đối với chủ đề được đề cập (tích cực, tiêu cực); và (ii) phân loại chủ đề (Topic Classification), nhằm xác định lĩnh vực hoặc nội dung chính của văn bản (ví dụ: thể thao, kinh tế, giáo dục...).

Chúng tôi tiến hành thực nghiệm trên các tập dữ liệu tiếng Anh, tiến hành tiền huấn luyện mô hình, và áp dụng kỹ thuật fine-tuning với tầng phân loại đơn giản trên token đặc biệt [CLS]. Ngoài ra, chúng

tôi tiến hành phân tích hiệu năng mô hình thông qua các chỉ số như độ chính xác, F1-score, đồng thời so sánh với các phương pháp khác.

Kết quả cho thấy BERT đạt hiệu suất vượt trội trong cả hai bài toán, khẳng định vai trò quan trọng của các mô hình Transformer trong việc phân tích ngữ nghĩa trong câu. Những phát hiện này mở ra tiềm năng ứng dụng BERT trong các hệ thống phân tích văn bản thực tế như phân tích dư luận xã hội, lọc nội dung, chăm sóc khách hàng tự động và đề xuất thông tin.

2. Các nghiên cứu liên quan

BERT (Bidirectional Encoder Representations from Transformers), được giới thiệu bởi Devlin và cộng sự [**devlin2018bert**], là một bước tiến quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên. BERT sử dụng kiến trúc Transformer hai chiều và được huấn luyện trước trên tập dữ liệu lớn, cho phép mô hình hiểu ngữ cảnh của từ theo cả hai chiều trái và phải. Kể từ khi ra đời, BERT đã đạt được kết quả vượt trội trên nhiều tác vụ NLP như phân tích cảm xúc, phân loại văn bản, trả lời câu hỏi và nhận dạng thực thể.

Nghiên cứu của Sun và cộng sự [**sun2019fine**] đã tập trung vào việc tìm hiểu cách fine-tuning BERT một cách hiệu quả cho các bài toán phân loại văn bản. Tác giả đã tiến hành đánh giá ảnh hưởng của các siêu tham số như kích thước batch, tốc độ học và số epoch đến hiệu quả của mô hình, đồng thời đề xuất các cấu hình tối ưu để đạt kết quả tốt nhất.

Trong khuôn khổ nghiên cứu này, chúng tôi thực hiện lại quá trình fine-tuning được đề xuất bởi Sun và cộng sự, áp dụng cho hai tác vụ phổ biến là phân tích cảm xúc và phân loại chủ đề trên dữ liệu tiếng Anh. Mục tiêu là kiểm chứng khả năng tái tạo kết quả và đánh giá mức độ tổng quát hóa của phương pháp trong bối cảnh khác với nghiên cứu gốc.

3. Tiền huấn luyện mô hình và tinh chỉnh mô hình BERT

Mô hình BERT được phát triển dựa trên kiến trúc Transformer và được huấn luyện qua hai giai đoạn: tiền huấn luyện mô hình (pre-training) và tinh chỉnh (fine-tuning).

3.1. Tiền huấn luyện mô hình (Pre-training)

Trong giai đoạn huấn luyện sơ cấp, BERT học các biểu diễn ngữ nghĩa tổng quát từ văn bản lớn thông qua hai nhiệm vụ chính:

- **Masked Language Modeling (MLM):** Một tỷ lệ nhất định các token (thường là 15%) trong câu đầu vào được che đi (mask), và mô hình được yêu cầu dự đoán các token bị che đó dựa trên ngữ cảnh hai chiều.
- **Next Sentence Prediction (NSP):** Cho hai câu A và B, mô hình dự đoán liệu câu B có phải là câu tiếp theo của câu A trong văn bản gốc hay không. Mục tiêu là giúp mô hình học được mối quan hệ giữa các câu.

Giai đoạn pre-training của BERT sử dụng các tập dữ liệu lớn như Wikipedia và BookCorpus, giúp mô hình học được kiến thức tổng quát về ngôn ngữ.

3.2. Tinh chỉnh mô hình (Fine-tuning)

Sau khi được huấn luyện sơ cấp, BERT có thể được tinh chỉnh cho các tác vụ cụ thể như phân loại văn bản, phân tích cảm xúc, hoặc trả lời câu hỏi.

Quá trình tinh chỉnh thường bao gồm việc:

- Gắn thêm một tầng đầu ra phù hợp với tác vụ (ví dụ: softmax cho phân loại).
- Huấn luyện lại toàn bộ mô hình với một tập dữ liệu có nhãn.
- Sử dụng các siêu tham số thích hợp như learning rate thấp, batch size nhỏ và số epoch vừa phải.

Với cách tiếp cận này, BERT đạt được hiệu suất rất cao chỉ với một lượng nhỏ dữ liệu huấn luyện trong giai đoạn fine-tuning.

4. Phương pháp nghiên cứu

Trong nghiên cứu này, chúng tôi thực hiện hai nhiệm vụ chính: đầu tiên, tiến hành tiền huấn luyện mô hình ngôn ngữ trên tập dữ liệu Wikipedia tiếng Anh để học các biểu diễn ngữ nghĩa tổng quát; sau đó, chúng tôi áp dụng phương pháp fine-tuning dựa trên mô hình đã tiền huấn luyện cho hai tác vụ cụ thể: phân tích cảm xúc và phân loại chủ đề văn bản. Tập dữ liệu dùng cho tác vụ phân loại chủ đề được xây dựng dựa trên tập dữ liệu Yahoo Answers do Xiang Zhang và cộng sự phát triển, từng được sử dụng trong nghiên cứu về *Character-level Convolutional Networks for Text Classification*, công bố tại Hội thảo về Hệ thống xử lý thông tin thần kinh (Neural Information Processing Systems - NIPS) năm 2015 [**zhang2015character**].

4.1. Dữ liệu và tiền xử lý

Trong nghiên cứu này, chúng tôi sử dụng ba tập dữ liệu tiếng Anh bao gồm hai tập để huấn luyện và đánh giá mô hình phân loại, và một tập dùng để tiền huấn luyện mô hình Transformer từ đầu:

Dữ liệu tiền huấn luyện: Wikipedia tiếng Anh (20220301)

Dữ liệu tiền huấn luyện được sử dụng trong nghiên cứu này là tập Wikipedia tiếng Anh phiên bản tháng 3 năm 2022, được cung cấp thông qua nền tảng Hugging Face tại địa chỉ `https://huggingface.co/datasets/wikipedia`. Đây là bản trích xuất đã được xử lý trước, bao gồm nội dung văn bản thuần từ các bài viết trên Wikipedia, loại bỏ các thẻ đánh dấu định dạng, hình ảnh, và siêu liên kết, giúp thuận tiện cho việc sử dụng trong huấn luyện mô hình ngôn ngữ. Tập dữ liệu này bao gồm hàng triệu bài viết với độ dài và chủ đề phong phú, phản ánh đa dạng các lĩnh vực từ khoa học, công nghệ đến văn hóa và xã hội.

Dữ liệu phân tích cảm xúc: Amazon Review Polarity

Tập dữ liệu được xây dựng từ các đánh giá sản phẩm trên Amazon. Các đánh giá 1 sao và 2 sao được gán nhãn tiêu cực (label = 1), trong khi các đánh giá 4 sao và 5 sao được gán nhãn tích cực (label = 2). Những đánh giá 3 sao bị loại bỏ để đảm bảo tính phân cực rõ ràng trong phân tích cảm xúc. Mỗi nhãn gồm khoảng 1,8 triệu mẫu huấn luyện và 200.000 mẫu kiểm tra. Dữ liệu được lưu dưới dạng `.csv` với ba cột chính: `Class Index` biểu thị nhãn phân loại (1 hoặc 2), `Review Title` là tiêu đề ngắn gọn của đánh giá, và `Review Text` chứa nội dung đầy đủ của đánh giá.

Dữ liệu phân loại chủ đề: Yahoo Answers

Tập dữ liệu Yahoo Answers bao gồm các câu hỏi người dùng được phân loại vào 10 chủ đề chính: Văn hóa – Xã hội, Khoa học – Toán học, Sức khỏe, Giáo dục – Tham khảo, Máy tính – Internet, Thể thao, Kinh doanh – Tài chính, Giải trí – Âm nhạc, Gia đình – Mối quan hệ, và Chính trị. Mỗi chủ đề có khoảng 140.000 mẫu huấn luyện và 6.000 mẫu kiểm tra. Dữ liệu được lưu ở định dạng `.csv` với bốn cột: `Class Index` là nhãn chủ đề từ 1 đến 10, `Question Title` là tiêu đề câu hỏi, `Question Content` chứa nội dung chi tiết của câu hỏi, và `Best Answer` là câu trả lời được đánh giá cao nhất. Chỉ những câu trả lời tốt nhất được giữ lại nhằm đảm bảo chất lượng của tập dữ liệu.

Tiền xử lý dữ liệu

Vì giới hạn về tài nguyên, tập dữ liệu chúng tôi sử dụng cho bài toán này chỉ là một phần nhỏ so với tập dữ liệu gốc. Cụ thể, đối với dữ liệu dùng cho quá trình

Tham số	Tiền huấn luyện	Cảm xúc	Văn bản
Batch size	32	32	32
Hidden size	256	256	256
Số lớp ẩn	6	6	6
Số đầu attention	8	8	8
Kích thước trung gian	1024	1024	1024
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Số epoch	4	3	4
Optimizer	Adam	Adam	Adam

Bảng 1: **Tham số huấn luyện cho các giai đoạn: tiền huấn luyện, phân loại cảm xúc và phân loại văn bản.**

tiền huấn luyện mô hình, chúng tôi chỉ lấy khoảng 1% so với tập dữ liệu Wikipedia phiên bản ngày 2022-03-01 tiếng Anh (`20220301.en`). Đối với dữ liệu dùng cho tinh chỉnh mô hình, chúng tôi sử dụng khoảng 1/15 kích thước tập dữ liệu gốc cho bài toán phân tích cảm xúc và ...

Toàn bộ văn bản từ các tập dữ liệu đều được xử lý bằng tokenizer `bert-base-uncased` của BERT từ thư viện Hugging Face Transformers. Các văn bản được mã hóa thành chuỗi token, thêm token đặc biệt, padding đến độ dài cố định (max length = 512), nhằm chuẩn hóa đầu vào cho mô hình.

Đối với tập dữ liệu Wikipedia, chúng tôi tiền huấn luyện mô hình Transformer với mục tiêu Masked Language Modeling (MLM) – một kỹ thuật học biểu diễn ngữ nghĩa sâu bằng cách che giấu ngẫu nhiên một số token trong văn bản và yêu cầu mô hình dự đoán lại chúng dựa trên ngữ cảnh hai chiều. Đây là bước quan trọng nhằm giúp mô hình học được các đặc trưng ngôn ngữ tổng quát trước khi chuyển sang giai đoạn tinh chỉnh cho các tác vụ cụ thể như phân tích cảm xúc và phân loại chủ đề.

4.2. Kiến trúc mô hình

Tiền huấn luyện mô hình

Mô hình được xây dựng dựa trên kiến trúc `BertForSequenceClassification` với tokenizer `bert-base-uncased`. Mặc dù BERT ban đầu sử dụng hai mục tiêu huấn luyện là MLM và Next Sentence Prediction (NSP) [**devlin2019bert**], các nghiên cứu sau đó, điển hình là RoBERTa [**liu2019roberta**], đã chứng minh rằng NSP không đóng góp đáng kể vào hiệu quả mô hình và có thể được loại bỏ mà vẫn đạt hiệu suất cao. Do đó, việc tiền huấn luyện thường tập trung vào tối ưu hóa MLM trước khi tinh chỉnh mô hình cho bài toán phân loại cụ thể. Trong giai đoạn tiền huấn luyện, chúng tôi đã quyết định sử dụng mô hình tuền huấn luyện với mục tiêu Masked Language Modeling (MLM) để học các biểu diễn ngữ nghĩa tổng quát từ ngữ cảnh hai chiều của văn bản.

Tinh chỉnh mô hình

Sau khi tiền huấn luyện, mô hình tiếp tục được tinh chỉnh (fine-tuning) cho tác vụ phân loại cụ thể. Tokenizer và kiến trúc mô hình vẫn giữ nguyên là `bert-base-uncased` và `BertForSequenceClassification` để đảm bảo tính nhất quán trong xử lý dữ liệu và ánh xạ token embedding.

4.3. Cấu hình huấn luyện

Toàn bộ quá trình huấn luyện, bao gồm tiền huấn luyện và tinh chỉnh (fine-tuning), đều được thực hiện trên GPU (Kaggle). Mô hình sử dụng kiến trúc

`BertForSequenceClassification` cùng tokenizer `bert-base-uncased`.

Ở giai đoạn tiền huấn luyện, mô hình được huấn luyện với mục tiêu Masked Language Modeling (MLM) để học các biểu diễn ngữ nghĩa từ dữ liệu đầu vào. Sau khi tiền huấn luyện, mô hình tiếp tục được tinh chỉnh (fine-tuned). Quy trình tinh chỉnh tuân theo đề xuất trong [**sun2019fine**]. Cấu hình tham số của các mô hình được thể hiện trong bảng 1.

Để đánh giá mô hình,chúng tôi dùng 10% dữ liệu huấn luyện được sử dụng làm tập test. Mô hình được đánh giá bằng độ chính xác (accuracy) và F1-score tuỳ theo yêu cầu của từng tác vụ.

5. Kết quả

5.1. Bài toán phân tích cảm xúc

Label	Prec	Rec	F1
Tiêu cực	0.8852	0.8677	0.8763
Tích cực	0.8699	0.8872	0.8784
Accuracy		0.8774	
Macro avg	0.8775	0.8774	0.8774
Weighted avg	0.8775	0.8774	0.8774

Bảng 2: **Kết quả đánh giá mô hình phân tích cảm xúc trên tập dữ liệu Amazon.**

Mô hình phân tích cảm xúc được huấn luyện trên tập dữ liệu amazon đạt độ chính xác tổng thể là 87.74%, cho thấy khả năng phân loại cảm xúc tích cực và tiêu cực hiệu quả. Cụ thể, đối với lớp nhãn 0, *Precision* đạt 88.52%, *Recall* là 86.77% và *F1-score* là 87.63%. Trong khi đó, lớp nhãn 1 ghi nhận *Precision* là 86.99%, *Recall* đạt 88.72% và *F1-score* đạt 87.84%. Các chỉ số trung bình như *Macro average* và *Weighted average* đều ở mức xấp xỉ 87.74%, phản ánh độ cân bằng trong hiệu suất của mô hình giữa các lớp. Nhìn chung, mô hình đã học được các đặc trưng biểu cảm trong văn bản và có khả năng tổng quát hóa tốt trên tập kiểm tra.

5.2. Bài toán phân loại chủ đề văn bản

graphicx

Class	Prec	Rec	F1
Xã hội & Văn hóa	0.5787	0.6371	0.6065
Khoa học & Toán học	0.7279	0.7492	0.7384
Sức khỏe	0.7316	0.8000	0.7643
Giáo dục & Tham khảo	0.5548	0.5340	0.5442
Máy tính & Internet	0.8505	0.8361	0.8432
Thể thao	0.8678	0.8603	0.8640
Kinh doanh & Tài chính	0.6273	0.4641	0.5335
Giải trí & Âm nhạc	0.6961	0.7167	0.7063
Gia đình & Mối quan hệ	0.7240	0.7688	0.7457
Chính trị & Chính phủ	0.7637	0.7666	0.7652
Accuracy			0.7137
Macro Avg	0.7123	0.7133	0.7111
Weighted Avg	0.7124	0.7137	0.7115

Bảng 3: **Kết quả đánh giá mô hình phân loại chủ đề văn bản trên tập dữ liệu Yahoo.**

Bảng 3 trình bày kết quả đánh giá mô hình phân loại chủ đề văn bản trên tập dữ liệu Yahoo với 10 lớp. Mô hình đạt độ chính xác tổng thể (*Accuracy*) là **71.37%**, phản ánh khả năng phân loại tương đối ổn định. Các chỉ số trung bình như *Macro F1-score* và *Weighted F1-score* lần lượt là **0.7111** và **0.7115**, cho thấy mô hình duy trì hiệu suất khá đồng đều giữa các lớp, mặc dù vẫn tồn tại sự chênh lệch nhất định.

Một số lớp có hiệu suất rất cao, nổi bật như: *Thể thao* (F1 = 0.8640), *Máy tính & Internet* (F1 = 0.8432), và *Chính trị & Chính phủ* (F1 = 0.7652). Điều này cho thấy mô hình có khả năng nhận diện tốt các chủ đề có đặc trưng ngôn ngữ rõ rệt.

Tuy nhiên, vẫn còn một số lớp có hiệu suất thấp hơn, chẳng hạn như: *Kinh doanh & Tài chính* (F1 = 0.5335), *Giáo dục & Tham khảo* (F1 = 0.5442), và *Xã hội & Văn hóa* (F1 = 0.6065). Nguyên nhân có thể đến từ sự giao thoa trong ngữ nghĩa giữa các chủ đề, hoặc sự mất cân bằng dữ liệu trong tập huấn luyện.

6. Kết luận

Bài viết này trình bày quá trình fine-tuning mô hình BERT cho hai tác vụ quan

trọng trong xử lý ngôn ngữ tự nhiên: phân tích cảm xúc và phân loại chủ đề. Dựa trên hai bài báo nền tảng, chúng tôi đã tái hiện lại các thí nghiệm và xác nhận hiệu quả mạnh mẽ của BERT trong việc hiểu ngữ nghĩa và cấu trúc ngữ cảnh.

Việc đạt được kết quả cao trên cả hai tập dữ liệu chứng minh rằng fine-tuning BERT là một phương pháp hiệu quả, không chỉ trong môi trường ban đầu của nghiên cứu mà còn có thể áp dụng linh hoạt cho các bài toán phân loại văn bản khác nhau.

Trong tương lai, chúng tôi dự định mở rộng nghiên cứu sang các mô hình ngôn ngữ đa ngôn ngữ (như mBERT hoặc XLM-R) và áp dụng cho các tập dữ liệu tiếng Việt, nhằm đánh giá mức độ thích nghi của mô hình với ngôn ngữ không phải tiếng Anh.

Kết quả cho thấy mô hình BERT fine-tuned đạt độ chính xác cao ở cả hai tác vụ, gần tương đương hoặc cao hơn một chút so với kết quả được báo cáo trong nghiên cứu gốc. Điều này xác nhận tính khả tái lập của phương pháp và khả năng tổng quát hóa tốt của mô hình.

Tài liệu

[1] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805, 2018.

[2] Sun, Chi, Qiu, Xipeng, Xu, Yige, and Huang, Xuanjing. *Fine-tuning BERT for Text Classification with Modified Representations*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[3] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805, 2019.

[4] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692, 2019.

[5] Zhang, Xiang, Zhao, Junbo, and LeCun, Yann. *Character-level convolutional networks for text classification*. In *Advances in Neural Information Processing Systems*, volume 28, 2015.