


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://www.youtube.com/watch?v=Q5AcCu43ze0>
(ví dụ: <https://www.youtube.com/watch?v=AWq7uw-36Ng>)
- Link slides (dạng .pdf đặt trên Github):
(ví dụ: <https://github.com/mynameuit/CS2205.APR2023/TenDeTai.pdf>)
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Vi Minh Huy• MSSV: 220101025 	<ul style="list-style-type: none">• Lớp: CS2205.CH181• Tự đánh giá (điểm tổng kết môn): 9.0/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 1• Link Github:
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT TRIỂN MÔ HÌNH BERT CHO VIỆC HUẤN LUYỆN MÔ HÌNH MÃ HÓA HAI CHIỀU TỪ TRANSFORMER TRONG XỬ LÝ NGÔN NGỮ

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

BERT: PRE-TRAINING OF DEEP BIDIRECTIONAL TRANSFORMERS FOR LANGUAGE

TÓM TẮT *(Tối đa 400 từ)*

Biểu diễn ngôn ngữ là một bước quan trọng trong xử lý ngôn ngữ tự nhiên, giúp máy tính có thể hiểu và xử lý ngôn ngữ con người. Hiện nay có nhiều mô hình biểu diễn ngôn ngữ khác nhau, mỗi mô hình có ưu và nhược điểm riêng. Hai mô hình biểu diễn ngôn ngữ tiêu biểu là ELMo [1] và OpenAI GPT [2]. Mô hình ELMo rất hiệu quả trong việc hiểu ý nghĩa của các từ trong văn bản, mô hình này đã tiếp cận nhiều khía cạnh của ngữ cảnh, nhưng chưa khai thác hoàn toàn ngữ cảnh hai chiều. Còn mô hình OpenAI thì rất hiệu quả trong việc sinh văn bản và có khả năng tạo ra văn bản trôi chảy và tự nhiên nhưng OpenAI chưa có khả năng hiểu mối liên hệ hai chiều giữa các từ.

Như vậy hạn chế của các mô hình hiện tại, thứ nhất là chưa khai thác đầy đủ ngữ cảnh hai chiều. Cả ELMo và GPT đều có những hạn chế trong việc hiểu mối quan hệ giữa các từ trong văn bản. Thứ hai là hạn chế trong việc hiểu sâu các mối quan hệ ngữ nghĩa. Do không khai thác đầy đủ ngữ cảnh, các mô hình hiện tại gặp khó khăn trong việc hiểu ý nghĩa sâu sắc của văn bản.

Do đó vấn đề cấp thiết là cần phát triển các mô hình biểu diễn ngôn ngữ mới có khả năng khai thác đầy đủ ngữ cảnh hai chiều, hiểu sâu các mối quan hệ ngữ nghĩa giữa các từ trong văn bản.

➤ Tính mới của đề tài

Đề tài đề xuất phát triển mô hình BERT (Bidirectional Encoder Representations from Transformers), là mô hình mới giúp giải quyết các hạn chế của các mô hình trước đó bằng cách tích hợp sâu ngữ cảnh từ cả hai hướng trong quá trình biểu diễn đối tượng ngôn ngữ. Đề xuất này dựa vào kiến trúc Transformer [3] cung cấp nền tảng cho việc phát triển các kỹ thuật mới như Masked Language Model (MLM) và Next Sentence Prediction (NSP), để cung cấp cách tiếp cận và xử lý đầy đủ hơn trong việc mô hình hóa ngôn ngữ.

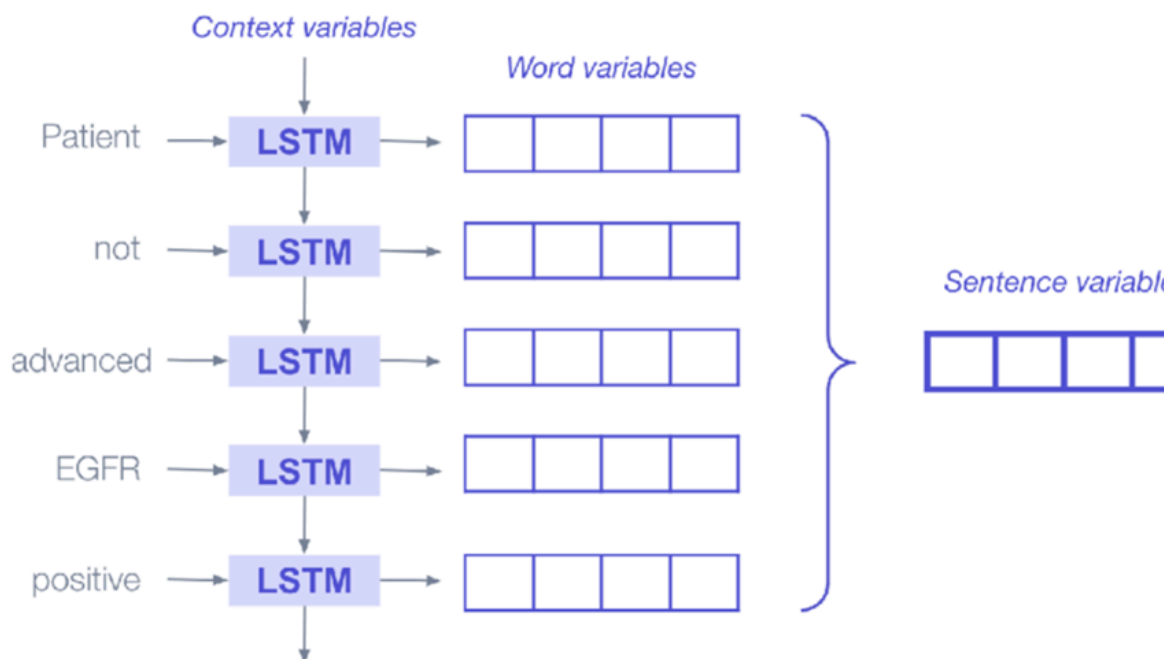


Figure 2: Illustration of bidirectional context [5]

➤ Lợi ích khoa học và thực tiễn khi vấn đề được giải quyết

Việc phát triển BERT sẽ đóng góp quan trọng về lý thuyết và thực tiễn cho lĩnh vực khoa học máy tính và xử lý ngôn ngữ tự nhiên.

➤ Lợi ích khoa học

BERT sẽ mở ra một hướng nghiên cứu mới trong việc xây dựng các mô hình ngôn ngữ có khả năng hiểu ngữ cảnh một cách toàn diện. Mô hình BERT sẽ cung cấp một khung sườn mới cho việc phát triển các mô hình biểu diễn ngôn ngữ tiếp theo, đặc biệt là trong việc xử lý các ngôn ngữ có cấu trúc phức tạp.

➤ Lợi ích thực tiễn

BERT sẽ có khả năng cải thiện đáng kể hiệu suất của các hệ thống NLP trong nhiều ứng dụng từ hỗ trợ trực tuyến, hỏi đáp tự động, đến các hệ thống khai thác thông tin. Ngoài ra, khả năng hiểu ngôn ngữ chính xác và sâu hơn sẽ làm tăng chất lượng tương tác giữa người dùng và máy tính, mở ra các cơ hội mới trong việc thiết kế các giao diện người dùng thông minh hơn và các hệ thống đề xuất nội dung cá nhân hóa.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Phát triển mô hình biểu diễn ngôn ngữ BERT có khả năng khai thác đầy đủ ngữ cảnh hai chiều sử dụng cấu trúc Transformer.
- Đánh giá hiệu suất của BERT trên các tiêu chuẩn như GLUE và SQuAD.
- Tinh chỉnh mô hình BERT để tối ưu hóa cho các tác vụ NLP cụ thể như phân tích cảm xúc, suy diễn ngôn ngữ, hỏi đáp tự động.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

1. **Tiền huấn luyện BERT:** Nội dung này với mục tiêu là phát triển và tiền huấn luyện mô hình BERT sử dụng kiến trúc Transformer với kỹ thuật Masked Language Model (MLM) và Next Sentence Prediction (NSP). Và nội dung này sẽ được thực hiện bằng các phương pháp sau:
2. **Thu thập dữ liệu:** Thu thập một tập hợp dữ liệu văn bản lớn không nhãn từ các

nguồn trực tuyến để tiền huấn luyện.

3. **Áp dụng kiến trúc Transformer:** Sử dụng MLM để ngẫu nhiên che đi một số token và dạy mô hình dự đoán token bị che đó dựa trên ngữ cảnh xung quanh.
4. **Sử dụng NSP :** Để cải thiện khả năng hiểu mối quan hệ giữa các câu.
5. **Tinh chỉnh BERT cho các tác vụ cụ thể:** Nội dung này với mục tiêu là tinh chỉnh mô hình BERT tiền huấn luyện cho các tác vụ NLP cụ thể như phân tích cảm xúc, suy diễn ngôn ngữ, và trả lời câu hỏi. Và nội dung này sẽ được thực hiện bằng các phương pháp:
6. **Sử dụng dữ liệu gán nhãn:** Sử dụng các tập dữ liệu gán nhãn cụ thể cho mỗi tác vụ để tinh chỉnh mô hình.
7. **Áp dụng kỹ thuật tinh chỉnh:** Áp dụng các kỹ thuật tinh chỉnh như điều chỉnh tốc độ học và bổ sung lớp phân lớp cụ thể cho từng tác vụ.
8. **Phân tích đánh giá hiệu quả toàn diện của mô hình:** Nội dung này với mục tiêu là với mục tiêu là đánh giá toàn diện hiệu quả và hạn chế của mô hình BERT qua nhiều tác vụ và ngữ cảnh khác nhau. Và nội dung này sẽ được thực hiện bằng các phương pháp sau đây:
9. **So sánh hiệu suất:** Phân tích hiệu suất của mô hình trên các bộ dữ liệu chuẩn và so sánh với các mô hình tiên tiến khác.
10. **Phân tích nhạy cảm:** Thực hiện phân tích nhạy cảm và kiểm tra độ lỗi để xác định hạn chế.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

Tiền huấn luyện BERT

Vượt trội các mô hình hiện tại trên GLUE (F1-score > 80%), SQuAD 2.0 (F1-score > 83%), WMT 2014 English-to-German translation (BLEU > 30%).

Tinh chỉnh BERT cho các tác vụ cụ thể:

Vượt trội các mô hình chuyên dụng cho từng tác vụ:

- Phân tích cảm xúc (SST-2): F1-score > 85%.
- Suy diễn ngôn ngữ (MNLI): Accuracy > 80%.
- Trả lời câu hỏi (SQuAD 2.0): EM > 70%.

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

[0]Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers) (pp. 4171-4186). Association for Computational Linguistics. ACL Anthology

[1]. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227-2237. Association for Computational Linguistics.

[2]. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3737-3746. Curran Associates, Inc.

[3]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998-6008. Curran Associates, Inc.

[4]. Huang, Z., Xu, S., Hu, M., Hu, M., Wang, X., Qiu, J., Fu, Y., Zhao, Y., Peng, Y., & Wang, C. (2020). Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems. IEEE Access, PP(99), 1-1. DOI:

10.1109/ACCESS.2020.2988903.

[5]. Adamson, B., Waskom, M., Blarre, A., Kelly, J., Krismer, K., Nemeth, S., ... Cohen, A. B. (2023). Approach to Machine Learning for Extraction of Real-World Data Variables from Electronic Health Records. Preprint at bioRxiv, March 2023. DOI: 10.1101/2023.03.02.23286522.