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

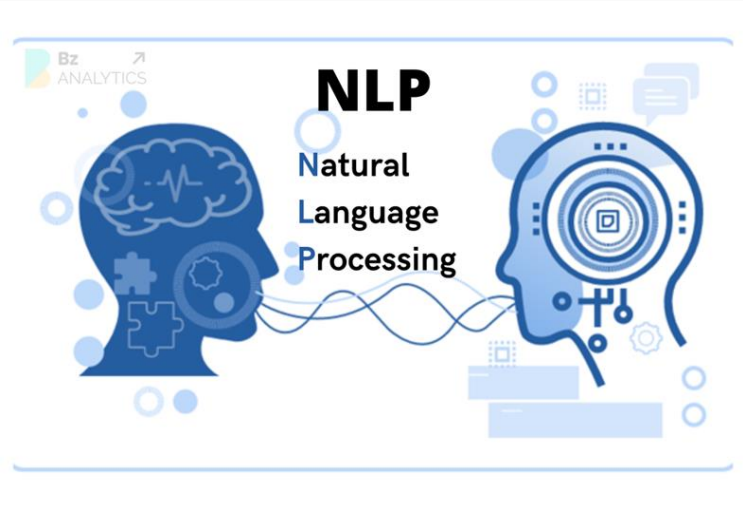
Vi Minh Huy - 220101025

Tóm tắt

- Lớp: CS2205.CH181
- Link Github: https://github.com/HuyVi/CS2205.CH181_PPLNCKH
- Link YouTube video: https://www.youtube.com/watch?v=Q5AcCu43ze0
- Họ và Tên: Vi Minh Huy

Giới thiệu

- Hai mô hình biểu diễn ngôn ngữ tiêu biểu là ELMo [1] và OpenAI GPT [2], nhưng vẫn còn một số hạn chế.
- BERT (Bidirectional Encoder Representations from Transformers) được đề xuất tích hợp sâu ngữ cảnh từ cả hai hướng trong quá trình biểu diễn đối tượng ngôn ngữ.

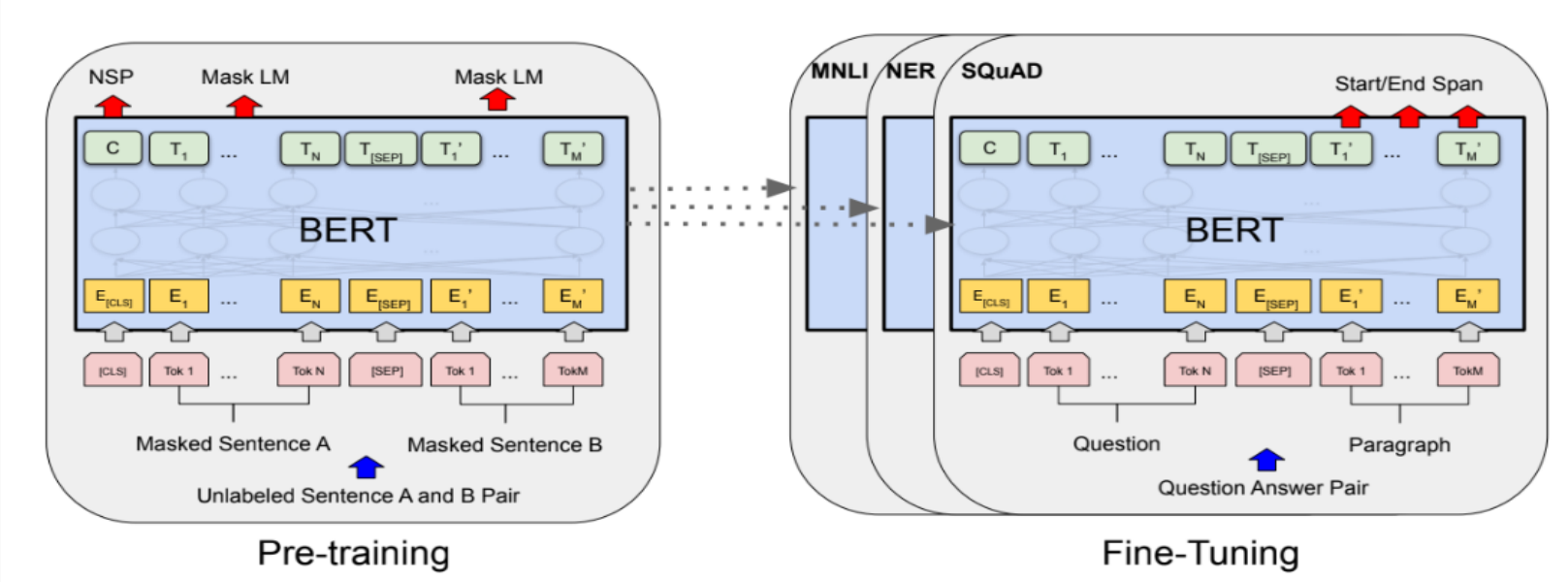


Mục tiêu

- 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

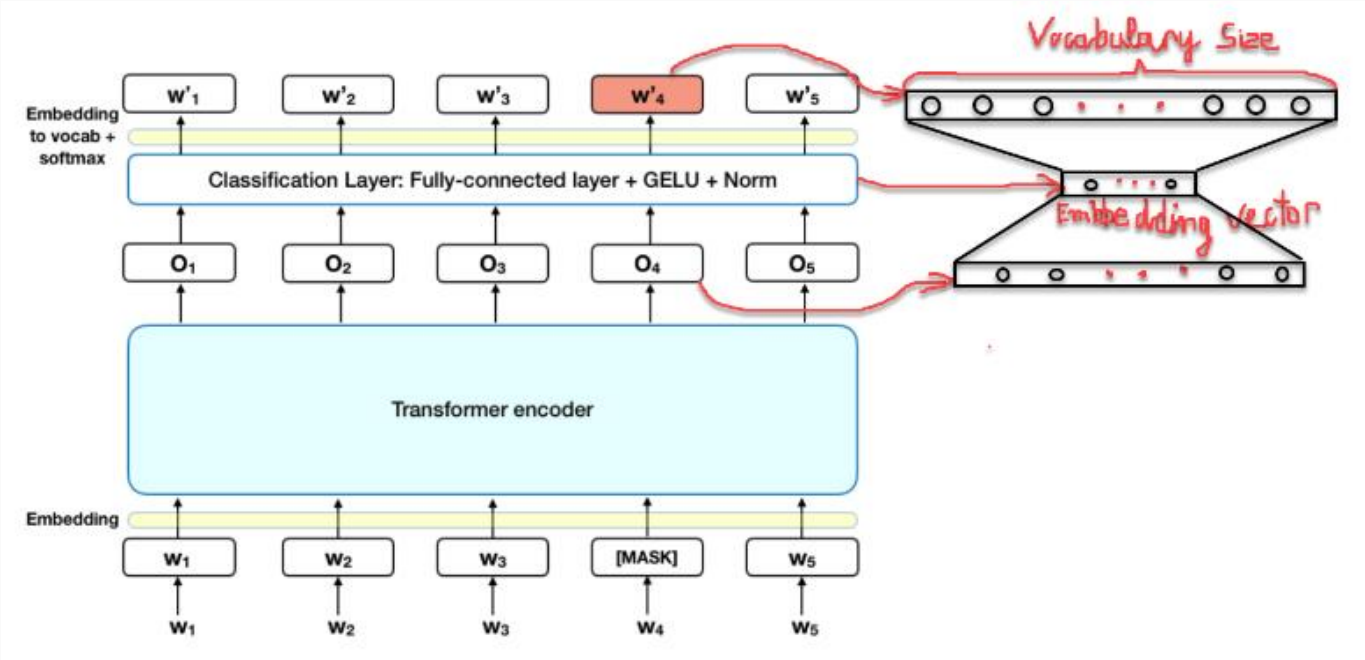
Fine-tuning model BERT



Minh họa: Overall pre-training and fine-tuning procedures for BERT
Nguồn: https://arxiv.org/pdf/1810.04805v2

Nội dung và Phương pháp(tt)

Masked ML (MLM)



Minh họa: Sơ đồ kiến trúc BERT cho tác vụ Masked ML.

Nội dung và Phương pháp(tt)

Next Sentence Prediction (NSP)

Đây là một bài toán phân loại học có giám sát với 2 nhãn (hay còn gọi là phân loại nhị phân). Input đầu vào của mô hình là một cặp câu (pair-sequence) sao cho 50% câu thứ 2 được lựa chọn là câu tiếp theo của câu thứ nhất và 50% được lựa chọn một cách ngẫu nhiên từ bộ văn bản mà không có mối liên hệ gì với câu thứ nhất. Nhãn của mô hình sẽ tương ứng với IsNext khi cặp câu là liên tiếp hoặc NotNext nếu cặp câu không liên tiếp

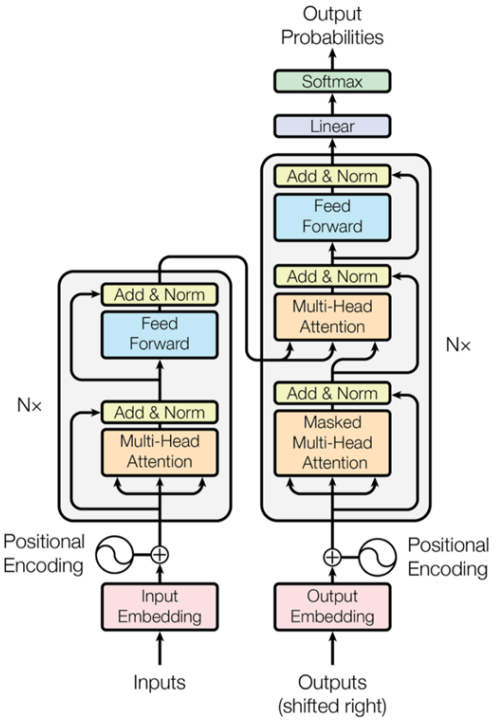


Figure 1: The Transformer - model architecture.

Minh họa: Sơ đồ kiến trúc BERT cho tác vụ NSP

Kết quả dự kiến

- 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:
  - 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

- [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.