Report paper: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Giáp Ngọc Hiệu

**Tóm tắt nội dung**

Bài report này là phần tóm tắt lại ý tưởng, nội dung chính, những điều làm được của paper

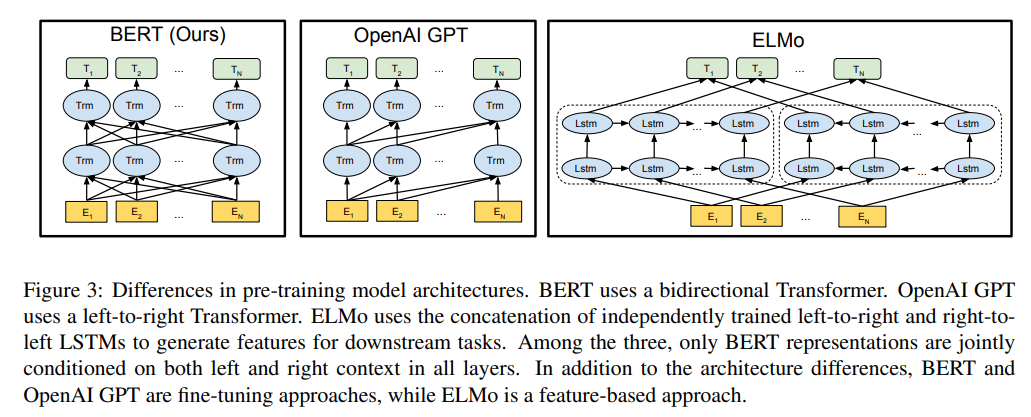
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

1. **Ý tưởng**

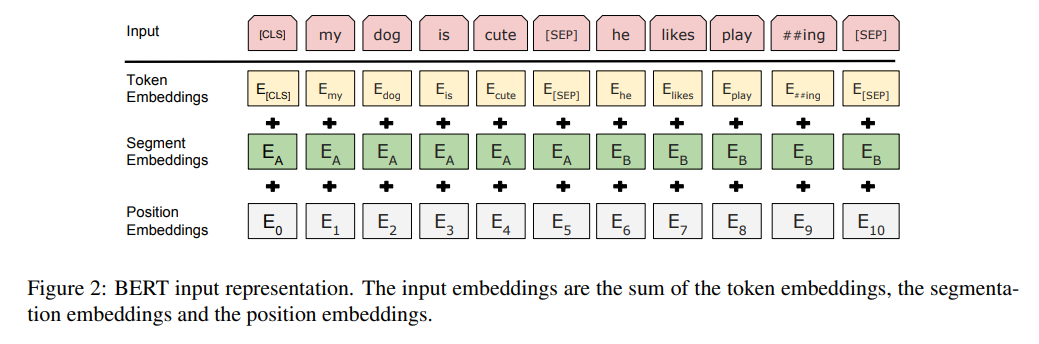
Tác giả giới thiệu một mô hình biểu diễn ngôn ngữ mới có tên BERT (Bidirectional Encoder Representations from Transformers). BERT được thiết kế để pre- train deep bidirectional representations từ văn bản không được gắn nhãn bằng cách điều hòa chung cả ngữ cảnh bên trái và bên phải trong tất cả các lớp. Điều này trái ngược với những nghiên cứu trước đây đã xem xét một chuỗi văn bản từ trái sang phải hoặc kết hợp đào tạo từ trái sang phải và phải sang trái. Kết quả là, mô hình BERT được đào tạo trước có thể được tinh chỉnh chỉ với một lớp đầu ra bổ sung để tạo ra các mô hình hiện đại cho nhiều nhiệm vụ, chẳng hạn như trả lời câu hỏi và suy luận ngôn ngữ, mà không cần nhiều sửa đổi kiến ​​trúc nhiệm vụ cụ thể.

1. **Phương pháp**
   1. **BERT Training Steps**

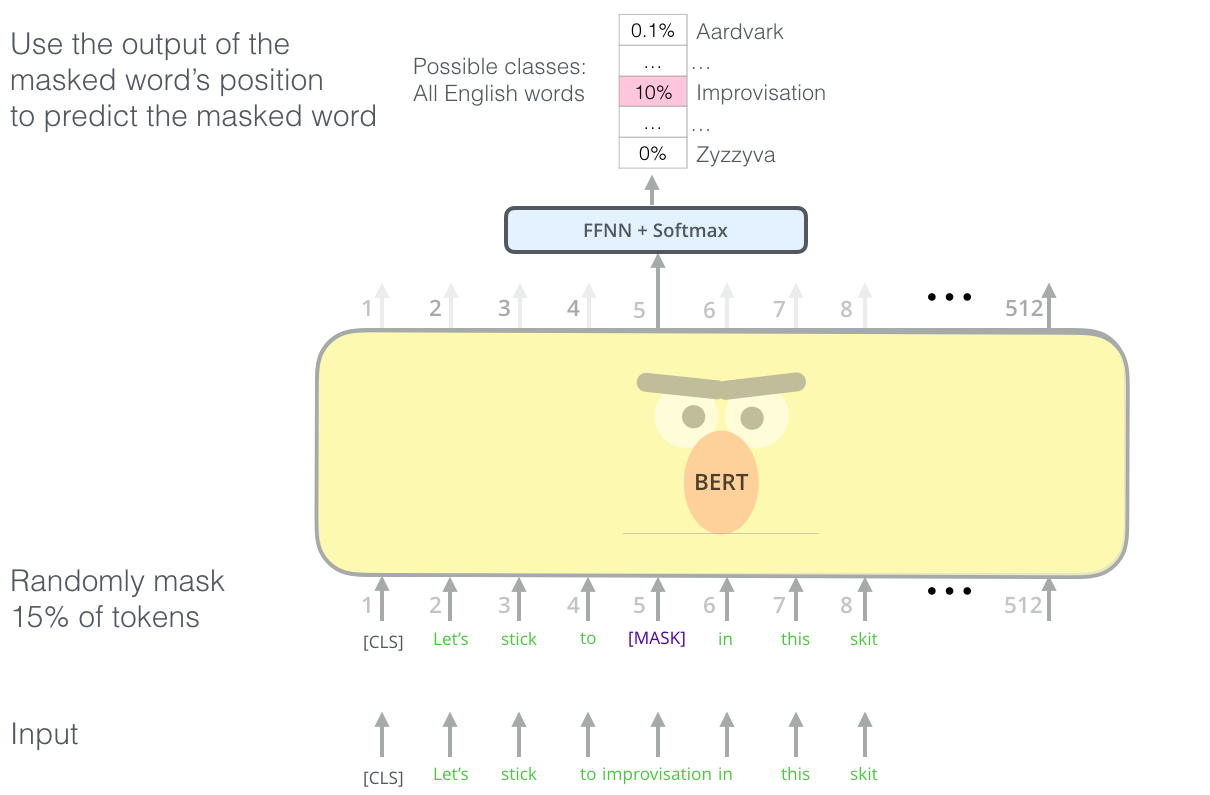
* Có hai bước: pre-training và fine-tuning.
* Trong quá trình pre-training, mô hình được đào tạo trên dữ liệu chưa được gắn nhãn qua các nhiệm vụ pre-training khác nhau.
* Để fine-tuning, mô hình BERT trước tiên được khởi tạo với các pre-trained parameters và tất cả các tham số được fine-tuning bằng cách sử dụng dữ liệu được gắn nhãn từ các downstream tasks.
  1. **BERT Model Architecture**

****

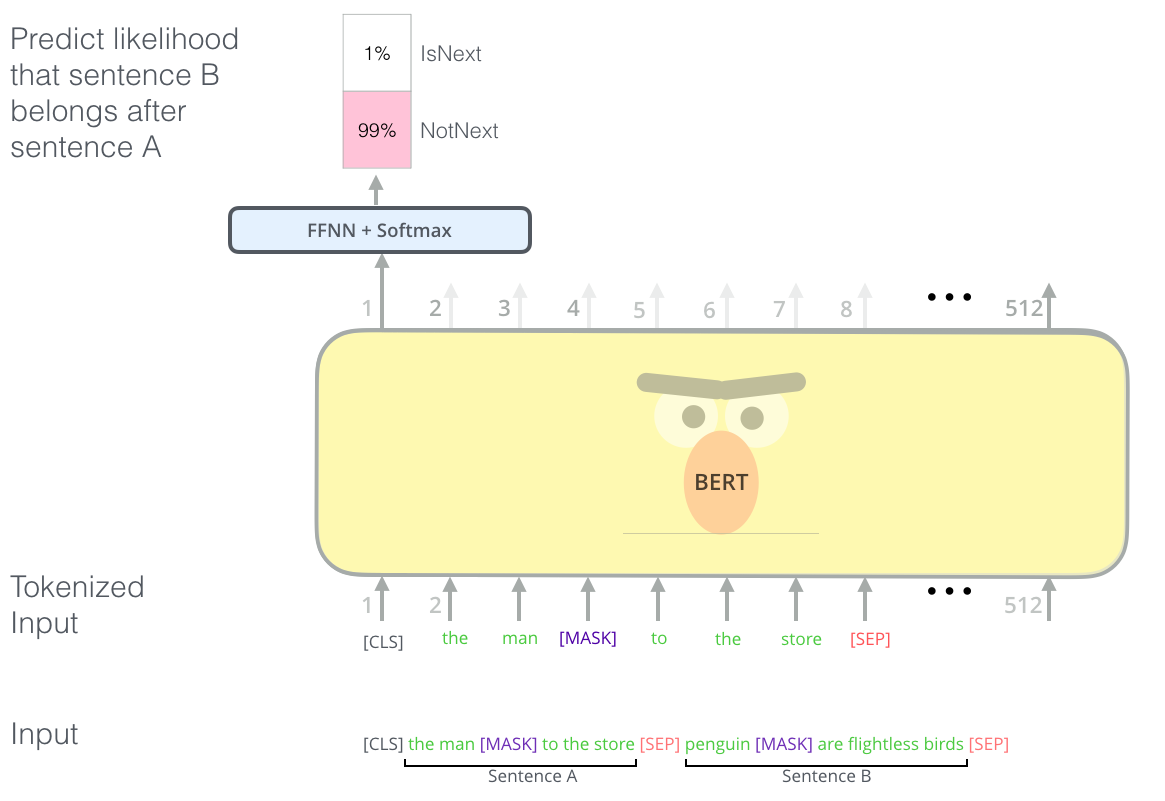
* Kiến trúc mô hình của BERT là một multi-layer bidirectional Transformer encoder dựa trên Transformer. Việc thực hiện gần như giống hệt với bản gốc.
* Số layer (tức là khối Transformer) được ký hiệu là L, hidden size được ký hiệu là H và số lượng self-attention heads là A^3.
* Hai kích thước mô hình được đánh giá:
  + BERTBASE (L=12, H=768, A=12, Total Parameters=110M), có cùng kích thước mô hình với OpenAI GPT nhằm mục đích so sánh. BERT Transformer sử dụng bidirectional self-attention, trong khi GPT Transformer sử dụng constrained self-attention trong đó mọi token chỉ có thể chú ý đến ngữ cảnh bên trái của nó.
  + BERTLARGE (L=24, H=1024, A=16, Total Parameters =340M).
  1. **Input Representation**

****

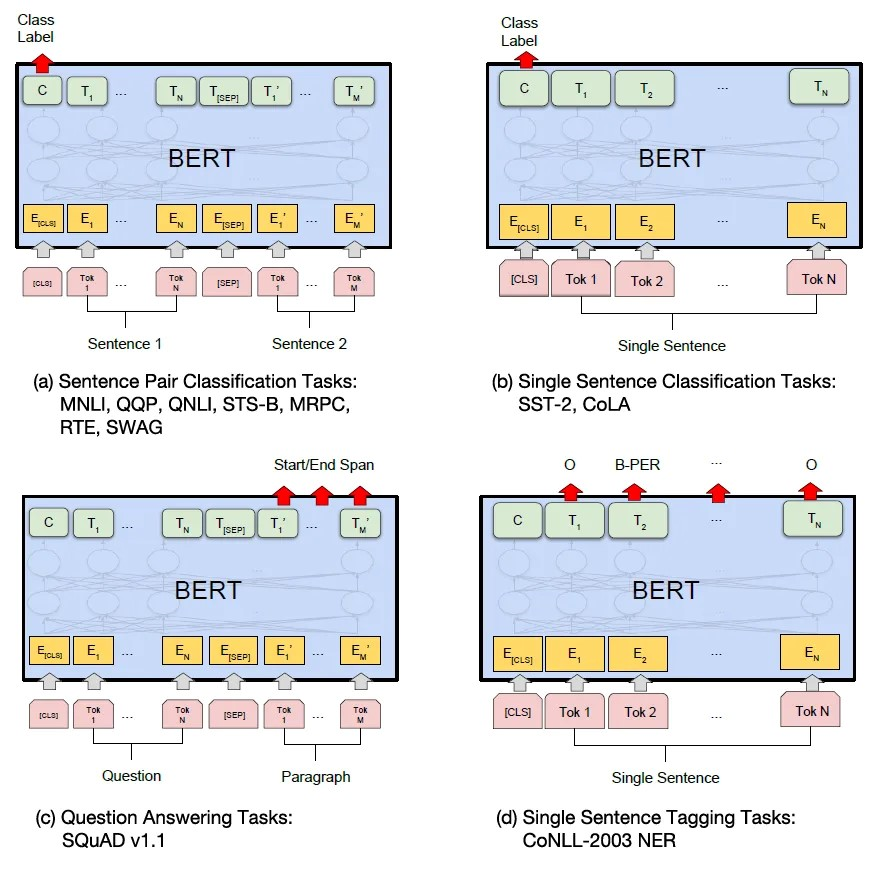
* Để làm cho BERT xử lý nhiều down-stream tasks khác nhau, input representation có thể biểu diễn rõ ràng cả một câu đơn và một cặp câu (ví dụ: <Question, Answer>) trong một chuỗi mã thông báo.
* WordPiece embeddings (Wu et al., 2016) với 30,000 token vocabulary, được sử dụng.
* Token đầu tiên của mỗi chuỗi luôn là token phân loại đặc biệt ([CLS]).
* Token ([SEP]) được sử dụng để phân tách.
* Đối với một token nhất định, input representation của nó được xây dựng bằng cách tính tổng token, segment, and position embeddings tương ứng.
  1. **Pretraining BERT**
     1. **Masked LM (MLM)**



* Một số phần trăm input tokens được mask một cách ngẫu nhiên và sau đó các mã thông báo bị mask đó sẽ được dự đoán.
* 15% tổng số tokens WordPiece được mask ngẫu nhiên trong mỗi chuỗi.
* Chỉ những từ bị mask mới được dự đoán thay vì xây dựng lại toàn bộ đầu vào. Nếu token thứ i được chọn, token thứ i được thay thế bằng (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged i-th token 10% of the time. Sau đó, Ti sẽ được sử dụng để dự đoán original token với cross entropy loss.
  + 1. **Next Sentence Prediction (NSP)**

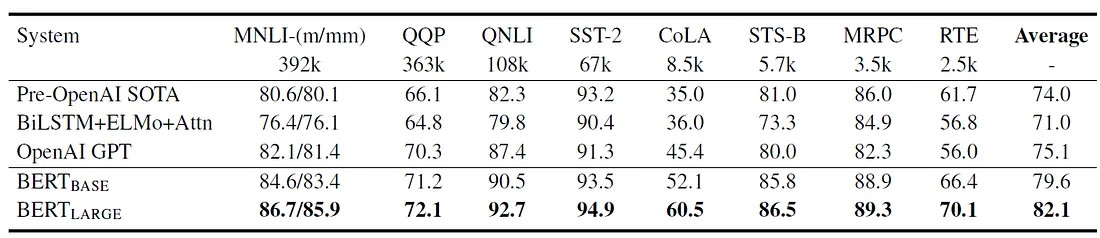


* Downstream tasks như Question Answering (QA) và Natural Language Inference (NLI) dựa trên việc hiểu mối quan hệ giữa hai câu, điều này không được mô hình hóa ngôn ngữ nắm bắt trực tiếp. Để đào tạo một mô hình hiểu được mối quan hệ của câu, tác giả pre-train cho một binarized next sentence prediction task có thể được tạo ra từ bất kỳ kho ngữ liệu đơn ngữ nào.
* Cụ thể, khi chọn câu A và B cho mỗi ví dụ huấn luyện trước, 50% time B là câu tiếp theo thực tế theo sau A (được gắn nhãn là IsNext) và 50% time là một câu ngẫu nhiên từ kho ngữ liệu (được gắn nhãn dưới dạng NotNext).
  + 1. **Pretraining Dataset**
* Đối với pre-training corpus, BooksCorpus (800M từ) và English Wikipedia (2.500M từ) được sử dụng.
* Đối với Wikipedia, chỉ các đoạn văn bản được trích xuất và danh sách, bảng và tiêu đề bị bỏ qua.
  1. **Fine-Tuning BERT**

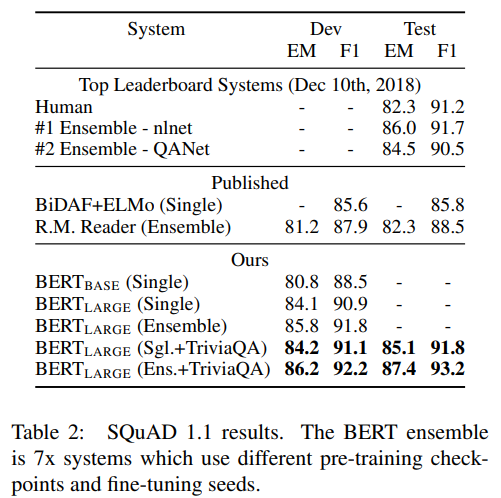


* Đối với mỗi task, đầu vào và đầu ra dành riêng cho task chỉ cần được cắm vào BERT và tất cả các tham số đều được fine-tune từ đầu đến cuối.
* Ở đầu vào, câu A và câu B trong quá trình pre-training tương tự như (1) các cặp câu trong diễn giải, (2) các cặp giả thuyết-tiền đề trong phần dẫn chứng, (3) các cặp câu hỏi-đoạn văn trong phần trả lời câu hỏi, và (4) cặp văn bản-∅ suy biến trong phân loại văn bản hoặc sequence tagging.
* Ở đầu ra, các token representations được đưa vào output layer cho các token-level tasks, chẳng hạn như sequence tagging hoặc question answering, và [CLS] representation được đưa vào output layer để phân loại, chẳng hạn như sentiment analysis.
* So với pre-training, việc fine-tune tương đối ít tốn kém. Tất cả các kết quả trong bài báo có thể được sao chép trong tối đa 1 giờ trên một Cloud TPU hoặc vài giờ trên GPU.

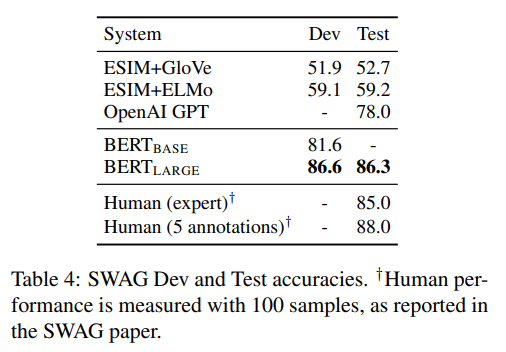
1. **Experimental & Results**
   1. **GLUE**



* Cả BERTBASE và BERTLARGE đều vượt trội hơn trong mọi nhiệm vụ với mức chênh lệch đáng kể, đạt được mức cải thiện độ chính xác trung bình tương ứng là 4,5% và 7,0% so với mô hình hiện đại trước đó.
  1. **SQuAD 1.1.**

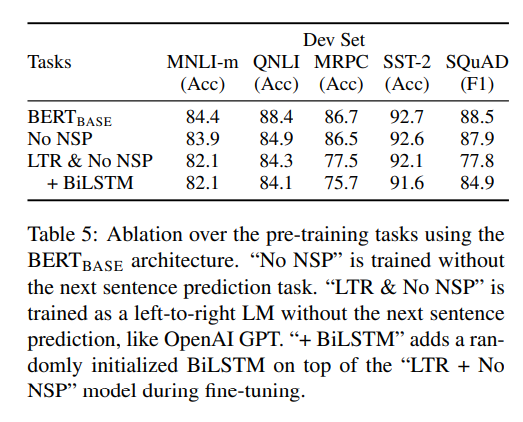
****

* BERT tốt nhất vượt trội hơn hệ thống bảng xếp hạng hàng đầu với +1,5 F1 trong tập hợp và +1,3 F1 dưới dạng một single system. Trên thực tế, mô hình BERT đơn lẻ vượt trội hơn ensemble system hàng đầu về điểm F1.
  1. **SWAG**

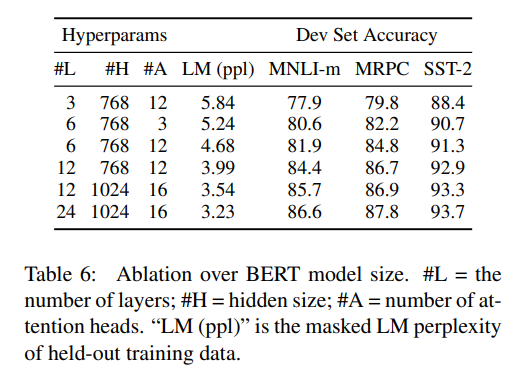
****

* BERTLARGE vượt trội hơn hệ thống ESIM+ELMo ở mức +27,1% và OpenAI GPT là 8,3%.

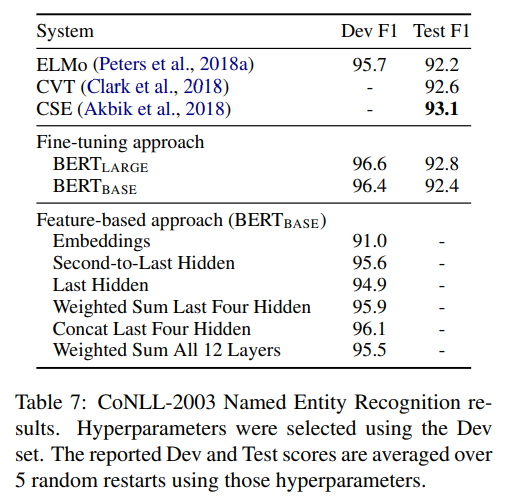
1. **Ablation Study**
   1. **Effect of Pre-Training Tasks**

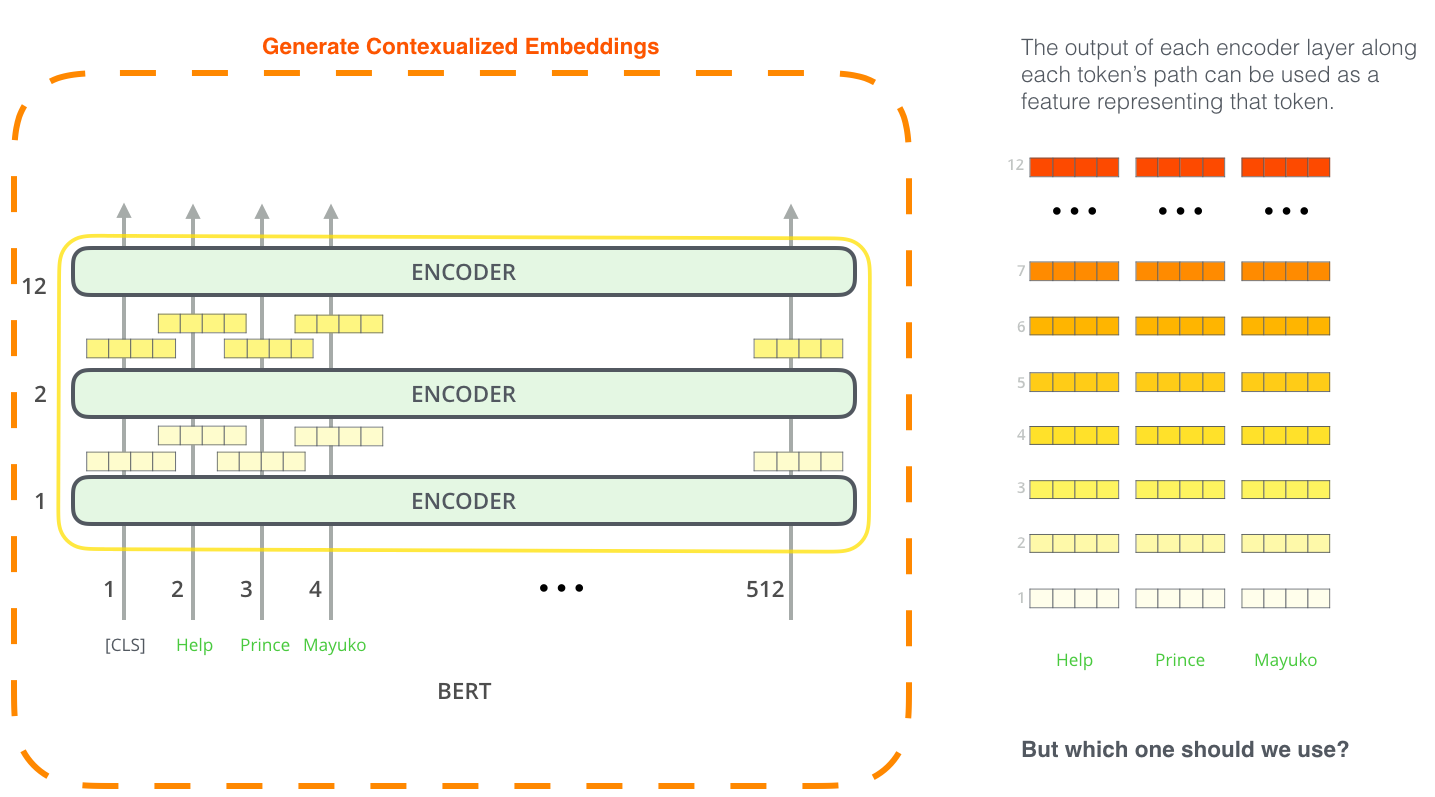


* “No NSP” được huấn luyện mà không có next sentence prediction task.
* “LTR & No NSP” được đào tạo thành LM từ trái sang phải mà không có dự đoán câu tiếp theo, giống như OpenAI GPT.
* “+ BiLSTM” thêm BiLSTM được khởi tạo ngẫu nhiên lên trên mô hình “LTR + No NSP” trong quá trình tinh chỉnh.
* Việc xóa NSP sẽ ảnh hưởng đáng kể đến hiệu suất trên QNLI, MNLI và SQuAD 1.1. Mô hình LTR hoạt động kém hơn mô hình MLM trên tất cả các tác vụ, với mức giảm lớn trên MRPC và SQuAD.
  1. **Effect of Model Size**

****

* Các mô hình lớn hơn dẫn đến sự cải thiện độ chính xác nghiêm ngặt trên cả bốn bộ dữ liệu, ngay cả đối với MRPC chỉ có 3.600 ví dụ đào tạo được gắn nhãn và khác biệt đáng kể so với các pre-training tasks.
* Đây là công trình đầu tiên chứng minh một cách thuyết phục rằng việc mở rộng quy mô đến kích thước mô hình cực lớn cũng dẫn đến những cải tiến lớn đối với các nhiệm vụ ở quy mô rất nhỏ, miễn là mô hình đã được pre-train đầy đủ.
  1. **Feature-based Approach with BERT**

****



* Tất cả các kết quả BERT được trình bày cho đến nay đều sử dụng fine-tuning approach, trong đó một classification layer được thêm vào pre-trained model và tất cả các tham số đều được fine-tune chung trong nhiệm vụ tiếp theo.
* Tuy nhiên, với feature-based approach, trong đó các feature cố định được trích xuất từ ​​pre-trained model, có những ưu điểm nhất định.
  + Đầu tiên, không phải tất cả các tasks đều có thể được biểu diễn dễ dàng bằng Transformer encoder architecture.
  + Thứ hai, những lợi ích tính toán chính có thể được tính toán trước.
* Feature-based approach được áp dụng bằng cách trích xuất các kích hoạt từ một hoặc nhiều lớp mà không cần tinh chỉnh bất kỳ tham số nào của BERT.
* Các contextual embeddings được sử dụng làm đầu vào cho BiLSTM hai lớp 768 chiều được khởi tạo ngẫu nhiên trước classification layer.
* Phương pháp hoạt động tốt nhất kết hợp các token representations từ bốn lớp ẩn trên cùng của pre-trained Transformer, chỉ kém 0,3 F1 sau khi tinh chỉnh toàn bộ mô hình.
* Điều này chứng tỏ rằng BERT có hiệu quả cho cả finetuning và feature-based approaches.

1. **Note**

* BERT có thể được sử dụng trong nhiều nhiệm vụ ngôn ngữ khác nhau:
  + Có thể xác định mức độ đánh giá tích cực hay tiêu cực về bộ phim. (Sentiment Analysis) Giúp chatbot trả lời câu hỏi của bạn. (Question answering)
  + Dự đoán văn bản của bạn khi viết email (Gmail). (Text prediction)
  + Có thể viết một bài về bất kỳ chủ đề nào chỉ với vài câu đầu vào. (Text generation)
  + Có thể tóm tắt nhanh các hợp đồng pháp lý dài hạn. (Summarization)
  + Có thể phân biệt các từ có nhiều nghĩa (như ‘bank’) dựa trên văn bản xung quanh. (Polysemy resolution)
* Ưu điểm:
  + Task Flexibility: Khả năng tinh chỉnh của BERT cho phép nó thích ứng với nhiều nhiệm vụ NLP, chẳng hạn như phân loại văn bản, nhận dạng thực thể được đặt tên và phân tích tình cảm.
  + Bidirectional Context: Bằng cách xem xét toàn bộ ngữ cảnh của câu, BERT nắm bắt được mối quan hệ giữa các từ theo cả hai hướng. Điều này làm cho nó có giá trị đối với những nhiệm vụ trong đó việc hiểu ý nghĩa và ngữ cảnh của các từ liên quan đến toàn bộ câu là rất quan trọng.
  + Contextual Word Representations: BERT tạo ra các cách biểu diễn từ sensitive theo ngữ cảnh, cho phép hiểu sâu hơn về ngữ nghĩa ngôn ngữ và các mối quan hệ từ.
* Nhược điểm:
  + BERT yêu cầu tính toán nhiều hơn cũng như chi phí đào tạo do kích thước của nó.