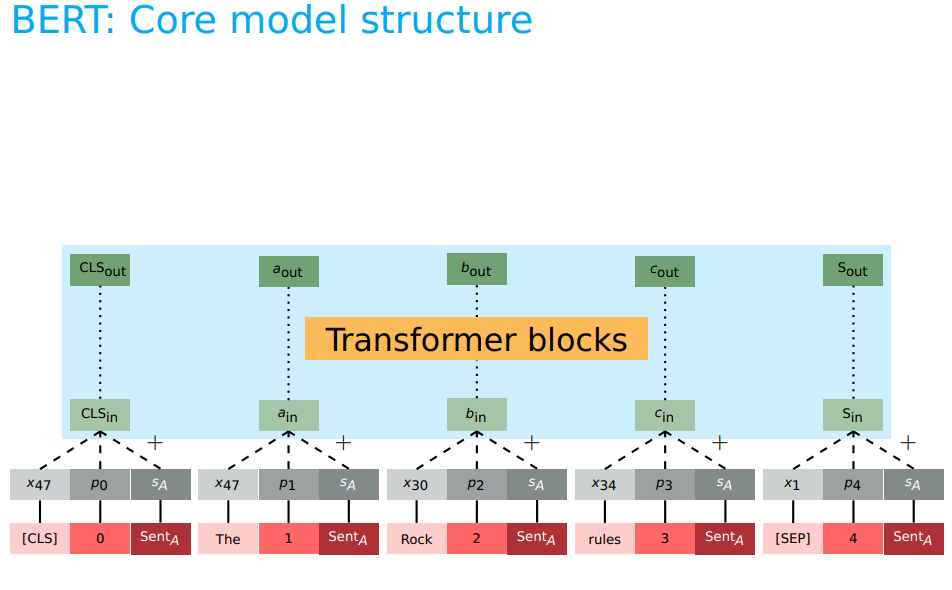
**Báo cáo tiến trình học tập tuần 8**

1. **BERT**
2. **Core model structure**

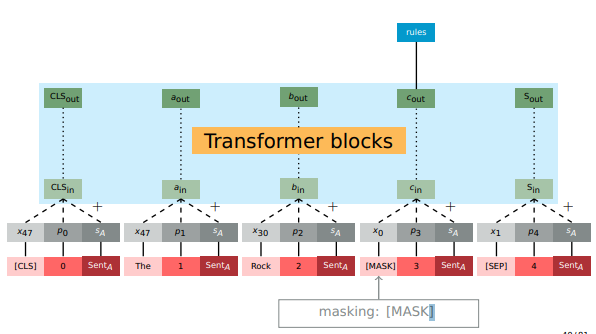
****

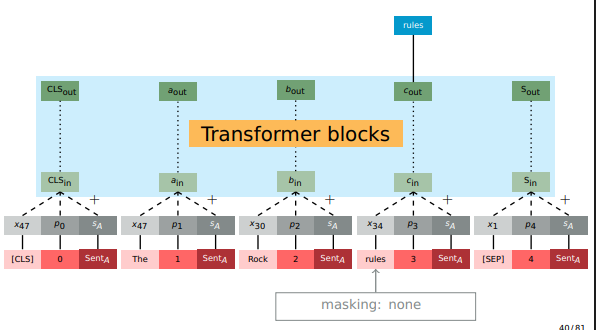
Trong đó:

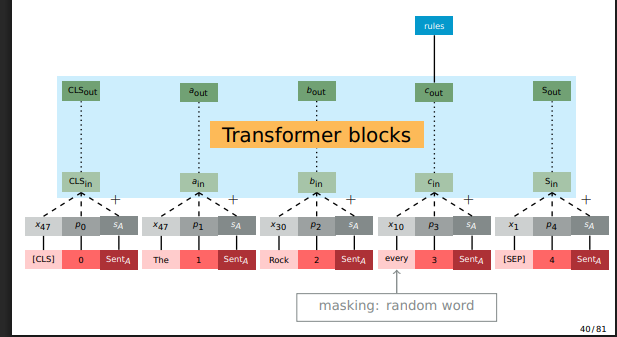
* + **CLS:** là một token đặc biệt, lưu toàn bộ nội dung của chuỗi đầu vào.
  + **SentA:** là một lable đánh dấu từ đó thuộc câu nào (trong trường hợp này là thuộc câu A).
  + **SEP:** dùng để ngăn tách giữa các chuỗi.
* Sau đó tất cả được embedding và gộp thành một embedding duy nhất, cho qua các khối Transformer, thu được các vector đầu ra đã được cập nhật.

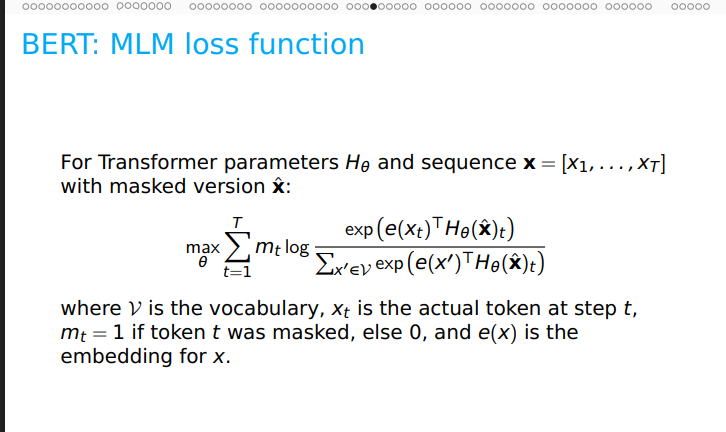
1. **Masked language modeling (MLM)**

* Đây là một thủ thuật của BERT: trong 15% token chọn để mask, thì:
  + 80% → thay bằng [MASK]
  + 10% → thay bằng từ ngẫu nhiên
  + 10% → giữ nguyên từ gốc
* Điều này giúp mô hình **không quá phụ thuộc vào token [MASK]** vốn không xuất hiện khi fine-tune thực tế.

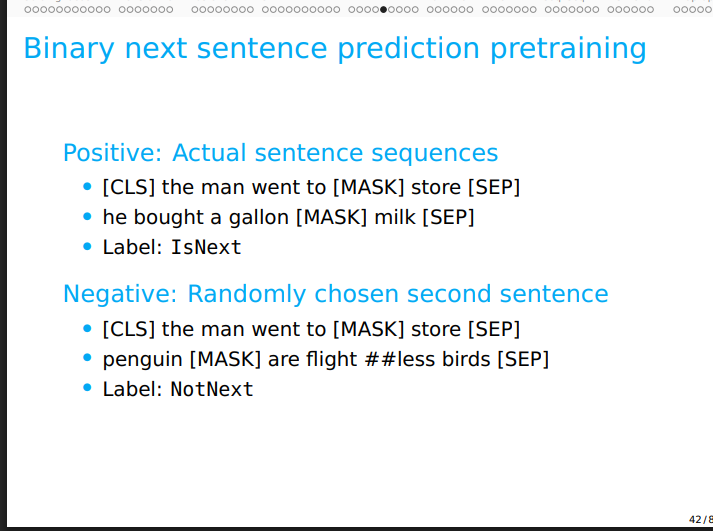




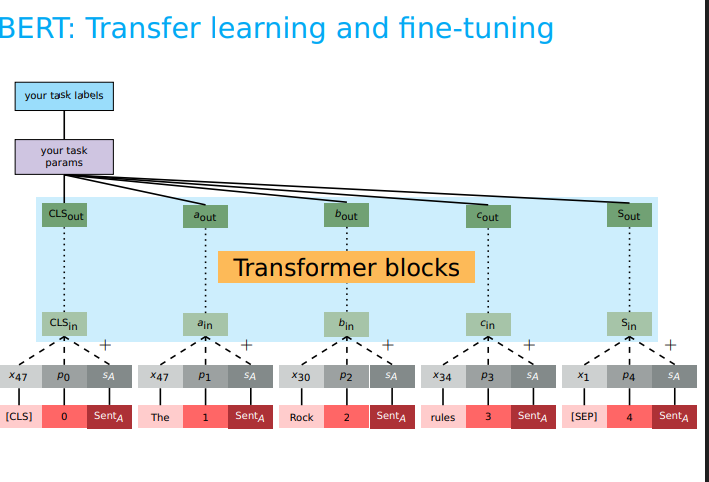


****

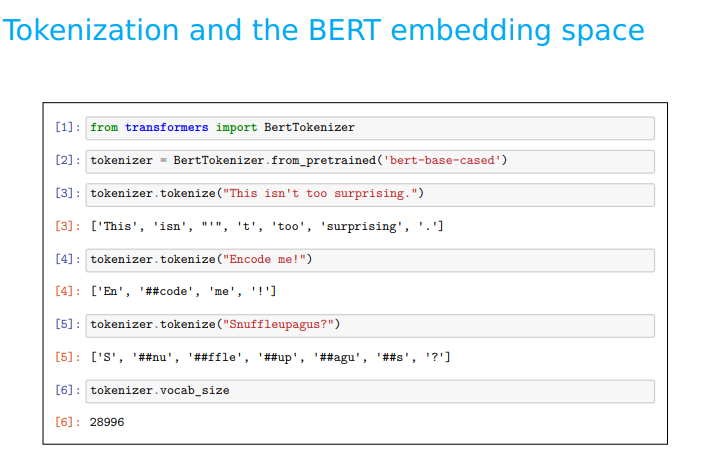
* Hàm loss của BERT đo xác suất (mức độ liên quan) giữa từ bị che ​ và vector hidden ​ mà mô hình sinh ra từ ngữ cảnh.



* Token [CLS] ở đầu câu sẽ sinh ra một **vector đặc biệt** sau khi qua Transformer — gọi là **pooled output**.
* Vector này được đưa qua **một lớp phân loại (classification head)** để dự đoán:
  + IsNext → hai câu liên tiếp thật.
  + NotNext → hai câu ngẫu nhiên.



* your task params : phần mô hình nhỏ được gắn thêm để chuyển BERT thành mô hình cho bài toán cụ thể.
* Your task lables: là nhãn thật để tính loss khi finetune.



* Bert dùng WordPiece để chuẩn bị cho embedding space

1. **Robustly optimized BERT approach (RoBERTa)**

|  |  |  |  |
| --- | --- | --- | --- |
| Thành phần | BERT (2018) | RoBERTa (2019) | Giải thích / ý nghĩa |
| Masking | Static masking | Dynamic masking | Ở BERT, token bị che (mask) cố định suốt quá trình huấn luyện — nghĩa là mỗi câu luôn bị che ở cùng vị trí. RoBERTa thay đổi vị trí bị che mỗi epoch → giúp mô hình học mạnh hơn vì thấy nhiều ngữ cảnh khác nhau. |
| Input cấu trúc | Hai câu ghép (sentence A + sentence B) | Một chuỗi câu liên tiếp, có thể vượt biên giới đoạn | BERT luôn lấy 2 câu để phục vụ bài toán NSP. RoBERTa bỏ giới hạn này → cho phép mô hình hiểu văn bản dài, không bị ràng buộc bởi ranh giới đoạn. |
| Next Sentence Prediction (NSP) | Có NSP | Không dùng NSP | RoBERTa chứng minh rằng NSP không giúp gì nhiều cho downstream tasks, thậm chí còn làm mô hình học kém hơn. |
| **Batch size** | 256 | 2,000 | Batch lớn hơn giúp mô hình học ổn định và hội tụ nhanh hơn khi kết hợp với learning rate phù hợp (gradient noise giảm). |
| Tokenizer | WordPiece | Byte-Pair Encoding (BPE) ở mức ký tự | BPE linh hoạt hơn với từ mới, lỗi chính tả hoặc ngôn ngữ khác nhau (multi-lingual robustness). |
| Dữ liệu huấn luyện | BooksCorpus + Wikipedia (~16GB) | Wikipedia + CC-News + OpenWebText + Stories (~160GB) | Dữ liệu lớn hơn gấp 10 lần → giúp mô hình hiểu ngôn ngữ tự nhiên phong phú hơn, giảm bias. |
| Số bước huấn luyện (steps) | 1M | ~500k (nhưng batch lớn hơn, tổng tokens nhiều hơn) | RoBERTa huấn luyện “ngắn hơn” về số step nhưng xử lý nhiều tokens hơn do batch size tăng. |
| Chiều dài chuỗi (sequence length) | Huấn luyện ngắn trước, sau đó dài hơn | Huấn luyện toàn bộ với full-length (512 tokens) | BERT giai đoạn đầu chỉ huấn luyện trên chuỗi ngắn để tiết kiệm tài nguyên. RoBERTa bỏ qua — học full context từ đầu để không bị “quen” với văn bản ngắn. |