**Embedding in BERT**

1. Introduction to BERT

BERT (Bi-directional Encoder Representations from Transformers) được phát hành vào năm 2018 bởi 1 team ở Google AI Language. Như tên của nó, BERT được tạo nên bằng cách xếp chồng nhiều encoders của transformer lên nhau. Kiến trúc BERT được thiết kế để học các biểu diễn 2 chiều một cách sâu sắc từ các văn bản không được gán nhãn.

Do đó, pre-trained BERT model có thể được fine-tuned với chỉ một lớp output để tạo nên các SOTA models cho một loạt các tasks như question answering, language inference mà không phải sửa sửa kiến trúc từng task cụ thể.

2. Pretraining of BERT

Mục tiêu của pretraining BERT là làm cho nó hiểu thể nào là ngôn ngữ và thế nào là ngữ cảnh. Điều này đạt được bằng cách sử dụng Masked Language Modeling (MLM) và Next Sentence Prediction (NSP).

2.1 Masked Language Modeling (MLM)

* BERT lấy các câu với các khoảng trống làm đầu vào và mục tiêu là điền vào những chỗ trống đó
* Điều này giúp BERT hiểu ngữ cảnh 2 chiều của ngôn ngữ

2.2 Next Sentence Prediction (NSP)

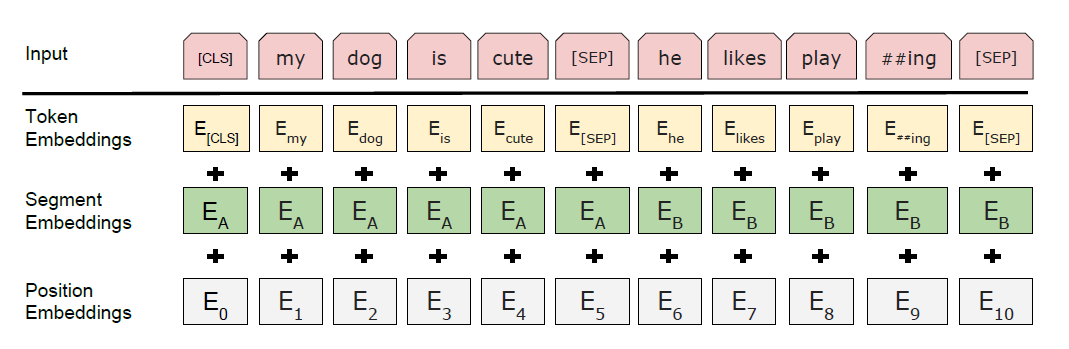
* BERT lấy 2 câu và kiểm tra xem câu thứ 2 có phải là câu phía sau của câu thứ 1 hay không
* Điều này giúp BERT hiểu ngữ cảnh qua các câu khác nhau

Do vậy sử dụng cả MLM và NSP giúp BERT hiểu ngôn ngữ một cách sâu sắc.

3. Embeddings in BERT

Embeddings là các vector biểu diễn từ, các từ tương tự nhau có các thành phần gần bằng nhau trong các vectors của chúng. Ví dụ: king và queen

Tại sao chúng ta cần embeddings? Lý do chính của việc chuyển đổi từ words sang embeddings là để model có thể dễ dàng làm việc. Khi words được chuyển sang embeddings, model có thể hiểu mức độ quan trọng về ngữ nghĩa của 1 từ ở dạng số và do vậy mô hình có thể thực hiện các phép toán trên embeddings



Một số tokens đặc biệt:

* CLS: Viết tắt của từ classification. Nó được thêm vào để biểu diễn sự phân loại câu. Nó thường được thêm vào đầu chuỗi tokens để đại diện cho toàn bộ câu.
* SEP: Được sử dụng trọng NSP. Nó là 1 token phân tách giữa 2 câu để model biết đâu là câu tiếp theo. Trong trường hợp chỉ có 1 câu, SEP được đặt ở cuối chuỗi tokens.
* MASK: Được sử dụng trong MLM. Nó được sử dụng để biễu diễn các từ bị che đi.

3.1 Token embeddings

Token embeddings là pretrained embeddings cho các từ khác nhau.

Để tạo ra những pretrain token embeddings này, 1 phương pháp gọi là WordPiece Tokenization được sử dụng để tokenize văn bản. Đây là chiến lược tokenization hướng dữ liệu, nó cố gắng cân bằng vocab size và out-of-vocab words.

Mỗi wordpiece token sẽ được chuyển đổi sang 1 vector 768 chiều bởi Token Embeddings layer.

3.2 Segment Embeddings

Segment embeddings đơn giản là số lượng câu được mã hóa thành vector.

Model phải biết 1 token cụ thể thuộc câu A hay câu B trong BERT. Thực hiện điều này bằng cách sinh ra các tokens cố định, 1 token cho câu A, 1 token cho câu B.

Chỉ có 2 vectors trong Segment Embeddings layer. Tất cả token thuộc câu 1 được gán cho vector đầu tiên , tất cả token thuộc câu 2 được gán cho vector thứ 2.

3.3 Position Embeddings

Position embeddings là vị trí của từ trong câu được mã hóa thành vector

3.4 Combining these embeddings

Input embeddings là tổng của token embeddings, segmentation embeddings và position embeddings. Cả 3 embeddings được tổng hợp để tạo ra biểu diễn có cỡ (1, n, 768). Đây là đầu vào cho encoder.

4. Conclusion

Embedding của BERT là lý do chính tạo nên tốc độ và hiệu năng đáng kinh ngạc của model.

https://iq.opengenus.org/embeddings-in-bert/#:~:text=Segment%20embeddings%20are%20basically%20the,and%20one%20for%20sentence%20B.