



Processamento de Linguagem Natural com Transformers

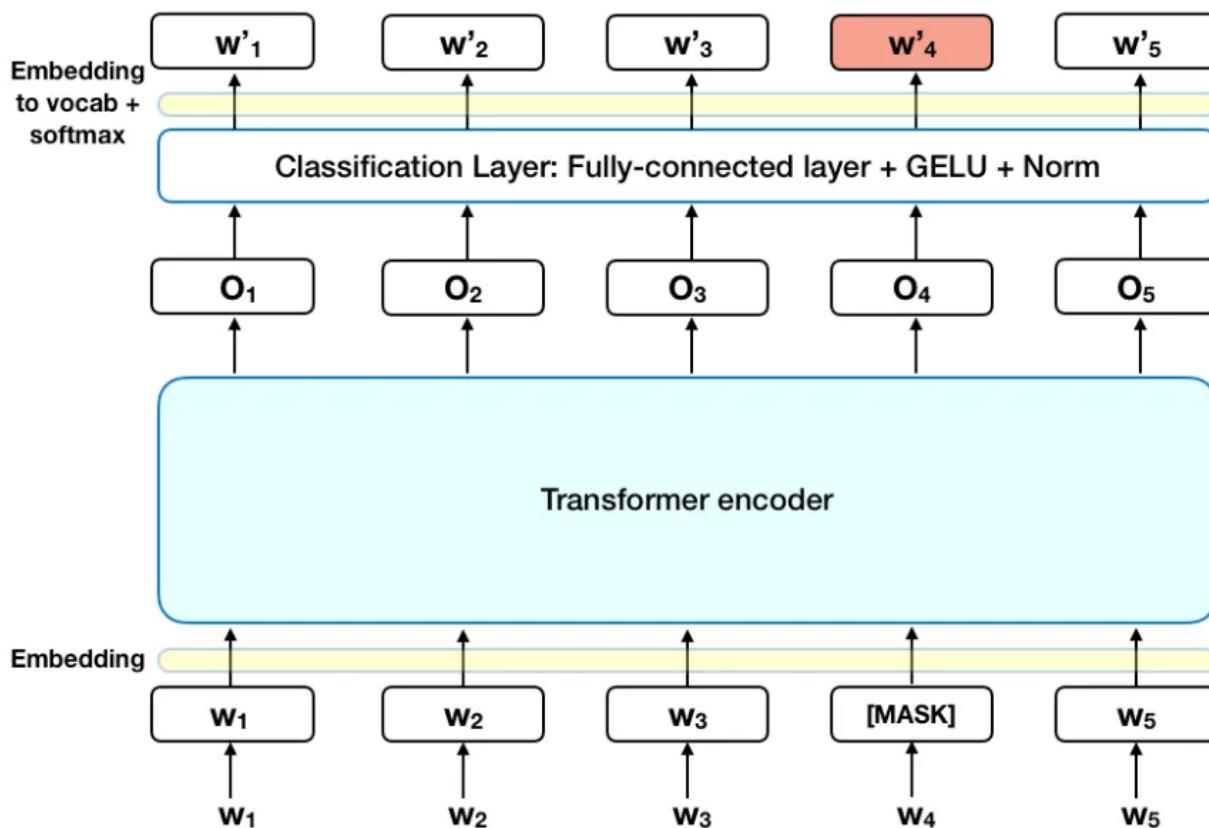
Processamento de Linguagem Natural com Transformers

Masked LM (MLM) e
Next Sentence Prediction (NSP)

O BERT faz uso do Transformer, um mecanismo de atenção que aprende relações contextuais entre palavras em um texto. Em sua forma padrão, o Transformer inclui dois mecanismos separados — um codificador que lê a entrada de texto e um decodificador que produz uma previsão para a tarefa. Como o objetivo do BERT é gerar um modelo de linguagem, apenas o mecanismo do codificador é necessário.

Ao contrário dos modelos direcionais, que leem a entrada de texto sequencialmente (da esquerda para a direita ou da direita para a esquerda), o codificador Transformer lê toda a sequência de palavras de uma só vez. Portanto, é considerado bidirecional, embora seja mais correto dizer que não é direcional. Essa característica permite que o modelo aprenda o contexto de uma palavra com base em todos os seus arredores (esquerda e direita da palavra).

A imagem abaixo é uma descrição de alto nível do codificador Transformer. A entrada é uma sequência de tokens, que são primeiro incorporados em vetores e depois processados na rede neural. A saída é uma sequência de vetores de tamanho H, em que cada vetor corresponde a um token de entrada com o mesmo índice.



Ao treinar modelos de linguagem, há um desafio de definir uma meta de previsão. Muitos modelos prevêem a próxima palavra em uma sequência (por exemplo, “A criança voltou para casa de ___”), uma abordagem direcional que inherentemente limita o aprendizado de contexto. Para superar esse desafio, o BERT utiliza duas estratégias de treinamento:

Masked LM (MLM)

Antes de alimentar as sequências de palavras no BERT, 15% das palavras em cada sequência são substituídas por um token [MASK]. O modelo então tenta prever o valor original das palavras mascaradas, com base no contexto fornecido pelas outras palavras não mascaradas na sequência. Em termos técnicos, a previsão das palavras de saída requer:

- 1- Adicionar uma camada de classificação na parte superior da saída do codificador.
- 2- Multiplicar os vetores de saída pela matriz embedding, transformando-os na dimensão do vocabulário.
- 3- Calcular a probabilidade de cada palavra no vocabulário com softmax.

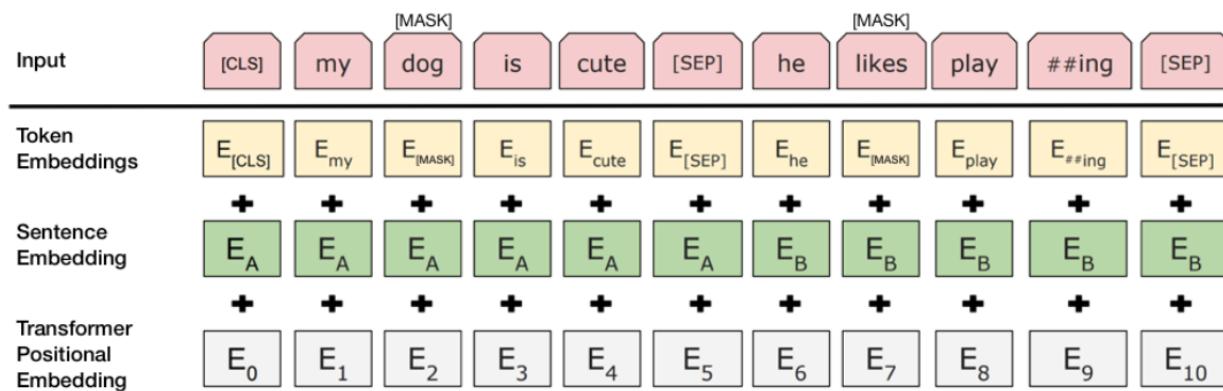
A função de perda no modelo BERT leva em consideração apenas a previsão dos valores mascarados e ignora a previsão das palavras não mascaradas. Como consequência, o modelo converge mais lentamente do que os modelos direcionais, uma característica que é compensada por sua maior percepção do contexto.

Next Sentence Prediction (NSP)

No processo de treinamento BERT, o modelo recebe pares de sentenças como entrada e aprende a prever se a segunda sentença do par é a sentença subsequente no documento original. Durante o treinamento, 50% das entradas são um par em que a segunda sentença é a sentença subsequente no documento original, enquanto nos outros 50% uma sentença aleatória do corpus é escolhida como a segunda sentença. A suposição é que a sentença aleatória será desconectada da primeira sentença.

Para ajudar o modelo a distinguir entre as duas sentenças em treinamento, a entrada é processada da seguinte maneira antes de entrar no modelo:

- 1- Um token [CLS] é inserido no início da primeira frase e um token [SEP] é inserido no final de cada frase.
- 2- Uma embedding de frase indicando a Sentença A ou a Sentença B é adicionada a cada token. As embeddings de sentença são semelhantes em conceito às embeddings de token com um vocabulário de 2.
- 3- Uma embedding posicional é adicionada a cada token para indicar sua posição na sequência. O conceito e a implementação da embedding posicional são apresentados no artigo Transformer.



Para prever se a segunda frase está realmente conectada à primeira, as seguintes etapas são executadas:

- 1- Toda a sequência de entrada passa pelo modelo Transformer.
- 2- A saída do token [CLS] é transformada em um vetor em forma de 2×1 , usando uma camada de classificação simples (matrizes aprendidas de pesos e vieses).
- 3- Calculando a probabilidade de IsNextSequence com softmax.

Ao treinar o modelo BERT, Masked LM e Next Sentence Prediction são treinados juntos, com o objetivo de minimizar a função de perda combinada das duas estratégias.