



Processamento de Linguagem Natural com Transformers

Processamento de Linguagem Natural com Transformers

Representação de Sentenças com BERT

BERT, que é a sigla para "Bidirectional Encoder Representations from Transformers", é um modelo de linguagem que gera representações densas de texto, conhecidas como embeddings. Essas representações capturam uma variedade de propriedades semânticas e sintáticas do texto.

Para entender como BERT representa as sentenças, primeiro precisamos entender a estrutura de entrada que o BERT utiliza. Uma entrada para o BERT é uma sequência de tokens, que são basicamente pedaços de texto. Esses tokens podem ser palavras ou partes de palavras. O BERT utiliza uma técnica chamada WordPiece, que divide palavras em subpalavras se a palavra não estiver em seu vocabulário.

Além dos tokens do texto de entrada, a entrada do BERT inclui os seguintes tokens especiais:

[CLS]: Este token é adicionado ao início de cada sentença e é usado quando precisamos de uma representação agregada da sentença, por exemplo, para classificação de sentença.

[SEP]: Este token é usado para separar duas sentenças quando o BERT é usado para tarefas que requerem duas sentenças como entrada.

Então, para representar uma sentença, o BERT gera embeddings para cada token na sentença. Esses embeddings são gerados com base não apenas no token em si, mas também em seu contexto, ou seja, os outros tokens que estão ao seu redor. Por isso, dizemos que o BERT é bidirecional - ele olha para os tokens antes e depois de um token específico para gerar seu embedding.

Se você quiser uma representação única para toda a sentença, você pode usar o embedding do token [CLS]. Durante o treinamento, o BERT é projetado para fazer com que o token [CLS] contenha uma representação agregada de toda a sentença.

A representação de uma sentença com o BERT envolve a geração de embeddings para cada token na sentença, com cada embedding sendo influenciado pelo contexto em que o token aparece. Para representações de sentença, o token [CLS] é comumente utilizado.

Veremos isso na prática no Estudo de Caso a seguir.