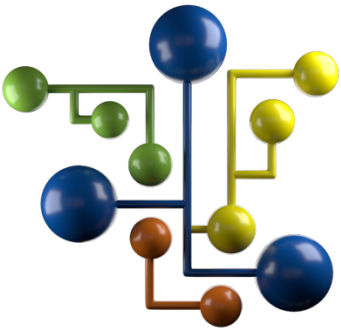


Data Science Academy

www.datascienceacademy.com.br

# Processamento de Linguagem Natural

# Como o BERT Funciona?

Em Visão Computacional, os pesquisadores mostraram repetidamente o valor do Transfer Learning - pré-treinando um modelo de rede neural com um dataset conhecido, por exemplo, ImageNet e, em seguida, realizando o ajuste fino - usando a rede neural treinada como base de um novo modelo específico. Nos últimos anos, os pesquisadores têm demonstrado que uma técnica semelhante pode ser útil em muitas tarefas de linguagem natural.

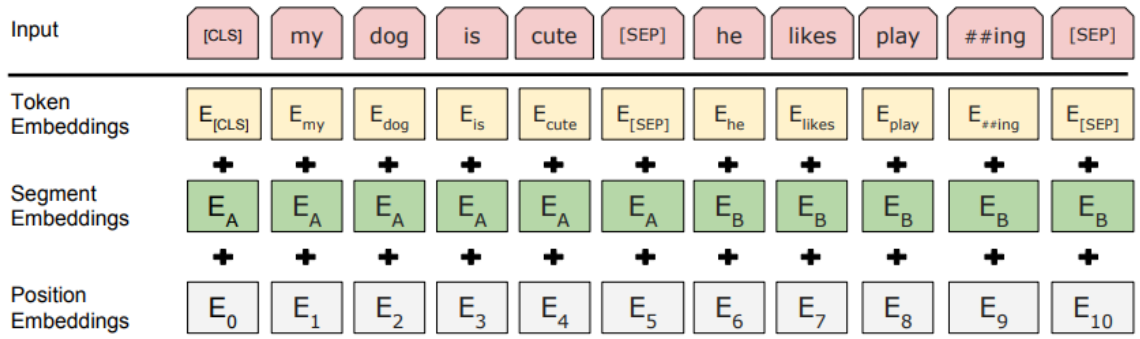
Uma abordagem diferente, também popular nas tarefas de PLN e exemplificada no recente artigo do ELMo, é o treinamento baseado em recursos. Nesta abordagem, uma rede neural pré-treinada produz embeddings de palavras que são usadas como recursos nos modelos de PLN. Essa é a ideia fundamental por trás do BERT.

**Como o BERT Funciona?**

O BERT faz uso do Transformer, um mecanismo de atenção que aprende relações contextuais entre palavras (ou sub-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 preciso 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 tabela abaixo é uma descrição de alto nível do codificador Transformer. A entrada é uma sequência de tokens, que são incorporados primeiro aos vetores e depois processados na rede neural. A saída é uma sequência de vetores, na qual 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 preveem a próxima palavra em uma sequência (por exemplo, "A mulher chegou na casa de ___"), com uma abordagem direcional que limita inerentemente a aprendizagem de contexto.

Para superar esse desafio, o BERT usa duas estratégias de treinamento: Masked LM (MLM) e Next Sentence Prediction (NSP), que veremos a seguir.

Aqui tem uma página incrível com um exemplo completo de funcionamento do BERT:

http://jalammar.github.io/illustrated-bert/