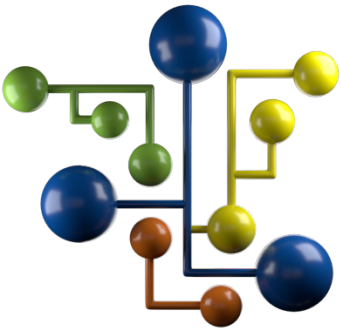

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Conhecendo o Modelo BERT

O BERT (Bidirectional Encoder Representations from Transformers) é o resultado de um paper recente publicado por pesquisadores do Google AI Language. O BERT causou alvoroço na comunidade de Machine Learning, apresentando resultados de ponta em uma ampla variedade de tarefas de PLN, incluindo respostas a perguntas, inferência de linguagem natural, entre outras atividades de PLN.

A principal inovação técnica do BERT é a aplicação do treinamento bidirecional do Transformer, um modelo de atenção, à modelagem de linguagem. Isso contrasta com os esforços anteriores, que analisavam uma sequência de texto da esquerda para a direita ou combinavam treinamento da esquerda para a direita e da direita para a esquerda. Os resultados da pesquisa mostram que um modelo de linguagem treinado bidirecionalmente pode ter um senso mais profundo do contexto e fluxo da linguagem do que os modelos de linguagem de direção única. No paper, os pesquisadores detalham uma nova técnica chamada Masked LM (MLM), que permite treinamento bidirecional em modelos nos quais era anteriormente impossível. Link do paper logo abaixo.

O modelo BERT foi construído sobre representações contextuais. O que o torna único é que é a primeira representação de linguagem profundamente bidirecional e sem supervisão, pré-treinada usando apenas um corpus de texto simples. Como é de código aberto, qualquer pessoa com conhecimento de aprendizado de máquina pode criar facilmente um modelo de PLN sem a necessidade de fornecer conjuntos de dados maciços para treinar o modelo, economizando tempo, energia, conhecimento e recursos.

O BERT é descendente direto do GPT (Generalized Language Models), e superou vários modelos de PLN fornecendo os melhores resultados no Question Answering (SQuAD v1.1), Natural Language Inference (MNLI) e outras estruturas.

Por fim, o BERT é pré-treinado em um grande corpus de texto não rotulado, que inclui toda a Wikipedia (cerca de 2.500 milhões de palavras) e um corpus de livros (800 milhões de palavras).

Aqui você encontra o paper da pesquisa, detalhes do projeto e o SQuAD, onde o BERT superou até o desempenho humano:

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/pdf/1810.04805.pdf

Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing
https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

SQuAD: 100,000+ Questions for Machine Comprehension of Text
https://arxiv.org/abs/1606.05250