Nome: Fabio Grassiotto RA: 890441
IA024 - Aula 4_5
Leitura do Artigo “BERT Pre-training of Deep Bidirecional Transformers for Language Understanding”

Resumo do Artigo:

Este artigo apresenta o modelo de representação de linguagem BERT (Bidirectional Encoder Representations from Transformers) baseado em Transformers bidirecionais.
Os autores justificam a necessidade de se implementar um modelo bidirecional dado que as técnicas existentes naquele momento restringiam a capacidade de representações pré-treinadas, especialmente quando os modelos são *fine-tuned* para tarefas *downstream* específicas.
O artigo explora trabalhos relacionados, ressaltando as propostas de treinamento não-supervionado baseado em features e *fine-tuning* já existentes no estado da arte assim como técnicas de transferência de aprendizado.
Na seção principal do artigo os autores apresentam o modelo, detalhando sua implementação e processos de pré-treinamento e *fine-tuning*.

- BERT é um modelo baseado na implementação original da arquitetura Transformer conforme descrito por Vaswani et al (2017).
- O modelo utiliza o tokenizador WordPiece com marcadores específicos para delimitação dos pares de sentença.
- O pré-treinamento do BERT não seguiu técnicas tradicionais, mas sim utilizando duas tarefas não supervisionadas, *MaskedLLM* e *Next Sentence Prediction (NSP)*. A primeira tarefa, MaskedLLM, elimina por através de máscaras cerca de 15% dos tokens de entrada de forma aleatória, enquanto que a segunda tarefa, NSP, envolve treinar um modelo para prever a próxima frase em uma sequência de texto dada a frase anterior.
- Os dados de pré-treinamento utilizados foram do BookCorpus, com 800 milhões de palavras, e a Wikipedia em inglês, com 2500 milhões de palavras.
- O processo de fine-tuning é simplificado pelo fato do BERT fazer o uso do mecanismo de auto-atenção da arquitetura Transformer. Para cada tarefa, os autores simplesmente utilizaram as entradas e saídas específicas e treinaram todos os parâmetros do início até o fim.

Após a apresentação do modelo, os autores mostram que o BERT supera os métodos de última geração anteriores em tarefas de NLP (GLUE, SQUAD v1.1 e 2.0, SWAG) demonstrando sua eficácia na captura do contexto e do significado das palavras nas frases.
Na conclusão, os autores estabelecem que a utilização de *transfer learning* com modelos de linguagem demostram que o uso do aprendizado não supervisionado é essencial para os sistemas de entendimento de linguagens, e que a contribuição principal do artigo foi a generalização dessa conclusão para arquiteturas profundas bidirecionais permitindo que modelos pré-treinados consigam lidar com tarefas amplas de NLP.