Comentário crítico: *Language Models are Unsupervised Multitask Learners*

Aluno: Maria Fernanda Bosco
RA: 183544

O artigo introduz o GPT-2, que aposta em uma escala massiva de dados e parâmetros. Diferente do BERT, que introduziu o paradigma bidirecional, o GPT-2 retoma o paradigma unidirecional e mostra que apenas a tarefa de *next token prediction* já é suficiente para gerar capacidades emergentes em diversas tarefas, caracterizando um *unsupervised multitask learning*.

A primeira preocupação dos autores foi em definir um conjunto de dados de treinamento grande e que tivesse maior diversidade de estilo e linguagem. Para isso, construíram o *WebText*, que foi formado a partir de links externos do *Reddit* que já tivessem sido verificados e avaliados pelos usuários da plataforma. Ou seja, textos que de alguma forma já foram considerados interessantes por algum ser humano. O *WebText* acabou sendo formado por 8 milhões de documentos com um total de 40GB de texto.

Além de uma escolha promissora do *dataset*, também escolheram uma abordagem diferente para a representação dos inputs, o *Byte Pair Enconde* (BPE). Diferente dos artigos que já tínhamos lido anteriormente, nesta abordagem ao invés de *tokenizar* as palavras uma a uma, cada string é decomposta em bytes (0–255), e o BPE aprende a agrupar sequências de bytes mais frequentes em tokens até formar um vocabulário de tamanho desejado (neste caso ficou em torno de 50k). Assim, palavras comuns tendem a ser representadas por um único token, enquanto palavras raras ou símbolos (incluindo emojis) são decompostos em sub-words, eliminando o problema de *out-of-vocabulary*.

Arquiteturalmente, o GPT-2 é um *Transformer decoder-only* com máscara causal em suas camadas de autoatenção. A versão XL possui 48 blocos, cada um composto por: *LayerNorm* (pré-subcamada), *Masked Multi-Head Self-Attention* (com máscara causal triangular e 25 heads), uma *Feed-Forward Network* (duas camadas lineares com ativação GELU) e conexões residuais. Esse modelo atinge 1.5B de parâmetros.

A máscara causal é um componente fundamental: ela impede vazamento de informação futura durante o treino, garantindo consistência entre o objetivo autoregressivo (prever um token dado somente os últimos) e a inferência (onde o futuro ainda não existe). Isso força o modelo a internalizar padrões de dependência causais da linguagem — sintaxe, gramática, estilo e coerência — e o torna altamente eficaz em prever a próxima palavra em tarefas de *language modeling*.