Destaques da leitura de **Attention Is All You Need**

Nome: Caio Petrucci dos Santos Rosa **RA:** 248245

Os destaques da leitura do artigo foram os seguintes:

- A utilização de *self-attention* para gerar novos *embeddings* para cada *token* no contexto utilizando os outros *tokens* do próprio contexto. Isso faz com que o modelo seja capaz de utilizar das outras informações no contexto para criar representações (*embeddings*) melhores dos *tokens* do contexto. Por exemplo, esse mecanismo de auto-atenção possibilita com que o modelo consiga aproximar palavras como “apple” de palavras de frutas em certos contextos e de palavras relacionadas a marcas de tecnologias em outros;
- No fundo, a arquitetura dos blocos propostos no artigo seguem uma estrutura relativamente simples.
  Na parte do *encoder*, as maiores diferenças estão na proposição da sub-camada de *multi-head attention* e da adição de um *positional embedding* que adiciona uma característica de posição aos *embeddings* do *input*. Pelo o que entendi, essa parte da arquitetura tem como objetivo realizar o melhor *encoding* possível do contexto de *input*, fazendo relações entre os diferentes *tokens* para extrair o máximo de significado semântico possível.
  Já na parte do *decoder*, além das proposições feitas para o *encoder*, foi proposto uma camada de *masked multi-head attention* que, até onde entendi, serve para tornar possível o paralelismo do treinamento dado que a característica auto regressiva da predição do próximo *token* continua sendo mantida através da definição dos pesos de conexões “ilegais” da atenção como -∞;
- A discussão sobre os usos de redes neurais recorrentes e convolucionais para lidar com a característica auto regressiva de sequência de textos e os trabalhos prévios que utilizaram essas abordagens. Os autores também apontam que alguns desses trabalhos já utilizaram mecanismos de atenção, porém podemos interpretar que, talvez pela dificuldade de criar representações latentes boas (como os *embeddings*) através dessas técnicas, o uso da atenção associado a recorrência de RNNs ou convoluções (que dão uma influência enorme para o posicionamento das informações) leva a resultados piores do que a auto-atenção sozinha;
- O uso de *residual connections* (ou *skip-connections*) para evitar com que as informações originais dos *embeddings* do *input* não seja totalmente perdida e misturada no empilhamento de camadas de auto-atenção;
- A utilização de *multi-head attention* para criar diferentes representações de “subespaços” de relacionamento entre *tokens*. Porém, não entendi muito bem porque essa abordagem soluciona um problema que os autores descreveram a partir do seguinte: “With a single attention head, averaging inhibits this.” no fim da página 4;
- As análises de performance e custo feitas, utilizando a complexidade assintótica, para comparar diferentes abordagens de arquitetura;
- Achei um pouco estranho que, na seção 5.4 Regularization, os autores mencionam que utilizaram 3 técnicas de regularização, porém listam apenas duas: **Residual Dropout** e **Label Smoothing**.