Nome: Rafael Claro Ito (R.A.: 118430)
Resumo do artigo: Attention Is All You Need

Os principais modelos usados para tarefas de tradução de texto até antes deste paper ser publicado eram basicamente baseados em redes neurais convolucionais ou redes recorrentes, tendo sempre como característica um *encoder* e um *decoder*. Este artigo propõe uma nova arquitetura denominada *Transformer*, sem convoluções ou recorrências, mas baseado puramente em mecanismos de atenção. Além de atingir resultados melhores, essa nova arquitetura pode ser treinada em um tempo bem menor do que os modelos anteriores.

As redes recorrentes eram bastante utilizadas devido a sua característica de dinâmica, por relacionar saídas/entradas anteriores à atual. No *Transformer*, as dependências globais entre entrada e saídas são obtidas através do mecanismo de atenção. Isso traz como principal vantagem a possibilidade de maior paralelização, algo mais difícil de se conseguir com as redes recorrentes devido a sua própria construção.

A arquitetura do modelo é dividida em *encoder* e *decoder*, sendo usado no modelo final um empilhamento de seis camadas de cada. No *encoder*, temos duas subcamadas. Inicialmente as palavras são convertidas em *embeddings* e somadas com *embeddings* posicionais. Em seguida, temos a primeira subcamada, denominada atenção *multi-head*, e por fim uma camada denominada *feed forward*. Ao final de cada subcamada tem-se uma conexão residual seguida de uma camada de normalização.

A subcamada de atenção *multi-head* funciona de maneira similar ao cálculo da auto-atenção padrão (*scaled dot-product*), com a diferença de agora haver uma projeção linear diferente para os valores de *queries*, *keys* e *values*. Além disso, são calculadas *h* camadas paralelas de atenção, tendo seus valores concatenados ao final. É importante notar que o custo computacional do *multi-head* é semelhante ao de um *single-head* de dimensionalidade completa, devido ao fato de cada cabeça ter suas projeções em uma dimensão reduzida (D/h).

Os *embeddings* posicionais são calculados a partir de funções senos e cosenos. Também é possível aprendê-los durante o treinamento, mas a Tabela 3 do artigo mostra que o resultado final fica quase que invariante. A subcamada de *feed-forward* nada mais é do que uma MLP de duas camadas com função de ativação ReLU.

No *decoder*, temos três subcamadas. Inicialmente temos os *embeddings* das palavras de saída (deslocados uma posição para direita) somados aos *embeddings* posicionais. Em seguida também temos uma camada de auto-atenção, com a diferença de aqui haver uma máscara que garante que predições de uma determinada posição só tenha conhecimento de saídas passadas àquela posição. Em seguida, temos uma nova subcamada de auto-atenção, mas que desta vez leva em conta as saídas do *encoder*. Por fim, tem-se uma subcamada *feed forward*. Assim como no *encoder*, ao final de cada subcamada temos a aplicação de uma soma residual seguida de uma normalização.

O treinamento da tarefa de tradução do inglês para alemão foi realizada no dataset WMT 2014 (*English-German*), assim como a tradução de inglês para francês (*English-French*), tendo durado 12 horas nos modelos mais simples e 3.5 dias no modelo final. Foram utilizados três métodos de regularização: *dropout* na camada de *embedding*, *dropout* na camada residual e *label smoothing*. Uma coisa que achei interessante foi o fato de os autores usarem um *learning rate* crescente durante os primeiros 4000 passos, para só depois disso começarem a decrescê-lo.

Os resultados apresentados são impressionantes. Para a tarefa de tradução do inglês para alemão o modelo atingiu 28.4 BLEU, superando todos modelos anteriores, incluindo *ensembles*, estabelecendo portanto um novo *SOTA* (*state-of-the-art*). Para a tarefa a tradução de inglês para francês o modelo obteve 41.0 BLEU, sendo este o melhor resultado de modelos únicos (sem ser *ensembles*) e tendo demorado menos que um quarto do tempo para treiná-lo!

Em seguida, o artigo mostra as diversas experimentações que foram feitas nos modelos treinados, com a tentativa de se analisar a contribuição dos valores de cada hiperparâmetro na métrica final. Por fim, os autores mostram-se bastante otimistas com o futuro dos modelos baseados em atenção, pretendendo aplicá-lo em outras tarefas além de texto, como em imagens, áudio e vídeo.