Attention Is All You Need

Este trabalho propõe uma arquitetura simples, chamada de Transformer, que é baseada inteiramente em mecanismos de atenção, prescindindo inteiramente da utilização de outras estruturas, como exemplo redes recorrentes ou convolucionais, apresentando resultados com alta qualidade, menos tempo de treinamento e capacidade de paralelização. Redes mais tradicionais como as citadas anteriormente, o número de operações necessárias para relacionar sinais de inputs e outputs arbitrariamente, cresce com a distância entre eles, os mecanismos de atenção podem paralelizar suas tarefas não dependendo das distâncias das entradas e saídas do sistema. A autoatenção relata diferentes posições de uma única sentença para a representação de toda sentença.

Neste modelo, o mapa transcreve uma sequência de símbolos para uma outra representação contínua, o decodificador então gera as saídas. A arquitetura Transformer utiliza esta sequência, empilhando mecanismos de autoatenção. O codificador é constituído por 6 camadas idênticas, cada uma delas com duas subcamadas. As camadas são cabeças variadas de autoatenção, redes totalmente conectadas de posicionamento. Estas são conectadas por mecanismos residuais e normalizações. O decodificador, também formado por 6 camadas, assim como o codificador, com a adição de uma terceira sub-camada, responsável pelo mecanismo de cabeças variadas de atenção para a saída da pilha do codificador. A autoatenção é modificada nestas camadas para evitar erros de sequência.

O mecanismo de atenção em si, é composto por três mapas, query, key e value, todos vetores. Os valores são computados como a soma com pesos dos valores, e os pesos para cada valor é atribuído de acordo com a função de query compatível. Para a atenção de escala do produto escalar, computa-se o produto escalar das queries com as keys divididas por $\sqrt{d_k}$ (dimensão de Key) e uma aplicação de softmax para obtenção dos pesos. Este mecanismo é mais ágil e eficiente que outros conhecidos, devido sua implementação com operações de matrizes. Para o mecanismo de múltiplas cabeças de atenção, ao invés de fazer uma única operação, projeta-se linearmente as queries, keys e values, h vezes com diferentes projeções e realiza-se às operações em paralelo, finalmente os resultados são concatenados e projetados para o resultado final. Este mecanismo permite que o modelo una as informações de atenção de diferentes subespaços e posições.

Foram utilizados embeddings para conversão das entradas, assim como a função de softmax para conversão das saídas em probabilidades. Visto que o modelo não utiliza redes recorrentes ou convolucionais, foram injetadas informações de posicionamento relativos e absolutos chamados de codificação posicional, qual diferentes possibilidades de escolha. A autoatenção foi utilizada principalmente por três motivos: a menor complexidade existente por camada, a quantidade de cálculo computacional requerida e a dependência de distância dentro da rede, comentado anteriormente, conectando todos os componentes por um custo computacional constante. Os modelos foram treinados nas bases de dados WMT 2014 English-German e English-French. Utilizou-se o otimizador Adam e learning rate adaptativo. Quanto a regularizações, aplicou-se dropouts residuais e Label Smoothing.

Entre os resultados obtidos, para a tradução inglês - alemão, resultou resultados superiores ao melhores reportados anteriormente, tornando-se estado da arte para o escore BLEU, sendo uma fração do custo dos modelos anteriores, em inglês - francês, obteve-se resultados superiores a todos as outras publicações de modelos únicos com ¼ do custo. Variantes também foram testadas, entre as principais observações, tem-se que uma única cabeça de atenção têm 0.9 BLEU abaixo da melhor e com muitas cabeças também têm-se uma pior qualidade. A redução da dimensão k piora o modelo, como esperado modelos maiores são melhores e o mecanismo de dropout é muito útil. Para a tarefa de parsing do inglês obteve-se bons resultados, inferiores apenas à RNN Grammar.

Apresenta-se um modelo baseado totalmente em mecanismos de atenção, treinados mais rápidos e estado da arte. Planeja-se agora aplicar em outras tarefas.