Comentário - Artigo "Attention Is All You Need"(2017)

Maria Fernanda Bosco

Agosto 2025

O artigo *Attention is All You Need* representa um marco na evolução dos modelos de linguagem. Em trabalhos anteriores, como *A Neural Probabilistic Language Model* (Bengio et al., 2003), foram introduzidos os embeddings e o uso de redes feed-forward para modelagem de linguagem, mas com limitação a contextos fixos. Na década seguinte, surgiram as RNNs/LSTMs, capazes de lidar com sequências de tamanho variável, mas ainda restritas por seu processamento sequencial e dificuldade em capturar dependências longas. Em seguida, modelos baseados em CNNs trouxeram paralelização e maior eficiência, mas exigiam muitas camadas para conectar palavras distantes. O Transformer rompe com essas limitações ao ser a primeira arquitetura baseada apenas em atenção, permitindo que todas as palavras se relacionem diretamente e de forma paralela.

A arquitetura proposta segue uma estrutura encoder–decoder. O encoder possui 6 camadas, cada uma com duas subcamadas: uma de *multi-head self-attention*, em que cada palavra pode "olhar" para todas as outras, e outra de *feed-forward*, com duas camadas densas aplicadas a cada posição. O decoder também possui 6 blocos, mas com três subcamadas: uma de *masked self-attention*, que limita cada palavra a olhar apenas para as anteriores, garantindo a geração auto-regressiva; uma de *encoder–decoder attention*, que conecta entrada e saída; e uma de *feed-forward*. Em todas as subcamadas há ainda conexões residuais seguidas de normalização, o que facilita o treinamento.

O mecanismo de atenção utilizado, denomidado *scaled dot-product attention*, calcula a similaridade entre pares de palavras a partir de matrizes de *queries, keys e values* (Q, K, V). O resultado é uma ponderação dos valores de cada palavra, misturando informações conforme sua relevância contextual. Para enriquecer as representações, o Transformer emprega multi-head attention: no modelo base, 8 cabeças de atenção paralelas, cada uma focando em aspectos distintos da sequência.

Como o modelo não possui nenhuma estrutura recorrente ou convolucional que preserve ordem, foi necessário introduzir o *positional encoding*. Esses vetores de posição, baseados em funções seno e cosseno de diferentes frequências, são somados aos *embeddings* das palavras. Dessa forma, a rede consegue distinguir posições absolutas e relativas, aprendendo a considerar tanto o significado das palavras quanto sua ordem na frase.

O treinamento foi realizado em tarefas de tradução inglês–alemão (4,5M pares de frases) e inglês–francês (36M pares). Foram definidos dois modelos: o base (512 dimensões de embedding, 8 cabeças de atenção) e o big (1024 dimensões, 16 cabeças). Avaliados pela métrica BLEU, padrão em tradução automática, os modelos alcançaram resultados superiores ao estado da arte, com destaque para o Transformer big (28.4 BLEU em EN–DE e 41.0 BLEU em EN–FR). Além da qualidade, os tempos de treino foram significativamente menores: 12h para o modelo base e 3,5 dias para o big, contrastando com semanas de treinamento necessárias em arquiteturas anteriores.

Assim, o artigo introduz o Transformer, primeiro modelo seq2seq baseado apenas em atenção. Ele trouxe maior paralelização, eficiência e desempenho em tradução automática, além de abrir caminho para aplicações além do texto (como imagens, áudio e vídeo). Essa proposta inaugurou a era dos LLMs modernos, que têm nos Transformers sua base arquitetural.