Resumo – Attention Is All You Need

O artigo propõe uma nova arquitetura de rede neural, o Transformer, para tarefas de transdução de sequência. Os modelos então dominantes eram baseados em redes recorrentes ou convolucionais que incluíam um codificador e um decodificador. O Transformer se baseia exclusivamente em mecanismos de atenção, dispensando blocos de recorrência ou convolução.

Redes recorrentes trabalham com sequências de estados escondidos, onde um estado é baseado no estado anterior para a mesma posição. Essa natureza sequencial impede a paralelização durante o treinamento do modelo e cria restrições no uso de memória. Mecanismos de atenção foram incorporados a redes que contêm recorrências, modelando melhor a dependência entre termos das sequências, independente da distância entre eles, garantindo assim melhor desempenho, mas contendo ainda as desvantagens citadas. Transformers permitem paralelização muito maior, atingindo novo estado da arte em tradução.

A arquitetura é composta por seis codificadores idênticos empilhados e seis decodificadores também idênticos entre si e empilhados. O codificador mapeia uma sequência de representações simbólicas de entrada $X = (x_1, ..., x_n)$ para uma sequência de representações contínuas $Z = (z_1, ..., z_n)$. Cada camada de codificador é formada por duas subcamadas, uma executando um mecanismo de auto-atenção multi-cabeça, a outra sendo uma rede densa tradicional, com conexão residual em torno de cada uma seguida de normalização. O decodificador recebe Z e gera uma sequência de símbolos de saída $Y = (y_1, ..., y_n)$ um elemento por vez. A cada passo o modelo usa os símbolos gerados anteriormente como entrada para gerar cada novo símbolo. Cada camada de decodificador é similar à de codificador, mas contendo uma terceira subcamada no meio, que executa auto-atenção multi-cabeça sobre a saída da pilha de codificadores. Além disso, a primeira sub-camada de atenção é modificada de forma a prevenir posições de atentarem para posições subsequentes, garantindo a auto-regressão.

O mecanismo de atenção (scaled dot-product attention) mapeia uma query Q e um conjunto de pares chave-valor (K e V) para a saída. Recebe com entrada vetores de dimensão d_k e valores de dimensão d_v, e aplica a função: $\text{Attention}(Q; K; V) = \text{softmax}(QK^T / \sqrt{d_k})V$. A normalização por raiz de d_k é empregada para evitar produtos crescerem muito no caso de valores altos de d_k. A atenção multi-cabeça aplica h projeções lineares aprendidas distintas sobre queries, chaves e valores, em paralelo, que depois são concatenadas e projetadas novamente, resultando na saída da sub-camada.

Cada sub-camada densa é composta por duas camadas tradicionais entremeadas por uma ativação ReLU: $FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$. A saída do Transformer é uma camada formada por V neurônios, onde V é o tamanho do vocabulário, seguida por uma softmax, fornecendo como resultado um vetor de probabilidades de cada token do vocabulário a cada passo. Para que o modelo capture a ordem dos tokens de entrada, na entrada ele soma ao encoding de conteúdo um encoding de posição, que pode ser fixo (uma função sinusoidal) ou aprendido.

O conjunto de dados WMT 2014 English-German padrão, que consiste de 4,5 milhões de pares de sentenças, bem como o WMT 2014 English-French, com 36 milhões de pares de sentenças, foram utilizados no experimento. Cada lote de treinamento consistiu de pares de sentenças contendo aproximadamente 25000 tokens em cada língua. Rodou em uma máquina com 8 GPUs NVIDIA P100. Cada passo de treino em modelos base demorou em média 0,4 segundos, totalizando 100 mil passos em 12 horas. Nos modelos maiores, 1 segundo por passo, 300 mil passos, totalizaram 3,5 dias. Otimização Adam com $\beta_1=0,9$, $\beta_2=0,98$ e $\varepsilon=10^{-9}$, e learning rate em rampa, e ao final decrescendo. Aplicado dropout com taxa de 0,1 ao final de cada subcamada, antes da soma residual e normalização, e label smooting de 0,1.

A avaliação foi realizada em escore BLEU. No inglês-alemão, atingiu escore 28,4, dois pontos abaixo do melhor resultado anterior. No inglês-francês, atingiu escore 41,0, também recorde, com ¼ do custo em relação a modelos prévios. Foram testadas variações em hiper-parâmetros do modelo, conseguindo melhor desempenho o big model.

De fato, desde então, os modelos Transformers, baseados em mecanismos de atenção, em diversas variações, têm dominado o cenário de aplicações em linguagem natural desde então.