Principais contribuições do artigo “Attention is all you need” (seções 1 a 3)

Aluno: Leandro Carísio Fernandes

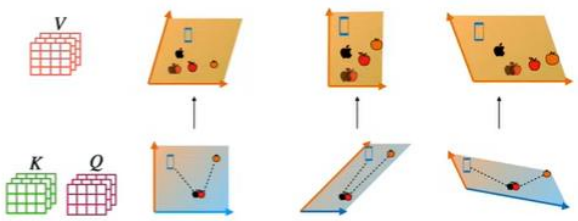
- A principal proposta é apresentar os Transformers, novo modelo de rede neural que depende inteiramente do mecanismo de auto-atenção. Diferentemente dos modelos anteriores (RNN), permite paralelizar a entrada (em vez de ser um processamento serial).
- O modelo de auto-atenção é definido como:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Q, K e V são gerados a partir de três matrizes (WQ, WK e WV) de tamanho d_k (a dimensão dos embeddings). Assim, a ideia três matrizes desse tamanho que serão treinadas.

O fluxo do mecanismo de auto-atenção é:

1. A rede recebe w, um batch de tamanho (B, L), onde B é a quantidade de amostras no batch e L é o tamanho do contexto (total de palavras).
2. Para cada palavra do batch, seus embeddings são extraídos, gerando x, uma matriz de tamanho (B, L, d_k) => cada palavra é representada por um vetor de embeddings d_k-dimensional.
3. São geradas WQ, WK e WV de tamanho (d_k, d_k)
4. É calculado Q = xWQ, K = xWK e V = xWV (operação matricial). Assim, Q, K e V tem a dimensão (B, L, d_k)
5. É calculado QK^T. Note que, como estamos em batch, é necessário executar tudo dentro do mesmo batch. Assim, a operação feita é com as dimensões (B, L, d_k) e (B, d_k, L). A saída da operação (e entrada do softmax) tem dimensão (B, L, L).
6. É feito o softmax na última dimensão (_, _, AQUI) e o resultado é multiplicado por V. Assim, multiplica-se (B, L, L) por (B, L, d_k), para cada batch, obtendo-se (B, L, d_k).
7. Esse resultado tem as mesmas dimensões da entrada. O que é feito é uma transformação linear dos embeddings para obter melhores embeddings. Isso é ilustrado no vídeo [1]:



[1] https://www.youtube.com/watch?v=UPtG_38Oq8o&ab_channel=Serrano.Academy