Summary of the Transformer in **Attention Is All You Need**

Huaqing Mao

1 Paper Information

- **Title: Attention Is All You Need**
- **Authors: Ashish Vaswani et al.**
- **Journal/Conference: 31st NIPS**
- **Year of Publication: 2017**

2 Summary

2.1 Embedding and positional encoding

We use a single sentence as input for illustration. Suppose we have a sentence S, 'This is an sentence.', we tokenize S and get a sequence of tokens 'This', 'is', 'a', 'sentence', $<\mathrm{EOS}>$, $<\mathrm{PAD}>$,...,$<\mathrm{PAD}>$, where $<\mathrm{EOS}>$ is the end-of-sentence token, and $<\mathrm{PAD}>$ the padding to make the sequence have the fixed size L we specify, assuming the input and output both have fixed length L. Apply a dictionary mapping tokens to indices, so now we have a sequence of integers that represents the sentence S. We convert the sequence of shape (L) to (L, d_{model}) by learned/learnable embeddings of dimension d_{model}.

To encode the information of relative positions, positional encoding is used

$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{\text{model}}}) \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{\text{model}}}) \end{aligned} \tag{1}$$

Then we add the 'positional encodings' Eq. 1 to the input embeddings to get Q, K and V of size (L, d_{model}) and pass them into the encoder and decoder stacks.

2.2 Multi-head attention

For the i_{th} head of a multi-head attention specified by $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, we pass QW_i^Q, KW_i^K and VW_i^V, in attention

Eq. 2, Q,K,V are the same tensor of size (L, d_{model}). Hence, QW_i^Q, KW_i^K and VW_V^Q are of shape (L, d_k), (L, d_k) and (L, d_v).

$$\mathrm{Attention(A, B, C) = softmax(\frac{AB^T}{\sqrt{d_b}})C} \tag{2}$$

We concatenate all attention head $\mathrm{Attention(QW_i^Q, KW_i^K, VW_i^V)}$ of shape (L, d_v), and multiply with $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$, $\mathrm{Concat(head_1, ..., head_h)W^O}$. The resulting multi-head attention has the shape of (L, d_{model})

2.3 Position-wise Feed-Forward Networks

After the Add&Norm sublayers, we feed the tensor of shape (L, d_{model}) to the fully connected layer,

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

The first dimension is not touched. Tensors of size d_{model} in all positions go through the same linear transformations and Relu activation $(d_{\text{model}}) \rightarrow (d_{\text{ff}}) \overset{\text{Relu}}{\rightarrow} (d_{\text{model}})$