Lecture 5 Transformer Appendix

Multi-Head Self Attention

In self-attention, query:

$$q = a \cdot W^q$$

$$a: T \times d_{model}$$

$$W^q: d_{model} \times d_{model}$$

In multi-head self-attention, the query is split into two heads, number of heads = 2

Head length $d_k = \frac{d_{model}}{num.\ of\ heads}$

Query is now calculated in Head 1 and Head 2 as:

$$q^1 = a \cdot W^{q,1}$$

$$q^2 = a \cdot W^{q,2}$$

where $a: T \times d_{model}$ and

$$W^{q,1}: d_{model} \times d_k$$

$$W^{q,2}: d_{model} \times d_k$$

Following the same logic,

$k: T \times d_{model}$ is decomposed into $k^1: T \times d_k$ and $k^2: T \times d_k$.

$v: T \times d_{model}$ is decomposed into $v^1: T \times d_k$ and $v^2: T \times d_k$.

Two heads represent two types of relevance.

In Head 1

$$\alpha^1 = (q^1 \cdot k^{1^T})/\sqrt{d\_k}$$

where $\alpha^1: T \times T$, because $q^1: T \times d_k$ and $k^1: T \times d_k$ whose transpose $k^{1^T}: d_k \times T$.

$$\hat{\alpha}^1 = softmax(\alpha^1)$$

Output from the Head 1 is

$$b^1 = \hat{\alpha}^1 \cdot v^1$$

where $b^1: T \times d_k$ because $\hat{\alpha}^1: T \times T$ and $v^1: T \times d_k$.

Repeat the above process in Head 2, the output from the multi-head self-attention layer is

$$b = [b^1 \quad b^2]$$

where $b: T \times d_{model}$ because $b^1: T \times d_k$ and $b^2: T \times d_k$