Lecture 5 Transformer Appendix

Layer Normalization

Take Transformer – Encoder as example,

Given N = 6, when layer n = 1

$$residual = a + b^0$$

$a: T \times d_{model}$ is the input (i.e. output from the positional encoding step) fed into the multi-head self-attention in layer n = 1. T is the sequence length (the number of words in a sentence)

$b^0: T \times d_{model}$ is the output from the multi-head self-attention in layer n = 1.

To apply Layer Normalization to $residual$:

$$residual = \begin{bmatrix} residual_{1,1} & \dots & residual_{1,d\_model} \\ \vdots & \vdots & \vdots \\ residual_{T,1} & \cdots & residual_{T,d\_model} \end{bmatrix}$$

Calculate μ_1 to μ_T as follows:

$$\mu_1 = \frac{1}{d_{model}} \sum_{i=1}^{d\_model} residual_{1,i}$$

$$\dots$$

$$\mu_T = \frac{1}{d_{model}} \sum_{i=1}^{d\_model} residual_{T,i}$$

Calculate σ_1 to σ_T as follows:

$$\sigma_1^2 = \frac{1}{d_{model}} \sum_{i=1}^{d\_model} \left(residual_{1,i} - \mu_1\right)^2$$

$$\dots$$

$$\sigma_T^2 = \frac{1}{d_{model}} \sum_{i=1}^{d\_model} \left(residual_{T,i} - \mu_T\right)^2$$

For any word t in a sequence, its ith embedding dimension after layer normalization is:

$$c_{t,i} = \frac{residual_{t,i} - \mu_t}{\sigma_t}$$

Note: each embedding dimension is treated as a feature in machine learning. Layer normalization is thus taken along each timestamp and across all embedding dimensions. Batch normalization is thus taken along each embedding dimension and across all timestamps.