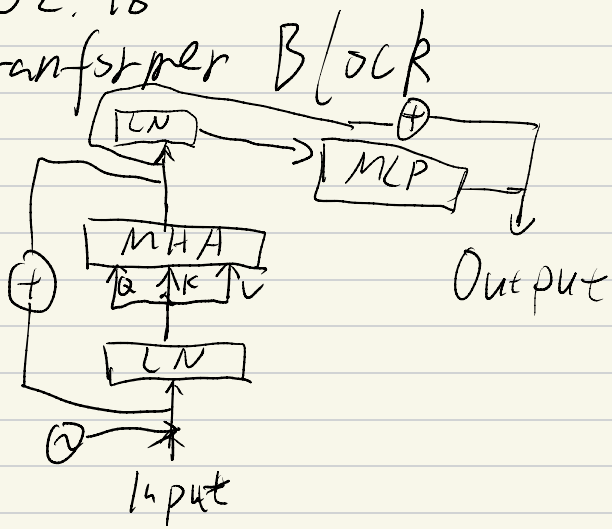
25.02.18

Transformer Block



① MHA : Parallelization

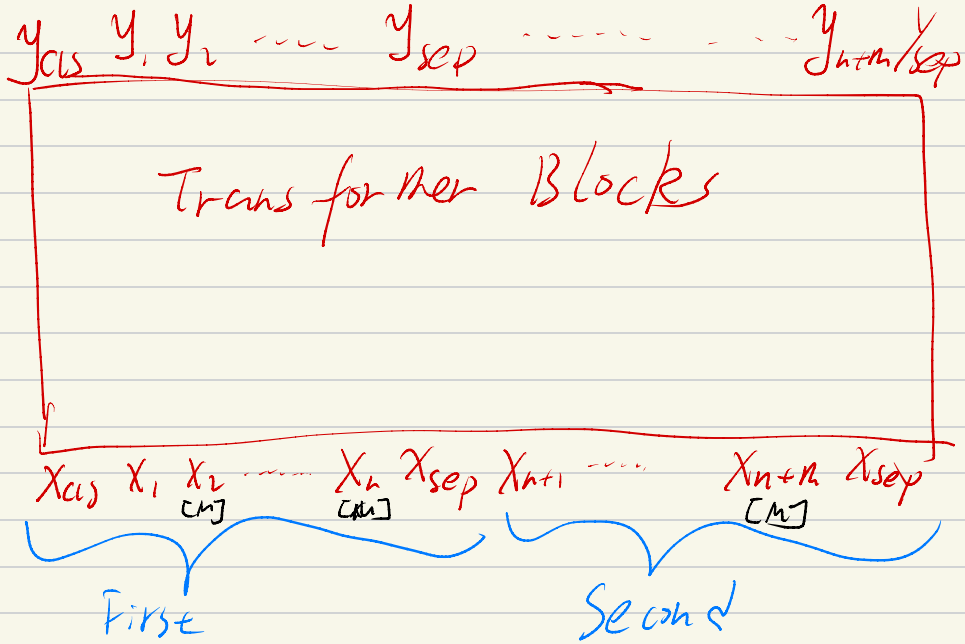
② ResNet / LN : Better Optimization

③ MLP : Non-linearity Information Compression

④ Position Encoding Sequence Info

⑤ Future - Sequence Masked Attention

Out



MLM: For all [M], [r], [n]

we use y_2 to predict.

$$z_2 = Softmax(A_{MLM} \cdot y_2 + b_{MLM})$$

$A_{MLM} \in \mathbb{R}^{|V| \cdot d_{model}}$

Loss: $-\log(z_2|_{\text{Actual Token of } x_2})$

NSP: $\text{Sigmoid}(A_{NSP} \cdot y_{cls} + b_{cls})$

$A_{NSP} \in \mathbb{R}^{d_{model}}$

Loss: Cross-Entropy