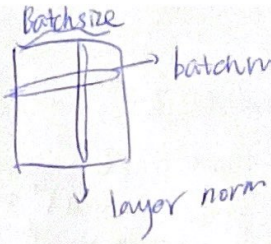
Transformer 的 summarizer

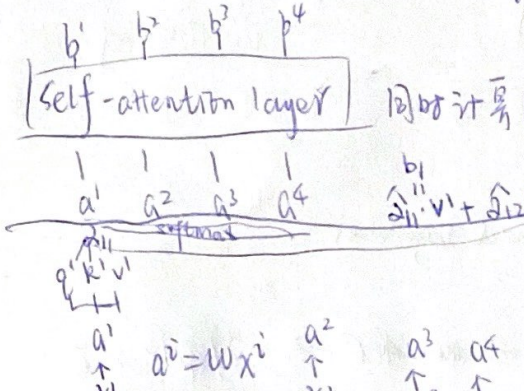
seq2seq model with self-attention

RNN 问题：hard to parallel



所以有人提出CNN代替RNN (Filter in higher layer can consider longer sequence)

但低层的只能考虑long sequence



$a^i = W x^i$

query (to match others): $q^i = W^q a^i$

key (to be matched): $k^i = W^k a^i$

value (information to be extracted): $v^i = W^v a^i$

对每个 q 去对每个 k 做 attention

Scaled dot-product attention: $\alpha_{1,2} = q^1 \cdot k^2 / \sqrt{d}$

d: dim of q and k

$A = K^T Q \rightarrow \hat{A}$: A softmax columnwise, $\rightarrow O = V\hat{A}$

input I

output O

$Q = W^q \cdot I$

$K = W^k \cdot I$

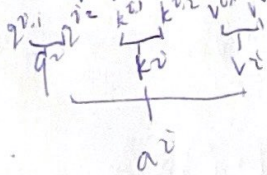
$V = W^v \cdot I$

$A = K^T Q \rightarrow \hat{A} \rightarrow O = V\hat{A}$

变形：multi-head self-attention

多个head关注local or 长时间 (global)

2-head



$q^{i,1} = W^{q,1} q^i$

$q^{i,2} = W^{q,2} q^i$

多个head各自独立做 attention，得到 $b^{i,1}$, $b^{i,2}$

head1 head2

可 $\begin{pmatrix} b^{i,1} \\ b^{i,2} \end{pmatrix}$ 乘矩阵得 $b^i = W^0 \begin{pmatrix} b^{i,1} \\ b^{i,2} \end{pmatrix}$

No position information in self-attention：缺点想要加位置 positional encoding

positional vector e^i fixed not learned

$e^i + a^i$：把 x^i 与 one-hot position vector 做 append

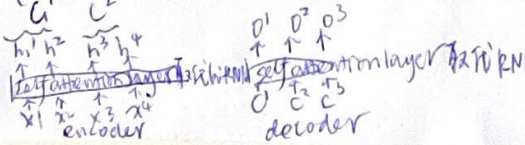
$$(W^I \; W^P)\begin{pmatrix} x^i \\ p^i \end{pmatrix} = W^I x^i + W^P p^i$$

$W^I x^i \rightarrow a^i$

$W^P p^i \rightarrow e^i$

可以learn，但是人设计

seq2seq with attention



Transformer 结构