$\boxed{\text{Self-Attention}}$ encodes a sequence into a form that accounts for the other respective data in the sequence

a
b
c

→ $\boxed{\text{Self-Attention}}$ →

$a' \mid b,c$
$b' \mid a,c$
$c' \mid a,b$

* each element in new form is aware of surrounding data

---

① .

k_1 v_1 k_2 v_2
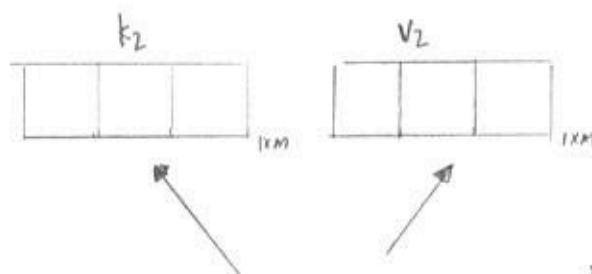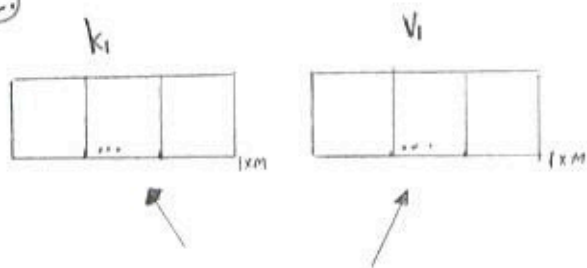
Input 1   Input 2

Q_1 Q_2

$W = [K_{n \times m}, V_{n \times m}, Q_{n \times m}]$

~ $n \Rightarrow$ input dim
~ $m \Rightarrow$ output dim

Ex: $\begin{bmatrix} \text{Input 1} \\ \text{Input 2} \end{bmatrix}_{2 \times n} K_{n \times m}$

$= \begin{bmatrix} k_{1, (1 \times m)} \\ k_{2, (1 \times m)} \end{bmatrix}_{2 \times m}$

$\begin{bmatrix} \text{Input 1} \\ \text{Input 2} \end{bmatrix}_{2 \times n} V_{n \times m} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_{2 \times m}$

$\begin{bmatrix} \text{Input 1} \\ \text{Input 2} \end{bmatrix}_{2 \times n} Q_{n \times m} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}_{2 \times m}$

---

② . Repeat for each input

For Input 1:

Ⓐ. $S_{1, (1 \times 1)} = Q_1 k_1^T$
$S_{2, (1 \times 1)} = Q_1 k_2^T$
$\hat{S}_1 = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$

→ Ⓑ. $\text{softmax}(\hat{S}_1) \to \hat{\hat{S}}_1$

→ Ⓒ. $\hat{v}_1 = v_1 \odot \hat{\hat{S}}_1$
$\hat{v}_2 = v_2 \odot \hat{\hat{S}}_2$

→ Ⓓ. $\text{output\_1}_{(1 \times m)} = \sum_{i=1}^{2} \hat{v}_i$

---

③ .

$\boxed{\text{Input 1}}_{1 \times n}$ $\boxed{\text{Input 2}}_{1 \times n}$ → $\boxed{\text{Self-Attention}}$ → $\boxed{\text{output\_1}}_{1 \times m}$ $\boxed{\text{output\_2}}_{1 \times m}$

$W = [K_{n \times m}, V_{n \times m}, Q_{n \times m}]$

Multi-head attention

$$X = \begin{array}{|c|} \hline a \\ \hline b \\ \hline \end{array}$$

Self-attention - 1 → $\begin{array}{|c|} \hline a'_1, 1b \\ \hline b'_1, 1a \\ \hline \end{array}$

o
o
o

Self-attention - n → $\begin{array}{|c|} \hline a'_n, 1b \\ \hline b'_n, 1a \\ \hline \end{array}$

→ Concatenate → Dense → X^*

Transformers

Encoder

#-blocks N_x

Add & Norm
Feed Forward
Add & Norm
Multi-Head Att. ⟵ doesn't care about order

Positional Encoding ⊕

Input Embedding ⟵ push to care more about position

Inputs
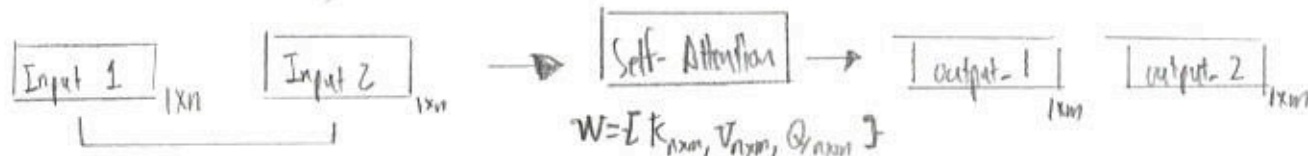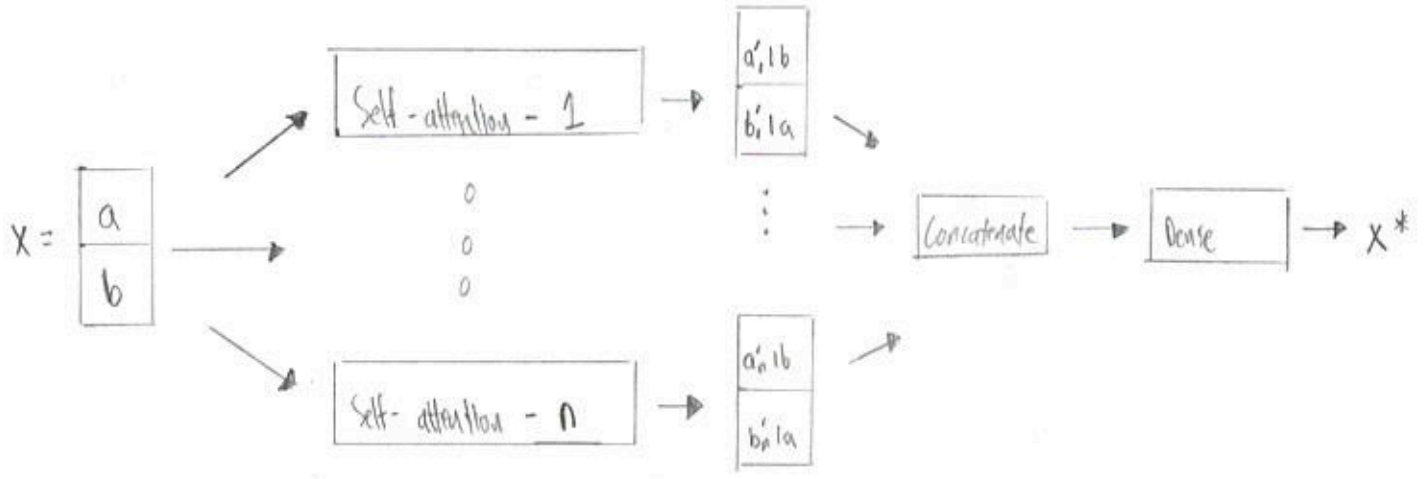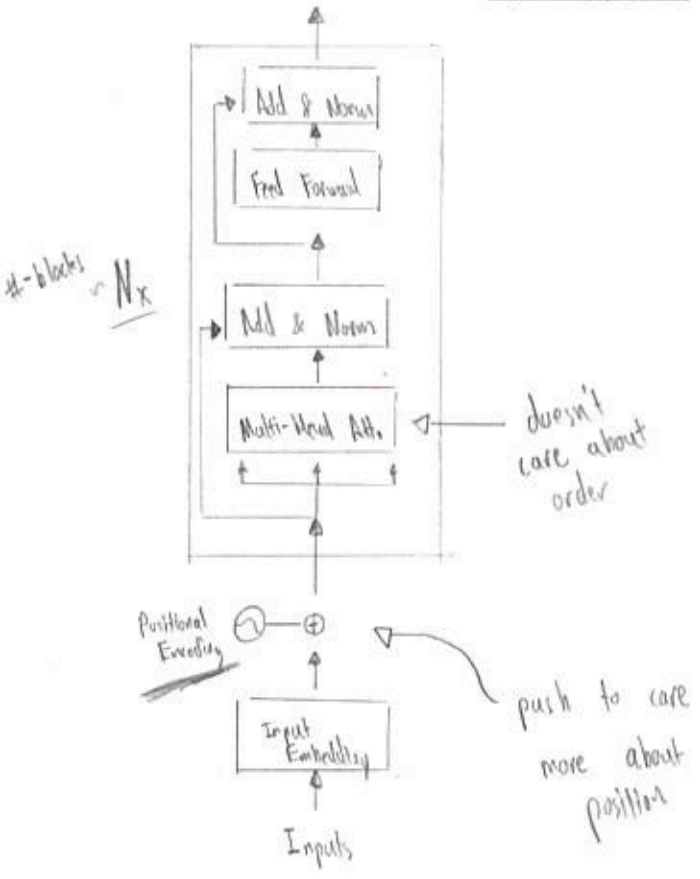
Recurrent Network vs MH-attention
- RN can't be parallized, MH-att can
- RN not numerically stable