

# Transformer

## Sequence to Sequence Model

Context

Input

### 1. Word Embedding

- Word ID 512
- nn.Embedding

### 2. Positional Encoding

- Transformer 用 RNN
- Positional
  - 位置向量
  - 位置嵌入

## Encoder

Encoder block 6 個

block

### 1. Multi-Head Self Attention

- Q, K, V  $\rightarrow$  “”

### 2. Feed Forward Network

- MLP
- Layer Norm

Layer Norm

- Layer Norm

### Layer Normalization

## Decoder

Input

- Decoder
- Auto-regressive

Output

Decoder 3 層 + LayerNorm

- Masked Multi-Head Self-Attention
- Encoder-Decoder Attention

### 1. Masked Multi-Head Self-Attention

- Masked Multi-Head Self-Attention
- Decoder “”

### 2. Encoder-Decoder Attention

- Decoder Encoder

- **Query**  $\otimes$  **Decoder-Key**  $\otimes$  **Value**  $\otimes$  **Encoder**  $\otimes$
- Decoder “ $\otimes$ ” Encoder  $\otimes$  Decoder “ $\otimes$ ” Encoder  $\otimes$  Decoder “ $\otimes$ ” Encoder  $\otimes$

### 3. Feed Forward Network

- $\text{ffn}(\text{input}) + \text{ffn}(\text{ReLU/GeLU})(\text{input})$

### 4. $\text{ffn}$ + LayerNorm

LayerNorm

- $\text{LayerNorm}(\text{output} = \text{input} + \text{Sublayer}(\text{input}))$
- LayerNorm  $\otimes$

- $\text{ffn}(\text{input}) + \text{ffn}(\text{ReLU/GeLU})(\text{input}) + \text{LayerNorm}(\text{input})$



## Train

### 1. Encoder

- $\text{Encoder}(\text{input})$
- $\text{Encoder}(\text{input})$

### 2. Decoder

- $\text{Decoder}(\text{input}, \text{label})$
- $\text{Decoder}(\text{input}, \text{label})$

### 3. Teacher Forcing

- $\text{Decoder}(\text{input}, \text{token})$  “ $\text{token}$ ”  $\otimes$   $\text{Decoder}(\text{input}, \text{token})$
- $\text{Decoder}(\text{input})$
- $\text{Decoder}(\text{input})$

## Attention

- **Decoder Self-Attention Mask**  $\otimes$   $\text{Decoder}(\text{input})$
- **Encoder-Decoder Attention**  $\otimes$   $\text{Decoder}(\text{input})$  “ $\otimes$ ”  $\text{Encoder}(\text{input})$

Attention

- $\text{Attention} \rightarrow \text{softmax} \rightarrow \text{Attention}$
- $\text{Attention}$  one-hot  $\otimes$   $\text{Decoder}(\text{input})$   $\rightarrow$  **Cross Entropy Loss**

Attention

Attention

Attention

## Teaching Forcing

$\text{Decoder}(\text{input})$  “ $\otimes$ ”  $\text{Teacher Forcing}$   $\otimes$   $\text{Decoder}(\text{input})$

## Decoder の構造

- 入力の最初の要素は  $\langle \text{BOS} \rangle$  である
- 各時間ステップで “ $\langle \text{EOS} \rangle$ ” が発生する
- 各時間ステップで “ $\langle \text{UNK} \rangle$ ”
- …となる

↑ これは Decoder の構造を示す図

## Teacher Forcing

### Residual Connection

これは

逐次的解釈における “ $\langle \text{EOS} \rangle$ ” の発生を “ $\langle \text{UNK} \rangle$ ” とする “ $\text{UNK}$ ” の発生を “ $\langle \text{EOS} \rangle$ ” とする

これは

- 逐次的解釈の  $y = F(x) + x$
- 逐次的解釈の  $\partial L / \partial x$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial x} = \frac{\partial L}{\partial y} \cdot (\frac{\partial y}{\partial F(x)} + 1)$$

- 逐次的解釈の  $\frac{\partial L}{\partial x}$  は  $1$  である

### Layer Normalization

逐次的解釈における Layer Normalization

- BatchNorm は逐次的解釈における Layer Normalization
- NLP における BatchNorm は逐次的解釈における batch によるもの
- Layer Normalization は逐次的解釈における feature dimension による Batch Normalization
- feature dimension による

### Explanation

#### 逐次的解釈の self-attention

これは



- 逐次的解釈における self-attention の構造
- 逐次的解釈における self-attention の構造

```
$\$ \begin{aligned} Q &= \begin{bmatrix} q^1 & q^2 & q^3 & q^4 \end{bmatrix} \\ \begin{bmatrix} a^1 & a^2 & a^3 & a^4 \end{bmatrix} W^q &\quad K = \begin{bmatrix} k^1 & k^2 & k^3 & k^4 \end{bmatrix} \\ K^T &= \begin{bmatrix} v^1 & v^2 & v^3 & v^4 \end{bmatrix} \\ V &= \begin{bmatrix} a^1 & a^2 & a^3 & a^4 \end{bmatrix} W^v \end{aligned} \$\$
```

- 逐次的解釈における attention score
- ```
$\$ \begin{aligned} A &= \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} \\ \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} \\ \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & \alpha_{4,4} \end{bmatrix} \\ &\quad Q \cdot A^T = Q \cdot \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} \\ \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} \\ \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & \alpha_{4,4} \end{bmatrix} \end{aligned} \$\$
```

3.  $\text{softmax}(d_k) \cdot \text{key}/\text{querey}$

4. □□□□□□□□□□

```
 $$ [b^1,b^2,b^3,b^4] = \begin{bmatrix} \alpha^1 & \alpha^1 & \alpha^1 & \alpha^1 \\ \alpha^1 & \alpha^2 & \alpha^2 & \alpha^2 \\ \alpha^1 & \alpha^2 & \alpha^3 & \alpha^3 \\ \alpha^1 & \alpha^2 & \alpha^3 & \alpha^4 \end{bmatrix} \cdot \begin{bmatrix} v^1 \\ v^2 \\ v^3 \\ v^4 \end{bmatrix} $$

```

□ \$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d\_k}})V\$

5

□□□□□□□□□□□□□□□□\$W^{\{q,1\}},W^{\{q,2\}}\$



A horizontal row of fifteen empty square boxes, intended for children to write their names in, likely as part of a classroom activity.

- \$A^i,A^j,A^m,A^n\$ \$W^q W^k W^v\$ \$Q K V\$
  - \$W^q W^k W^v\$ \$Q K V\$

```
$$ Q = \begin{bmatrix} a^i & a^j & a^m & a^n \end{bmatrix} W^q \quad K = \begin{bmatrix} a^i \\ a^j \\ a^m \\ a^n \end{bmatrix} W^k \quad V = \begin{bmatrix} a^i & a^j & a^m & a^n \end{bmatrix} W^v \quad $$
```

- 

```
$Q\equiv K\otimes V\$ \begin{array}{l} Q_1 = QW^{\{q,1\}} \\ Q_2 = QW^{\{q,2\}} \\ K_1 = KW^{\{q,1\}} \\ K_2 = KW^{\{q,2\}} \end{array} \begin{array}{l} V_1 = VW^{\{q,1\}} \\ V_2 = VW^{\{q,2\}} \end{array} \end{array} \$
```

- 

```
 $$ \begin{aligned} Q_1 &= \begin{bmatrix} q^{i,1} & q^{j,1} & q^{m,1} & q^{n,1} \\ q^{i,2} & q^{j,2} & q^{m,2} & q^{n,2} \end{bmatrix} \\ Q_2 &= \begin{bmatrix} q^{i,1} & q^{j,1} & q^{m,1} & q^{n,1} \\ q^{i,2} & q^{j,2} & q^{m,2} & q^{n,2} \end{bmatrix} \end{aligned} $
```

```
\end{align} $$
```

```
 $$ K_1 = \begin{bmatrix} k^{i,1} & k^{j,1} & k^{m,1} & k^{n,1} \end{bmatrix} K_2 = \begin{bmatrix} k^{i,2} & k^{j,2} & k^{m,2} & k^{n,2} \end{bmatrix} $$
```

```

$$ V_1 = \begin{bmatrix} v^{i,1} & v^{j,1} & v^{m,1} & v^{n,1} \end{bmatrix} V_2 = \begin{bmatrix} v^{i,2} & v^{j,2} & v^{m,2} & v^{n,2} \end{bmatrix} $$

• ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်

$$ \text{head\_1} = \text{softmax}(\frac{\text{Q}_1 \text{K}_1^T}{\sqrt{d_k}}) \text{V}_1 \text{ head\_2} = \text{softmax}(\frac{\text{Q}_2 \text{K}_2^T}{\sqrt{d_k}}) \text{V}_2 $$

• ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်

$$ \text{multihead} = \begin{bmatrix} \text{head\_1} & \text{head\_2} \end{bmatrix} \text{W}^O $$

• ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်

```

## Self-attention

ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်  
**FC**များကိုသွင်းဆောင်ရန်

### Sequence Labeling

ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်  
• ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန် **FC**များကိုသွင်းဆောင်ရန်



ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်  
• ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်

- ဗိုလ်ချုပ်(\$Query, Q)\$
- ဗိုလ်ချုပ်(\$Key, K)\$
- ဗိုလ်ချုပ်(\$Value, V)\$

ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်

- ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန် \$Q \otimes K \otimes V\$ မှာ
  - ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန်
    - \$Q = X \times W^Q\$
    - \$K = X \times W^K\$
    - \$V = X \times W^V\$

• ဗိုလ်ချုပ်များကိုသွင်းဆောင်ရန် (Q) မှာ အတွက် (K) မှာ အတွက် **attention score** မှာ ဖြစ်ပါသည်

- dot product မှာ အတွက် (Q) မှာ အတွက် (K) မှာ အတွက် " " \* " " \* " "
  - အတွက် (Q) မှာ အတွက် (K) မှာ အတွက် " " \* " " \* " "
    - အတွက် (K) မှာ အတွက် (Q) မှာ အတွက် " " \* " " \* " "

- $\text{softmax}^{**"Q\text{-Key"}**"**\text{Q-Value"}**}$
  - $\text{softmax}(\text{Q-Value})$  softmax( $\text{Q-Value}$ )
  - $\text{softmax}(\text{Q-Value}) \cdot \text{softmax}(\text{Q-Value})$
  - $\text{softmax}(\text{Q-Value}) \cdot \text{softmax}(\text{Q-Value})$  attention score
  - $\text{softmax}(\text{Q-Value})$  softmax( $\text{Q-Value}$ )
  - $\text{Q-K}^{**"Q\text{-Key"}**"**\text{Q-Value"}**}$   $\text{Q-K}^{**"Q\text{-Key"}**"**\text{Q-Value"}**}$
  - $\text{V}^{**"Q\text{-Key"}**"**\text{Q-Value"}**"**\text{Q-Value"}**}$   $\text{V}^{**"Q\text{-Key"}**"**\text{Q-Value"}**"**\text{Q-Value"}**}$

## Multi-head Self-attention

██████████ \$Q\$ ██████████

- $\text{Q}^T \mathbf{a}^T \mathbf{i} \mathbf{S} \mathbf{B} \mathbf{A} \mathbf{B}^T \mathbf{S}^T \mathbf{B}^T$ 
    - $\mathbf{Q}^T \mathbf{b}^T \{i, 1\} \mathbf{S}$ 
      - $\mathbf{Q}^T \mathbf{q}^T \{i, 1\} \mathbf{B} \mathbf{k}^T \{i, 1\} \mathbf{B}^T \mathbf{b}^T \{i, 1\} \mathbf{S}$
    - $\mathbf{Q}^T \mathbf{b}^T \{i, 2\} \mathbf{S}$ 
      - $\mathbf{Q}^T \mathbf{q}^T \{i, 2\} \mathbf{B} \mathbf{k}^T \{i, 2\} \mathbf{B}^T \mathbf{b}^T \{i, 2\} \mathbf{S}$
    - $\mathbf{Q}^T \mathbf{b}^T \{i\} \mathbf{S}$ 
      - $\mathbf{Q}^T \mathbf{b}^T \{i, 1\} \mathbf{S} \mathbf{B}^T \mathbf{b}^T \{i, 2\} \mathbf{S} \mathbf{B} \mathbf{A} \mathbf{B}^T \mathbf{S}^T \mathbf{B}^T \mathbf{b}^T \{i\} \mathbf{S}$  **attention score**
    - $\mathbf{Q}^T \mathbf{b}^T \{i, 3\} \mathbf{S}$



## Masked Multi-Head Self-Attention(遮罩多頭自我注意力)

A decorative horizontal bar consisting of a series of small, evenly spaced rectangular boxes, likely a graphic element or a separator line.

"A B C D" မြန်မာစာတမ်း token များ

**K\begin{bmatrix} 1 & \text{masked} & \text{masked} & \text{masked} \\ 1 & 1 & \text{masked} & \text{masked} \end{bmatrix}**

masked\ 1 &1& 1 &masked\ 1&1&1&1 \end{bmatrix} \$\$ 

## Word Embedding

## One-Hot Encoding

- 
  - 

## Positional Encoding

- **Self-attention** は「**自己注意**」の意で、各単語が他の単語と関連性をもつていて、その関係性を学習する機能です。
  - 例として、「apple banana orange」の3つの単語を入力すると、各単語が他の2つの単語と関連性があることを学習します。

11 / 11

1. 位置向量 **positional vector** 位置向量
  2. 语义向量

### **Final\_embedding = Token\_embedding + Positional\_encoding**

Positional Encoding

Positional Encoding

Positional Encoding

Positional Encoding

$\text{PE}_{\{(pos, 2i)\}} = \sin(pos/10000^{2i/d_{\text{model}}})$   $\text{PE}_{\{(pos, 2i+1)\}} = \cos(pos/10000^{2i/d_{\text{model}}})$

1.  $\text{pos}$
  2.  $2i \leq d_{\text{model}}$
  3.  $d_{\text{model}}$
- I am a kid
    - I pos=0
      - I pos=0
      - am pos=1
    - am pos=1
      - i=0
        - $\text{PE}_{\{(1,0)\}} = \sin(1/10000^{0/d_{\text{model}}})$
        - $\text{PE}_{\{(1,1)\}} = \cos(1/10000^{0/d_{\text{model}}})$
      - i=1
        - $\text{PE}_{\{(1,2)\}} = \sin(1/10000^{2/d_{\text{model}}})$
        - $\text{PE}_{\{(1,3)\}} = \cos(1/10000^{2/d_{\text{model}}})$

Positional Encoding

- $w_i = 10000^{\frac{2i}{d_{\text{model}}}}$ 
  - Positional
  - $\text{pos}/w_i$  10000 Positional Encoding
  - Positional LSTM
- Positional
  - $d_{\text{model}} w_i$   $10000^{\frac{2i}{d_{\text{model}}}}$   $\text{pos}/10000^{\frac{2i}{d_{\text{model}}}}$  Positional
  - Positional

Positional Encoding

$\underbrace{\begin{bmatrix} \text{PE}_{\{(pos + \Delta, 2i)\}} & \text{PE}_{\{(pos + \Delta, 2i+1)\}} \end{bmatrix}}_{\text{Final embedding}} = \underbrace{\begin{bmatrix} \cos(\Delta \theta_i) & \sin(\Delta \theta_i) \\ -\sin(\Delta \theta_i) & \cos(\Delta \theta_i) \end{bmatrix}}_{\text{Rotation matrix}} \underbrace{\begin{bmatrix} \text{PE}_{\{(pos, 2i)\}} & \text{PE}_{\{(pos, 2i+1)\}} \end{bmatrix}}_{\text{Positional Encoding}}$

$\underbrace{\begin{bmatrix} \cos(\Delta \theta_i) & \sin(\Delta \theta_i) \\ -\sin(\Delta \theta_i) & \cos(\Delta \theta_i) \end{bmatrix}}_{\text{Rotation matrix}} \underbrace{\begin{bmatrix} \text{PE}_{\{(pos, 2i)\}} & \text{PE}_{\{(pos, 2i+1)\}} \end{bmatrix}}_{\text{Positional Encoding}} = \begin{bmatrix} \cos(\Delta \theta_i) \text{PE}_{\{(pos, 2i)\}} + \sin(\Delta \theta_i) \text{PE}_{\{(pos, 2i+1)\}} & -\sin(\Delta \theta_i) \text{PE}_{\{(pos, 2i)\}} + \cos(\Delta \theta_i) \text{PE}_{\{(pos, 2i+1)\}} \end{bmatrix}$

```

\theta_i = \frac{1}{10000^{\frac{2i}{d_{model}}}} \\

\$\underbrace{ \begin{bmatrix} \cos(\Delta \theta_i) & \sin(\Delta \theta_i) \\ -\sin(\Delta \theta_i) & \cos(\Delta \theta_i) \end{bmatrix} }_{\Delta \theta_i} \$

\Delta \theta_i\$pos\\

```



```

\$ \$ \$ \begin{aligned} \begin{bmatrix} PE_{(pos + \Delta, 2i)} & PE_{(pos + \Delta, 2i+1)} \\ \end{bmatrix} &= \begin{bmatrix} \sin((pos + \Delta) \cdot \theta_i) & \cos((pos + \Delta) \cdot \theta_i) \\ \cos((pos + \Delta) \cdot \theta_i) & \sin((pos + \Delta) \cdot \theta_i) \end{bmatrix} \\ &= \begin{bmatrix} \sin(pos \cdot \theta_i) \cos(\Delta \cdot \theta_i) + \cos(pos \cdot \theta_i) \sin(\Delta \cdot \theta_i) & \cos(pos \cdot \theta_i) \cos(\Delta \cdot \theta_i) - \sin(pos \cdot \theta_i) \sin(\Delta \cdot \theta_i) \\ \cos(pos \cdot \theta_i) \cos(\Delta \cdot \theta_i) - \sin(pos \cdot \theta_i) \sin(\Delta \cdot \theta_i) & \sin(pos \cdot \theta_i) \cos(\Delta \cdot \theta_i) + \cos(pos \cdot \theta_i) \sin(\Delta \cdot \theta_i) \end{bmatrix} \\ &= \begin{bmatrix} \cos(\Delta \cdot \theta_i) & \sin(\Delta \cdot \theta_i) \\ -\sin(\Delta \cdot \theta_i) & \cos(\Delta \cdot \theta_i) \end{bmatrix} \begin{bmatrix} PE_{(pos, 2i)} & PE_{(pos, 2i+1)} \\ \end{bmatrix} \end{aligned} \\

```

\end{aligned} \\$

### Transformer\\$

```
\$ \$ \$ \begin{aligned}
```

```
w_i &= 10000^{2i/d_{model}} \theta_i &= pos / 10000^{2i/d_{model}} \theta_i &= pos / w_i
```

```
\end{aligned} \$
```

- $w_i = 10000^{2i/d_{model}} \theta_i = pos / 10000^{2i/d_{model}}$
- $w_i = pos / 10000^{2i/d_{model}} \theta_i = pos / 10000^{2i/d_{model}} \cdot 0.1 \Delta$   
 $pos = 10000 \frac{1000}{10000} = 0.1 \Delta$
- $w_i = 10000 \sin(pos/10000) / pos \approx 20000 \pi \approx 62831.85$

•

- $QK^T$
- $(query \cdot key)^*/(query \cdot query)^*/(key \cdot key)^*$

```

\$ q_i = W_q(x_i + p_i) \ k_j = W_k(x_j + p_j) \$\$

\$ q_i^T k_j = (x_i + p_i)^T W_q^T W_k (x_j + p_j) = \underbrace{x_i^T W_q^T W_k}_{\text{query}} \underbrace{x_j}_{\text{key}}
+ \underbrace{x_i^T W_q^T W_k p_j + p_i^T W_q^T W_k x_j}_{\text{cross}}
+ \underbrace{p_i^T W_q^T W_k p_j}_{\text{key-key}} \$\$
```

```

\$ q_i^T k_j = x_i^T W_q^T W_k x_j + \underbrace{p_i^T W_q^T W_k p_j}_{\text{key-key}} = x_i^T W_q^T W_k x_j
+ \underbrace{\beta_{i-j}}_{\text{cross}} \$\$ \beta_{i-j} = \frac{\text{dot product}}{\text{norm of query} \cdot \text{norm of key}}
```

**T5**

•

•

**ALib**

QK\$...\$m\$



m\_h=2^{-\frac{8 \times h}{n\\_head}}\\$ n\\_head \\$  
m=\begin{bmatrix} 2^{-\frac{8 \times 1}{n\\_head}} & 2^{-\frac{8 \times 2}{n\\_head}} & \dots & 2^{-\frac{8 \times n\\_head}{n\\_head}} \end{bmatrix} \\$\\$ D\_{ij} = -(i - j)

softmax(q\_i K^T + m \cdot \text{dot}[-(i-1), \dots, -2, -1, 0]) \\$\\$

## RoPE

QK

Attention\_score = q\_m \cdot k\_n^T \\$\\$ Attention\_score = q\_m \cdot k\_n^T \\$\\$ q^{\text{rot}} = q\_m \cdot R\_m \cdot k^{\text{rot}} = k\_n \cdot R\_n \\$\\$ \begin{aligned} q^{\text{rot}} \cdot k^{\text{rot}} &= q\_m R\_m R\_n k\_n^T \\ &= q\_m R\_m R\_{-n} k\_n^T \\ &= q\_m R\_{m-n} k\_n^T \end{aligned}

1. 旋转矩阵

2. 余弦矩阵

\$m, n\$ \$m, n\$ \$m-n\$

余弦矩阵

- 余弦矩阵 RoPE \$d/2\$ 余弦矩阵

$R_{\theta,m} = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & \dots & 0 & \sin m\theta_1 & \cos m\theta_2 & -\sin m\theta_2 & 0 & \dots & 0 & \sin m\theta_{m-n} & \cos m\theta_{m-n} & 0 & \dots & 0 & \sin m\theta_{m-n} & \cos m\theta_{m-n} \end{bmatrix}$

- 正弦矩阵

\$\$

**\begin{bmatrix} x\_{i}' & x\_{i+1}' \end{bmatrix}**

$\begin{bmatrix} \cos m\theta_i & -\sin m\theta_i \\ \sin m\theta_i & \cos m\theta_i \end{bmatrix} \begin{bmatrix} x_i & x_{i+1} \end{bmatrix}$

\$\$

\$(m-n)\$ 余弦矩阵

$\theta_i = \frac{1}{2} \cdot \frac{10000}{d_{\text{model}}} \cdot i$  \$\\$ R\_{m-n} = \begin{bmatrix} \cos((m-n)\theta\_i) & -\sin((m-n)\theta\_i) \\ \sin((m-n)\theta\_i) & \cos((m-n)\theta\_i) \end{bmatrix} \$\\$ R\_n^T R\_m = R\_{m-n}\$ \$\\$ R\_{-m} = R\_m^T

\$\$ (R\_nq)^{TR\\_mk} = q^{TR\_{\{m-n\}}k} \quad \text{and} \quad q\_n^{\wedge \{rot\}} = R\_nq \backslash k\_m^{\wedge \{rot\}} = R\_mk \quad \text{and} \quad QK\\$ \rightarrow

### KV Cache

- $QKV$
- $QKV$
- $QKV$
- $QKV$

token  $\rightarrow$   $QKV$   $\rightarrow$   $QKV$

QKV

- $QKV$
- $QKV$
- $QKV$
- $QKV$

QKV Cache  $\rightarrow$  “ $QKV$ ”  $\rightarrow$   $QKV$

### KV Cache



### KV Cache



## Pre-Norm | Post-Norm

Pre-Norm  $\rightarrow$  Post-Norm  $\rightarrow$  Transformer  $\rightarrow$  Layer Norm (LN) Layer  
Normalization

- $QKV$
- $QKV$   $\rightarrow$  Transformer  $\rightarrow$  LN  $\rightarrow$   $QKV$
- $QKV$