Transformer

□Sequence to Sequence Model□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□**Context**□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□

1. **Word Embedding**□□□□□□□

- □□□□□□□□□□ID□□□□□□□□□□□ 512□□
- □□□□□□ `nn.Embedding` □□□□

2. **Positional Encoding**□□□□□□

- □□ Transformer □□ **RNN** □□□□□□□□□□□□□□□□□□□
- □□□□□□
  - □□□□□□□□□□□□□□□
  - □□□□□□□□□□□□□□□□

Encoder□□□□

□□ Encoder □□□□□□□□□□□**block**□□□□□ 6 □□□

□□□□**block**□□□□□□□□□

1. □□□□□□□□**Multi-Head Self Attention**□
   - □□□□□□□ → □□“□ □ □□”□□□□□□□□□□
2. □□□□□□□**Feed Forward Network**□
   - □□□□□□□□□□□□□□□□□□□□ MLP
   - □□□□□□□□

□□□□□□□□□

- □□□□
- **Layer Normalization**

Decoder□□□□

□□□□□□□

- □□□□□□□□□□□□□□□□□□□□□□□□
- □**Auto-regressive**□□□□□□□

□□□□

□□□ Decoder □□ **3** □□□ + □□□□ **+ LayerNorm**□

- □□□
- □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

1. **Masked Multi-Head Self-Attention**□□□□□□□□□□
   - □□□□□□□□□□□□□“□□□□□□□□□”
   - □□□□□ Decoder □□□□“□□□□□” □ □□□□□□
2. **Encoder-Decoder Attention**□□□□□□□□
   - □□□□ Decoder □□□ Encoder □□□□□□□□

    - **Query □□ Decoder□Key □ Value □□ Encoder □□□□**

    - □ Decoder □“□□”□□□□□□□□□□□□□□□□□□□□□□ Decoder“□□”□□□□□□□□□□/□□/□□□□□□

3. **Feed Forward Network**□□□□□□□□

    - □□□□□□□□□ + □□□□□**ReLU/GeLU**□□

4. **□□□□ + LayerNorm**

   □□□□□□□□

    - □□□□□ `output = input + Sublayer(input)`
    - LayerNorm□□□□□□□□□□□□

- **□□□□□□□□□□n□□□□□□□□□□**

image-20250702225431058

Train□□□□□□

1. **Encoder**□

    - □□□□□□□□□□□□□□□□□□□□□□□□□□□□□

    - □□□□□□□□□□□□□□□□□□□□□

2. **Decoder**□

    - □□□□□□□□□□□□□□□□□□□□□□ label □□□□□□□

    - □□□□□□□□□□□□□□“□ □ ___”□□“□”□

3. **Teacher Forcing**□

    - □□□□□□□□□□□□□□□□□□□□□□□□□□□**token□□“□□□□”**□□□□□□□□□□□□token
    - □□□□Decoder □□□□□□□□□□□□□□
    - □□□□□□□□□□ □ □□□□□□□□□□□

Attention□□

- **Decoder □□□ Self-Attention□ Mask □□□□□□□□□□□□□□□□□□□□□**
- **Encoder-Decoder Attention**□ Decoder □□□□□□“□□” Encoder □□□□□□□□□□□□□□□□□□□□□□

□□□□

- **□□□□□□□ → softmax → □□□□□□□□□□**
- □□□□□ one-hot □□□□□ → □□ **Cross Entropy Loss**

□□□□

□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□

Teaching Forcing

□□□□Decoder □□□□□“□□□□□□□□”□□□□□□□“□□□□□□□□”□□□□□□□ `Teacher Forcing` □□□□□□□

□□ Decoder □□□□

- □□□□□□□□□□□□□□□ `<BOS>` □□□□□
- □□□□□ □□□□□□□□□□“□”□□□□□
- □□□□□□□□“□ □”
- …□□□□

□ □□ □□□□□□□□□□□□□□□□□□□□□□

□□□□□□ Teacher Forcing□

Residual Connection□□□□□□

□□□□

□□□□□□□□□□□□□□□□□□□"□□"□"□□□□□"□□□□□□□□□□x□□□□F(x)□□□□□□□□□□□□□x + F(x)

□□□□□□

- □□□□□□□□□ $y = F(x) + x$□□□□□□
- □□□□□□□□□□ ∂L/∂x □□□□□□□□□

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

- □□ $\frac{\partial{F(x)}}{\partial{x}}$ □□□□□1□□□□□□□□□□□□

Layer Normalization

□□□□□□□□□□□□□□□□□□□□□

- □ **BatchNorm** □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
- □ NLP □□□□□□ BatchNorm □□□□□□□□□□□□□□batch □□□□□□□
- **Layer Normalization□□□□□feature□□□dimension□□□□□□Batch Normalization□□□□□feature□□□□dimension□□□□□**

Explanation

□□□□□□□□□□**self-attention□□□□□**

□□□□□

image-20250427121123206

> 1. □□□□□□□□□□□□□□□□□□a^1,a^2,a^3,a^4
> 2. □□□□□□□□□□□□□□□
>
> $$ \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 \[1em] K = \begin{bmatrix}k^1 \ k^2 \ k^3 \ k^4\end{bmatrix} &= \begin{bmatrix}a^1 \ a^2 \ a^3 \ a^4\end{bmatrix}W^k \[1em] V = \begin{bmatrix}v^1 \ v^2 \ v^3 \ v^4\end{bmatrix} &= \begin{bmatrix}a^1 \ a^2 \ a^3 \ a^4\end{bmatrix}W^v \ \end{aligned} $$
>
> 2. □□***attention score*** $$ \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} &=Q \cdot K^T\ &=

> \begin{bmatrix}q^1 \ q^2 \ q^3 \ q^4\end{bmatrix} \cdot \begin{bmatrix}k^1 & k^2 & k^3 & k^4\end{bmatrix} \end{aligned} $$
>
> 3. □□$\sqrt{d_k}$□□□**softmax**□d_k□□□$key/querey$□□□□□□□
>
> $$ \begin{aligned} &\quad\quad\quad\quad\quad\quad\text{□□□□ } A \xrightarrow{\text{softmax}} \text{□□□□ } A^{'} \ &\frac{1}{\sqrt{d_k}}\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 \Rightarrow \quad \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} $$
>
> 4. □□□□□□□□□□□□
>
> $$ [b^1,b^2,b^3,b^4] = \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} \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 $$

□□□□□

□□□□□□□□□□□□□□□□□□□$W^{q,1},W^{q,2}$

image-20250510194255934

□□□□□□□□□□□□□□□□□□□□

- □□□□□□□a^i,a^j,a^m,a^n□□□□$W^q□W^k□W^v$□□□□□□□□□□□□□□□□□□$Q□K□V$
- □□□$W^q□W^k□W^v$□□□□□□□

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

- □□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□**$Q□K□V$**□□□□ $$ \begin{align} Q_1 = QW^{q,1} \quad Q_2 = QW^{q,2} \ K_1 = KW^{q,1} \quad K_2 = KW^{q,2}\ V_1 = VW^{q,1} \quad V_2 = VW^{q,2} \end{align} $$

- □□□□□□

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

\end{align} $$

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

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

- [illegible]

$$ head_1= softmax(\frac{Q_1K_1^T}{\sqrt{d_k}})V_1\ head_2= softmax(\frac{Q_2K_2^T}{\sqrt{d_k}})V_2 $$

- [illegible]

$$ multihead = \begin{bmatrix} head_1 & head_2 \end{bmatrix} $$

- [illegible]

$$ output = multihead\cdot W^O $$

Self-attention

[illegible]

[illegible]**FC**[illegible]**FC**[illegible]

Sequence Labeling[illegible]

[illegible]

- [illegible]**FC**[illegible]

image-20250418130432286

[illegible]

[illegible]

[illegible]

- [illegible]$(Query, Q)$
- [illegible]$(Key, K)$
- [illegible]$(Value, V)$

[illegible]

- [illegible]$Q[illegible]K[illegible]V$[illegible]
    - [illegible]
        - $Q = X × W^Q$
        - $K = X × W^K$
        - $V = X × W^V$

      [illegible]$W^Q[illegible]W^K[illegible]W^V$[illegible]

- [illegible](Q)[illegible](K)[illegible]**attention score**[illegible]
    - dot product[illegible]
        - [illegible](Q)[illegible]**"[illegible]"**
        - [illegible](K)[illegible]**"[illegible]"**

        - □□□□□□**"□□□□□□□□□□□□□□□□□□"**

- □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□**softmax**□□□□□□□□□□□□□□□□□

- □□softmax□□□□□□□RELU□□□□□□□□□□□□□□□

- □□□□□□□□□□□□□□□□□(V)□□□□□□□□□□attention score□□□□□□□□
    - □□□(V)□□□□□□□□□□□□□
    - Q-K□□□□□**"□□□□□□□□"**□□□□□□□
    - V□□□□□□□□**"□□□□□□□□□"□□□□□□**

Multi-head Self-attention□□□□□□□

□□□□□□□□$Q$□□□□□□□□□□□□□□□□□□

- □□$a^i$□□□□□□□□□
    - □□$b^{i,1}$
        - □□$q^{i,1}□k^{i,1}□k^{j,1}$□□□$b^{i,1}$
    - □□$b^{i,2}$
        - □□$q^{i,2}□k^{i,2}□k^{j,2}$□□□$b^{i,2}$
    - □□$b^{i}$
        - □□$b^{i,1}$□$b^{i,2}$□□□□□□□□□□□□$b^{i}$□□□**attention score**
    - □□□□□□□□□□

image-20250420112828589

Masked Multi-Head Self-Attention(□□□□□□□)

□□

□□ `"A B C D"` □□□□□□□□□□□token□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□K□V□□ $$ \begin{bmatrix} 1 & masked &masked&masked\ 1 & 1 &masked & masked\ 1 &1& 1 &masked\ 1&1&1&1 \end{bmatrix} $$ image-20250702191639250

Word Embedding□□□□□

□□□□□□□□□□□□□□**One-Hot Encoding**□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

- □□□□□□□□□□□□□□□□
- □□□□□□□□□□□□□□□□

Positional Encoding

- **Self-attention**□□□□□□□□□□□□"□□□□□□"□□□□□□□□□□□□□□□□□□□□□□□□□□
- □□

□□□□□□□□□□□

1. □□□□□□□**positional vector**□□□□□□□□□□□□□□
2. □□□□□□□□□□□□□□□□□□□□□□□□

> ***Final_embedding = Token_embedding + Positional_encoding***

□□□□□□

□□□□□□□□□□□□□□

□□□□□□□□□□□

□□□

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

1. $pos$□□□□□
2. $2i$□□□□□□$2i\le d_{model}$
3. $d_{model}$□□□□□□□□□□□□□□□□□□□□□$d_{model}$□□□□

- □□I am a kid
    - □□□□
        - I□pos□0
        - am□pos□1□□□□□□□
    - □□□□□□□am□pos=1□
        - i=0
            - □0□□$PE_{(1,0)} = sin(1/10000^{0/d_{model}})$
            - □1□□$PE_{(1,1)}=cos(1/10000^{0/d_{model}})$
        - i=1
            - □2□□$PE_{(1,2)}=sin(1/10000^{2/d_{model}})$
            - □3□□$PE_{(1,3)}=cos(1/10000^{2/d_{model}})$

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

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

□□□□□□□□□

$$ \underbrace{\begin{bmatrix} PE_{(pos + \Delta,2i)} \ PE_{(pos + \Delta,2i+1)} \end{bmatrix}}_{□□(pos + \Delta)□□□□}

\underbrace{ \begin{bmatrix} cos(\Delta \theta_i) & sin(\Delta \theta_i)\ -sin(\Delta \theta_i) & cos(\Delta \theta_i) \end{bmatrix}}*{□□□□□□} \underbrace{ \begin{bmatrix} PE*{(pos ,2i)} \ PE_{(pos ,2i+1)} \end{bmatrix}}_{□□pos□□□□} $$

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

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

□□□□□□□□□□□□□□□$\Delta\theta_i$□□□□□□□**pos**□□□

image-20250701224305593

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

\end{align} $$

Transformer□□□□□□□□□□□□□□□

□□□ $$ \begin{align}

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

\end{align} $$

- i□□$w_i$□□□□□□□□□□□□□□□$\theta_i$□pos□□□□□□□□□□□□
- i□□$w_i$□□□□□□□□□□□□□□□$\theta_i$□pos□□□□□□□□□□□□□□□□□□□□□□□□0□□□$\Delta pos=1000$□$\frac{1000}{10000} = 0.1$□□□□□□□□□
- □□$w_i$□10000□$sin(pos/10000)$□pos□□$20000\pi$□□

□□□□□□

- □□□□□□□□□□□□□□□□□□□□□□□□□□□□$QK^T$□□□

- □□□□□□□□□□□□□□□□□□□□□□□□**(query□key□□□□)□□□□□□□□□□□**□□□$Q$□$K$□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□$Q$□$K$□□□□□ $$ q_i = W_q(x_i+p_i)\ k_j=W_k(x_j+p_j) $$ □□□□□ $$ q_i^Tk_j = (x_i+p_i)^TW_q^TW_k(x_j+p_j)\ =\underbrace{x_i^TW_q^TW_kx_j}{□□□□□ □}+\underbrace{x_i^TW_q^TW_kp_j+p_i^TW_q^TW_kx_j}{□□-□□□□ □}+\underbrace{p_i^TW_q^TW_kp_j}_{□□□□□□} $$ □□□□□□□□□□□□□□□□□□□□□

$$ q_i^Tk_j =x_i^TW_q^TW_kx_j + \underbrace{p_i^TW_q^TW_kp_j}{□□□□□} \ = x_i^TW_q^TW_kx_j +\underbrace{\beta{i-j}}{□□□□□} $$ □$\beta{i-j}$□□□□□□□□□□□□□□□□□□□□□□□□□□□□□$head_h$□□□□□□ □□□$\beta_{i-j}^h$(why??/)

T5□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□

ALibi□□□□□□□□□□□

□□□$QK$□□□□□□□□□□□□□□□□□□□□□□□□□□□□□$m$

image-20250702192315616

□□□m□□□□□□□□□□□□$m_h=2^{-\frac{8\times h}{n_head}}$□$n_{head}$□□□□□□□□□□ $$ m=\begin{bmatrix}2^{-\frac{8\times 1}{n_head}}\2^{-\frac{8\times 2}{n_head}}\ \cdots\2^{-\frac{8\times n_{head}}{n_head}}\end{bmatrix} $$ □□□□□□□□□□□□□□□□□□□□□□□□□□□$D_{ij} = -(i - j)$

□□□□□ $$ softmax(q_iK^T+m\cdot[-(i-1),\dots,-2,-1,0]) $$

RoPE□□□□□□□□

□□□□□□□□□□□□□□□□□□□□Q□K□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□ $$ Attention_score = q_m\cdot k_n^T $$ □□□□□□□□□□□□□□q□k□□□□□□□□□ □□□ $$ q^{rot} = q_m\cdot R_m \ k^{rot} = k_n \cdot R_n $$ □□□ $$ \begin{aligned} q^{rot}\cdot k^{rot} &= q_mR_mR_n^Tk_n^T \ &= q_mR_mR_{-n}k_n^T \ &= q_mR_{m-n}k_n^T \end{aligned} $$ □□□□□□□□□□□□$Q□K$□□□□□□□□**RoPE**□□□□□□□□□□□□□□□□□□□□□□□□□□□

1. □□□□□□□

   □□

2. □□□□□

□□□□□□ $m,n$ □□□□□□□□□□□□□□□□$m-n$

□□□□□□□

- □□□□□d□□□□RoPE□□□□□□$d/2$□□□□□□□□□□□□□□□**d**□□□□□□

$$ R_{\theta,m} = \begin{bmatrix} \cos m\theta_{1} & -\sin m\theta_{1} & 0 & \cdots & 0 \ \sin m\theta_{1} & \cos m\theta_{1} & 0 & \cdots & 0 \ 0 & 0 & \cos m\theta_{2} & -\sin m\theta_{2} & 0 \ 0 & 0 & \sin m\theta_{2} & \cos m\theta_{2} & 0 \ \vdots & \vdots & \vdots & \ddots & \vdots \ \end{bmatrix} $$

- □□□□□□□□□□□□□□□□□□

$$

\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}

$$

□□□□□$(m-n)$□□□□□□□

□□$\theta_i = \frac{1}{10000^{\frac{2i}{d_{model}}}}$ $$ R_{m-n} = \begin{bmatrix} cos((m-n) \theta_i) & sin((m-n) \theta_i)\ -sin((m-n) \theta_i) & cos((m-n) \theta_i) \end{bmatrix} $$ □□□□ $$ R_n^TR_m = R_{m-n} $$ □□□□□□□□□□ $$ R_{-m} = R^T_m $$

□□□□□□□□ $$ (R_nq)^TR_mk=q^TR_{m-n}k $$ □□□ $$ q_n^{rot} = R_nq\ k_m^{rot} = R_mk $$ □□□□ □□□□□□□$Q□K$□□

KV Cache

- □□□□□□□□□
- □□□□□□□□
- □□□□□□□$Q□K□V$□□□□
- □□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□token□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□$K□V$□□□□□□□□

□□□

- □□
- □□□
- □□□□
- □□□□□

□□□□□KV Cache□□□□□□□□□□□□□“□□”□□□□□$K□V$

□□□KV Cache

image-20250702191639250

□□KV Cache

image-20250702191137599

Pre-Norm□Post-Norm

Pre-Norm□□□□□□□**Post-Norm**□□□□□□□Transformer□□□□□□□□□□□□□□□□□□□□□□LN□Layer Normalization□□□□□□□

- □□□
- □□□□□□□□□□□□□□□□□□□□□□□□□□□□□Transformer□□□□□□□□□□□□□□□□□□□□□□□□□□□□□**□□□□□ (warm-up)**□
- □□□□□□□□□□□□□□□□□□□□□□□□□