# LLM Notes

### A PREPRINT

## July 18, 2025

## Contents

1	Atte	ntion	2
	1.1	Self Attention	2
	1.2	Multi-Head Attention	3
	1.3	KV-Cache	4
2	Posi	tional Embedding	6

#### 1 Attention

#### 1.1 Self Attention

Given input  $X \in \mathbb{R}^{T \times D}$  where T is the sequence length and D is the embedding dimension, we have

$$Q = W_Q X \in \mathbb{R}^{T \times d_k} \quad K = W_K X \in \mathbb{R}^{T \times d_k} \quad V = W_V X \in \mathbb{R}^{T \times d_v}$$
 (1)

$$\operatorname{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \operatorname{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right) \boldsymbol{V} \tag{2}$$

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_{1} - \\ -\mathbf{x}_{2} - \\ \vdots \\ -\mathbf{x}_{T} - \end{bmatrix}$$

$$\mathbf{W}_{Q}$$

$$\mathbf{W}_{K}$$

$$\mathbf{W}_{V}$$

$$\mathbf{W}_{V}$$

$$\mathbf{W}_{V}$$

$$\mathbf{V} = \begin{bmatrix} -\mathbf{q}_{1} - \\ -\mathbf{q}_{2} - \\ \vdots \\ -\mathbf{q}_{T} - \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} -\mathbf{k}_{1} - \\ -\mathbf{k}_{2} - \\ \vdots \\ -\mathbf{k}_{T} - \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} -\mathbf{v}_{1} - \\ -\mathbf{v}_{2} - \\ \vdots \\ -\mathbf{v}_{T} - \end{bmatrix}$$

$$\begin{bmatrix} -\mathbf{q}_{1} - \\ -\mathbf{q}_{2} - \\ \vdots \\ -\mathbf{q}_{T} - \end{bmatrix} \begin{bmatrix} \mathbf{k}_{1}^{T} & \mathbf{k}_{2}^{T} & \cdots & \mathbf{k}_{T}^{T} \\ \mathbf{k}_{1}^{T} & \mathbf{k}_{2}^{T} & \cdots & \mathbf{k}_{T}^{T} \end{bmatrix} = \begin{bmatrix} \mathbf{q}_{1}\mathbf{k}_{1}^{T} & \cdots & \mathbf{q}_{1}\mathbf{k}_{T}^{T} \\ \vdots & \ddots & \vdots \\ \mathbf{q}_{T}\mathbf{k}_{1}^{T} & \cdots & \mathbf{q}_{T}\mathbf{k}_{T}^{T} \end{bmatrix}$$

$$\mathbf{w}_{cights} \text{ for token 1}$$

$$= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1T} \\ A_{21} & A_{22} & \cdots & A_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ A_{T1} & A_{T2} & \cdots & A_{TT} \end{bmatrix}$$

$$\mathbf{v}_{cights} \text{ for token 1}$$

$$= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1T} \\ A_{21} & A_{22} & \cdots & A_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ A_{T1} & A_{T2} & \cdots & A_{TT} \end{bmatrix}$$

$$\mathbf{v}_{cights} \text{ for token 1}$$

$$\mathbf{v}_{cights} \text{ for causal attention}$$

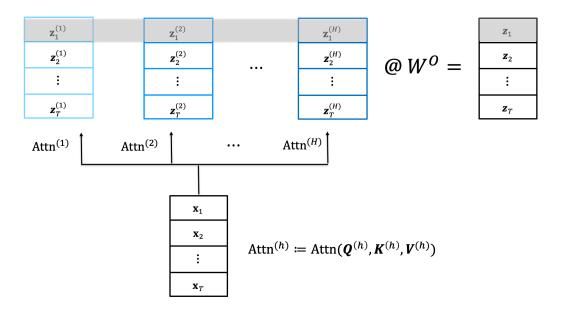
$$\mathbf{v}_{cights} \text{ for token 1}$$

$$\mathbf{v}_{cights} \text{ for causal attention}$$

#### 1.2 Multi-Head Attention

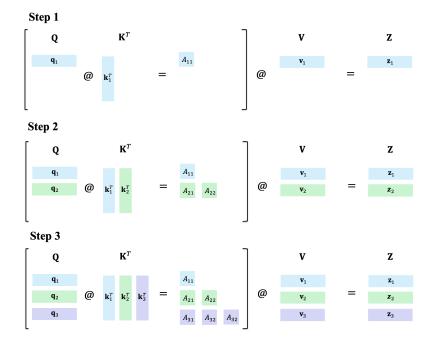
In multi-head attention, input X is first passed through H self-attention layer in parallel. Then, the output from each head is concatenated together and fused by a linear projection

$$\left[ \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(H)} \right] \mathbf{W}^{O} = \begin{bmatrix} \mathbf{z}_{1}^{(1)} & \mathbf{z}_{1}^{(2)} & \dots & \mathbf{z}_{1}^{(H)} \\ \mathbf{z}_{2}^{(1)} & \mathbf{z}_{2}^{(2)} & \dots & \mathbf{z}_{2}^{(H)} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{z}_{T}^{(1)} & \mathbf{z}_{T}^{(2)} & \dots & \mathbf{z}_{T}^{(H)} \end{bmatrix} \mathbf{W}^{O}$$
 (3)



#### 1.3 KV-Cache

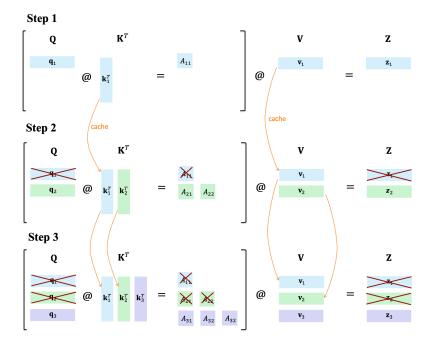
During inference we still use causal masking because this is how the model being trained. Let's look at a simple case where we only give the model a start token <s> and asks it to generate stuff:



KV-cache is built on this two observations:

- At each time step t, due to causal masking  ${m k}_{< t}$  and  ${m v}_{< t}$  will remain the same
- To predict  $\langle token_{t+1} \rangle$  we only need embedding of  $\langle token_t \rangle$ .

Therefore, we can make prediction efficiently by drop redundant and unnecessary computation



LLM Notes A PREPRINT

### Basically we have

Token 1:  $[K1, V1] \rightarrow Cache: [K1, V1]$ 

Token 2: [K2, V2]  $\rightarrow$  Cache: [K1, K2], [V1, V2]

...

Token n: [Kn, Vn]  $\rightarrow$  Cache: [K1, K2, ..., Kn], [V1, V2, ..., Vn]

## 2 Positional Embedding

## References