Multi-Head Attention

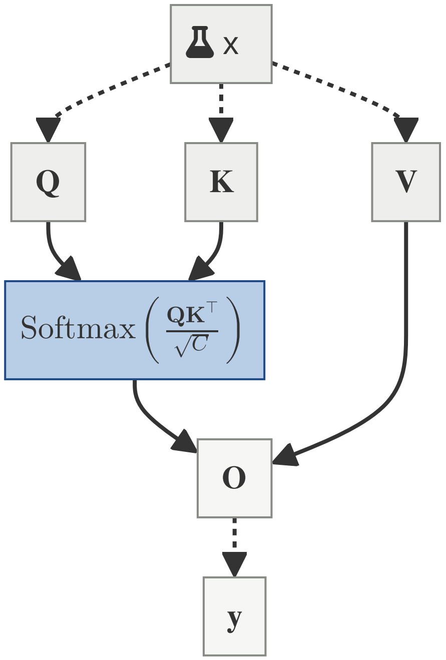
Recap: Attention

Attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}}\right)\mathbf{V}$$

Self-Attention $\mathbf{Q} = \mathbf{K} = \mathbf{V}$

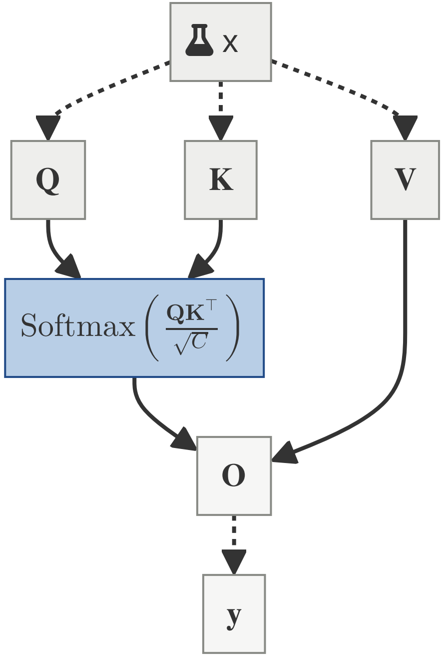
$$\text{Attention}(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{X}\mathbf{X}^\top}{\sqrt{C}}\right)\mathbf{X}$$



Issues With Self-Attention

$$\text{Attention}(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{X}\mathbf{X}^\top}{\sqrt{C}}\right)\mathbf{X}$$

$$= \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,N} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \alpha_{N,2} & \cdots & \alpha_{N,N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$



For any $(\mathbf{x}_i, \mathbf{x}_j)$ pair where $i \neq j$,

$$\begin{aligned} \mathbf{x}_i\mathbf{x}_j^\top &\leq \frac{1}{2}(\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{x}_j\mathbf{x}_j^\top) \\ &\leq \max(\mathbf{x}_i\mathbf{x}_i^\top, \mathbf{x}_j\mathbf{x}_j^\top) \\ \alpha_{i,j} &\leq \max(\alpha_{i,i}, \alpha_{j,j}) \quad \text{(softmax preserves order)} \end{aligned}$$

Issues With Self-Attention

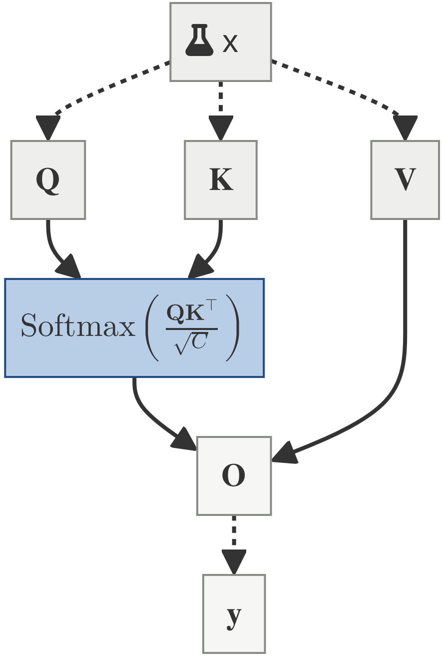
$$\text{Attention}(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{X}\mathbf{X}^\top}{\sqrt{C}}\right)\mathbf{X}$$

$$= \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,N} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \alpha_{N,2} & \cdots & \alpha_{N,N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$

$$\alpha_{i,j} \leq \max(\alpha_{i,i}, \alpha_{j,j})$$

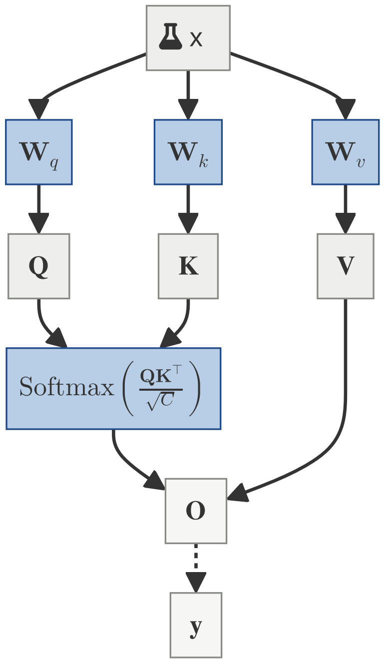
- diagonal **always be greater** than off-diagonals
- significantly limits expressive power

Solution: Apply weights to $\mathbf{Q}, \mathbf{K}, \mathbf{V}$



Attention With Weights

$$\begin{aligned}\text{Attention}(\mathbf{X};\mathbf{W}_Q,\mathbf{W}_K,\mathbf{W}_V) \\ &= \text{Attention}(\mathbf{X}\mathbf{W}_Q,\mathbf{X}\mathbf{W}_K,\mathbf{X}\mathbf{W}_V) \\ &= \text{Softmax}\left(\frac{\mathbf{X}\mathbf{W}_Q(\mathbf{X}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{X}\mathbf{W}_V\end{aligned}$$



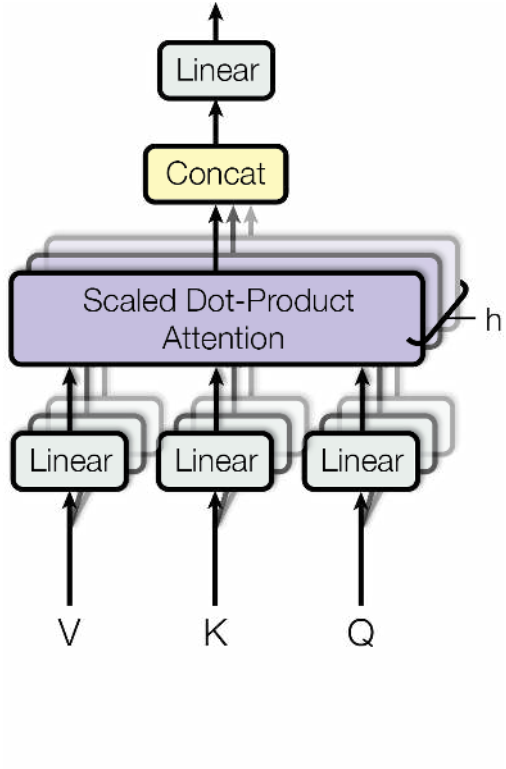
More expressive than linear projection ($\mathbf{X}\mathbf{W}_V$ or 1×1 2D-conv)

❌ All elements use the same attention

Solution: Multi-Head Attention

Simple concatenation of multiple attention layers ("heads")

- $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ vary per head



Multi-Head Attention

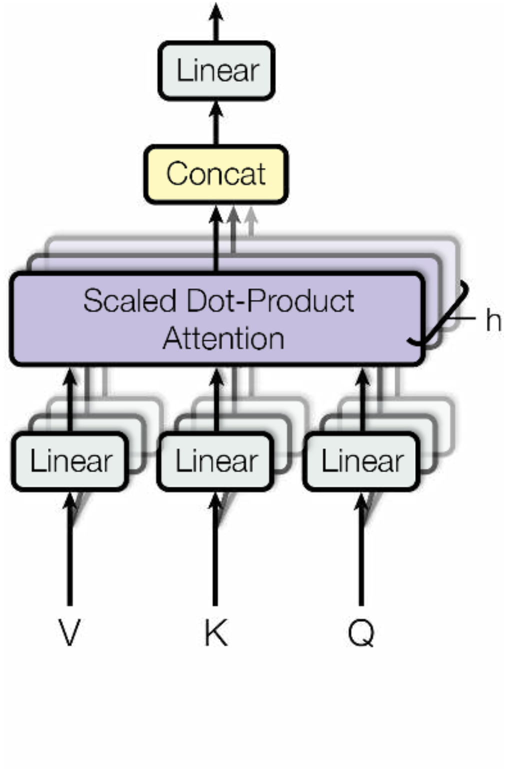
h heads, each with a set of linear projections

- $\mathbf{W}_{K,h} \in \mathbb{R}^{C \times d_k}$
- $\mathbf{W}_{Q,h} \in \mathbb{R}^{C \times d_k}$
- $\mathbf{W}_{V,h} \in \mathbb{R}^{C \times d_v}$

A final linear projection to map to output dimension

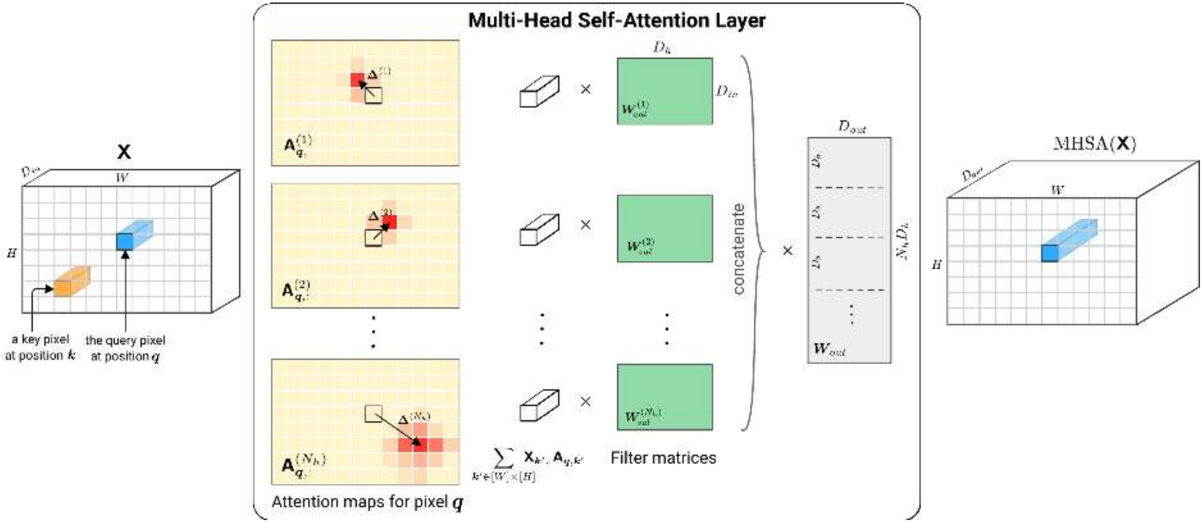
- $\mathbf{W}_O \in \mathbb{R}^{hd_v \times C}$

$$\begin{bmatrix} \text{Attention}(\mathbf{XW}_{Q,1}, \mathbf{XW}_{K,1}, \mathbf{XW}_{V,1}) \\ \vdots \\ \text{Attention}(\mathbf{XW}_{Q,h}, \mathbf{XW}_{K,h}, \mathbf{XW}_{V,h}) \end{bmatrix} W_O$$



Connection to Convolution

Multi-head attention with h heads is **more expressive** than a $\sqrt{h} \times \sqrt{h}$ 2D conv



Cordonnier *et al.* On the relationship between self-attention and convolutional layers. ICLR 2020. ↩️

Multi-Head Attention - TL;DR

Always use attention with weights

Self-attention with weights generalizes a 1×1 2D conv

Multi-head attention with h heads generalizes a $\sqrt{h} \times \sqrt{h}$ 2D conv

Always use **multi-head attention**