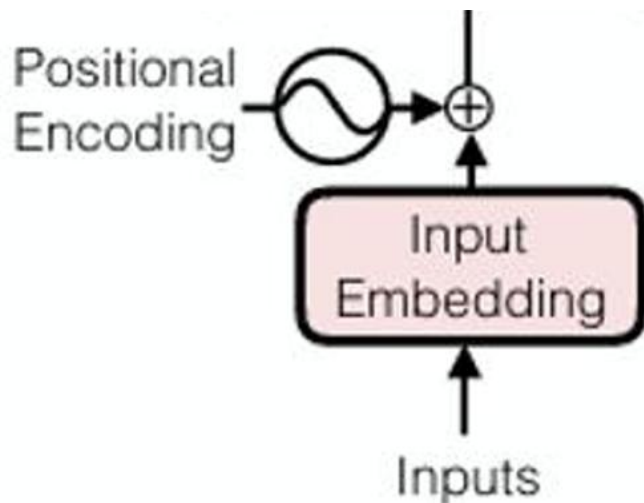


Input



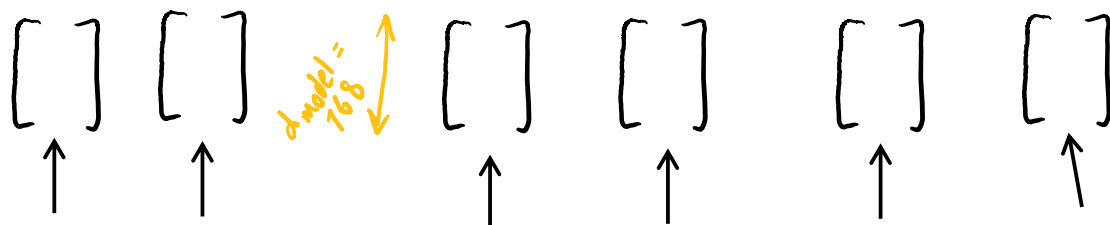
pos embedding

+

$[]_{n_{vocab} \times d_{model}}$

How? We

embeddings



token IDs
(one-hot vectors)

1080

45

$n_{vocab} = 50,000$

74

129

6704

13000

tokens

Artificial Intelligence is transforming the

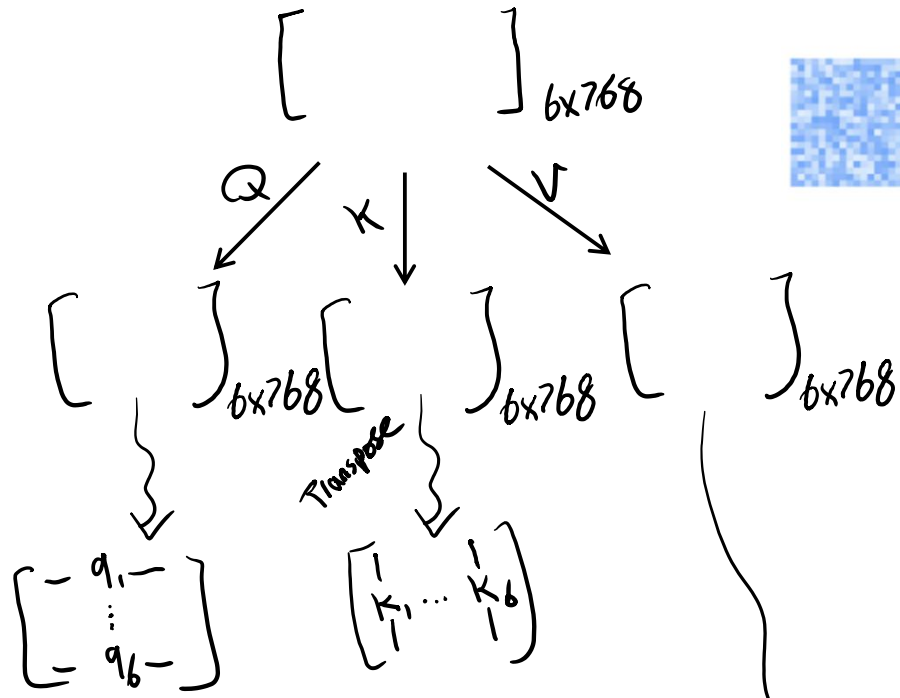
Artificial Intelligence is transforming the

$[]_{6 \times 768}$

$[]_{6 \times 50,000}$

Self Attention

QKV matrix



How?

dot product

$$\begin{bmatrix} q_1 k_1 & \dots & q_6 k_6 \end{bmatrix}$$

softmax

$$\Sigma = 1 \leftarrow \begin{bmatrix} 0.4 & 0.2 & 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

$$\begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_6 & - \end{bmatrix}$$

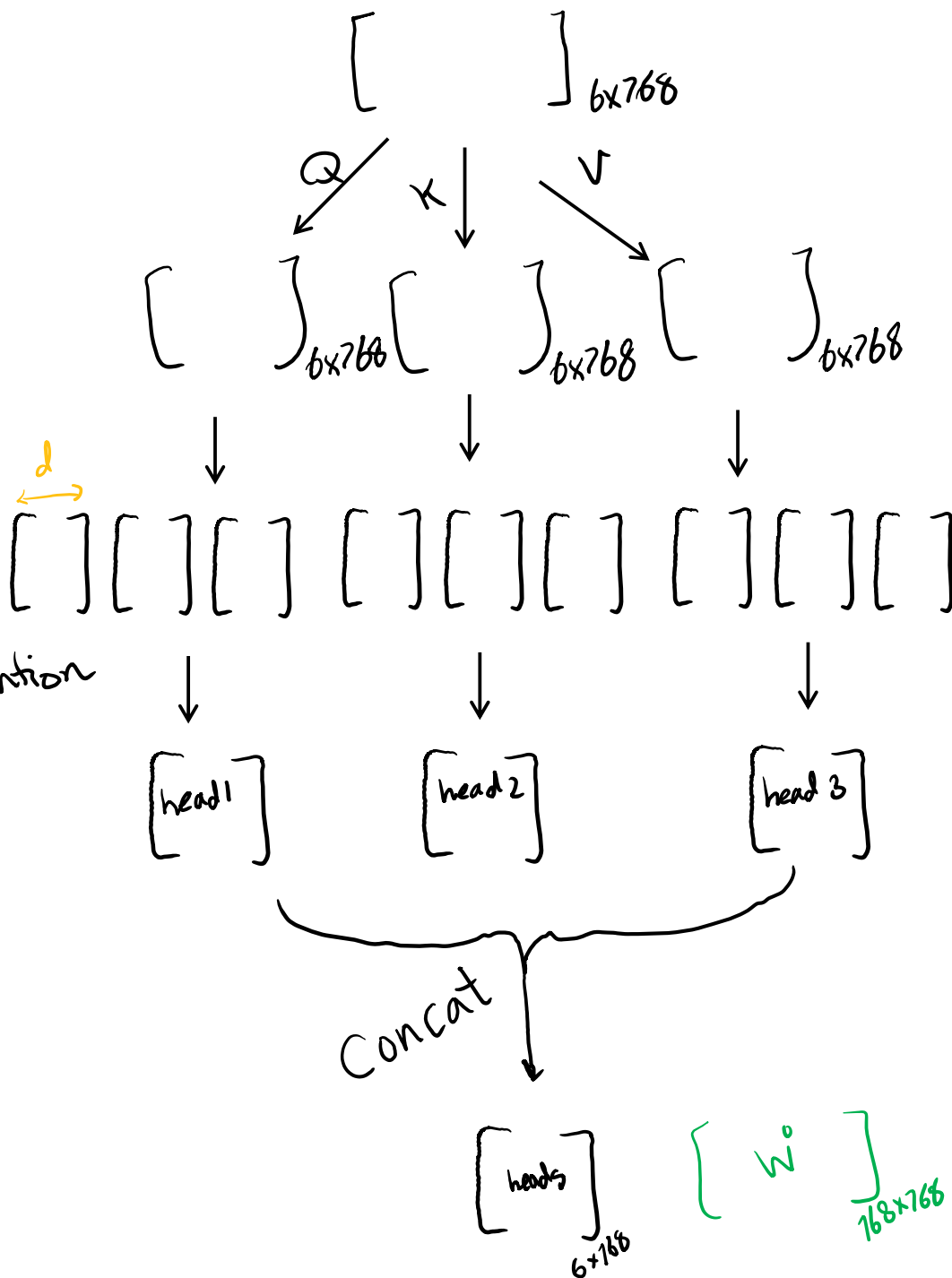
Multi-Head
Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O$$

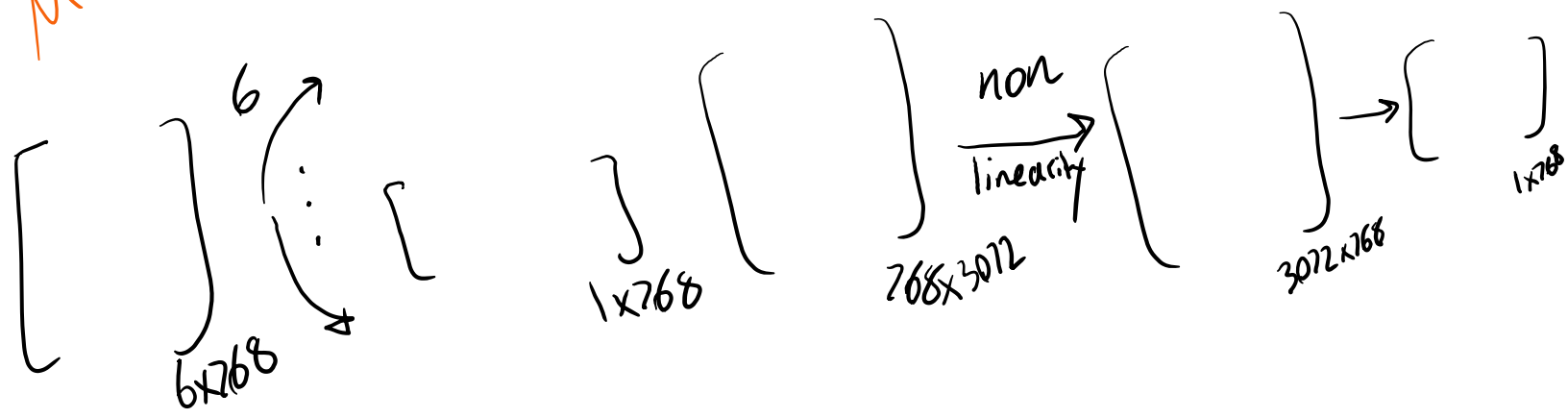
where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

$\frac{d_{\text{model}}}{h} = d$

$h=3$



MLP



Final linear layer

