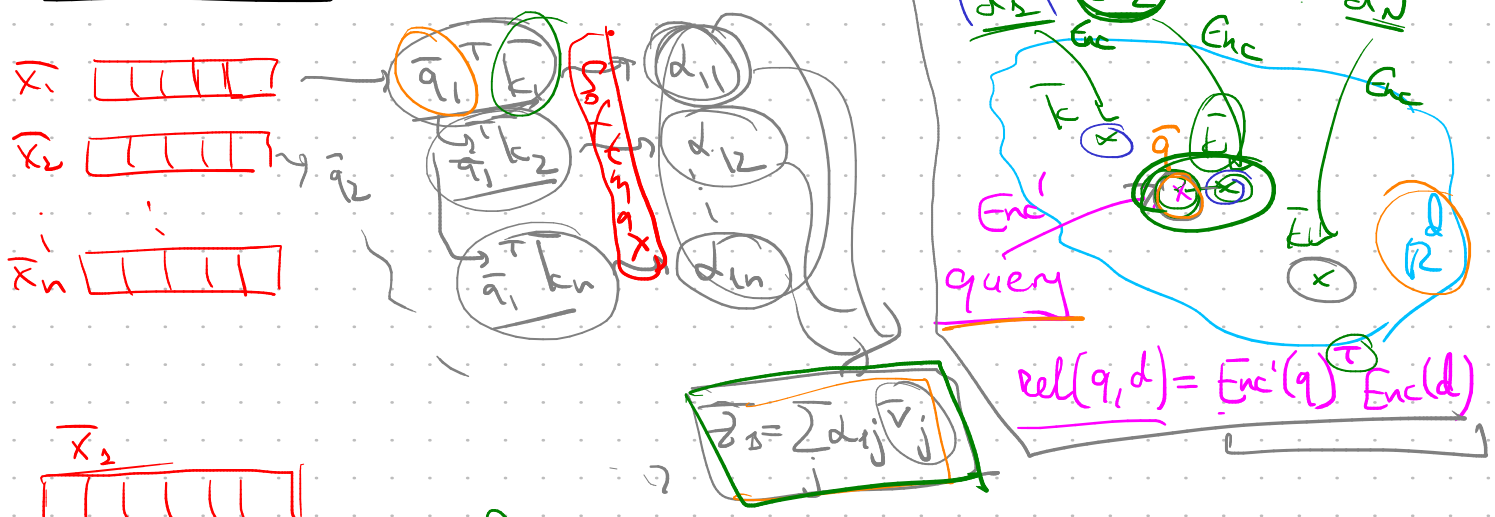
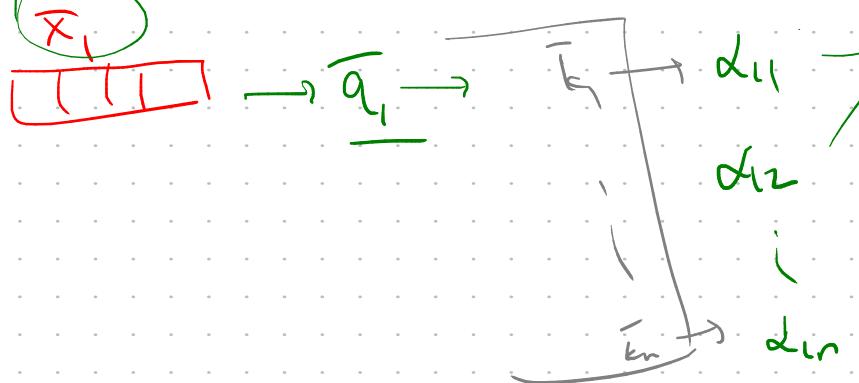


# TRANSFORMER



Input  $\bar{x}_1$  is processed to generate:

- $\bar{q}_1 = W^Q \bar{x}_1$
- $\bar{k}_1 = W^K \bar{x}_1$
- $\bar{v}_1 = W^V \bar{x}_1$



$$z_1 = \sum_j \alpha_{1j} \bar{v}_j$$

$$q_i, k_i \in \mathbb{R}^{d_k}$$

Flow diagram for the first input  $\bar{x}_1$  showing the attention mechanism:

$$\bar{x}_1 \rightarrow \bar{q}_1 \rightarrow \left[ \frac{1}{\sqrt{d_k}} \bar{q}_1^T \bar{k}_1, \frac{1}{\sqrt{d_k}} \bar{q}_1^T \bar{k}_2, \dots, \frac{1}{\sqrt{d_k}} \bar{q}_1^T \bar{k}_n \right] \xrightarrow{\text{Softmax}} \alpha_{1j} = \frac{e^{\frac{1}{\sqrt{d_k}} \bar{q}_1^T \bar{k}_j}}{\sum_s e^{\frac{1}{\sqrt{d_k}} \bar{q}_1^T \bar{k}_s}}$$

$$z_1 = \sum_j \alpha_{1j} \bar{v}_j = \sum_{j=1}^n \text{softmax}\left(\frac{1}{\sqrt{d_k}} \bar{q}_1^T \bar{k}_j\right) \bar{v}_j$$

$$Z = \text{softmax}\left(\frac{1}{\sqrt{d_k}} Q K^T\right) \cdot V, \text{ se } \begin{cases} Q = W^Q X \\ K = W^K X \\ V = W^V X \end{cases}$$

$n \times m$

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Multi-head attention

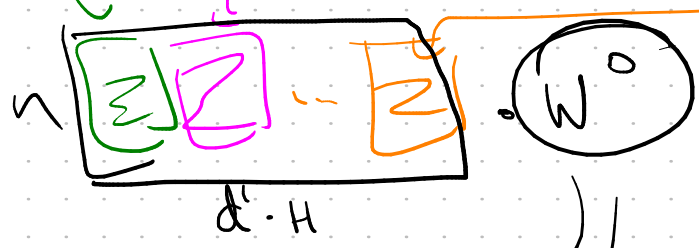
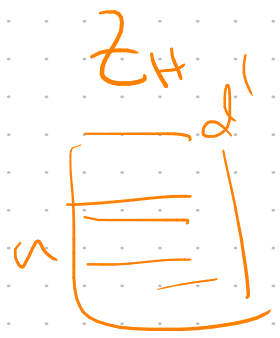
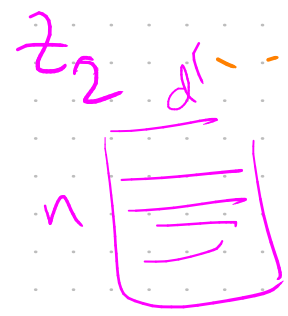
$W_1^Q, W_1^K, W_1^V$

$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$

$Z_1 = AV_1 \quad \sqrt{d_k} \in \mathbb{R}^{d'}$

$W_2^Q, W_2^K, W_2^V$

$W_H^Q, W_H^K, W_H^V$



Enc

$Z_m$

Decoder



$W^K \rightarrow K$

$W^V \rightarrow V$

