

$$a_{ij} = \text{softmax}_j \left(\frac{1}{\sqrt{d}} \bar{q}_i^T \bar{k}_j \right)$$

$$\bar{z}_i = \sum_{j=1}^K a_{ij} \bar{v}_j$$

Diagram showing the derivation of query, key, and value vectors from input vectors $\bar{x}_i \in \mathbb{R}^D$:

$\bar{q}_i \in \mathbb{R}^{D'} = W^Q \bar{x}_i$

$\bar{k}_i \in \mathbb{R}^{D'} = W^K \bar{x}_i$

$\bar{v}_i \in \mathbb{R}^{D'} = W^V \bar{x}_i$

Diagram showing the matrix operations for query, key, and value matrices:

$Q = W^Q X$ (Dimensions: $D \times d$)

$K = W^K X$ (Dimensions: $D \times d$)

$V = W^V X$ (Dimensions: $D \times d$)

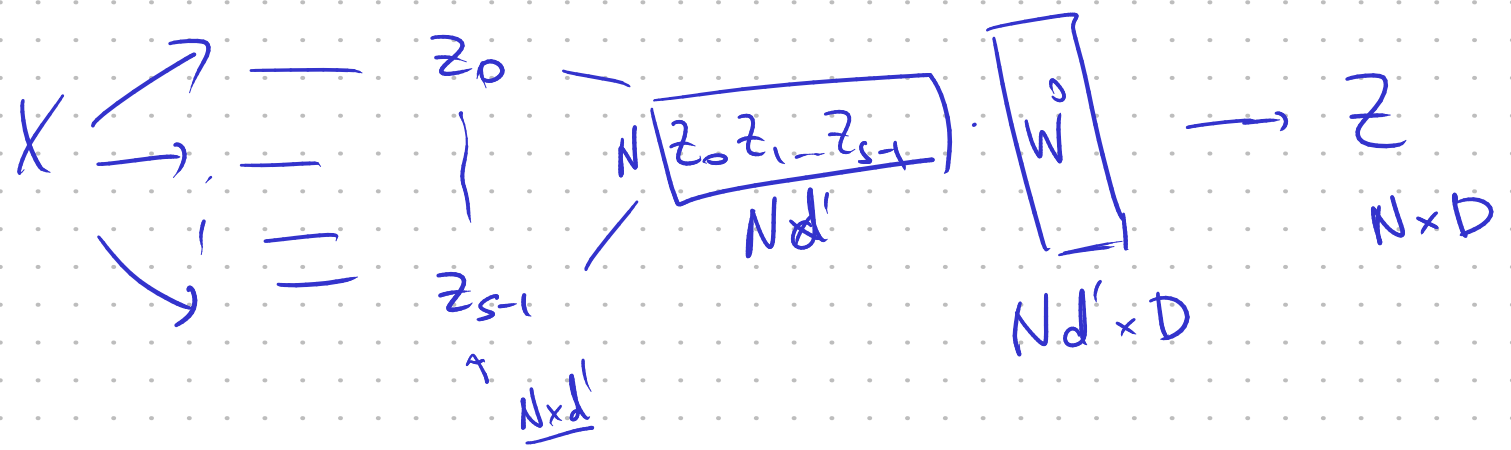
$$\bar{z}_i = \text{softmax} \left(\frac{1}{\sqrt{d}} Q K^T \right) \cdot V$$

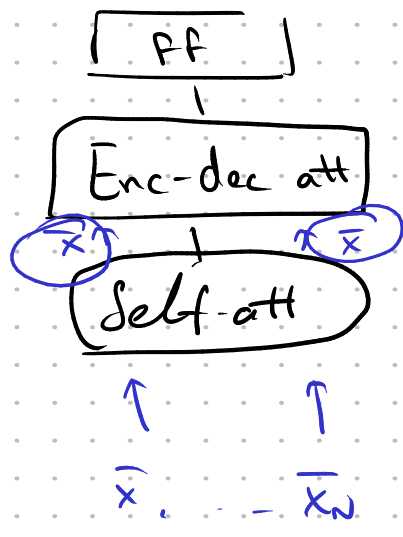
Diagram showing the number of weights for each matrix:

W^Q, W^K, W^V (Dimensions: $D \times d, D \times d, D \times d$)

W^0 (Dimensions: $N \times N$)

Number of queries: N





queries, keys

values

\bar{z}_1 - \bar{z}_n
Encoder