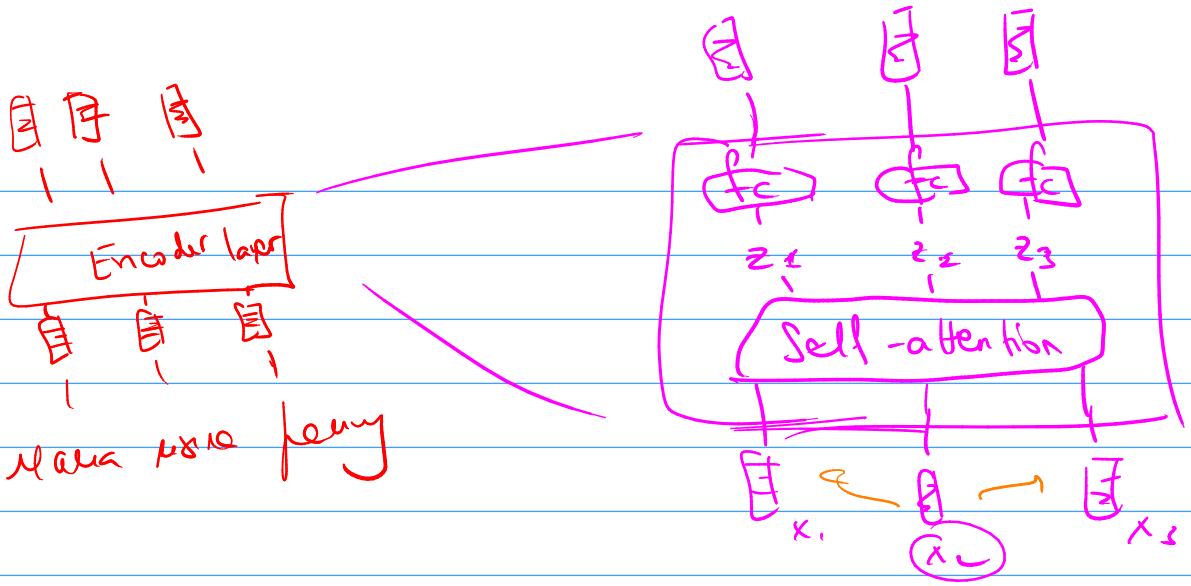
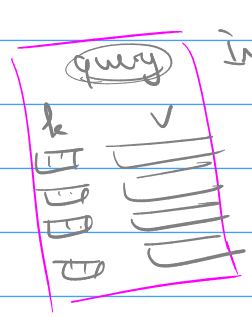
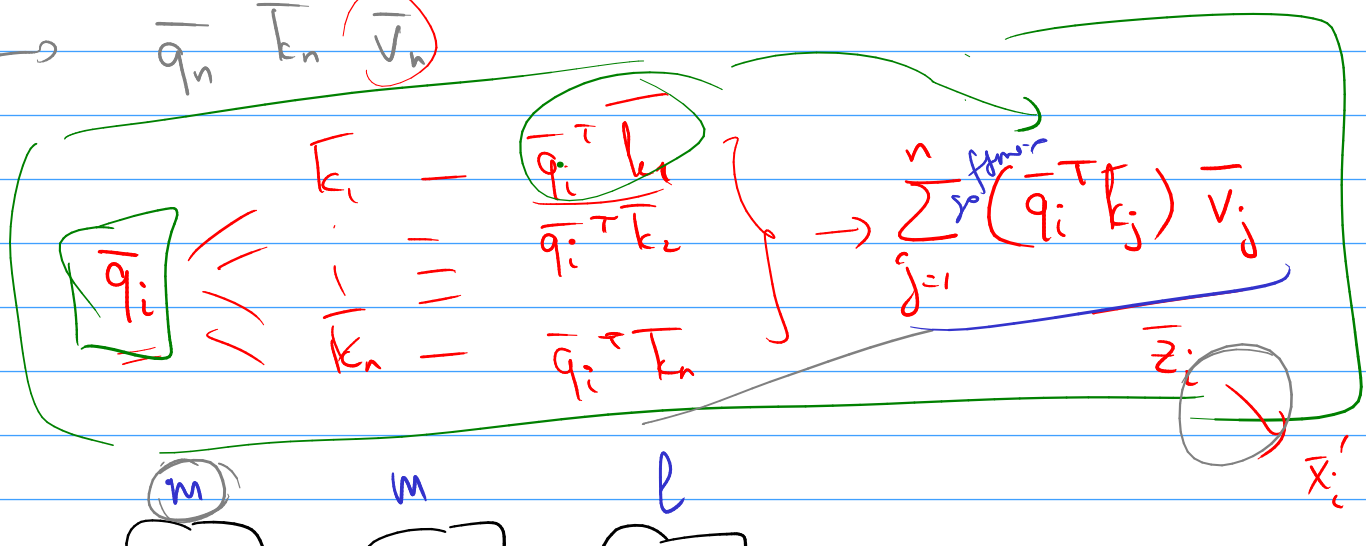
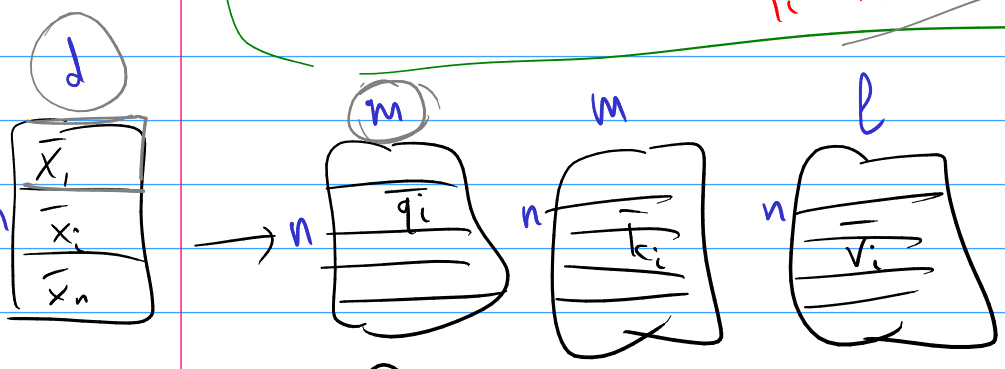








Attention head

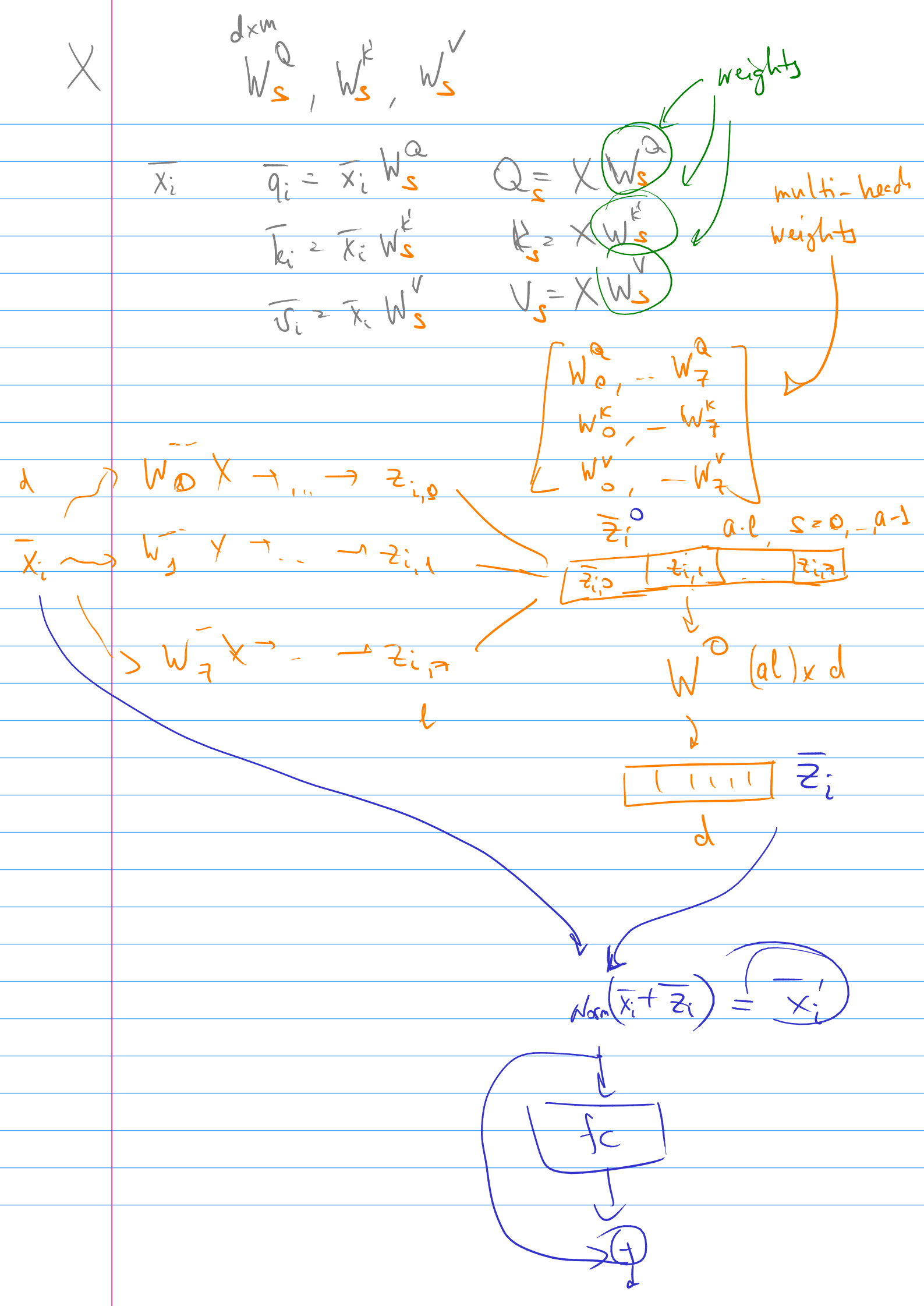
$$\text{softmax}\left(\frac{1}{\sqrt{m}}\left(QK^T\right)\right)V = \sum_{j=1}^{n} \text{softmax}\left(\bar{q}_i^T \bar{k}_j\right)\bar{v}_j \quad \begin{matrix} n \\ \\ l \end{matrix}$$

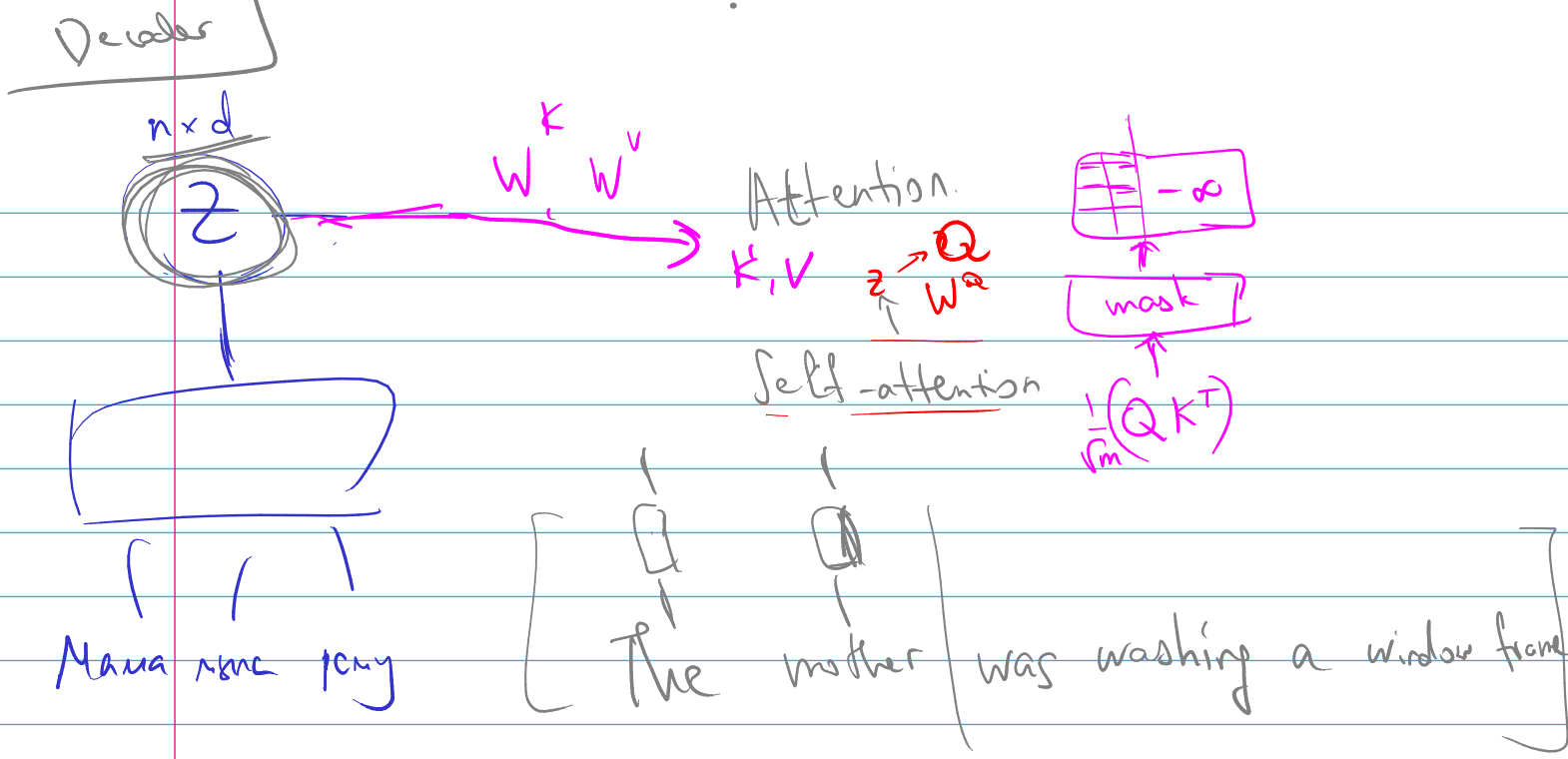
$n \times n$

$n \times l$

$s = 0..7$

Multi-head attention


X
$d \times m$
W_s^Q, W_s^K, W_s^V
weights
$\bar{x}_i$
$\bar{q}_i = \bar{x}_i W_s^Q$
$\bar{k}_i = \bar{x}_i W_s^K$
$\bar{v}_i = \bar{x}_i W_s^V$
$Q_s = X W_s^Q$
$K_s = X W_s^K$
$V_s = X W_s^V$
multi-head weights
$W_0^Q, \ldots W_7^Q$
$W_0^K, \ldots W_7^K$
$W_0^V, \ldots W_7^V$
$\bar{z}_i^0$
a·l, s = 0, ..., a-1
W^0 (al) x d
d
$\bar{z}_i$
$\text{Norm}(\bar{x}_i + \bar{z}_i) = \bar{x}_i'$
fc
+


Decoder
n × d
Z
W^K W^V
Attention
K, V
Q
Z
W^Q
Self-attention
$-\infty$
mask
$\frac{1}{\sqrt{m}}(QK^T)$
Мама мыла раму
[The mother | was washing a window frame]