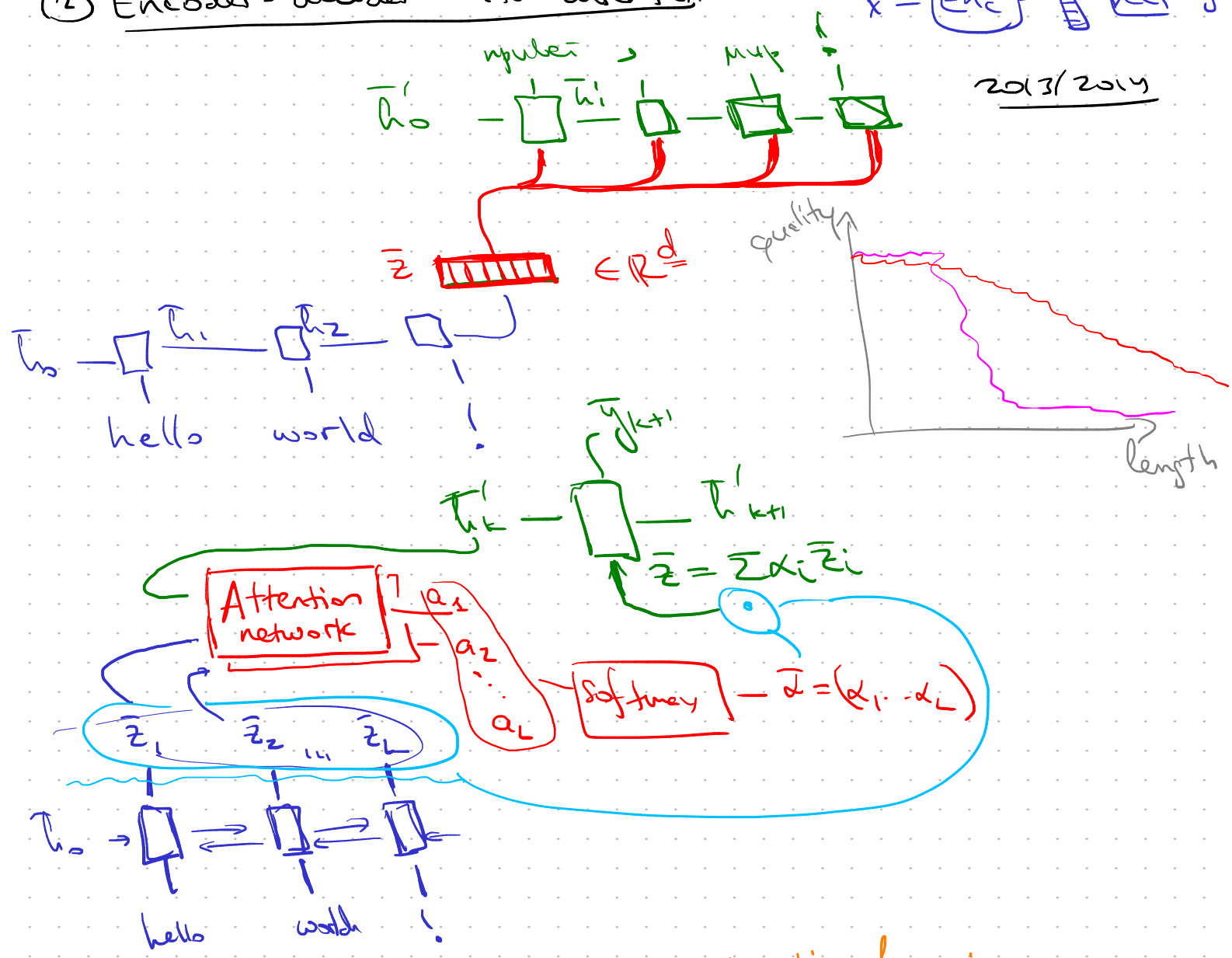
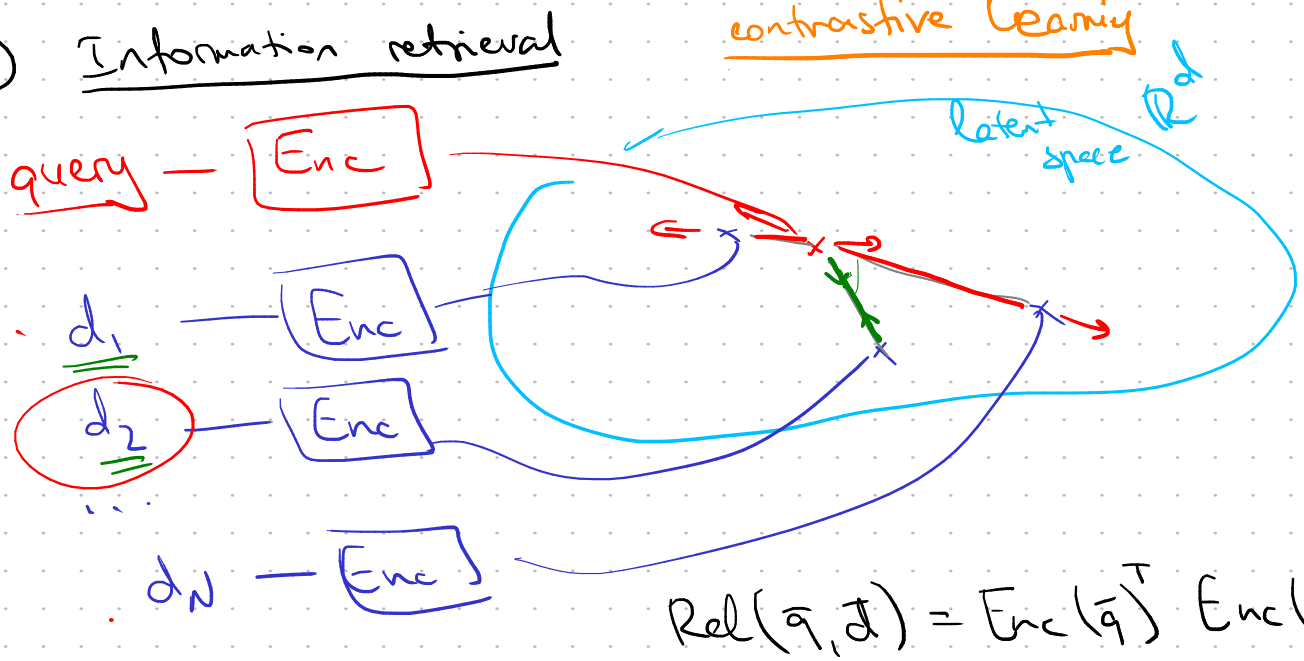
① Recurrent models of visual attention

② Encoder - decoder with attention

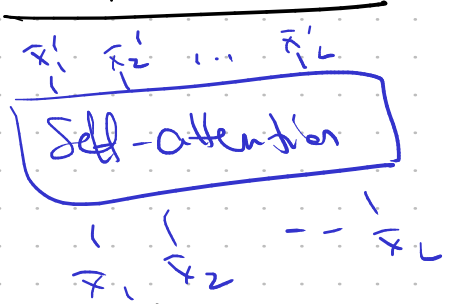


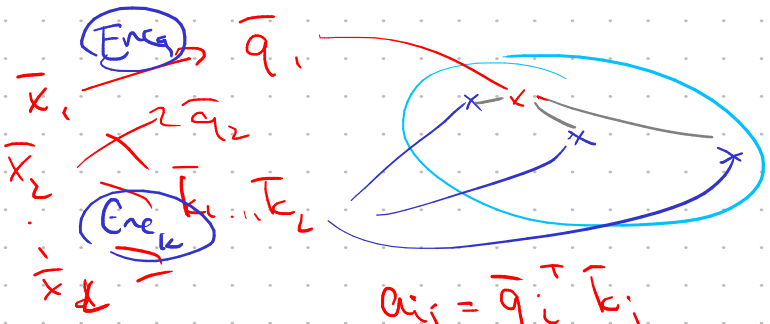

③ Information retrieval




$$Rel(q, d) = Enc(\bar{q})^T \, Enc(d)$$

# 4 Self-attention

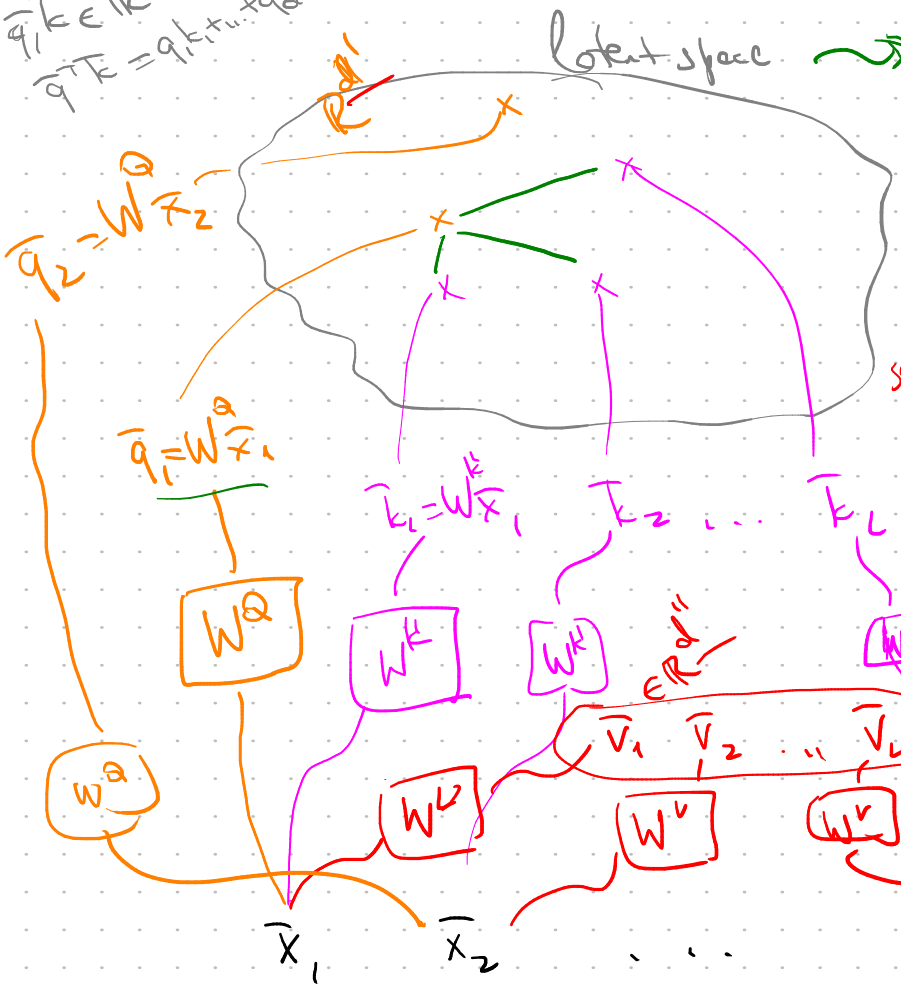



$a_{ij} = \bar{q}_i^T \bar{k}_j$

$\bar{q}_i, \bar{k} \in \mathbb{R}^{d'}$

$\bar{q}^T \bar{k} = q_1 k_1 + \dots + q_{d'} k_{d'}$




$a_{11} = \bar{q}_1^T \bar{k}_1, \; a_{12} = \bar{q}_1^T \bar{k}_2, \dots a_{1L} = \bar{q}_1^T \bar{k}_L$

$(a_{11}, a_{12}, \dots, a_{1L})$

$\downarrow$ $1/\sqrt{d'}$

$\downarrow$ softmax

self-attention weights

$$\alpha_{ij} = \frac{e^{\frac{1}{\sqrt{d'}} a_{ij}}}{\sum_s e^{\frac{1}{\sqrt{d'}} a_{is}}}$$

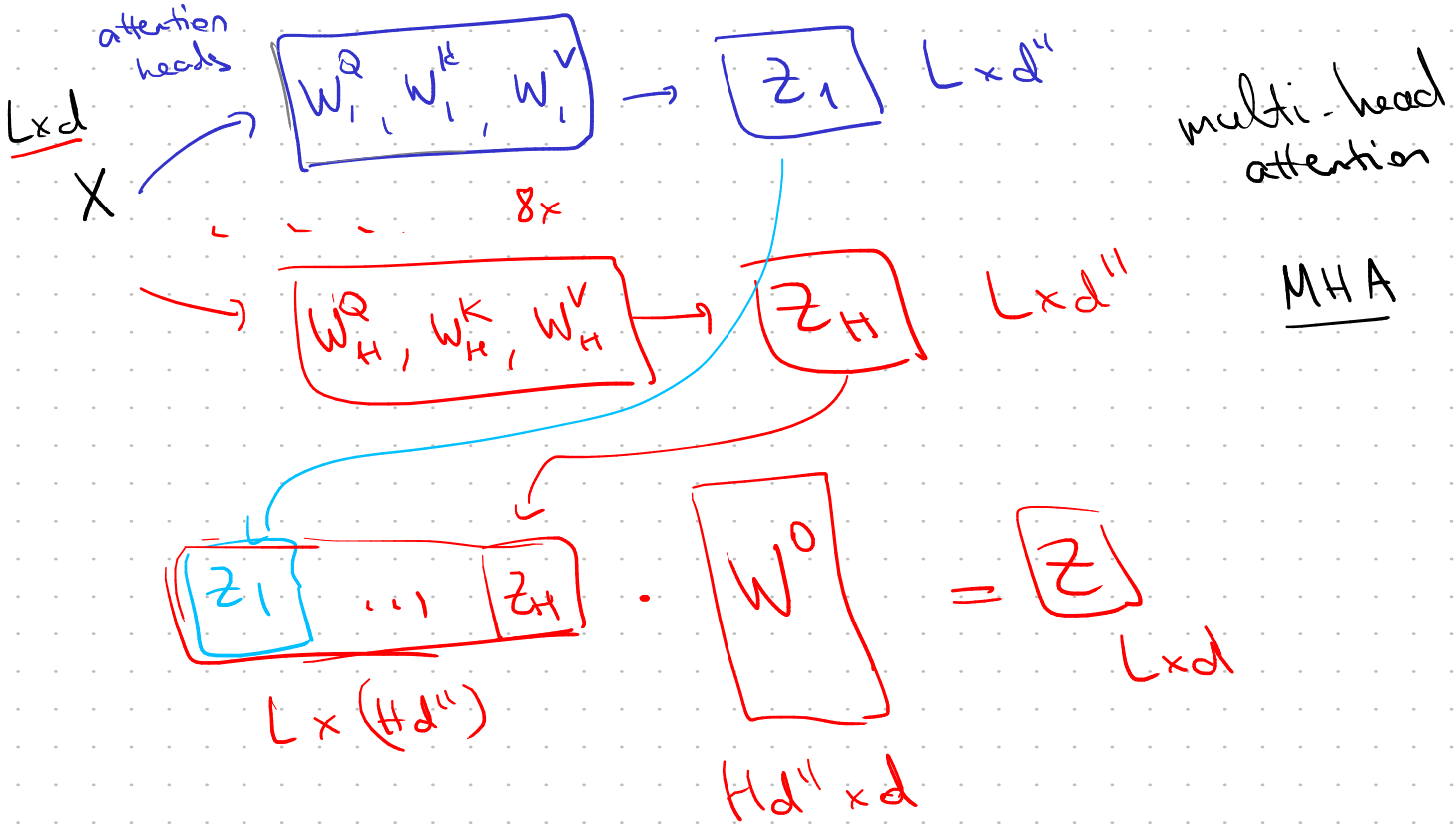
$$\bar{z}_1 = \sum_{j=1}^{L} \alpha_{1j} \bar{v}_j \in \mathbb{R}^{d''}$$

$$\bar{z}_2 = \sum_j \alpha_{2j} \bar{v}_j$$

$W^Q X = Q, \; W^K X = K, \; W^V X = V$

$$Z = \text{softmax}\left(\frac{1}{\sqrt{d'}} Q K^T\right) V$$

bottleneck | split-transform-merge


attention heads
L x d
X
$W_1^Q, W_1^K, W_1^V$
$Z_1$
L x d''
8x
multi-head attention
MHA
$W_H^Q, W_H^K, W_H^V$
$Z_H$
L x d''
$Z_1$ ... $Z_H$
L x (H d'')
$W^O$
H d'' x d
$Z$
L x d