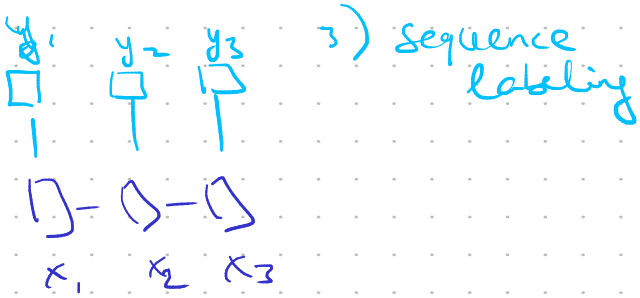
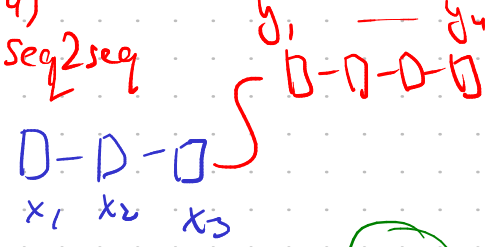
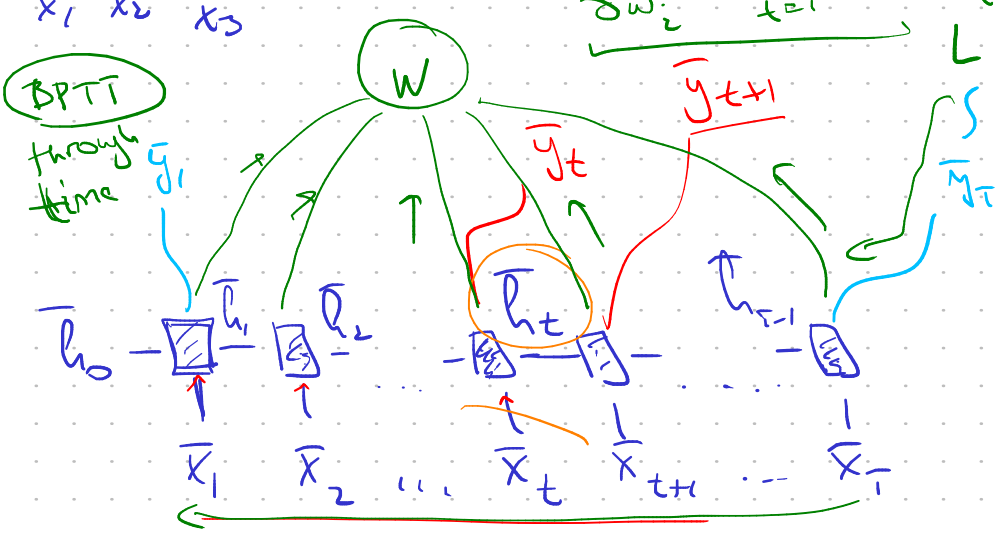
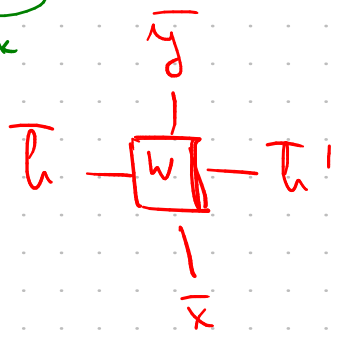
① Recurrent neural networks / sequences - text







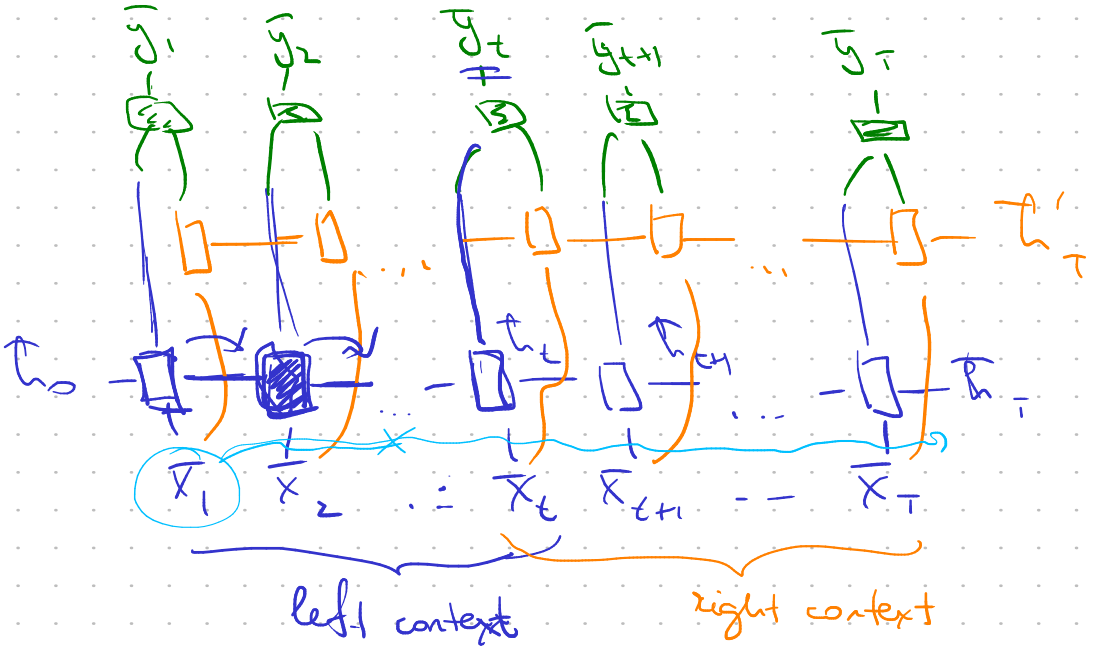




$$p(\bar{x}_1 \dots \bar{x}_n) = p(\bar{x}_1)\, p(\bar{x}_2|\bar{x}_1) \dots p(\bar{x}_n|\bar{x}_1 \dots \bar{x}_{n-1})$$

autoregressive $\quad \bar{y}_t = F(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_t; W)$

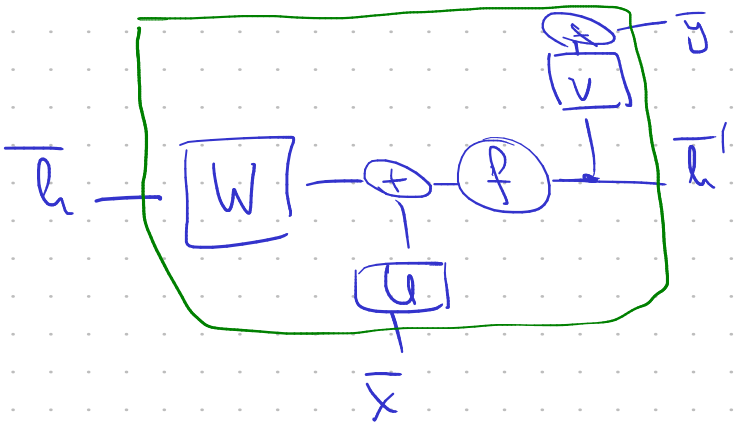
$$p(\bar{h}, \bar{y}|\bar{x}_1 \dots \bar{x}_T) = p(h_0)\, p(h_1, \bar{y}_1|h_0, \bar{x}_1)$$
$$p(h_2, \bar{y}_2|h_1, \bar{x}_2) \dots p(h_T, \bar{y}_T|h_{T-1}, \bar{x}_T)$$



bidirectional RNN

Elman

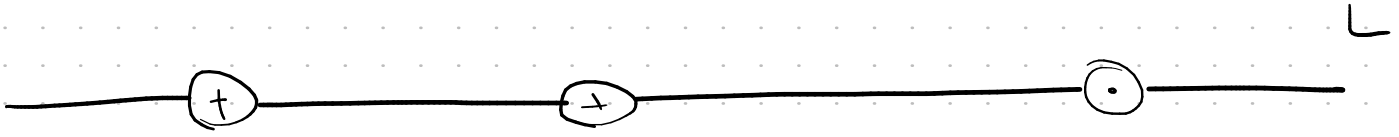
Jordan



$$\bar{h}' = f(W\bar{h} + U\bar{x})$$
$$\bar{y} = f(V\bar{h}')$$

$$h_1 = Wh_0 \,, \quad h_2 = W^2 h_0 \,, \quad \dots \quad h_t = W^t \cdot h_0$$

- exploding gradients: $|W^t| \to \infty$
- vanishing gradients: $|W^t| \to 0$

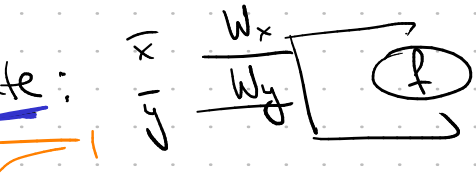


(2) LSTM – long short-term memory

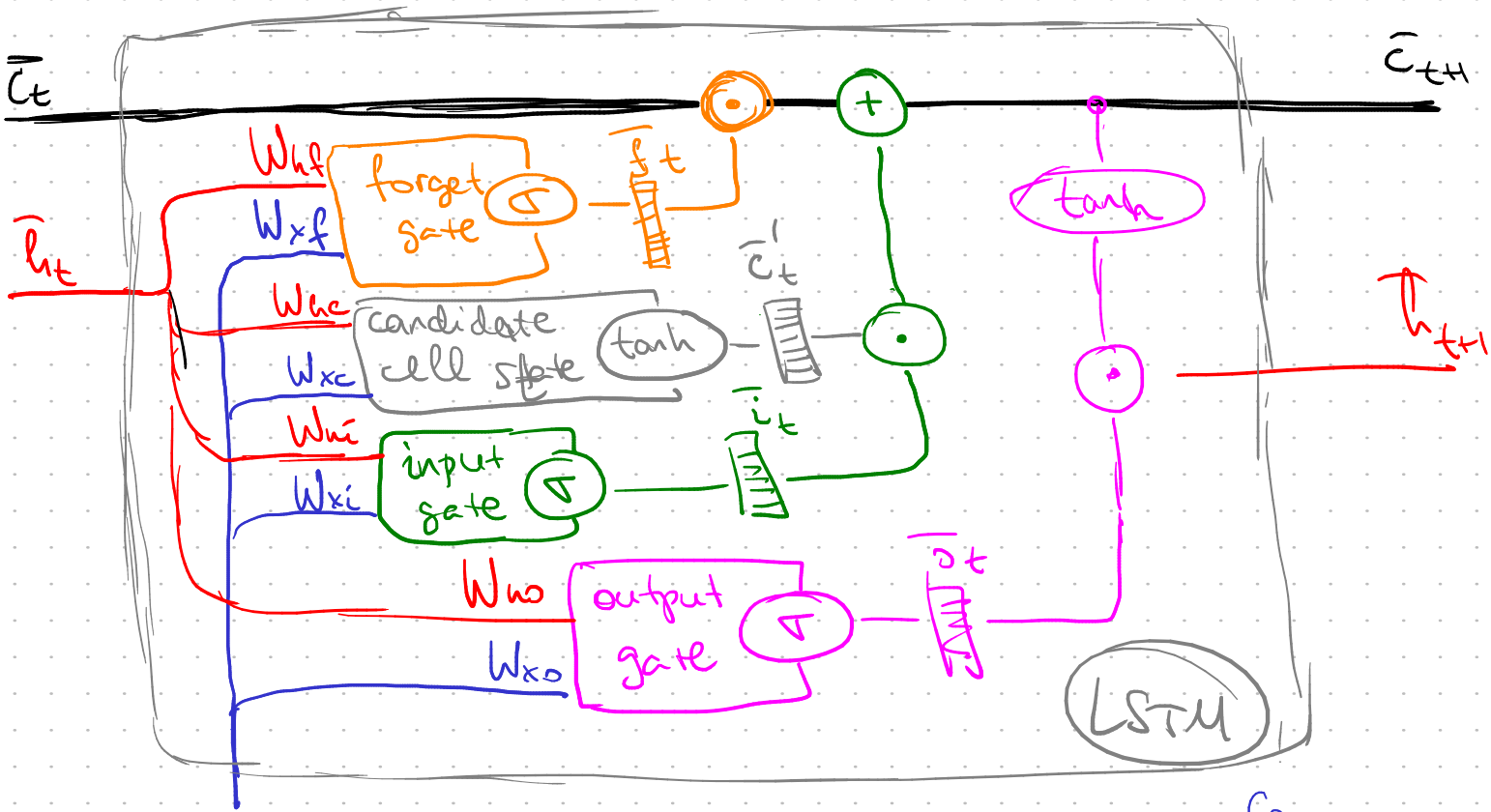
cell $(\bar{c}_t, \bar{h}_t)$

gate:





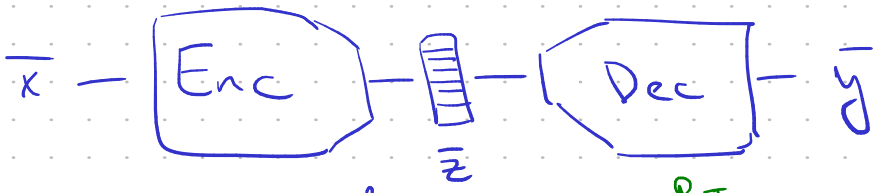
$$= f(W_x \bar{x} + W_y \bar{y} + \bar{w}_0)$$



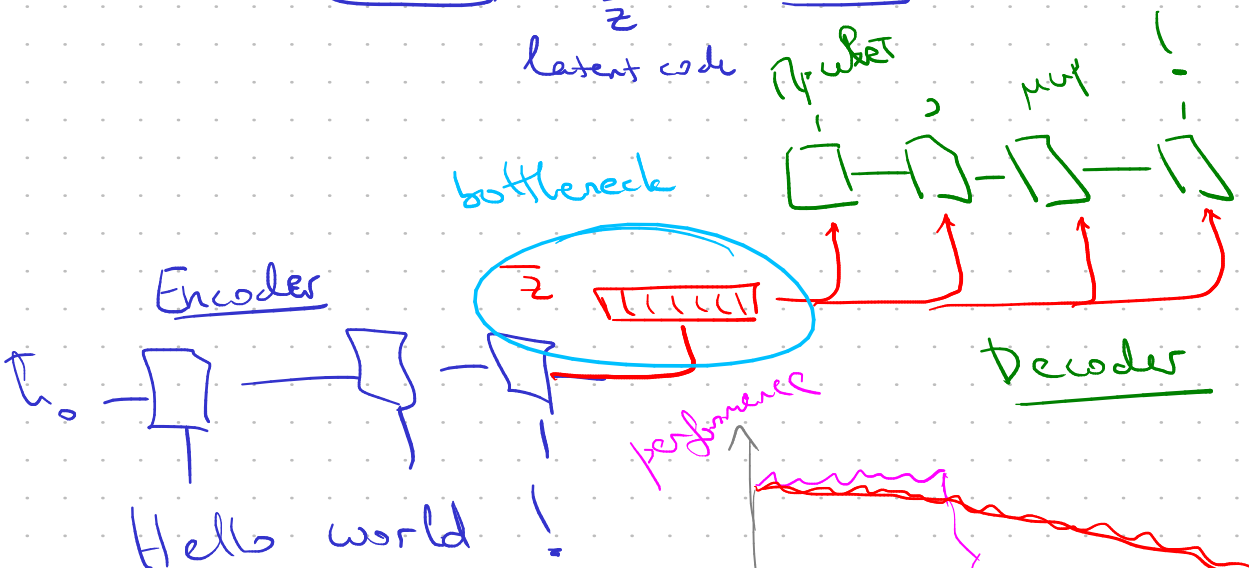
Hochreiter, Schmidhuber (1995), Gers et al. (2000)

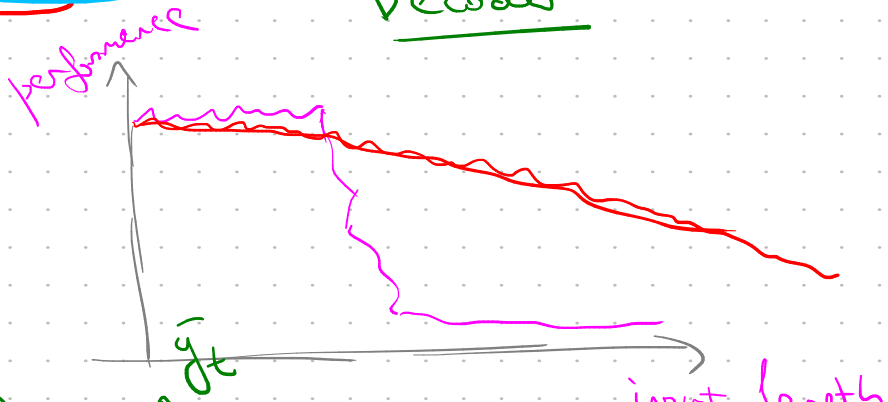
(3) Encoder-decoder with attention

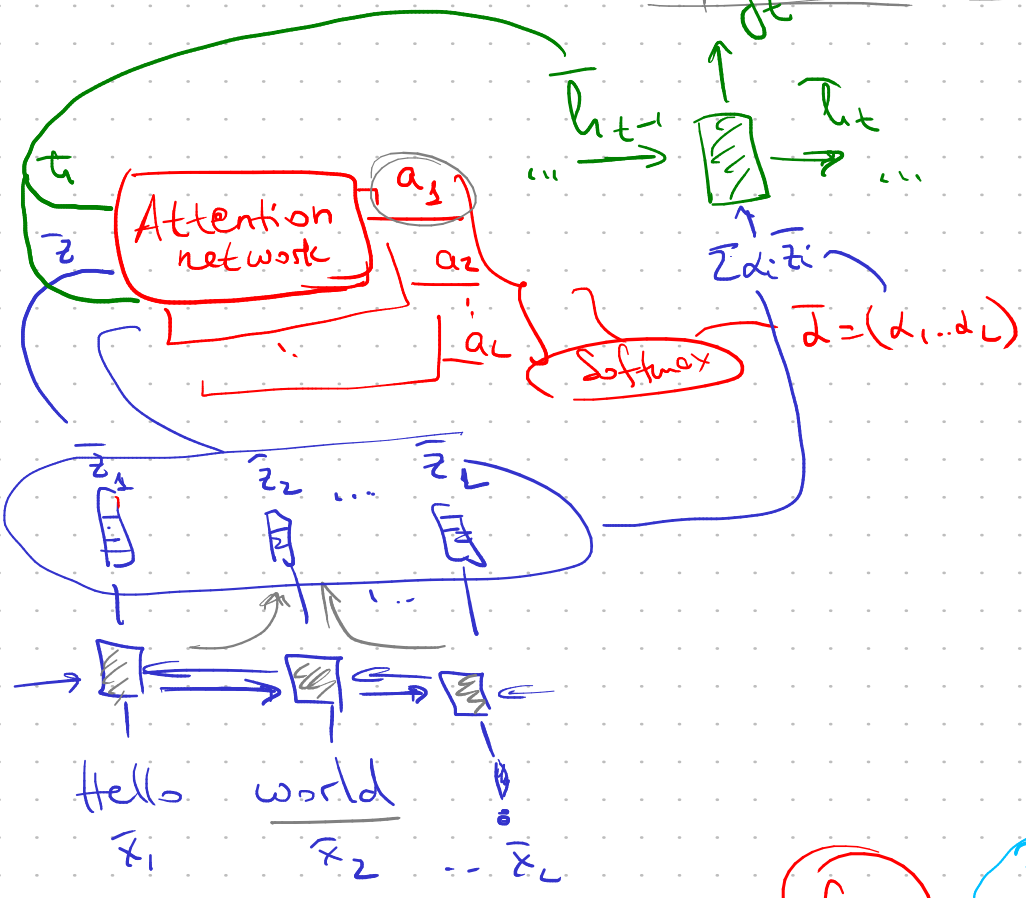
$\bar{y} = \hat{\bar{x}}$ – autoencoder



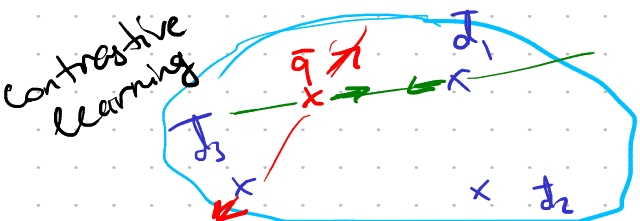
latent code

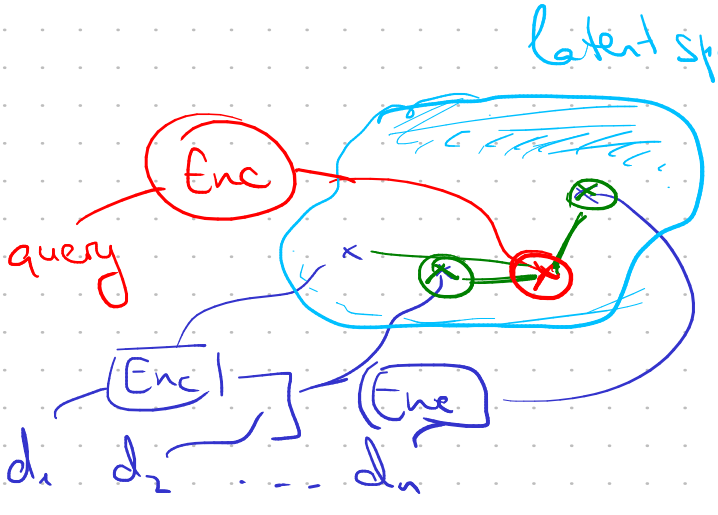






(4) Information retrieval



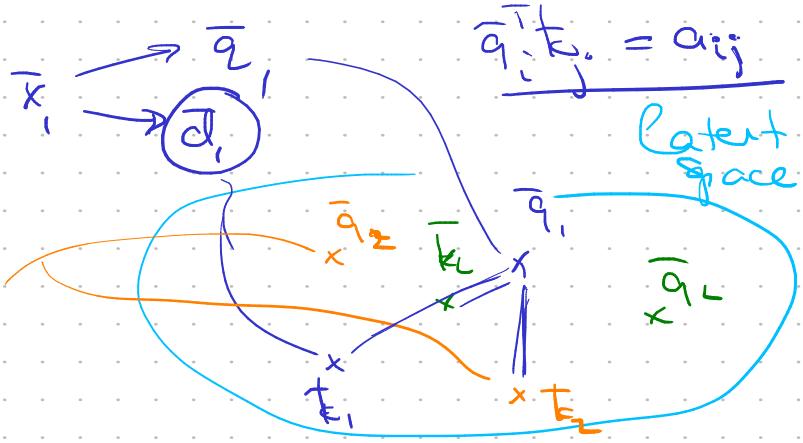


(5) Self-attention

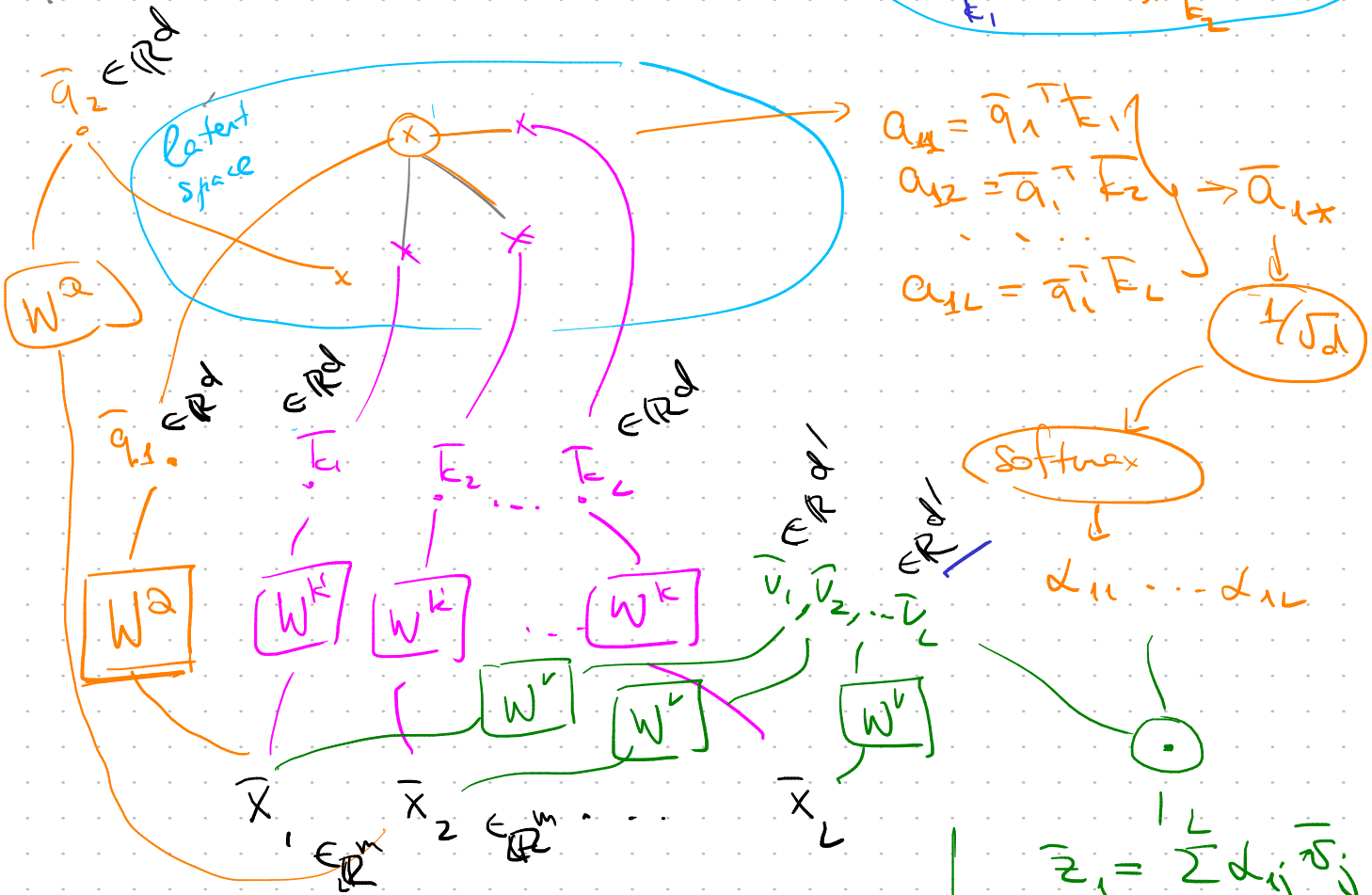
$\bar{x}_1, \bar{x}_2, \dots, \bar{x}_L$

$\bar{q}_i, \bar{k} \in \mathbb{R}^d$ $\quad \sigma \neq 1$

$|\sum a_i k_i| \approx O(\sqrt{d})$



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



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

$$Z = \underbrace{\text{Softmax}}_{O(L^2)}\left(\frac{1}{\sqrt{d}} \underbrace{QK^T}_{L \times L}\right) V$$

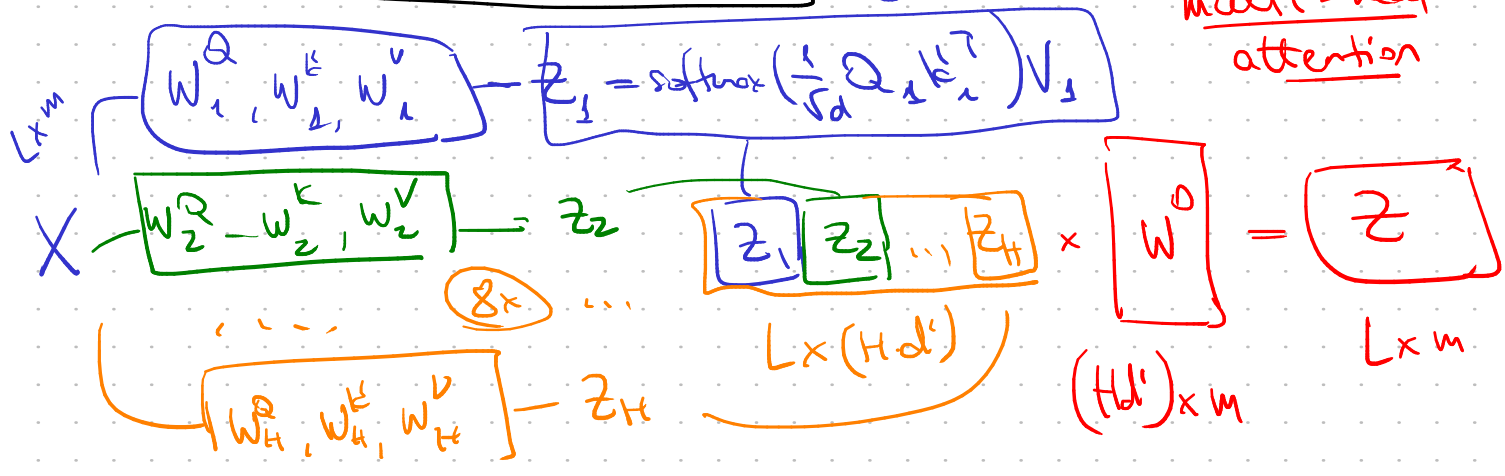
$Z \in \mathbb{R}^{L \times d'}$

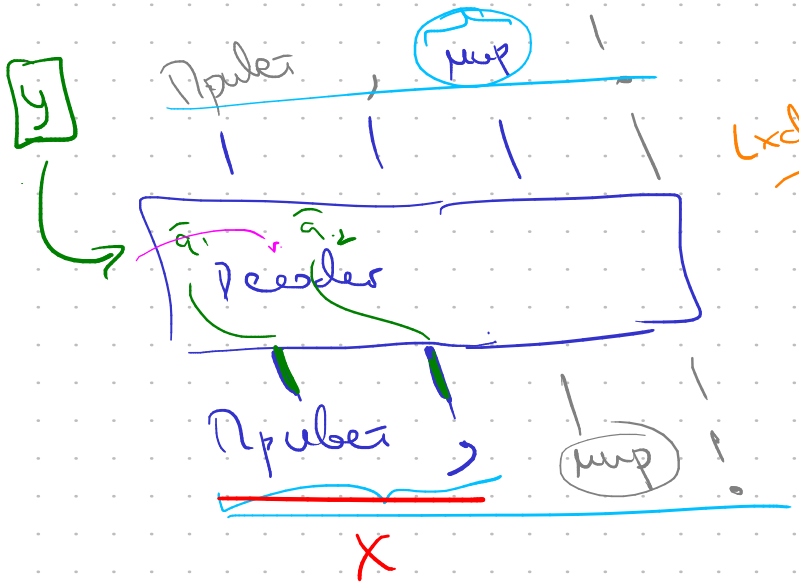
$\bar{z}_1 = \sum_{j=1}^{L} \alpha_{1j} \bar{v}_j$

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

$\bar{z}_L = \sum_j \alpha_{Lj} \bar{v}_j$

multi-head attention





Enc-dec attn:

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

$L \times d'$ $\quad$ $L \times L'$ $\quad$ $L' \times d'$

$$Q = W^Q X \qquad L \times m$$

$$K = W^K y \qquad V = W^V y$$

$L' \times m$ $\quad$ $L' \times m$