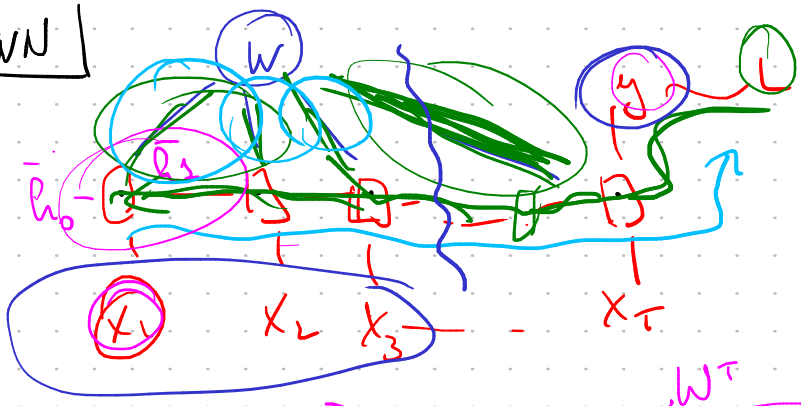
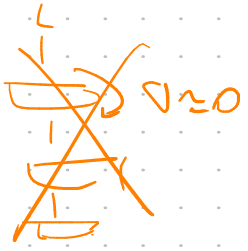
RNN



$\frac{\partial L}{\partial w_i}$

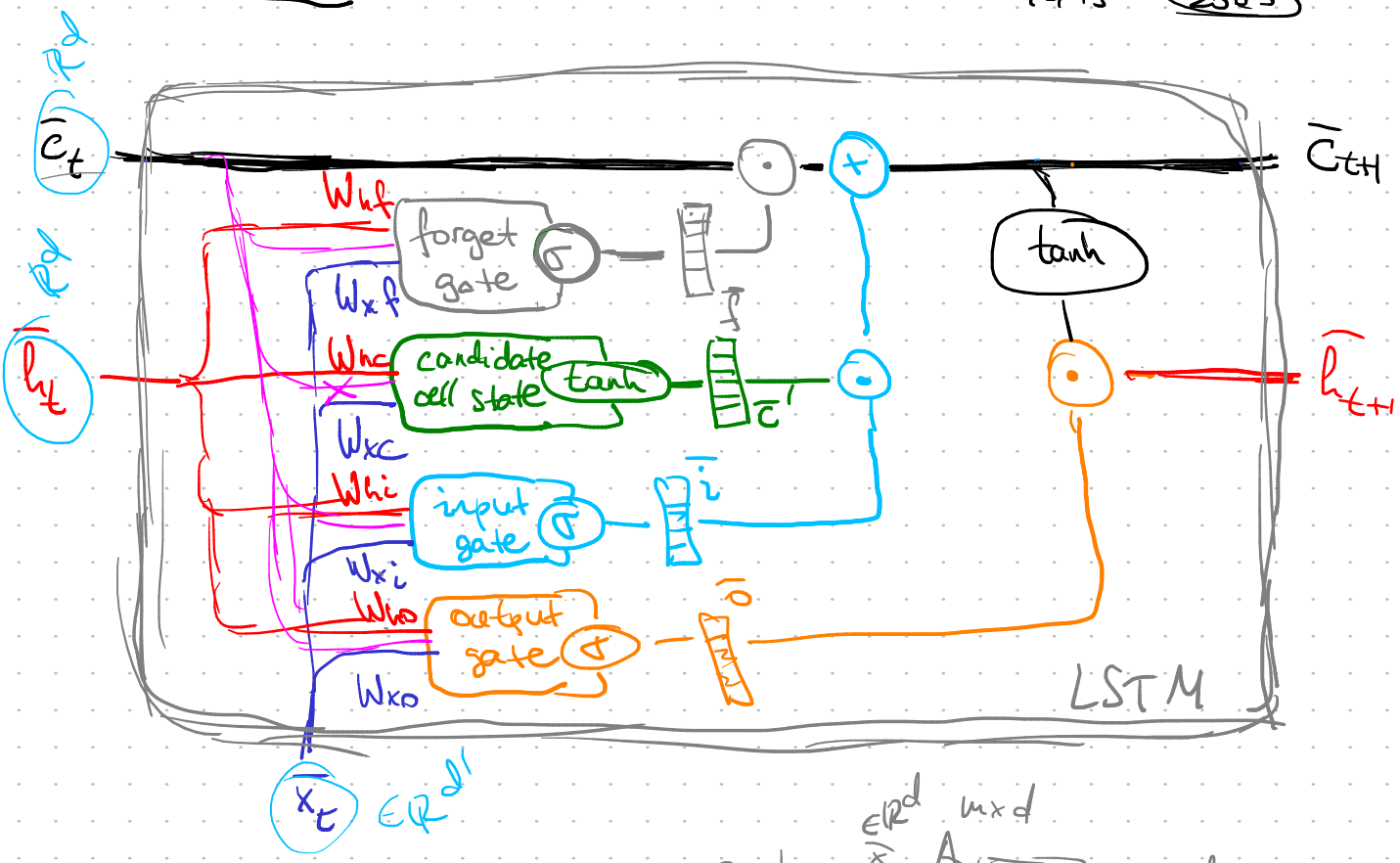


$$h_T = \overbrace{W \cdot W \cdot W \cdot W \cdots W}^{W^T} h_0$$

1) Exploding gradients $\quad \|W\| > 1 \quad \|\bar{x}\| < \|W^T \bar{x}\|$

2) Vanishing gradients $\quad \|W\| < 1 \quad \|W\bar{x}\| \approx \alpha \cdot \|\bar{x}\|$, $\alpha < 1$

LSTM – long short-term memory — Hochreiter / Schmidhuber 1995 (2005)

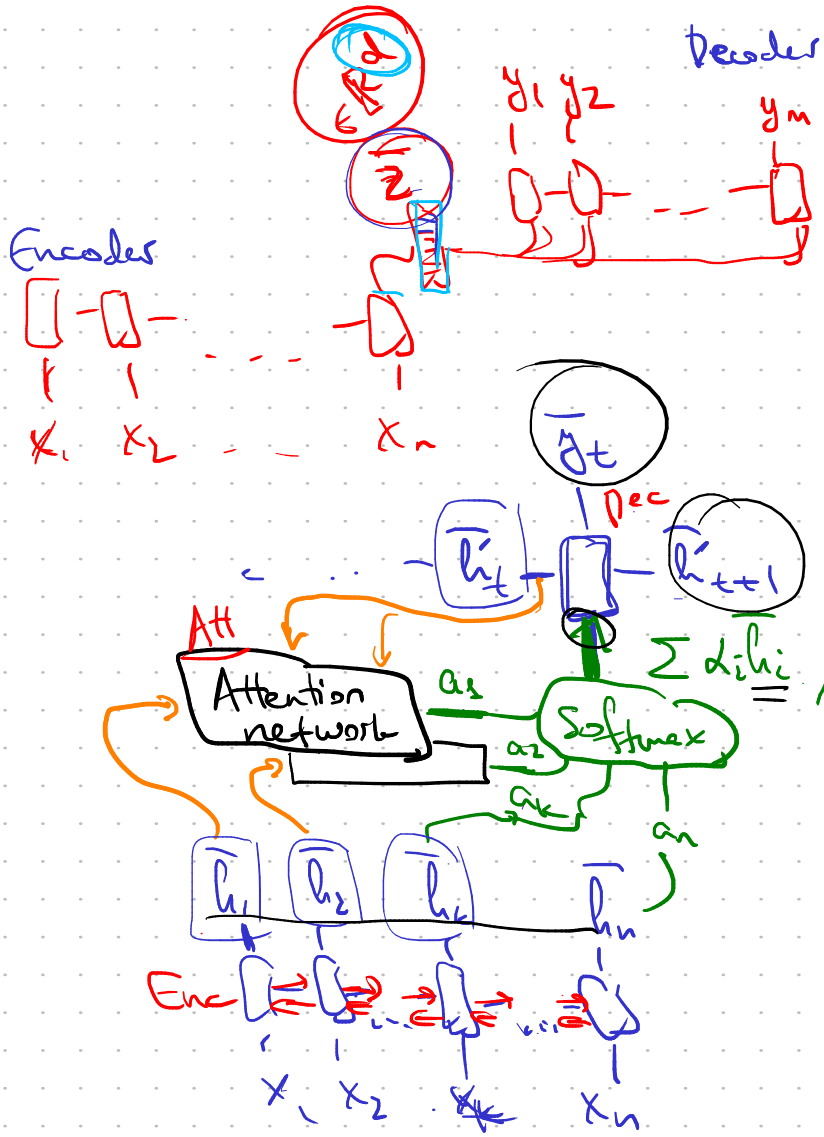


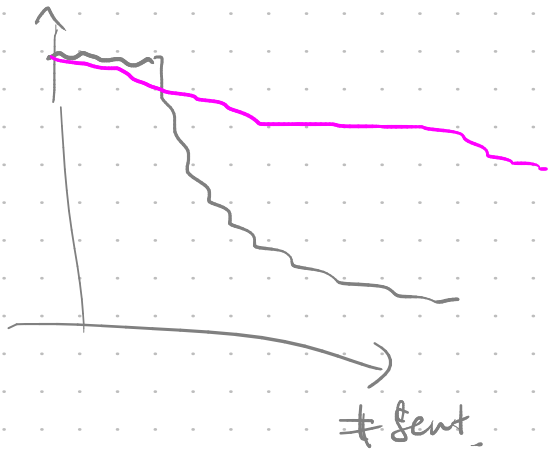
$W_{h*} : d \times d$

$W_{x*} : d \times d'$

gate: $\bar{x} \in \mathbb{R}^d$, A $m \times d$; $\bar{y} \in \mathbb{R}^{d'}$, B $m \times d'$; $f \to h \to f(\bar{x}, \bar{y})$

$h(A\bar{x} + B\bar{y} + \bar{c})$





$\sum \alpha_i h_i$, $\alpha_i \geq 0$, $\sum \alpha_i = 1$

$h_1 = Enc(x_1, s_1, s_1')$

$h_k = Enc(x_k, s_k, s_k')$

$a_k = Att_{(w)}(h_k, h_t')$

$\bar{a} = Att_{(w)}(H, h_t')$

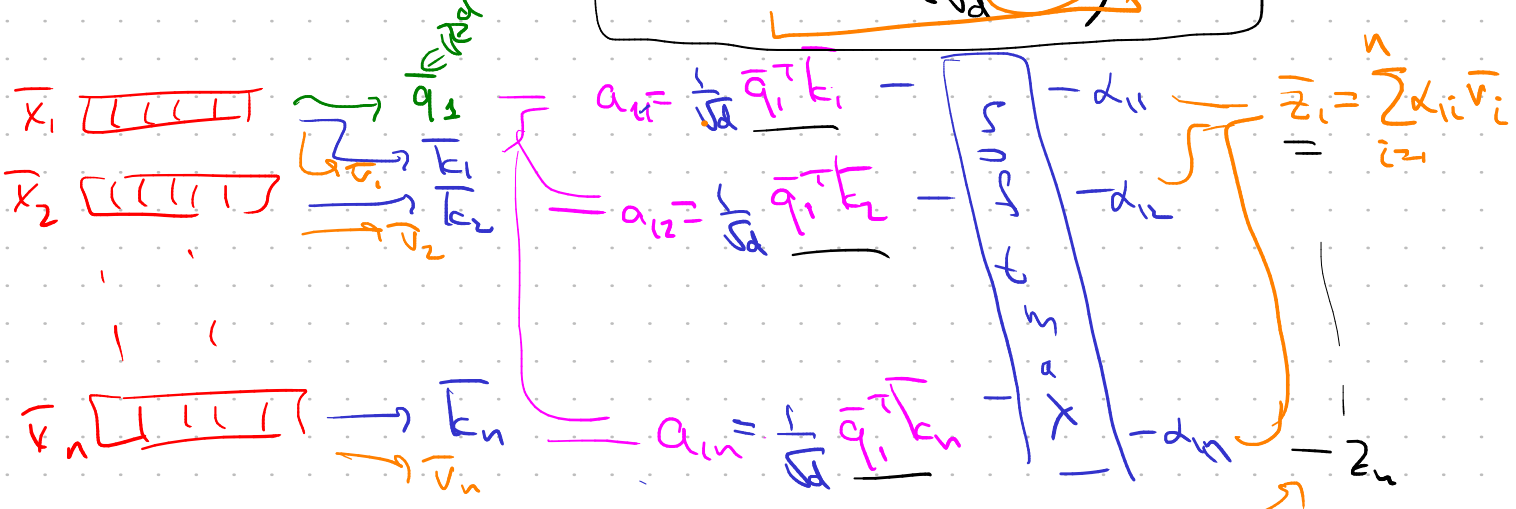
$\bar{\alpha} = softmax(\bar{a})$

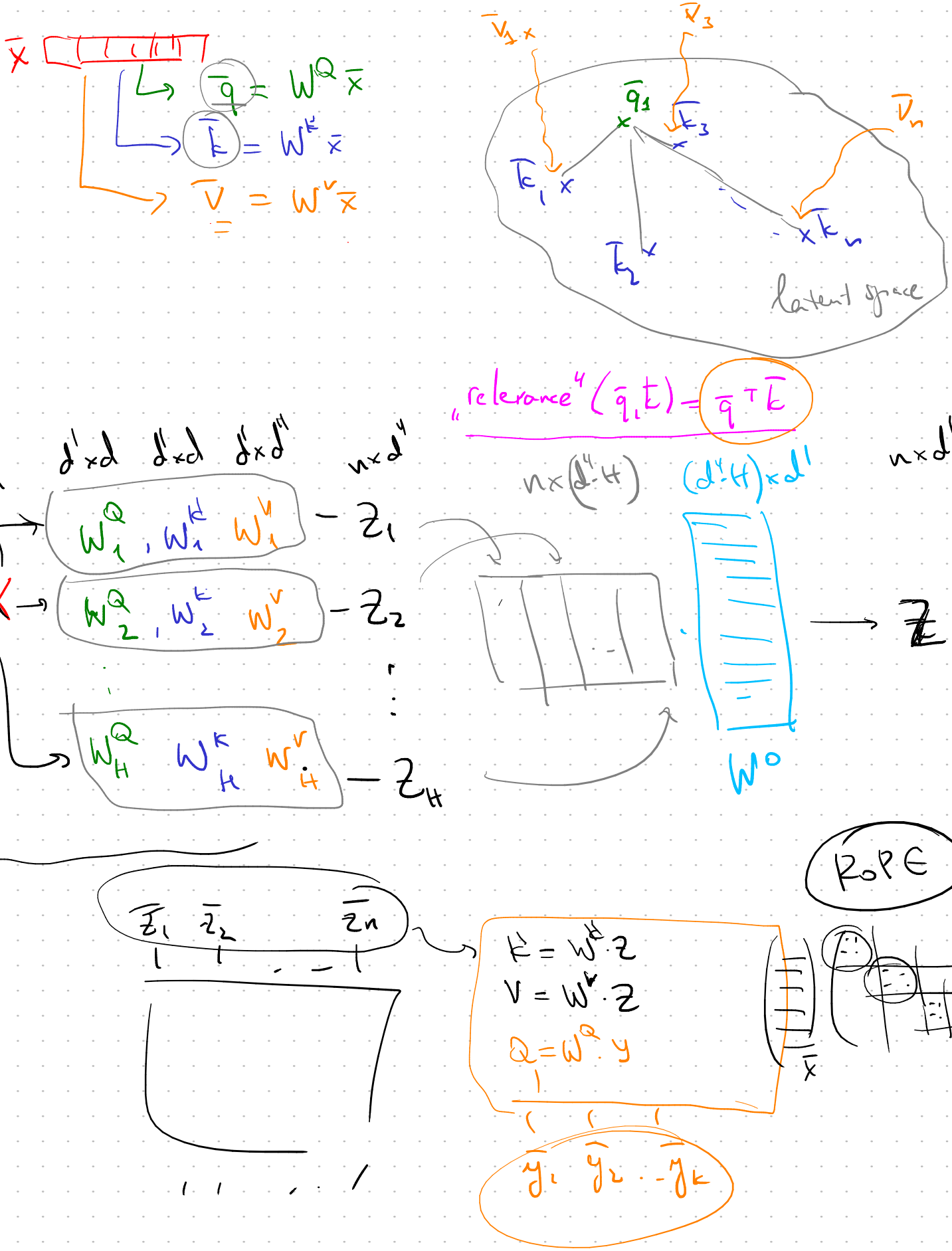
$\bar{y}_t = Dec(h_t', H \cdot \bar{\alpha})$

quadratic complexity

Self-attention

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




$\bar{x}$
$\bar{q} = W^Q \bar{x}$
$\bar{k} = W^K \bar{x}$
$\bar{v} = W^V \bar{x}$
$\bar{v}_1$ x
$\bar{v}_3$
$\bar{q}_1$
$\bar{k}_3$
$\bar{v}_n$
$\bar{k}_1$ x
$\bar{k}_2$ x
$\bar{k}_n$ x
latent space
"relevance" $(\bar{q}, \bar{k}) = \bar{q}^T \bar{k}$
$n \times d'$
$d' \times d$
$d \times d$
$d \times d''$
$n \times d''$
W_1^Q, W_1^K, W_1^V — Z_1
W_2^Q, W_2^K, W_2^V — Z_2
W_H^Q W_H^K W_H^V — Z_H
$n \times (d'' \cdot H)$
$(d'' \cdot H) \times d'$
$n \times d'$
W^O
Z
$\bar{z}_1$ $\bar{z}_2$ $\bar{z}_n$
$K = W^K \cdot Z$
$V = W^V \cdot Z$
$Q = W^Q \cdot y$
$\bar{y}_1$ $\bar{y}_2$ ... $\bar{y}_k$
RoPE
$\bar{x}$