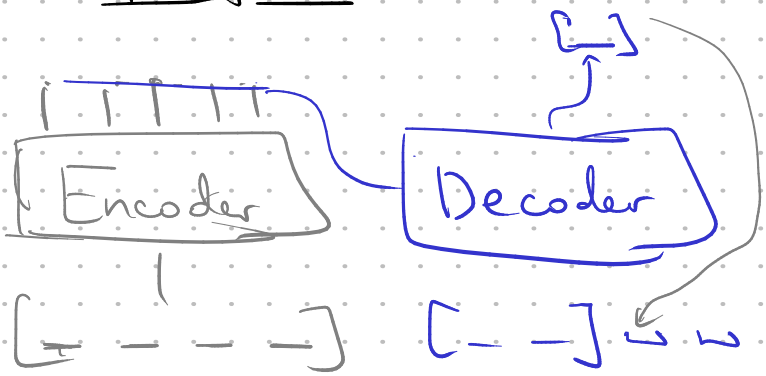
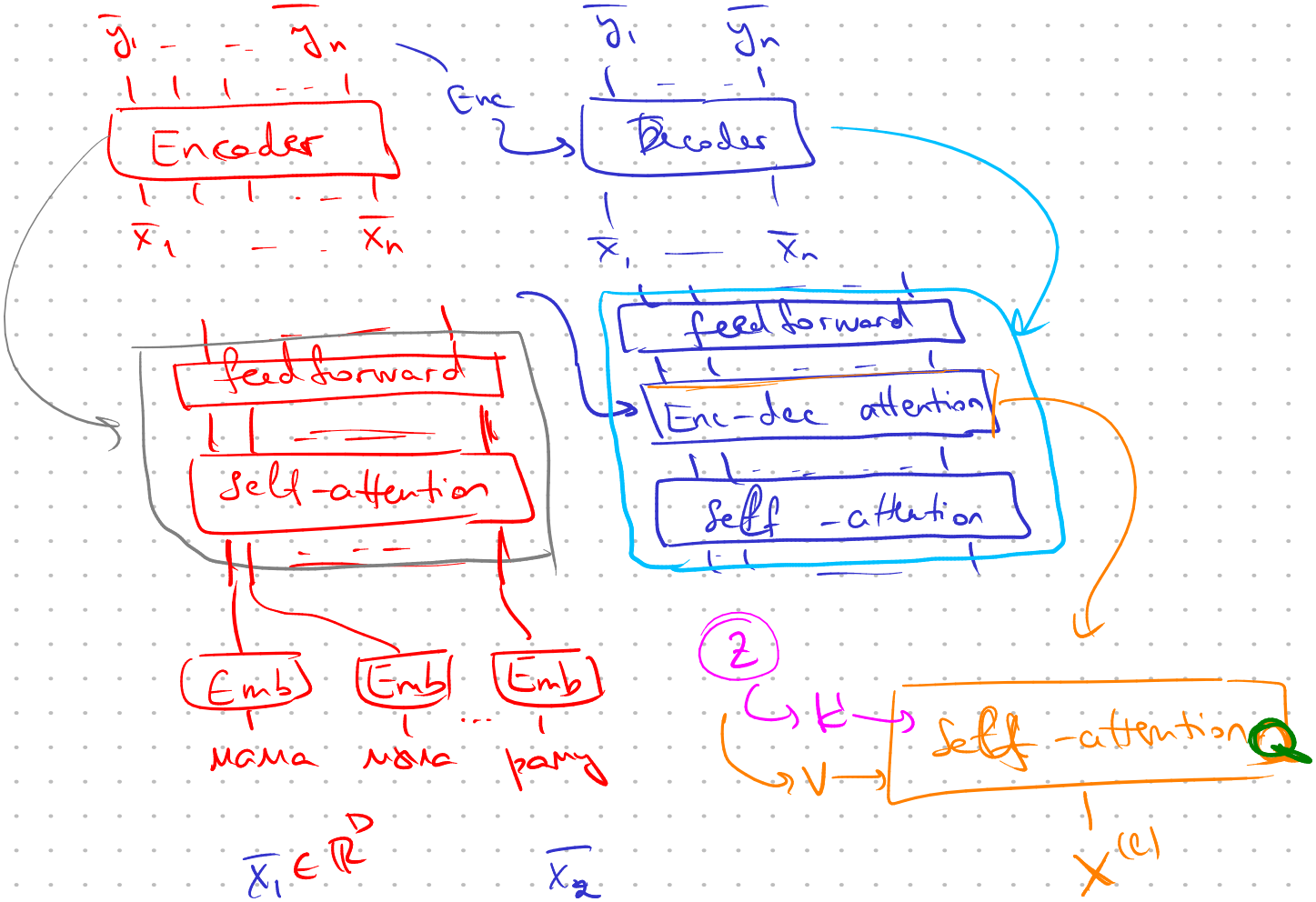
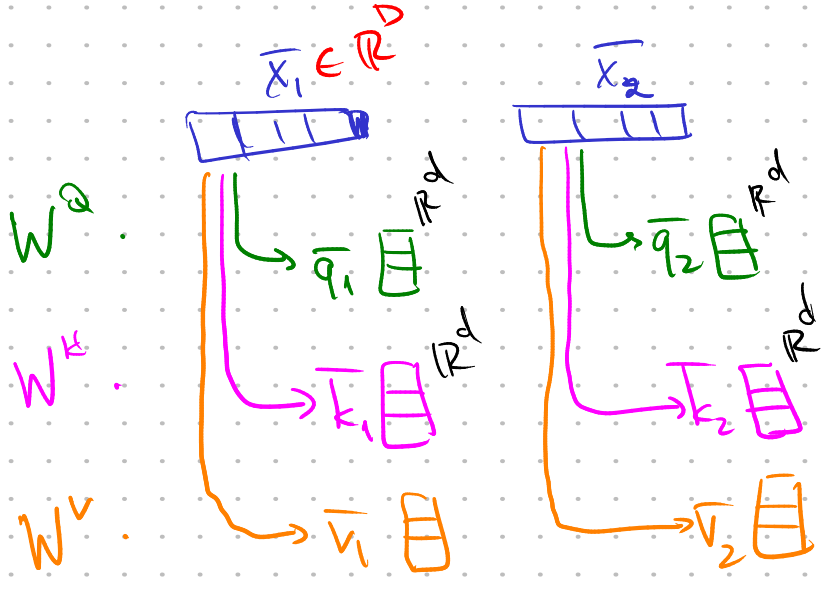
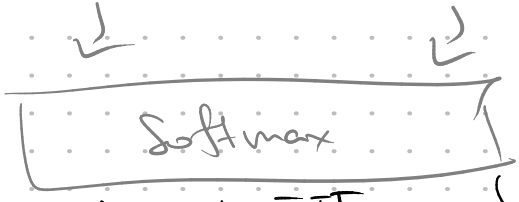
Transformer





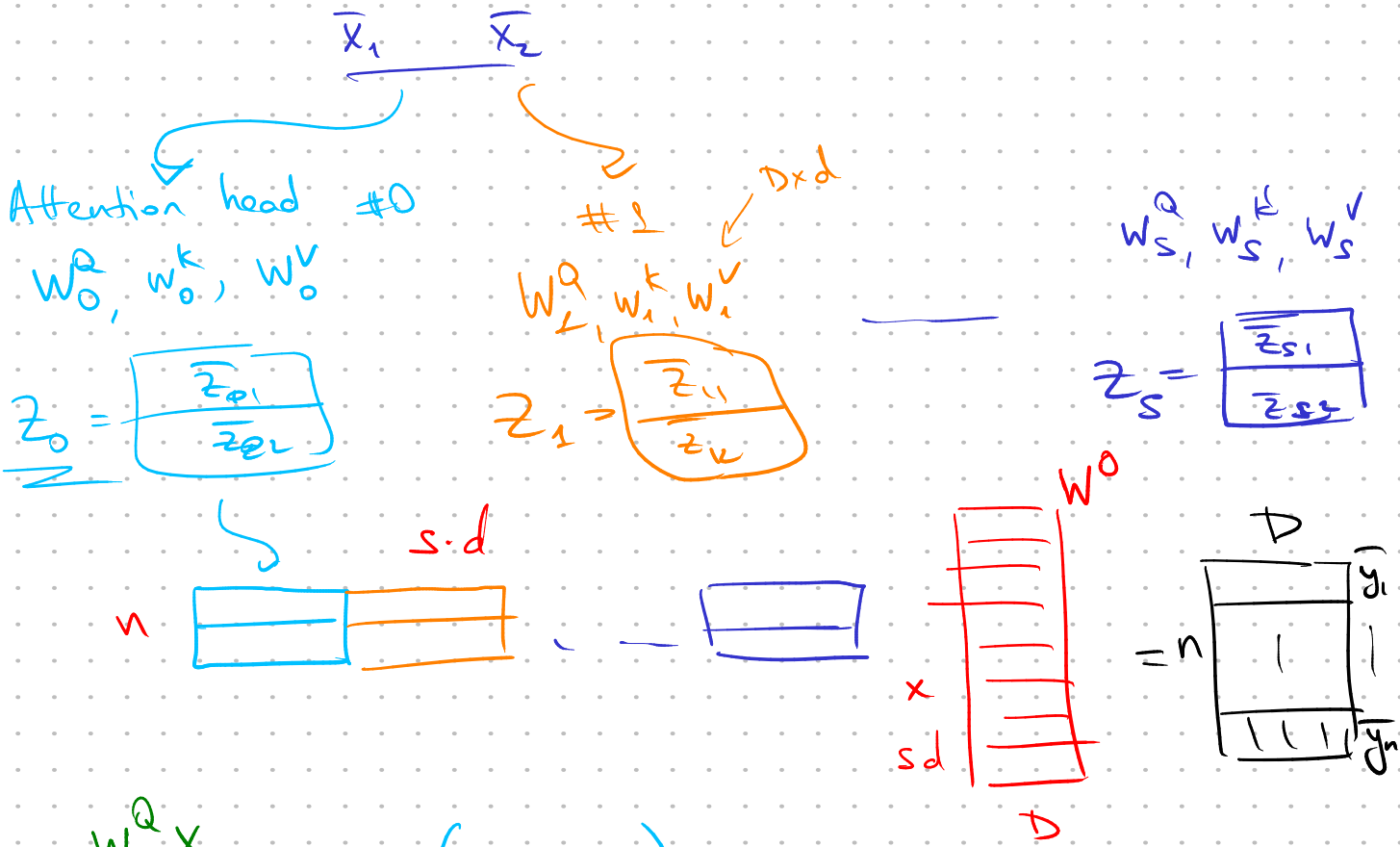




$$a_{1i} = \frac{e^{1/\sqrt{d}\, \bar{q}_1^T \bar{k}_i}}{\sum_l e^{1/\sqrt{d}\, \bar{q}_1^T \bar{k}_l}} \qquad a_{12} = \frac{\bar{q}_2^T}{\bar{q}_2^T}$$

$$\bar{z}_1 = \sum a_{1l} \bar{v}_l = \sum \text{softmax}(\bar{q}_1^T \bar{k}_l) \cdot \bar{v}_l$$

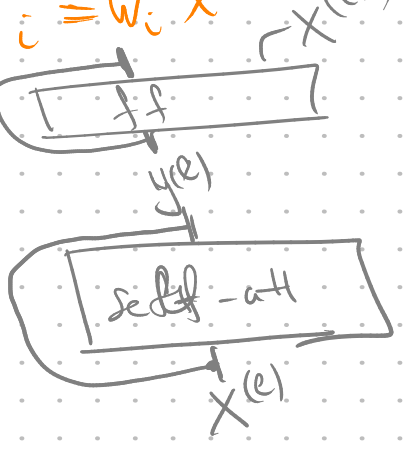
$$\bar{z}_2 = \sum a_{2l} \bar{v}_l \quad - \quad -$$

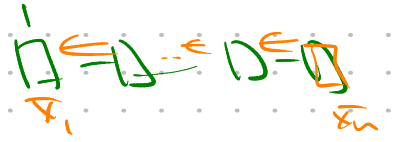


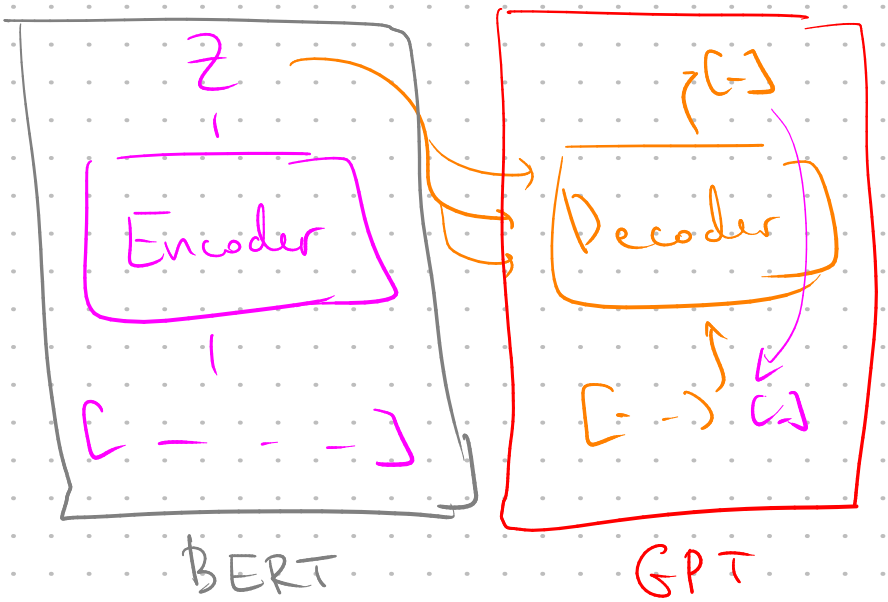
$$Q_i = W_i^Q X$$
$$K_i = W_i^K X$$
$$V_i = W_i^V X$$

$$\text{softmax}\left(\frac{1}{\sqrt{d}} Q_i K_i^T\right) \cdot V_i = Z_i$$

$$\text{concat}(Z_1 \dots Z_s) \cdot W^0 = Y$$







$\bar{u} \sim N(\bar{0}, \bar{I})$ ⤳ ⊙ ← × ← + ← $\bar{\mu}_x$ → $\bar{z} \sim N(\bar{\mu}_x, \bar{\sigma}_x)$

Gumbel-Max trick

$\bar{z} \sim Mult(\bar{\pi})$

Thm $z \sim Mult(\bar{\pi})$ ⟺ $z = \underset{i}{\operatorname{argmax}} (g_i + \log \pi_i)$,

where $g_i \sim Gumbel$

$p(g_i) = e^{-(g_i + e^{-g_i})}$, $F(g_i) = e^{-e^{-g_i}}$

Pf. $p(z=k) = p(g_k + \log\pi_k \geq g_j + \log\pi_j \ \forall j) =$

$= \int_{-\infty}^{\infty} \prod_{j \neq k} p(g_k + \log\pi_k \geq g_j + \log\pi_j \mid g_k) \cdot p(g_k)\, dg_k =$

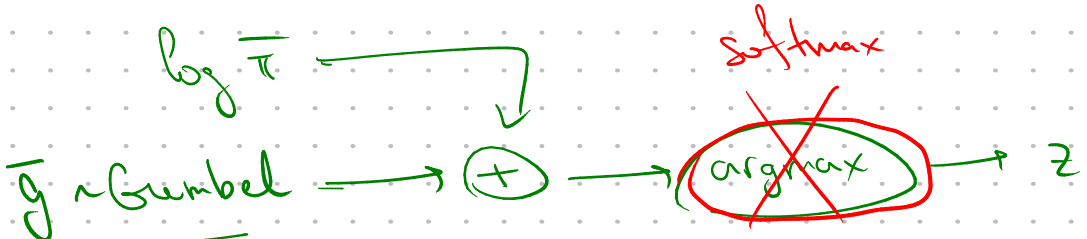
$= F(g_k + \log\pi_k - \log\pi_j)$

$= \int_{-\infty}^{\infty} \prod_{j \neq k} e^{-e^{-g_k - \log\pi_k + \log\pi_j}} \cdot e^{-g_k + e^{-g_k} + \log\pi_k - \log\pi_k}\, dg_k$

$= \int_{-\infty}^{\infty} e^{-\sum_{j \neq k} \pi_j \cdot e^{-g_k - \log \pi_k}} \cdot \pi_k e^{-(g_k + \log\pi_k + \pi_k e^{-g_k - \log\pi_k})}\, dg_k =$

$$= \pi_k \int_{-\infty}^{\infty} e^{-g_k - \log \pi_k} = e^{-g_k - \log \pi_k} \underbrace{\left(\pi_k + \sum_{j \neq k} \pi_j \right)}_{=1} \, dg_k = \pi_k$$

$$\underbrace{=1}$$



Gumbel-Softmax $\quad z = \underset{i}{\text{softmax}}\left(\frac{1}{\tau}(g_i + \log \pi_i) \right)$ $\quad \tau \to 0$