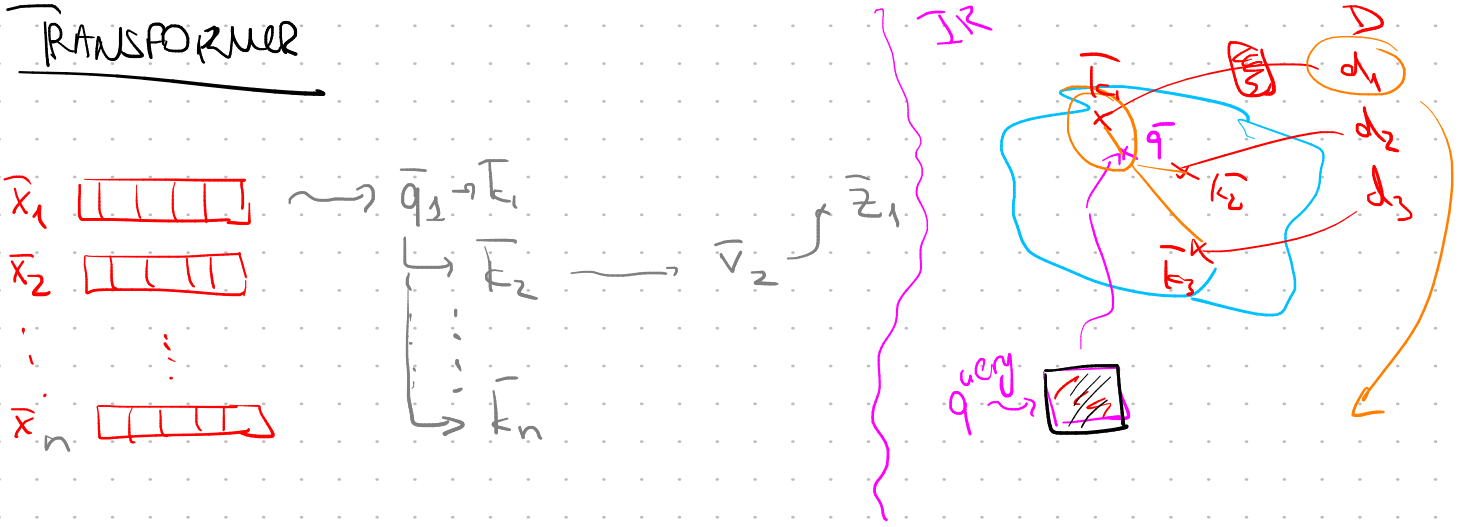
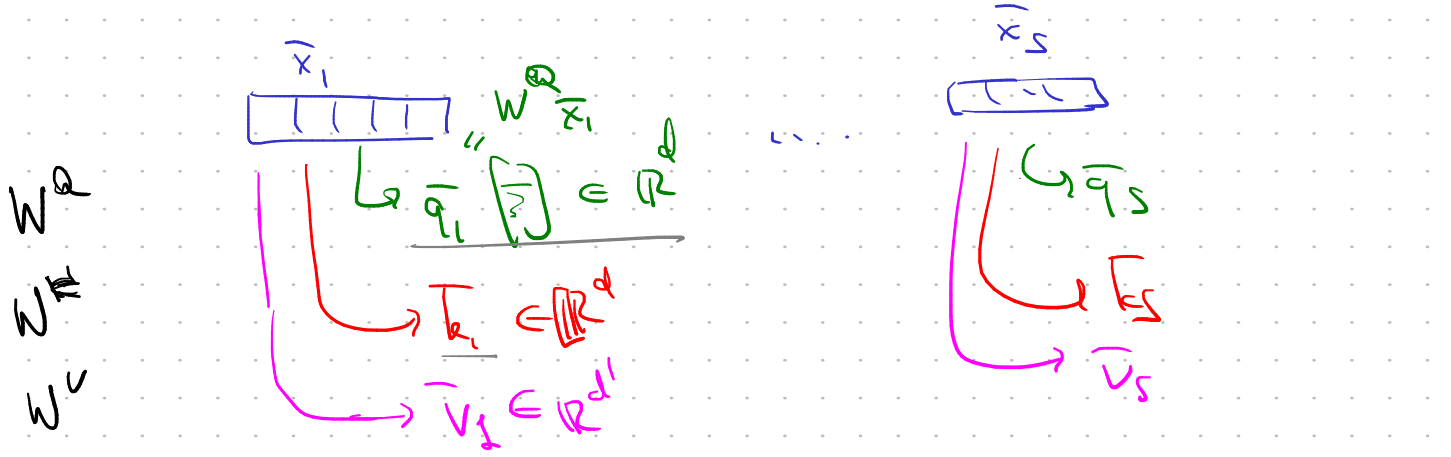
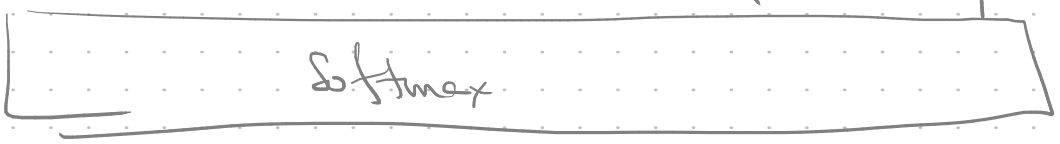
TRANSFORMER





$$\rightsquigarrow \quad \frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_1 \qquad \frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_2 \qquad \cdots \qquad \frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_s \qquad \cdots \qquad \frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_n$$



$$a_{11} \qquad a_{12} \qquad \cdots \qquad a_{1s} = \frac{e^{\frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_s}}{\sum_l e^{\frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_l}} \qquad \cdots \qquad a_{1n}$$

$$\boxed{\bar{z}_1 = \sum a_{1l}\bar{v}_l = \sum \text{softmax}\left(\frac{1}{\sqrt{d}}\bar{q}_1\cdot\bar{k}_l\right)\bar{v}_l}$$

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

$$\boxed{\bar{z}_n = \sum a_{nl}\bar{v}_l}$$

Multi-head attention

$\bar{x}_1 \; \bar{x}_2 \; \dots \; \bar{x}_n$ — X $n \times m$

Head #0 — W_0^Q, W_0^K, W_0^V

Head #1 — W_1^Q, W_1^K, W_1^V

Head #7 — W_7^Q, W_7^K, W_7^V

$$\bar{z}_{0i} = \sum_s \text{Softmax}\left(\frac{1}{\sqrt{d}} \bar{q}_{0i} \bar{k}_{0s}\right) \bar{v}_{0s}$$

$\bar{z}_{ni}$

$$Z_0 = \text{Softmax}\left(\frac{1}{\sqrt{d}} Q_0 K_0^T\right) V_0$$

$n \times d$ $\quad d \times n$ $\quad n \times d'$

Z_0 ($n \times d'$) $\quad \left(\frac{1}{\sqrt{d}} \bar{q}_i^T \bar{k}_j\right)$

Z_1 ($n \times d'$) $\quad Z_7$ ($n \times d'$)

$[Z_0 \; Z_1 \; \dots \; Z_7] \cdot W^O = Y$ $\quad n \times m$

$Z = \text{Concat}(Z_0 \dots Z_7)$

$\bar{x}$ — Enc — $\bar{\mu}$, $\bar{\sigma}$

$\bar{u} \sim N(\bar{u} | \bar{0}, I)$

$\bar{z} \sim N(\bar{z} | \bar{\mu}, \bar{\sigma})$ — Dec — $\hat{\bar{x}}$

$\bar{x}$ — Enc — $\bar{\pi} = (\pi_1 \dots \pi_K)$ $\quad \sum \pi_i = 1$

$\bar{g} \sim \text{Gumbel}$

$\tilde{\bar{z}} \sim \text{Mult}(\bar{z} | \bar{\pi})$ — Dec — $\hat{\bar{x}}$

$z_i = \text{Softmax}\left(\frac{1}{\tau}(g_i + \log \pi_i)\right)$

Gumbel-Max trick

$z \sim \text{Mult}(\pi) \iff z = \arg\max_i (g_i + \log \pi_i)$,

$\pi = (\pi_1 \dots \pi_k)$

где $g_i \sim$ Gumbel

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

Дано:

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

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

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

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

$$e^{-\pi_j \cdot e^{-g_k - \log \pi_k}}$$

$$= \int e^{-\sum_{j \neq k} \pi_j \cdot e^{-g_k - \log \pi_k}} e^{-g_k - e^{-g_k \pm \log \pi_k \pm \log \pi_k}} dg_k =$$

$$e^{-g_k - \log \pi_k + \log \pi_k} \cdot e^{-g_k - \log \pi_k}$$

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

$$= \pi_k \int e^{-e^{-g_k + \log \pi_k}} \cdot e^{-g_k + \log \pi_k} dg_k = \pi_k$$