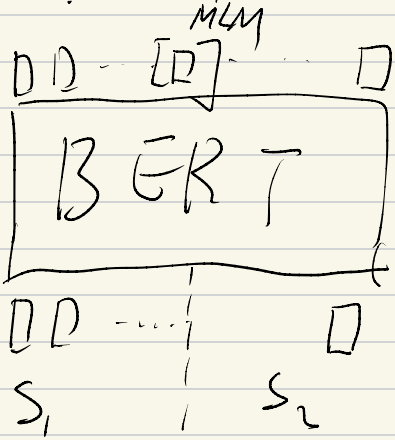
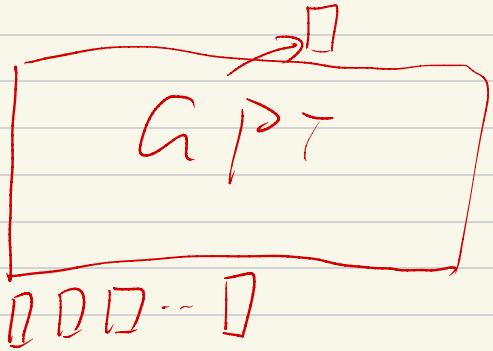
25. 02. 25



Encoder Only



Decoder Only

Easy for transforming texts into representations

KL Divergence

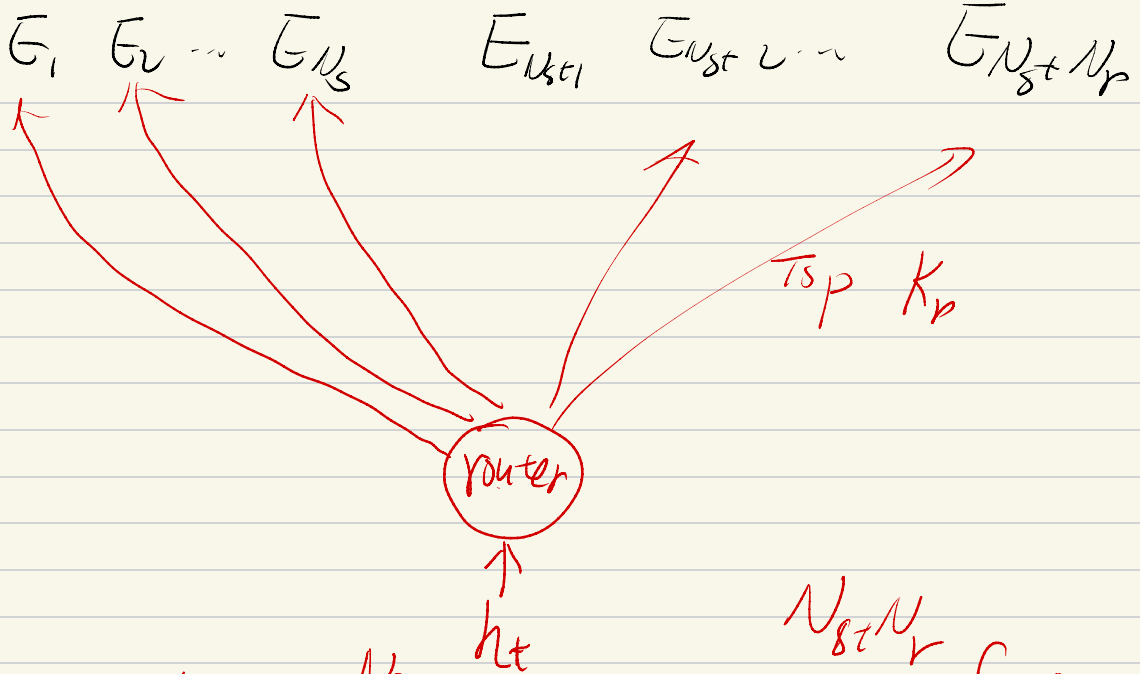
$KL(P|Q) = E_P \log[P/Q] = \sum_{i \in V} P_i \log P_i/Q_i$

$P, Q \in Dis(\{1, 2, \ldots |V|\})$

≥ 0

"=" holds iff $P = Q$

$KL(P|Q) \neq KL(Q|P)$



$$O_t = h_t + \sum_{i=1}^{N_s} f_i(h_t) + \sum_{i=N_{s+1}}^{N_s+N_r} g_i f_i(h_t)$$

f_i: MLP for expert i

$$g_i = \frac{g'_{it}}{\sum_{j=N_s}^{N_s+N_r} g'_{jt}}$$

$$g'_{jt} = \begin{cases} s_{i,t} & \text{top } k \\ 0 & \text{o.w.} \end{cases}$$

$$s_{i,t} = \text{Softmax}(h_t^T \cdot e_i) \qquad e_i: \text{Trainable}$$

$$g'_{it} = \begin{cases} s_{i,t} & \text{if } s_{i,t} + b_i \text{ is Top } k \\ 0 & \text{o.w.} \end{cases}$$