emb_dim

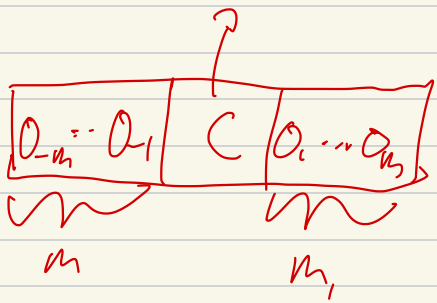
$|V|$ (Word_embedding_1)

$v_1 \in \mathbb{R}^{\text{emb\_dim}}$

v_2

$\vdots$

$v_{|V|}$



$$\hat{v} = \frac{1}{2m} \sum_{j=1}^{m} \left(v_{O_{-j}} + v_{O_j} \right)$$

emb_dim

$|V|$ (Word_emb_2)

u_1

u_2

$\vdots$

$u_{|V|}$

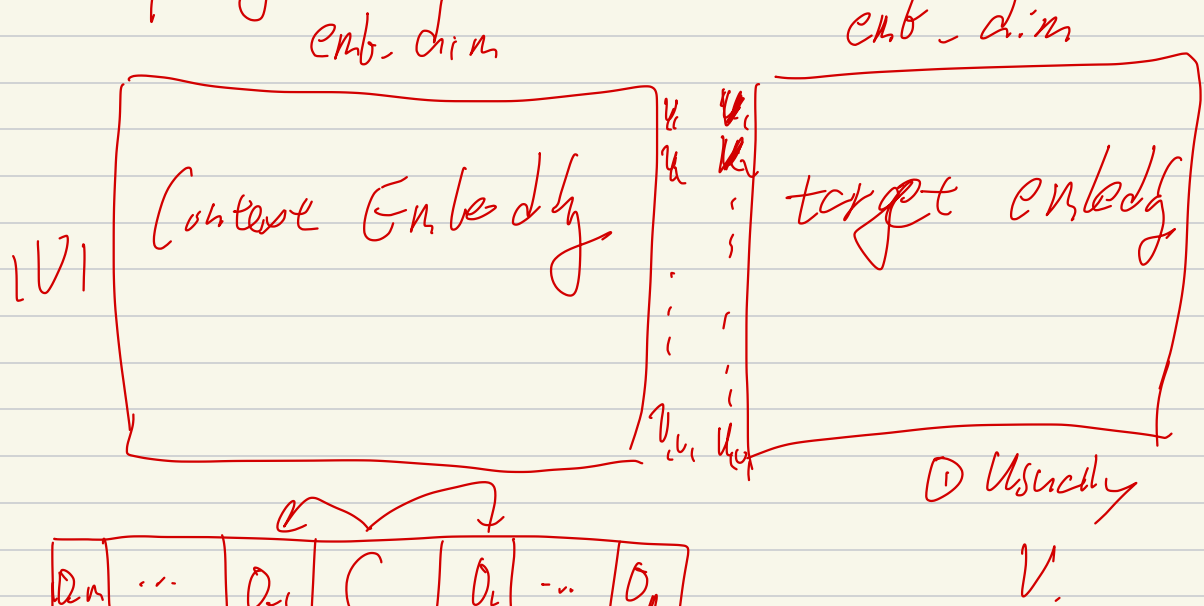
$$\Pr[C \mid O_{-m}\ O_{-(m-1)} \cdots O_{-1}, O_1, \cdots O_m]$$

$$= \frac{\exp(u_c^T \cdot \hat{v})}{\sum_{j=V} \exp(u_j^T \cdot \hat{v})}$$

$$\underset{v,u}{\operatorname{argmin}} -\sum_{c \in D} \log \left(\frac{\exp(u_c^T \cdot \hat{v})}{\sum_{j \in V} \exp(u_j^T \cdot \hat{v})} \right)$$

Output : V.

Skip gram



① Usually V.

② $\frac{u+v}{2}$

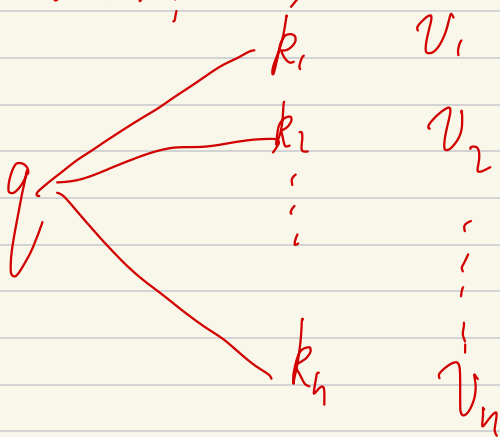
③ (u, v)

$$\min \sum_{CED} = \sum_{\substack{j=-m \\ j \neq 0}}^{m} \left(\log \frac{\exp(u_{o_j}^T \cdot v_c)}{\sum_{i \in V} \exp(u_i^T \cdot v_c)} \right)$$

$i \in V' \subseteq V$

Negative Sampling

Attention: key, value



$$\sum_{i=1}^{n} f(q, k_i) \cdot v_i$$

Transformer:

w_1 w_2 ... w_n

x_1 x_2 ... $x_n \in \mathbb{R}^d$

$W_q, W_k, W_v \in \mathbb{R}^{d \times d}$

$q_i = W_q x_i \in \mathbb{R}^d$, $k_i = W_k \cdot x_i$, $v_i = W_v x_i$

query key value

$$w'_{i,j} = \frac{q_i^T \cdot k_j}{\sqrt{d}} \xrightarrow{\text{softmax}} w_{ij} = \frac{\exp(w'_{ij})}{\sum_{j'=1}^{n} \exp(w'_{ij'})}$$

$$y_i = \sum_{j=1}^{n} w_{ij} \cdot v_j \in \mathbb{R}^d$$

$x_1\ x_2 \cdots \cdots\ x_n$

↓↓

$y_1\ y_2 \cdots\ y_n$

① Better Parallization ⇒ multi-head

② y is linear in v.

Add MLP $y \to MLP(y)$

③ Sequence information.

+ Position Encoding

input = x + Position Embedding

④ No future information.

masked Attention.

⑤ Optimization ?

Skip-connection.

Layer Normalization