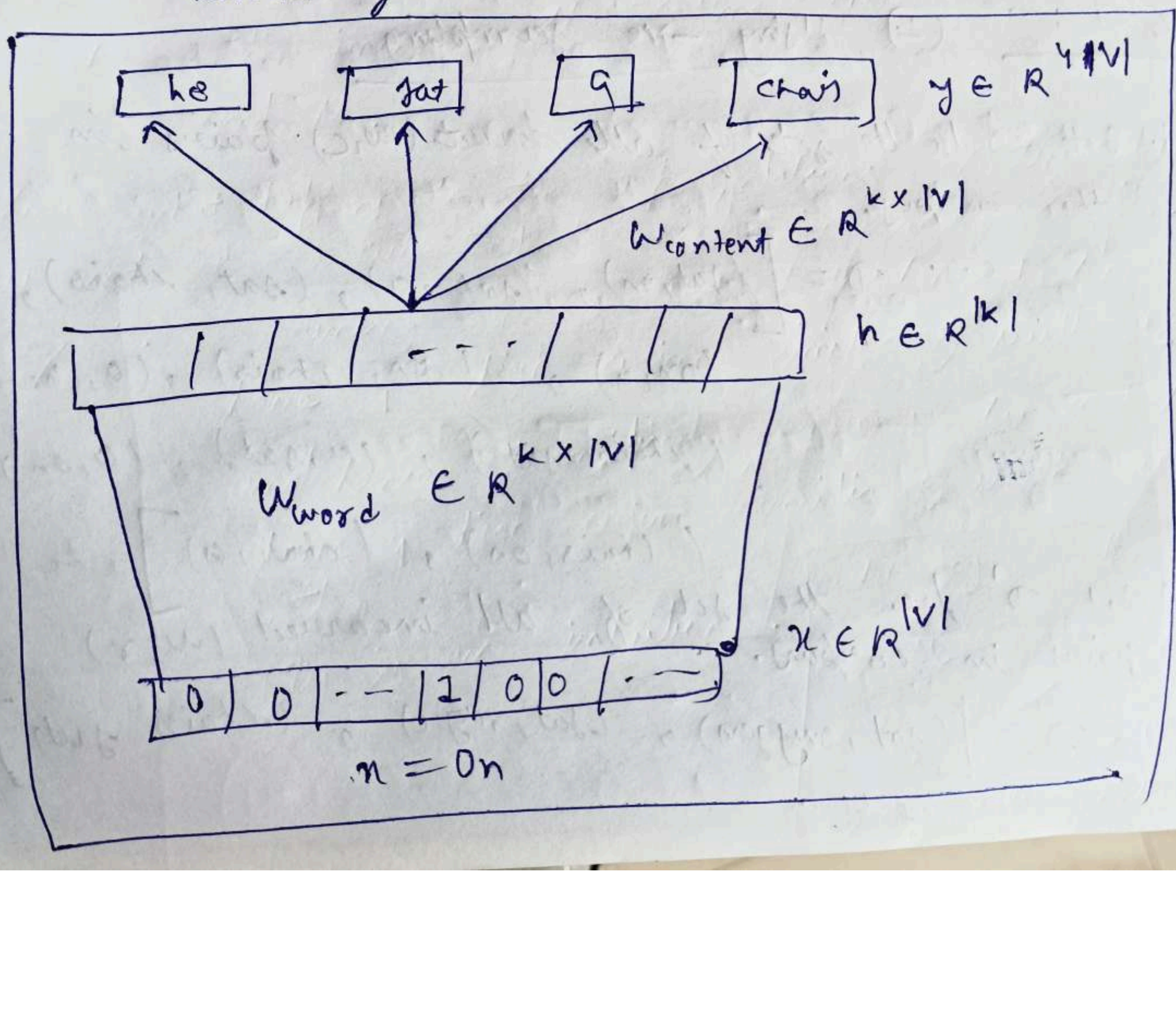
Skip - Gram - Model

→ ~~Previo~~ Previously, CBOW bag of words model predicted a word given context.

→ But, skip-gram model will predict context given a word.



→ Here the loss f^n will be sum of cross-entropies.

$$\boxed{\mathcal{L}(\theta) = -\sum_{i=1}^{d-1} \log \hat{y}_{w_i}}$$

<u>Problems</u> :-

(i.) Softmax f^n at the output is is computationally expensive.

$$\boxed{\hat{y}_w = \frac{e^{u_c \cdot v_w}}{\sum_{w' \in V} e^{u_c \cdot v_w}}}$$

u_c = c-th column of $W_{content}$

v_w = w-th column of W_{word}.

Sol^n :- ① <u>Using -ve sampling</u>

(i.) Let D be the set of all correct (w,c) pairs in the corpus.

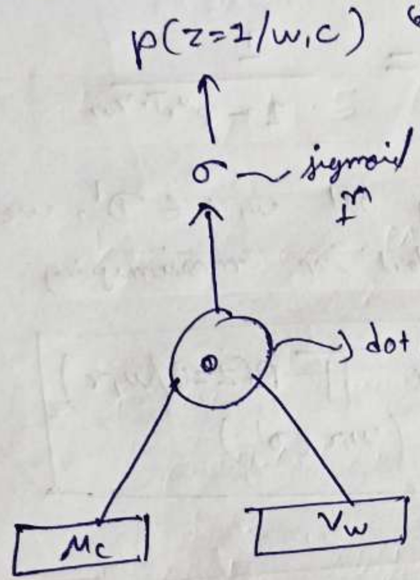
e.g. D = [(sat, on), (sat, a), (sat, chair), (on, a), (on, chair), (a, chair), (chair, sat), (a, sat), (a, on), (chair, on), (chair, a)] etc.

(ii.) Let D' be the set of all incorrect (w, r) pairs in corpus.

D' = [(sat, oxygen), (sat, magic), (chair, sad)] etc.

(iii.) D' can be constructed by randomly sampling a content word r which has never appeared with w a creating pair (w, r).

(iv.) As before let v_w be the representation of word w and μ_c be the representation of the content word c.



(*) For a given $(w,c) \in D$ we are interested in maximizing

$$p(z=1|w,c)$$

(*) Let us model this probability by

$$p(z=1/w,c) = \sigma(\mu_c^T v_w)$$

$$= \frac{1}{1 + e^{-\mu_c^T \cdot v_w}}$$

(*) Considering all $w,c \in D$, we are interested in

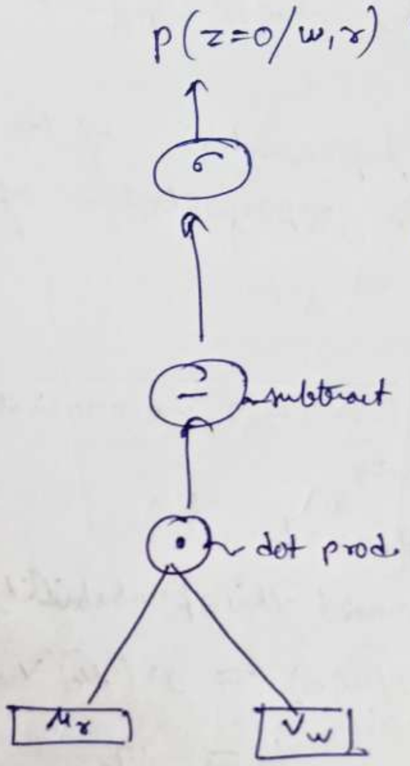
multiply

$$\boxed{\underset{\theta}{\text{maximise}} \prod_{(w,c) \in D} p(z=1/w,c)}$$

θ: is the word representation (v_w) and context representation (μ_c) for all words in our corpus.

we want to maximise the $p=1$ (probability) for each word in our corpus.

$\Rightarrow$ $\therefore$ multiplication (and).



(*) For $(w,r) \in D'$ we are interested in maximising

$$p(z=0/w,r)$$

(*) again we model this

$$p(z=0/w,r) = 1 - \frac{1}{1+e^{-\mu_r^T v_w}}$$

$$= \frac{1}{1+e^{\mu_r^T \cdot v_w}}$$

(*) considering all $w, r \in D'$, we are interested in maximising

$$\boxed{\underset{\theta}{\text{maximise}} \prod_{(w,r \in D')} p(z=0/w,r)}$$

~~Note~~ Note that we want $p(z=1/w,c)$ for each correct word $\in D$ and we also want $p(z=0/w,r)$ for each incorrect word $\in D'$. Both simultaneously i.e. AND.

Hence our final goal is

$$\underset{\theta}{\text{maximise}} \prod_{(w,c) \in D} p(z=1/w,c) \prod_{(w,r) \in D'} \left(p(z=0/w,r)\right)$$

By simplifying and taking log, we get.

$$\boxed{\underset{\theta}{\text{maximise}} \sum_{w,c \in D} \log(\sigma(u_c^T v_w)) + \sum_{(w,r) \in D'} \log(\sigma(-u_r^T v_w))}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Note, in the original research paper.

(i) Size of D' is k times the size of D.

(ii) random context word is drawn from a modified unigram distribution

$$\boxed{r \sim (P(r))^{3/4}}$$

where, $P(r) = \frac{\text{count}(r)}{N}$

N = total no. of words in corpus.