Word2vec

- Assign vectors to words.
- One approach
  - co-occurrence matrix.
  - SVD (Singular value decomposition)
  - Word2vec — CBOW (continuous bag of words model).
    - Skipgram.
- word2vec.

$$\underline{w_1 \cdots w_m}\; w_{m+1} \;\underline{w_{m+2} \cdots w_{2m+1}}$$

- Suppose we can assign vectors to words.
- Using those assignment we will define

CBOW

$$p(w_{m+1} \mid w_1 \cdots w_m, w_{m+2} \cdots w_{2m+1})$$

center word — context.

$$= \frac{\cancel{e^{u_c^T v_o}} \quad e^{v'^T_{w_{m+1}} V_{w_c}}}{\sum_{1}^{W}}$$

Skipgram

opposite.

V' — outside rep

V — inside rep.

Vocabulary size — V

Hidden layer size — N. (we want each word to be represented by a vector of size N.)

Input is a one-hot encoded vector.

Only one of $\{x_1 \dots x_V\}$ is 1 rest is 0.

Weights between input layer & hidden layer is a $V \times N$ matrix. W.

$W_{V \times N}$

N-dim vector rep of the word.

V

N

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_V \end{bmatrix} \qquad W^T \cdot X = \begin{bmatrix} h_1 \\ \vdots \\ h_N \end{bmatrix} \qquad h_{N \times 1}$$

When $x_k = 1$ & $x_{k'} = 0$ for $k' \neq k$.

$h = W^T X$ is the kth row of W written as a column

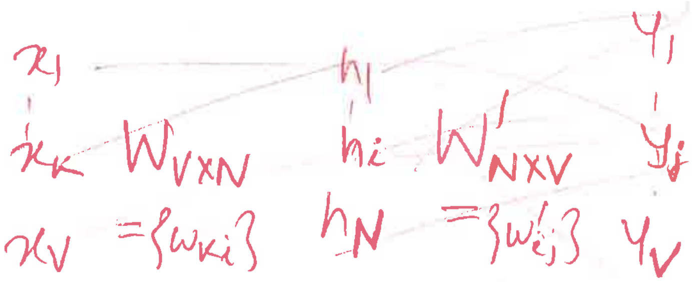
Use Property. x_k

V_{w_I} — vector rep of input word w_I

So to write it as a column we have $V_{w_I}^T$.

$$h = V_{w_I}^T$$

②



$V =$ Size of vocabulary.

$u_j = V'^{T}_{w_j} h$ where V'_{w_j} is the jth column of W' (Become jth row in W'^T)

= denotes the score of the jth word in the vocabulary.

$$p(w_j | w_I) = y_j = \frac{e^{u_j}}{\sum_{j'=1}^{V} e^{(u_{j'})}}.$$

$$= \frac{e^{V'^{T}_{w_j} V^{T}_{w_I}}}{\sum_{j'=1}^{V} e^{V'^{T}_{w_{j'}} V^{T}_{w_I}}}$$

V_w — input vector

$V_{w'}$ — output vector

(for one training sample)

Training Objective is to maximize the last equation:

③

---

the conditional probability of observing the actual word w_O given the input context word w_I

denote its index as j^*

$$\log p(w_O|w_I) = \log y_{j^*}$$

$$= \log \frac{e^{u_j^*}}{\sum_{j'=1}^{V} e^{u_{j'}}} = u_j^* - \log \sum_{j'=1}^{V} e^{u_{j'}}$$

$$= -E$$

So minimize $E = -\log p(w_O|w_I)$.

$$E = -u_j^* + \log \sum_{j'=1}^{V} e^{u_{j'}}$$

$$\frac{\partial E}{\partial u_j} = -1 + \frac{\partial \sum e^{u_{j'}} / \partial u_j}{\sum_{j'=1}^{V} e^{u_{j'}}} \quad \text{if } j^* = j$$

$$= 0 \quad + \quad " \quad j^* \neq j$$

$$= -t_j + \frac{e^{u_j}}{\sum_{j'=1}^{V} e^{u_{j'}}} = -t_j + y_j$$

$$= y_j - t_j = e_j$$

When $t_j = 1$ if $j = j^*$
$= 0$ o.w.

(4)

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w_{ij}}$$

$$= e_j \cdot \frac{\partial u_j}{\partial w_{ij}}$$

$$W_{V\times N} = \{w_{ki}\} = \begin{bmatrix} w_{11} & \cdots & w_{1i} & w_{1N} \\ w_{k1} & \cdots & w_{ki} & w_{kN} \\ \vdots & & & \\ w_{V1} & \cdots & w_{Vi} & w_{VN} \end{bmatrix} \quad W' = \begin{bmatrix} w'_{11} & \cdots & w'_{1j} & w'_{1V} \\ w'_{i1} & \cdots & w'_{ij} & w'_{iV} \\ \vdots & & & \\ w'_{N1} & \cdots & w'_{Nj} & w'_{NV} \end{bmatrix}$$

$$u_j = v'^{T}_{w_j} \cdot h = w'_{1j} h_1 + w'_{2j} h_2 + \cdots w'_{Nj} h_N$$

$$= \sum_{i=1}^{N} w'_{ij} h_i$$

So $\dfrac{\partial u_j}{\partial w'_{ij}} = h_i$

$$\frac{\partial u_j}{\partial h_i} = w'_{ij}$$

So $\dfrac{\partial E}{\partial w'_{ij}} = e_j \cdot h_i$

Using stochastic gradient $w'^{(new)}_{ij} = w'^{(old)}_{ij} - \eta \cdot e_j \cdot h_i$

or

$$v'^{(new)}_{w_j} = v'^{(old)}_{w_j} - \eta \cdot e_j \cdot \mathbf{h} \qquad \text{for } j = 1 \cdots V$$

$$\boxed{h = W^T \cdot x \quad \text{So } h_i = w_{1i} x_1 + w_{2i} x_2 + \cdots w_{Vi} x_V \quad h_i = \sum_{k=1}^{V} w_{ki} x_k \quad \frac{\partial h_i}{\partial w_{ki}} = x_k}$$

*4

Update equation for input → hidden weights

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i}\underbrace{\frac{\partial h_i}{\partial w_{ki}}}_{x_k}$$

$$E = -u_{j^*} + \underbrace{\log\left(e^{u_1} + e^{u_2} + \cdots e^{u_V}\right)}_{\text{page 3.}}$$

$\cancel{E = -u_{j^*} + \log(}$

each u_j has h_i's in them.

$$\frac{\partial E}{\partial h_i} = \frac{\partial E}{\partial u_1}\frac{\partial u_1}{\partial h_i} + \frac{\partial E}{\partial u_2}\frac{\partial u_2}{\partial h_i} + \cdots$$

$$= \sum_{j=1}^{V} \frac{\partial E}{\partial u_j}\frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{V} \underbrace{e_j}_{\text{page 3}} \cdot \underbrace{w_{ij}}_{\text{page 4.}} = EH_i$$

So $\dfrac{\partial E}{\partial w_{ki}} = EH_i \cdot x_k$ Gradient step is defined based on this.

Multi-word context: Defn of $h = v^T_{w_I}$ is changed to

$$h = \frac{1}{C}\left(v_{w_1} + v_{w_2} + \cdots v_{w_C}\right)^T$$

where C is the # words in the context.

Everything else stays same.