CBOW

Gerome Yoo

Department of Statistics, Sungkyunkwan University

2018

INDEX

CBOW

One-word Context

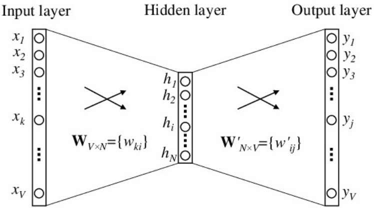

Figure: A Simple CBOW model with only one word in the context

- V : vocabulary size
- Condition : Adjacent layers are fully connected.
- Input : one-hot encoded vector of given context word
- W' is not transpose of W, is different weight matrix.

CBOW

One-hot encoding vector

- Dimension : $V \times 1$
- Note that using Vocabulary,
  $\iff$ no frequency information from corpus.

$$k\text{-th Word} \Rightarrow \begin{bmatrix} 0_1 \\ \vdots \\ 1_k \\ \vdots \\ 0_v \end{bmatrix}$$

CBOW

Weight Matrix : Input to Hidden

- Dimension : $V \times N$
  where N : hidden layer size

$$\mathbf{W}_{V \times N} = \{w_{ki}\} = \begin{bmatrix} \mathbf{v}_{w_1}^T \\ \vdots \\ \mathbf{v}_{w_k}^T \\ \vdots \\ \mathbf{v}_{w_v}^T \end{bmatrix}$$

- $\mathbf{v}_{w_k}$: k-th row of W in N-dimension vector representation.

CBOW

Hidden Layer

Given context (a word),
assuming $x_k = 1$ and $x_{k'} = 0$ for $k' \neq k$,

$$\mathbf{h} = \mathbf{W}^T \mathbf{x} = \mathbf{W}^T_{(k,\cdot)} := \mathbf{v}^T_{w_I} \rightarrow \mathbf{v}_{w_I} = \mathbf{v}_{w_k}??$$

which is essentially copying k-th row of W.

- Dimension : $N \times 1$
- $\mathbf{v}_{w_k}$ is vector representation of input word w_k
- Link(Activation) function of hidden layer units is simply *linear*, Directly passing its weighted sum of inputs to the next layer.

CBOW

Hidden Layer Computation

$$\mathbf{h} = \mathbf{W}^T \mathbf{x} = \begin{bmatrix} \mathbf{v}_{w_1} & \cdots & \mathbf{v}_{w_k} & \cdots & \mathbf{v}_{w_v} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 1_k \\ \vdots \\ 0_v \end{bmatrix} = \begin{bmatrix} \mathbf{v}_{w_k} \end{bmatrix}$$

CBOW

Weight Matrix : Hidden to Output

- Dimension : $N \times V$
  where N : hidden layer size

$$\mathbf{W'}_{N\times V} = \{w'_{ij}\} = \begin{bmatrix} \mathbf{v}'_{w_1} & \cdots & \mathbf{v}'_{w_j} & \cdots & \mathbf{v}'_{w_v} \end{bmatrix}$$

- $\mathbf{v}'_{w_j}$: j-th column of W' in N-dimension vector.
- Note that dimension V for output layer is arbitrarily chosen.

CBOW

Output Layer

$$u_j = \mathbf{v}'^{T}_{w_j}\mathbf{h}$$

- Dimension : $V \times 1$
- u_j is score for each word in vocabulary.
- Note that, u_j is not final element.
- To obatin distribution of words, we use softmax.

CBOW

Output Layer Computation

$$\mathbf{u} = \mathbf{W'}^T\mathbf{h} = \begin{bmatrix} \mathbf{v}'_{w_1} & \cdots & \mathbf{v}'_{w_j} & \cdots & \mathbf{v}'_{w_v} \end{bmatrix}^T \mathbf{h}$$

$$= \begin{bmatrix} \mathbf{v'}^T_{w_1} \\ \vdots \\ \mathbf{v'}^T_{w_j} \\ \vdots \\ \mathbf{v'}^T_{w_v} \end{bmatrix} \mathbf{h} = \begin{bmatrix} \mathbf{v'}^T_{w_1}\mathbf{h} \\ \vdots \\ \mathbf{v'}^T_{w_j}\mathbf{h} \\ \vdots \\ \mathbf{v'}^T_{w_v}\mathbf{h} \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_j \\ \vdots \\ u_v \end{bmatrix} \xRightarrow{softmax} \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_v \end{bmatrix}$$

CBOW

Applying Softmax to output layer

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

- log linear classification model.
- Use to obtain Posterior distribution of words.
- Multinomial distribution.
- y_j : j-th unit on output layer.

CBOW

Output layer

$$p(w_j|w_I) = y_j = \frac{exp(u_j)}{\sum_{j'=1}^{V} exp(u_{j'})} = \frac{exp(\mathbf{v'}^T_{w_j}\mathbf{h})}{\sum_{j'=1}^{V} exp(\mathbf{v'}^T_{w_{j'}}\mathbf{h})}$$

$$= \frac{exp(\mathbf{v'}^T_{w_j}\mathbf{v}_{w_I})}{\sum_{j'=1}^{V} exp(\mathbf{v'}^T_{w_{j'}}\mathbf{v}_{w_I})}$$

- Note $\mathbf{v}_w$ and $\mathbf{v'}_w$ are two representations of the word w.
- $\mathbf{v}_w$: **"input vector"**, comes from rows of $\mathbf{W}$.
- $\mathbf{v'}_w$: **"output vector"**, comes from columns of $\mathbf{W'}$.

CBOW

Update equation

$\Rightarrow$ Note that, actual computation is very impractical.

Training Objective

$$maximize\ p(w_O|w_I) = y_j = \frac{exp(\mathbf{v'}_{w_j}^T \mathbf{v}_{w_I})}{\sum_{j'=1}^{V} exp(\mathbf{v'}_{w_{j'}}^T \mathbf{v}_{w_I})}$$

- Given the input context W_I with regard to the weight, conditional probability of observing the actual output, Word W_O(=W_{j^*}).
- j^* : index of the actual output word in output layer.

CBOW

Loss Function

$$
\begin{aligned}
max\ p(w_O|w_I) &= max\ y_{j^*} = max\ log\ y_{j^*} \\
&= u_{j^*} - log \sum_{j'=1}^{V} exp(u_{j^*}) := -E \\
\iff E &= -log\ p(w_O|w_I) \\
\therefore max\ p(w_O|w_I) &= min\ E
\end{aligned}
$$

- Loss Function can be understood as a special case of the cross-entropy measurement between two probabilistic distributions.

Reference

- word2vec parameter learning explained
- 쉽게 쓰여진 word2vec
- Word2Vec으로 문장분류하기
- 한국어 Word2Vec
- word2vec 관련 이론 정리
- 한글 데이터 머신러닝 및 word2vec을 이용한 유사도 분석