Word Embeddings

Vocabulary size : N

1. One-hot representation of a word

Dimension : N

Euclidean distance

for any pair of words : $\sqrt{2}$

Cosine similarity : 0

2. Count based methods

Factorization based method : Word2Vec

Construction of count based method with factorization based method

$\Rightarrow$ GloVe

Count based method:

Corpus of text: size = L words

Vocabulary: size = N

$L \gg N$

A measure of word proximity:

Point-wise mutual information (PMI)

$$PMI(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)\, P(\text{word}_2)}$$

$$P(\text{word}_1) = \frac{n_{w_1}}{L}$$

$$P(\text{word}_2) = \frac{n_{w_2}}{L}$$

$$P(\text{word}_1, \text{word}_2) = \frac{n_{w_1 w_2}}{L-1}$$

Context of length k:

Context of length 4:

$$w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$$

$$\uparrow$$

$$\text{word}_t$$

Computation complexity is $O(N^2)$

Result: Cons: Words occurring on the context of a word

$$PMI(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(w_1)\, P(\text{word}_2)}$$

Result: Cons: Words occurring on the context of a word

Take the set of contexts: C^*

C : 10 k : 1

$$\underbrace{w_1\ w_2\ w_3\ w_4\ w_5}_{\text{context}}\ \underbrace{w_6\ w_7\ w_8\ w_9\ w_{10}}_{\text{context}}$$

$$10 = (2 \times 1 \times 5) \times 1$$

$$M = 2 \times L \times (2k \times 5) \times 1$$

Take the set of contexts: W_c

W_c: contexts in which word w_1 occurs

$$|W_c| = M_{w_1}$$

For the set of contexts of which word w_1 occurs, count the number of times word w_2 occurs

$$P(\text{word}_1, \text{word}_2) = \frac{n_{w_2}}{M}$$

$$P(w_2) = \frac{n_{w_2}}{M}$$

$$PMI(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(w_2)\, P(\text{word}_2)}$$

$$= \log_2 \frac{n_{w_1 w_2}}{n_{w_1}\, n_{w_2}}$$

Positive PMI:

$$PPMI(\text{word}_1, \text{word}_2) = \max(0, PMI(\text{word}_1, \text{word}_2))$$

Construct matrix:

$$X_{PPMI} = [\,PMI_{ij}\,]_{N \times N}$$

$$X_{ij} = PPMI(\text{word}_i, \text{word}_j)$$

Singular Value Decomposition (SVD):

$$\underset{\text{sparse, large}}{X_{PPMI}} = \underset{\text{dense, large}}{U}\ \underset{\text{diagonal}}{\Sigma}\ \underset{\text{dense, large}}{V^T}$$

left singular matrix

right singular matrix

$$X_{N \times N} = \underset{N \times d}{U}\ \underset{d \times d}{\Sigma}\ \underset{d \times N}{V^T}$$

U and V are orthonormal

Rank-k approximation of X

$$\underset{N \times N}{\hat{X}_{PPMI}} = \underset{N \times k}{U}\ \underset{k \times k}{\Sigma}\ \underset{k \times N}{V^T}$$

$$\hat{X}\, \hat{X}^T = (U \Sigma V^T)(V \Sigma U^T)$$

$$= U \Sigma\ \underbrace{V^T V}_{I}\ \Sigma U^T$$

$$= \underbrace{U \Sigma}_{N \times k}\ \underbrace{\Sigma^T U^T}_{k \times N}$$

$$\hat{X}\, \hat{X}^T = W W^T$$

$$W_{N \times k}\ W^T_{k \times N}$$

Every row of W corresponds to a word in vocabulary

$$d \ll N$$

Global representation of words

Prediction based method:

MILESTONE for prediction

Context: $w_{t-k} \ldots w_{t-1}\ w_{t+1} \ldots w_{t+k}$

$\uparrow$

word_t

Predict word w_t given its context

$$\text{Prediction word } w_t \text{ given context } w_{t-1}$$

$$\underset{\substack{\text{size}\\ \text{1 hot}}}{x_1} \quad \underset{\substack{\text{size}\\ \text{1 hot}}}{x_2} \quad \underset{\substack{\text{size}\\ \text{1 hot}}}{x_3}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{W_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{W_2} \quad \underset{\substack{\text{size}\\ \text{lower}}}{W_3}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{h_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{h_2} \quad \underset{\substack{\text{size}\\ \text{lower}}}{h_3}$$

Context word

W_1: Input weight connection with context words

W_2: Output weight connection with words in vocabulary

$$\underset{d \times 1}{h_1} = \underset{d \times N}{W_1^T}\ \underset{N \times 1}{x_1} = \underset{d \times 1}{u_1}$$

S : One-hot representation of a word

x_1 : word of vocabulary at S

$$S = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$U_{N \times N} \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} w_{13} \\ w_{23} \\ w_{33} \\ w_{43} \end{bmatrix}$$

$$h = W_1^T x_1$$

$$h = W_1^T S$$

$$h = W_1^T S = U S$$

$$h = U S$$

$$\text{Assume } \eta = 1$$

$$\Delta W_2 = -\eta \frac{\partial E}{\partial W_2}$$

$$= -\eta \frac{\partial}{\partial W_2} \left[-\ln y_j \right]$$

$$= -\eta \frac{\partial}{\partial W_2} \left[-\ln \frac{e^{u_j}}{\sum_{k=1}^{N} e^{u_k}} \right]$$

$$= -\eta \frac{\partial}{\partial W_2} \left[-\ln \left(\frac{e^{W_2^T h}}{\sum_{k=1}^{N} e^{W_k^T h}} \right) \right]$$

$$u_j = W_2^T h_{1 \times N}$$

$$u_{1 \times N} = W_2^T h_{N \times 1}$$

$$= -\eta \frac{\partial}{\partial W_2} \left[-\ln \left(\frac{e^{W_2^T h}}{\sum_{k=1}^{N} e^{W_k^T h}} \right) \right]$$

$$= -\eta \frac{\partial}{\partial W_2} \left[\ln \left(\sum_{k=1}^{N} e^{W_k^T h} \right) - \ln \left(e^{W_2^T h} \right) \right]$$

$$= -\eta \left[\frac{1}{\sum_{k=1}^{N} e^{W_k^T h}} \sum_{k=1}^{N} e^{W_k^T h} \frac{\partial}{\partial W_2} W_k^T h - \frac{\partial}{\partial W_2} W_2^T h \right]$$

$$= -\eta \left[\underbrace{\frac{1}{\sum_{k=1}^{N} e^{W_k^T h}} e^{W_2^T h}}_{y_j} \frac{\partial}{\partial W_2} W_k^T h - \frac{\partial}{\partial W_2} W_2^T h \right]$$

$$= -\eta \left[e - y_j \right] h$$

$$\Delta W_2 = \eta (1 - y_j)\, h$$

$$\Delta W_2 = -\eta \frac{\partial E}{\partial W_2}$$

$$= -\eta \sum_{j=1}^{N} \ln y_j$$

$$= -\eta \sum_{j=1}^{N} \ln \left(\frac{e^{W_j^T h}}{\sum_{k=1}^{N} e^{W_k^T h}} \right)$$

$$= -\eta \left[\sum_{j=1}^{N} \ln \left(e^{W_j^T h} \right) - \ln \left(\sum_{k=1}^{N} e^{W_k^T h} \right) \right]$$

$$\Delta W_2 = -\eta \left[W_j - \sum_{k=1}^{N} y_k W_k \right]$$

$$\text{Take } \frac{\partial}{\partial W_2}$$

Continuous Bag of Words (CBOW)

$$\underbrace{w_{t-2}\ w_{t-1}\ w_{t+1}\ w_{t+2}}_{\text{context}} \ldots \text{Predict } w_t$$

Word2Vec:

Network embedding with continuous Bag-of-words (CBOW)

$$\underset{\substack{\text{size}\\ \text{1 hot}}}{x_1} \quad \underset{\substack{\text{size}\\ \text{1 hot}}}{x_2} \quad \underset{\substack{\text{size}\\ \text{1 hot}}}{x_3}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{W_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{W_2} \quad \underset{\substack{\text{size}\\ \text{lower}}}{W_3}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{h_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{h_2} \quad \underset{\substack{\text{size}\\ \text{lower}}}{h_3}$$

$$U = W_2^T$$

output stream

$$\underset{\substack{\text{size}\\ \text{lower}}}{u_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{u_2} \quad \underset{\substack{\text{size}\\ \text{lower}}}{u_3}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{h} = \frac{U S}{\text{size lower}}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{h_{1 \times N}} = \underset{\substack{\text{size}\\ \text{lower}}}{U_{1 \times N}}\ \underset{\substack{\text{size}\\ \text{lower}}}{S_{N \times 1}}$$

$$h = U S$$

$$h = [0 \ 0 \ 1 \ 0] \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = W_3 - W_4$$

Context of length 4:

$$E = \frac{1}{2k} \sum_{j=1}^{N} \ln y_j \qquad u_j : \text{Value of } j^{th} \text{ unit in the output}$$

$$\Delta W_2 = \eta (1 - y_j) \frac{\partial}{\partial W_2} u_j$$

$$\Delta W_2 = \eta \left(\frac{\partial}{\partial W_2} W_2^T h - y_j \frac{\partial}{\partial W_2} W_2^T h \right)$$

$$\Delta W_2 = \eta \left(W_2 - y_j \frac{\partial}{\partial W_2} W_2^T h \right)$$

$$\Delta W_2 = \eta \left(W_2 - y_j \frac{\partial}{\partial W_2} W_k^T h \right)$$

$$\Delta W_2 = \eta \left(W_2 - \sum_{k=1}^{N} y_k W_k \right)$$

$$\Delta W_2 = \eta \left[W_j - \sum_{k=1}^{N} y_k W_k \right]$$

Word2Vec: Network Embedding with Skip-gram method

$$\text{Predict the context words by the word } w_t$$

$$\underset{\substack{\text{size}\\ \text{1 hot}}}{x} \quad \underset{\substack{\text{size}\\ \text{lower}}}{W_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{h} \quad \underset{\substack{\text{size}\\ \text{lower}}}{U} \quad \underset{\substack{\text{size}\\ \text{lower}}}{u_1} \quad \underset{\substack{\text{size}\\ \text{lower}}}{u_2} \quad \underset{\substack{\text{size}\\ \text{lower}}}{u_3} \qquad U = [0\ 0\ 0\ 1 \ldots]$$

Strategy not

Context word

Context word

$$\underset{\substack{\text{size}\\ \text{lower}}}{h} = \frac{U S}{\text{size lower}}$$

$$\underset{\substack{\text{size}\\ \text{lower}}}{h_{1 \times N}} = \underset{\substack{\text{size}\\ \text{lower}}}{U_{1 \times N}}\ \underset{\substack{\text{size}\\ \text{lower}}}{S_{N \times 1}}$$

$$h = U S$$

$$\Delta W_2 = -\eta \frac{\partial E}{\partial W_2}$$

$$\Delta W_2 = \eta (1 - y_j)\, h$$

$$\Delta W_2 = \eta \left(W_2 - y_j \frac{\partial}{\partial W_2} W_2^T h \right)$$

$$\Delta W_2 = \eta \left(W_2 - y_j \frac{\partial}{\partial W_2} W_k^T h \right)$$

$$\Delta W_2 = \eta \left[W_j - \sum_{k=1}^{N} y_k W_k \right]$$

Skip-gram method using negative sampling:

D : Set of correct pairs w, c

Training

$\bar{D}$: Set of incorrect pairs w, c

$$P(z=1 \mid w, c) = \frac{1}{1 + e^{-W_c^T W_w}} = \sigma(W_c^T W_w)$$

$$P(z=0 \mid w, c) = 1 - \frac{1}{1 + e^{-W_c^T W_w}}$$

$$P(z=1 \mid w, c) = 1 - \frac{1}{1 + e^{-W_c^T W_w}} = \frac{1 + e^{-W_c^T W_w} - 1}{1 + e^{-W_c^T W_w}} = \frac{1}{1 + e^{W_c^T W_w}} = \sigma(-W_c^T W_w)$$

$$\theta = \{ W_w, W_c \}_{w \in V}$$

$$\underset{\theta}{\text{Maximize}}\ \prod_{(w,c) \in D} \sigma(W_c^T W_w) \cdot \prod_{(w,c) \in \bar{D}} \sigma(-W_c^T W_w)$$

Hyperparameters:

- Context size: k
- Dimension of W_w, W_c : d
- M

For a word W_w, consider M random of (W_c, w) pairs

M is a hyperparameter

$M = 5, 10$

Global Vectors (GloVe) method:

Count based method and Prediction based method

Co-occurrence matrix: $X = [X_{ij}]_{N \times N}$

Cost : Want to minimize J_θ the cost of a word

$$P(\text{Cons} \mid \text{Words}) = \frac{P(\text{Cons}, \text{Words})}{P(\text{Words})} = \frac{X_{ij}}{X_i}$$

$$\log P(\text{Cons} \mid \text{Words}) = \log \frac{X_{ij}}{X_i}$$

$$\log X_{ij} - \log X_i = \log \frac{X_{ij}}{X_i}$$

$$\log X_{ij} - \log X_i = \log \frac{X_{ij}}{X_i}$$

$$\log X_{ij} - \log X_i = \log X_{ij} - \log X_i - \log X_j$$

$$\underset{\theta}{\text{Minimize}}\ \left(\log X_{ij} - \log X_i - \log X_j \right)^2$$

Obtained from Co-occurrence matrix