


# DiDL(Word2Vec)

☀ 상태	Done
👤 작성자	 지현 김
🕒 최종 편집 일시	@2023년 11월 28일 오후 7:04
☰ 태그	

## 15.1. Word Embedding (word2vec)

- 단어가 의미를 가지듯이, word vector는 단어를 표현하기 위해 사용되는 feature 벡터
- Word Embedding : word -> vector
- NLP(Natural Language Processing)의 기본

### 15.1.1. One-Hot Vectors Are a Bad Choice

- RNN 때는 One-Hot vector를 사용해서 단어를 표현
- 그러나 단어가 N개 있으면, 인덱스도 N개 있어야 하고, 벡터표현도 매우 어려움  
(0,0,...,1,...0) -> 매우 비효율적
- 또한 one hot vector는 Cosine similarity를 표현할 수 없음(내적이 0)

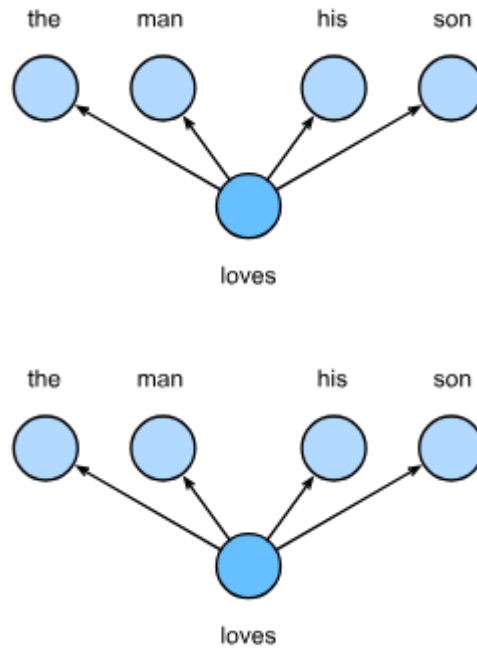
### 15.1.2. Self-Supervised word2vec

- 이 문제를 해결하기 위해 등장한 tool -> word2vec
- 각 단어에 fixed-length vector를 mapping -> similarity와 analogy relationship(비유 관계) 더 잘 파악함
- *skip-gram*, *continuous bag of words (CBOW)* 라는 두 모델을 포함함.
- Word2Vec 모델에서 의미 있는 표현을 얻기 위한 훈련은 조건부 확률에 의존하는데, 이는 말뭉치에서 단어들의 주변 단어를 사용하여 특정 단어를 예측하는 것으로 해석할 수 있습니다. 데이터에 레이블이 없이 수행되기 때문에 skip-gram과 continuous bag of words 모두 self-supervised models(라벨링 없이 input data에 대해 학습하는 머신 러닝 모델)로 간주됨.

### 15.1.3. The Skip-Gram Model

- Skip-gram model은 단어(center word)가 단어를 둘러싸고 있는 text sequence(context words)를 만들어 낼 수 있다고 가정함.

- "The man loves his son"의 예시.



- "loves"를 center word로 고르고 context window size를 2로 잡으면, skip-gram 모델은 center word인 "loves"를 기준으로 거리가 2 이내인 context words인 "the", "man", "his", "son"을 생성할 조건부 확률을 구하게 된다.

$$P(\text{"the"}, \text{"man"}, \text{"his"}, \text{"son"} \mid \text{"loves"}).$$

$$P(\text{"the"}, \text{"man"}, \text{"his"}, \text{"son"} \mid \text{"loves"}).$$

- 여기서 각 context words가 조건부 독립이라면 ("loves"에 의해 독립적으로 각각 생성된 context word라면) 확률은 다음과 같이 재정의된다.

$$P(\text{"the"} \mid \text{"loves"}) \cdot P(\text{"man"} \mid \text{"loves"}) \cdot P(\text{"his"} \mid \text{"loves"}) \cdot P(\text{"son"} \mid \text{"loves"}). \quad (15.1.3)$$

$$P(\text{"the"} \mid \text{"loves"}) \cdot P(\text{"man"} \mid \text{"loves"}) \cdot P(\text{"his"} \mid \text{"loves"}) \cdot P(\text{"son"} \mid \text{"loves"}). \quad (15.1.3)$$

- Skip-gram 모델에서, 각 단어는 조건부확률을 계산하기 위해 2개의 d차원 벡터 표현을 가진다. (ith word has  $v(i)$  -> center word,  $u(i)$  -> context word)
- 특정 Center word  $w_c$ 에 대해 context word  $w_o$ 를 생성할 조건부 확률은 softmax를 거쳐서 다음과 같다.

$$P(w_o | w_c) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)}, \quad ($$

$$P(w_o | w_c) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)}, \quad ($$

- word set  $V = \{0, 1, \dots, V-1\}$
- Text sequence 길이 =  $T$
- word at time step =  $t$ ( $t$ 번째 단어)
- context windows size =  $m$

라고 할 때, skip-gram 모델의 확률 함수란 어떠한 단어를 center word로 잡더라도 모든 context words를 생성할 수 있는 확률과 같다.

$$\prod_{t=1}^T \prod_{-m \leq j \leq m, j \neq 0} P(w^{(t+j)} | w^{(t)}),$$

### 15.1.3.1. Training

- Skip-gram 모델의 parameter = center word / 각 단어의 context word
- Training에서는 위에서 보았던 확률함수를 최대화하는 (= 다음의 확률 함수를 최소화하는) 모델 파라미터를 이끌어 낼 것이다,

$$-\sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} \log P(w^{(t+j)} | w^{(t)}).$$

- SGD를 이용하여 loss를 최소화하면, 각 반복마다 더 짧은 subsequence를 샘플링하여 gradient를 계산하여 모델 파라미터를 업데이트 할 수 있음.

- 이를 계산하기 위해서 center word vector와 context word vector에 대한 로그 조건부 확률의 gradient를 얻어야 함.

$$\log P(w_o | w_c) = \mathbf{u}_o^\top \mathbf{v}_c - \log \left( \sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c) \right).$$

$$\begin{aligned} \frac{\partial \log P(w_o | w_c)}{\partial \mathbf{v}_c} &= \mathbf{u}_o - \frac{\sum_{j \in \mathcal{V}} \exp(\mathbf{u}_j^\top \mathbf{v}_c) \mathbf{u}_j}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)} \\ &= \mathbf{u}_o - \sum_{j \in \mathcal{V}} \left( \frac{\exp(\mathbf{u}_j^\top \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)} \right) \mathbf{u}_j \\ &= \mathbf{u}_o - \sum_{j \in \mathcal{V}} P(w_j | w_c) \mathbf{u}_j. \end{aligned}$$

- 해당 계산은 wc를 중심으로 가진 단어들의 모든 조건부 확률을 필요로 함.
- 훈련 후에는 모든 word set 안의 단어가 v(i), u(i)를 얻게 됨