Word Embedding

강사 : 백병인

pi.paek@modulabs.co.kr

모두의연구소 Research Scientist

Language Representation to Neural Network

- 컴퓨터는 0과 1의 숫자밖에 모른다.

- 글자에 대응하는 숫자를 부여하면?
  - A = 1 = $01_{(2)}$
  - B = 2 = $10_{(2)}$
  - C = 3 = $11_{(2)}$

- A + B = C 의 관계가 있을까?
- A와 B의 거리는 A와 Z의 거리보다 가까울까?

One-Hot Encoding

- symbol을 vectorize하는 가장 단순한 방법
- 사전의 전체 단어 개수(N) 만큼의 차원을 가지는 vector로 표현
  - V_{hotel} = [0 0 0 0 0 0 1 0 0 0 0 0 0 0 0]
  - V_{motel} = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
  - V_{car} = [0 0 1 0 0 0 0 0 0 0 0 0 0 0 0]

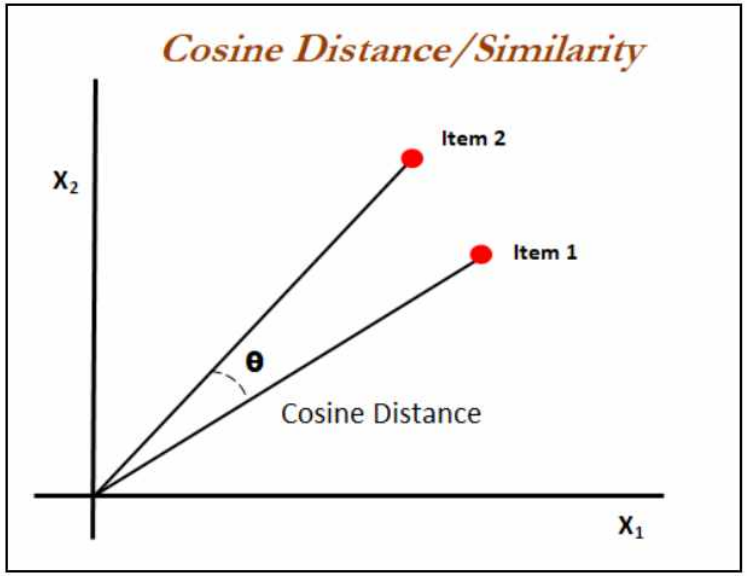
- 문제점
  - 너무 길다
  - 단어간 유사도가 전혀 표현되지 않는다.

Cosine Similarity

- 좋은 Vector Representation이 가져야 할 조건
  - 만약 language symbol의 vector representation을 만든다면,
  - 단어들 사이의 유사도를 정량적으로 표현할 수 있어야 한다.

- Cosine Similarity
  - 이 값이 클수록, 두 벡터는 유사한 개념을 표현한다고 보는 관점

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$



https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml

Word2Vec (Mikolov et.al. 2013)

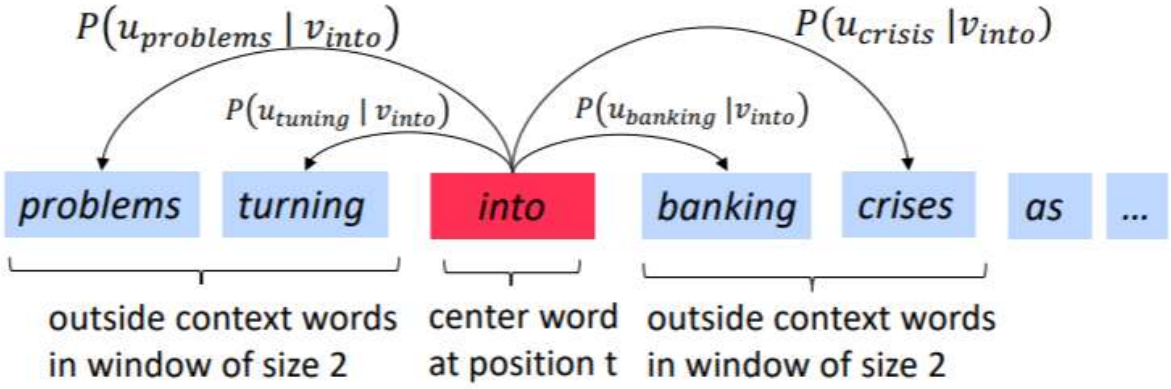
- 어떻게 word를 의미공간에 vector로 매핑할 수 있을까?
- 과연 word의 의미는 어떻게 발견되는 것일까?
- word(symbol)의 의미는 word 안에 없다.
- word가 사용되는 맥락(context)이 word의 의미를 규정한다.

- Distributional semantics
  - A word's meaning is given by the words that frequently appear close-by.
  - "You shall know a word by the company it keeps" (J. R. Firth 1957: 11)
  - One of the most successful ideas of modern statistical NLP!

(참고) http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf

Main Idea of Word2Vec (1)

- **Co-Occurence** Means Similarity of Meaning!!
- 한 문장 안에 함께 나타난 두 단어 사이에는 의미적 유사도가 있다는 가정



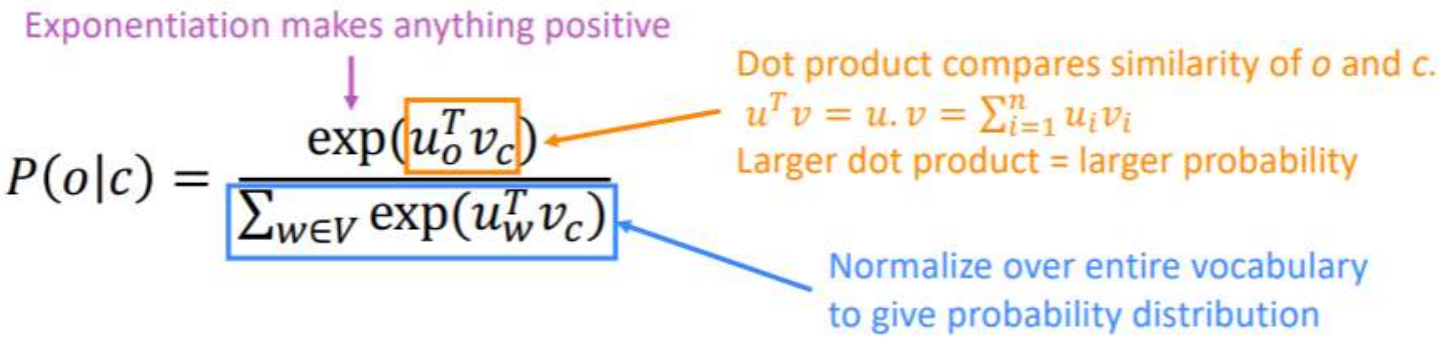
Word Vector Dictionary

$$\theta = \begin{bmatrix} v_{aardvark} \\ v_a \\ \vdots \\ v_{zebra} \\ u_{aardvark} \\ u_a \\ \vdots \\ u_{zebra} \end{bmatrix} \in \mathbb{R}^{2dV}$$

(참고) http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf

Main Idea of Word2Vec (2)

- 그렇다면, 함께 출현한 두 단어 사이의 cosine similarity를 높여 주는 쪽으로 딥러닝을 구현해 보면 어떨까?

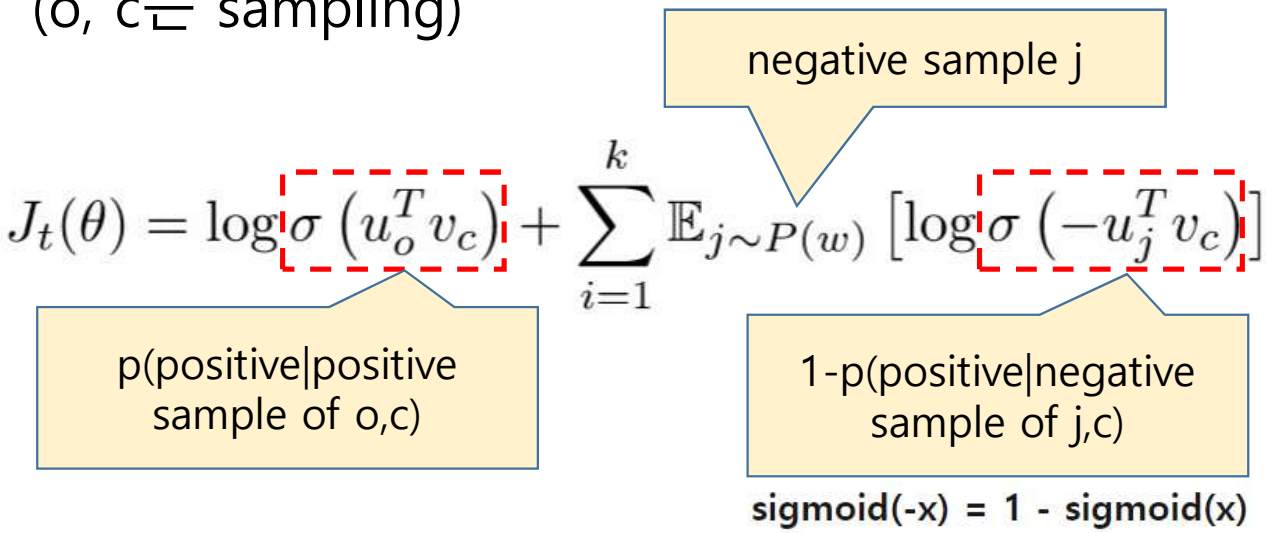


- 위 P(o|c)가 corpus 데이터를 가장 잘 반영하도록 Log Likelihood를 극대화해 보자!
- 그런데, vocabulary 전체에 대해 normalize를 한다면 계산량이 너무 많지 않을까?

(참고) http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf

Negative Sampling

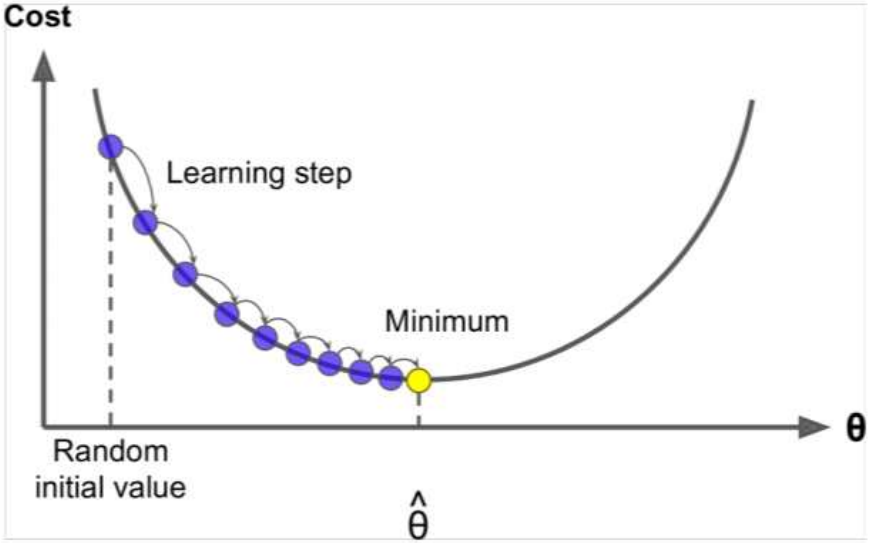
- ***모든 단어에 대해서 계산하지 말고 실제로 불가능한 예제(네거티브 샘플)에 대해서만 손실함수를 계산하자***
  - p(o|c) 확률분포를 근사하는 multi-classification 문제를
  - p(positive or negative|o,c) 를 근사하는 binary classification 문제로 바꾼다. (o, c는 sampling)



https://www.linkedin.com/pulse/heavy-softmax-use-nce-loss-shamane-siriwardhana/

Optimization with SGD

- 엄청나게 많은 데이터. 그러므로 샘플링을 통한 SGD로 학습.



$$\theta^{new} = \theta^{old} - \alpha \nabla_\theta J(\theta)$$

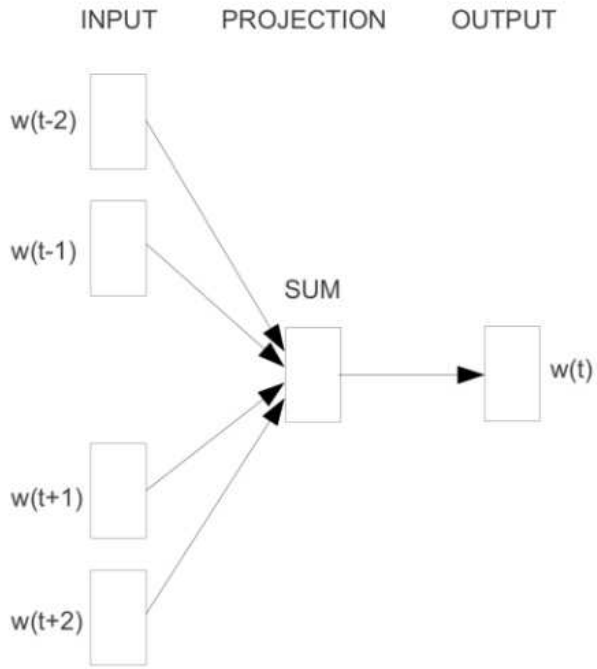
α = *step size* or *learning rate*

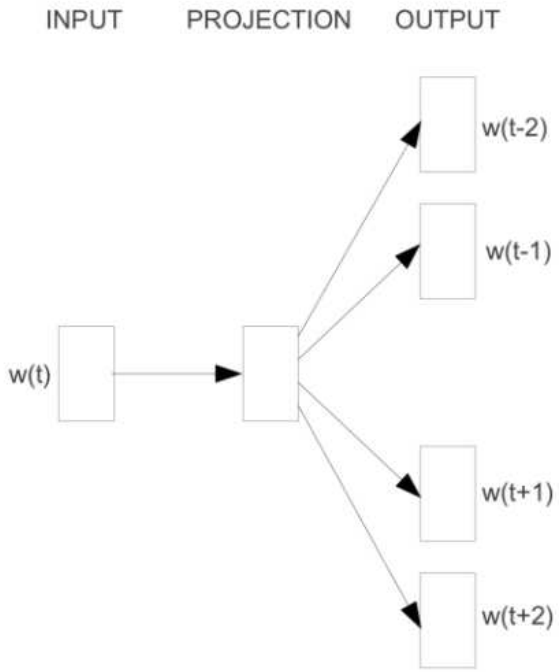
Algorithm:

```
while True:
    window = sample_window(corpus)
    theta_grad = evaluate_gradient(J,window,theta)
    theta = theta - alpha * theta_grad
```

(참고) http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf

CBoW & Skip-gram





CBow(Continuous Bag-of-words)

* 다음 빈칸에 알맞은 말은?

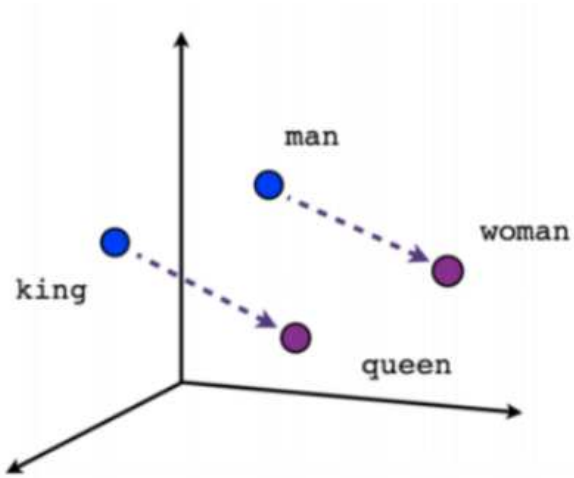
I go to _____ to study.

Skip-Gram

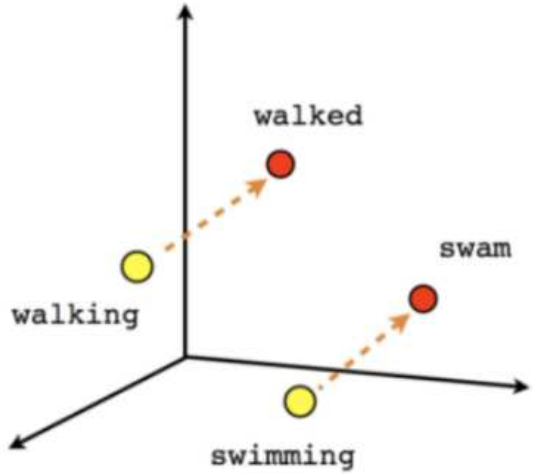
* 다음 단어로 짧은 글을 지어 보세요.

____ _____ school ____ _____

Word Vector can do Vector Sum.



Male-Female



Verb tense

Visualize Word Vectors (https://www.tensorflow.org/images/linear-relationships.png)

$$V_{king} - V_{man} + V_{woman} = V_{queen}$$

https://towardsdatascience.com/word-embedding-with-word2vec-and-fasttext-a209c1d3e12c

Semantic Distribution of Word Vector

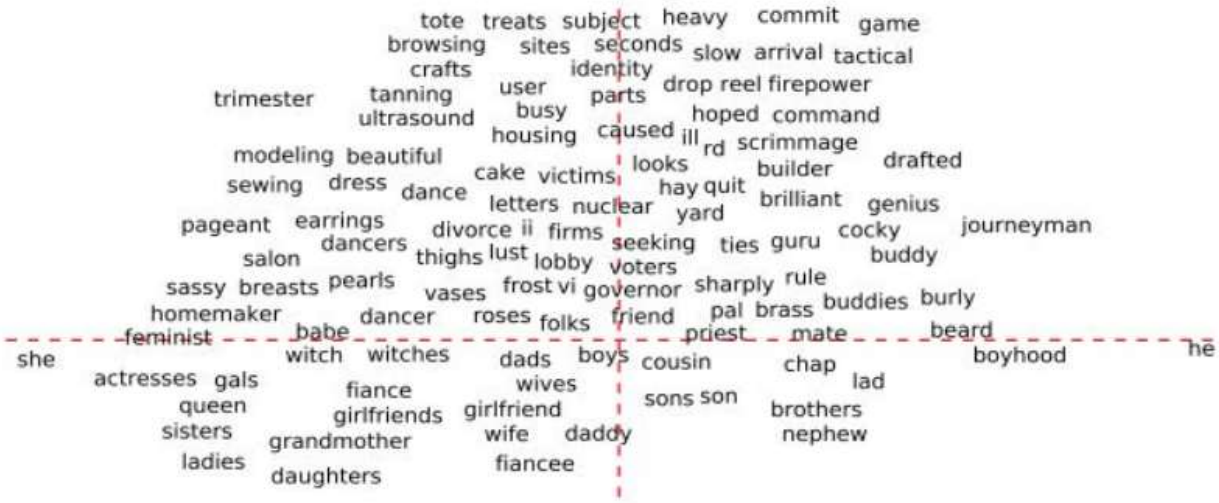


Figure 7: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

Any bias?? Why??

[Bolukbasi et.al. 2016] Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
https://arxiv.org/pdf/1607.06520.pdf