워드 임베딩

단어를 벡터로 표현

ex) "cat"과 "kitty"는 유사한 단어이므로 비슷한 벡터로 표현 · 짧은 거리
"hamburger"는 "cat", "kitty"와 유사하지 않으므로 다른 벡터로 표현 먼 거리

이런 식으로 공간에 좌표값을 부여

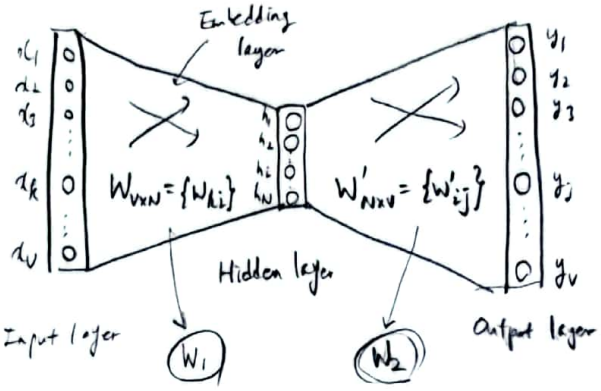
Word2Vec

같은 문장에서 나타난 인접한 단어들 간의 의미가 비슷할 것이라는 가정

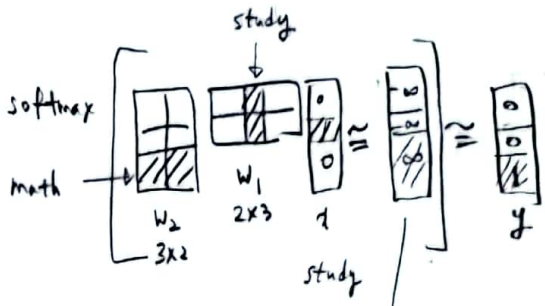
ex) (The cat purrs
This cat hunts mice

Idea: 한 단어가 주변에 등장하는 단어를 통해 의미를 알 수 있다
주어진 학습 데이터를 바탕으로 특정 단어 주변에 나타날 단어들의 확률 분포 예측

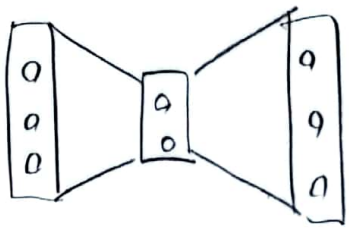
학습 방식:

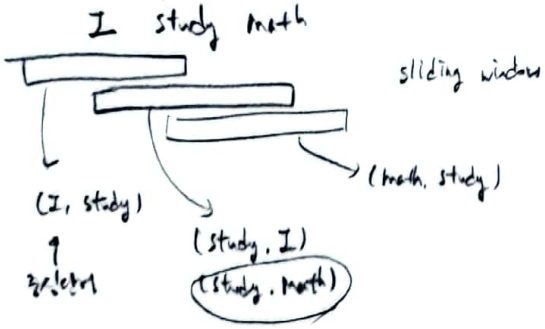


$$y = \text{softmax}(W_2 W_1 x)$$



ex) Sentence: "I study math"
Vocabulary: {"I", "study", "math"}
Input "study" [0,1,0] one-hot
Output: "math" [0,0,1]

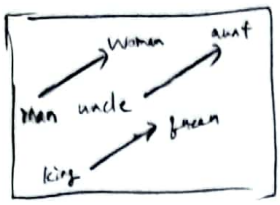


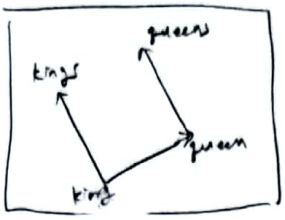


Input의 W_1 행렬의 벡터
Output의 W_2 행렬의 벡터의 내적이 최대한 커지도록 내적하는 작업이도록
(벡터가 유사)

학습된 Word2Vec은 단어들 간의 의미론적 관계를 벡터 연산에 적용할 수 있는 방식

같은 관계는 같은 벡터로 표현





ex) vec[queen] - vec[king] = vec[woman] - vec[man]

ex) vec[서울] - vec[한국] + vec[미국] = vec[워싱턴DC]

Word intrusion detection.

→ 단어들이 주어졌을 때 나머지 단어와 의미가 가장 상이한 단어는 무엇인가?

ex) staple hammer saw drill

유클리드 거리 계산 후 평균 거리가 가장 큰 것을 선택

Word2vec은 NLP의 다수 영역에서 성능 향상

- 단어 유사도
- 개체 번역
- part-of-speech (pos) 태깅
- Named entity recognition (NER, 고유명사 인식)
- 감정 분석
- 관계망
- semantic lexicon building ...

GloVe

Word2Vec과 같이 사전에 각 단어들이 함께 등장한 빈도수를 계산하여
단어 임베딩 벡터의 내적 값과 사전에 계산된 값의 차이를 줄이는 방법으로 학습

$$J(\theta) = \frac{1}{2}\sum_{i,j=1}^{N} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

학습 속도가 빠르고 작은 corpus에서도 잘 동작
(중복 계산을 줄임!)

f~