N422

단어를 벡터화하는 법.

One-hot Encoding
- 유사도 구할수 X

Embedding
- 단어를 고정 길이의 벡터로 (차원이 일정한 벡터)
- 연속적인 값 나옴. ex) hello, JH
one-hot) [0,1] [1,0]
embedding) [0.33] [0.21]

- **단어의 분산 표현(Distributed Representation)**
  - 원-핫 인코딩의 개념과 단점에 대해서 이해할 수 있습니다.
  - 분포 기반의 표현, 임베딩이 무엇인지 설명할 수 있습니다.
- **Word2Vec**
  - CBoW와 Skip-gram의 차이에 대해서 설명할 수 있습니다.
  - Word2Vec의 임베딩 벡터를 시각화한 결과가 어떤 특징을 가지는지 설명할 수 있습니다.
- **fastText**
  - OOV 문제가 무엇인지에 대해 설명할 수 있습니다.
  - 철자(Character) 단위 임베딩 방법의 장점에 대해 설명할 수 있습니다.

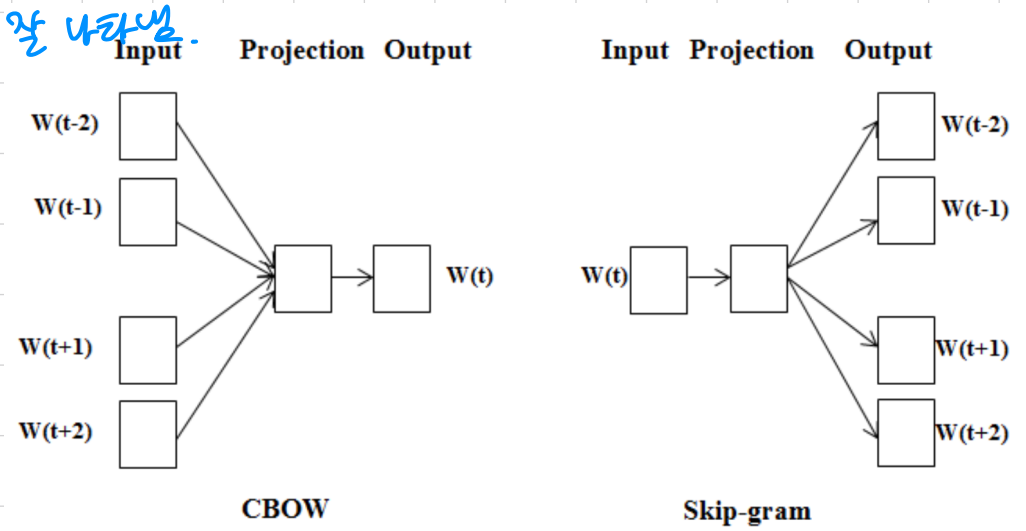
② 분포기반

1) Word2Vec ∴ 의미적, 문법적 관계 잘 나타냄.

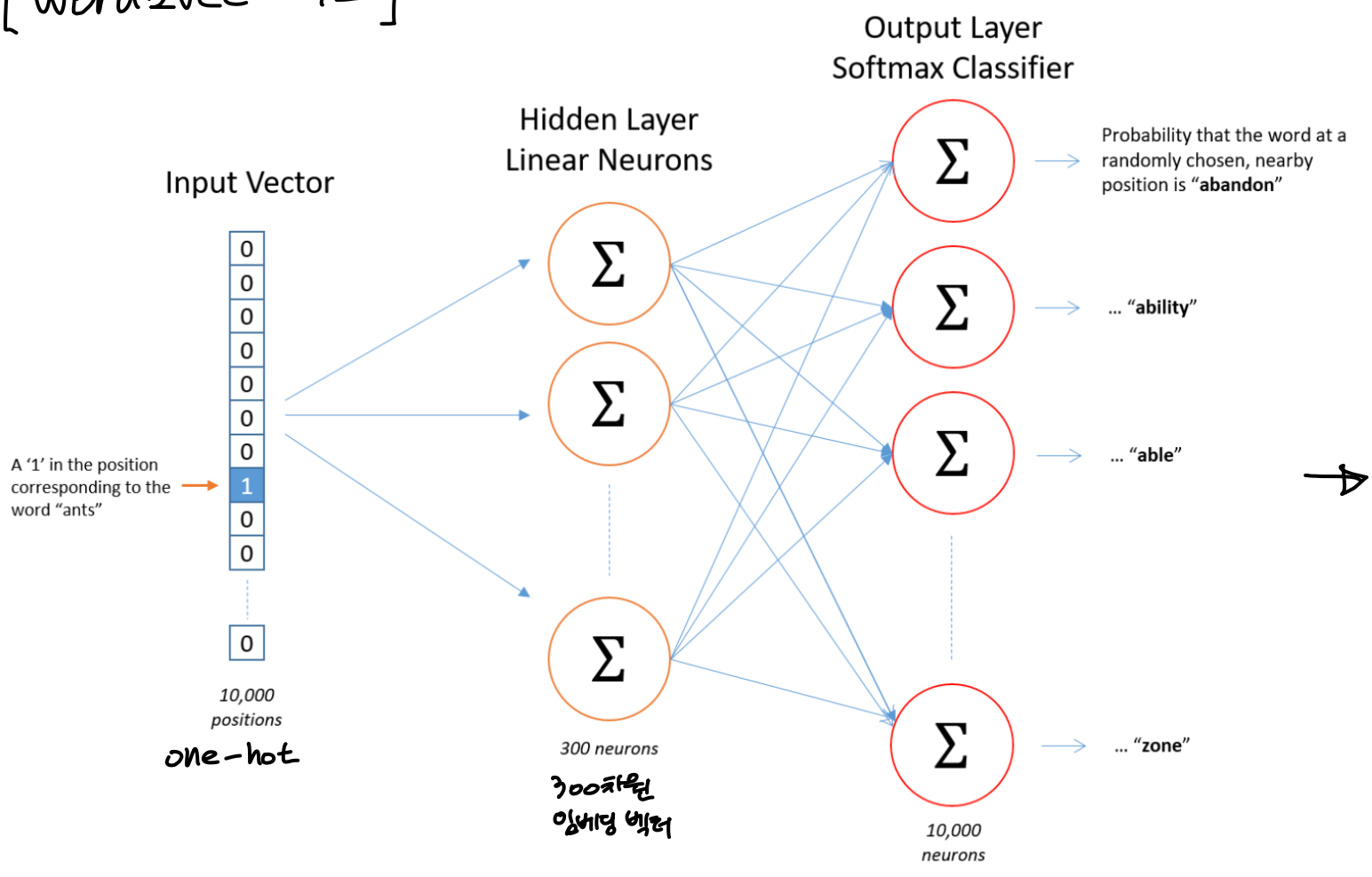
- 양옆에 있는 두 단어의 관계 활용
- 분포가설 (유유상종..) 잘 반영

a. CBoW (주변 → 하나)

b. skip-gram (하나 → 주변)

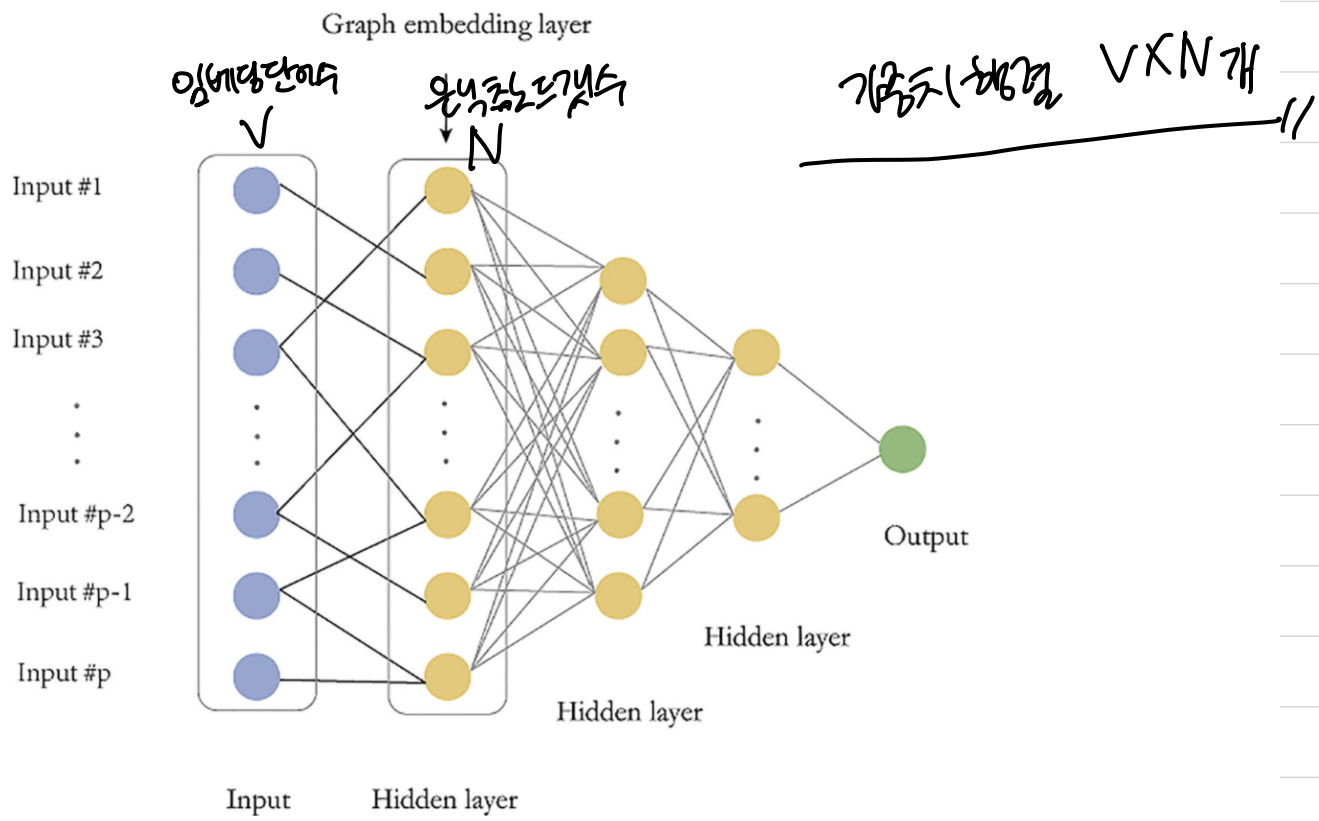


[word2vec 구조]



● Embedding Methods

NLP task를 수행하기 전, 단어를 벡터로 만드는 임베딩 작업을 케라스를 이용해서 하는 방법은 크게 두 가지가 있습니다.



- 케라스의 내장 함수인 Embedding()을 사용하기
- Pre-trained word embedding 가져와서 Embedding Layer에 주입하기

Reference

https://ebbnflow.tistory.com/154

https://www.youtube.com/watch?v=sY4YyacSsLc

가중치 행렬이란?

그럼 입력층과 은닉층을 잇는 가중치행렬 W을 좀 더 자세히 살펴보겠습니다. 위 그림과 아래 그림을 비교하면서 보시면 좋을 것 같은데요, V는 임베딩하려는 단어의 수, N은 은닉층의 노드 개수(사용자 지정)입니다. Word2Vec은 최초 입력으로 **one-hot-vector**를 받는데요, $1 \times V$ 크기의 one-hot-vector의 각 요소와 은닉층의 N개 각 노드는 1대1 대응이 이뤄져야 하므로 가중치행렬 W의 크기는 $V \times N$이 됩니다. 예컨대 학습 말뭉치에 단어가 1만개 있고 은닉층 노드를 300개로 지정했다고 칩시다. 그럼 가중치행렬 W는 좌측 하단의 오렌지색 행렬 형태가 됩니다.