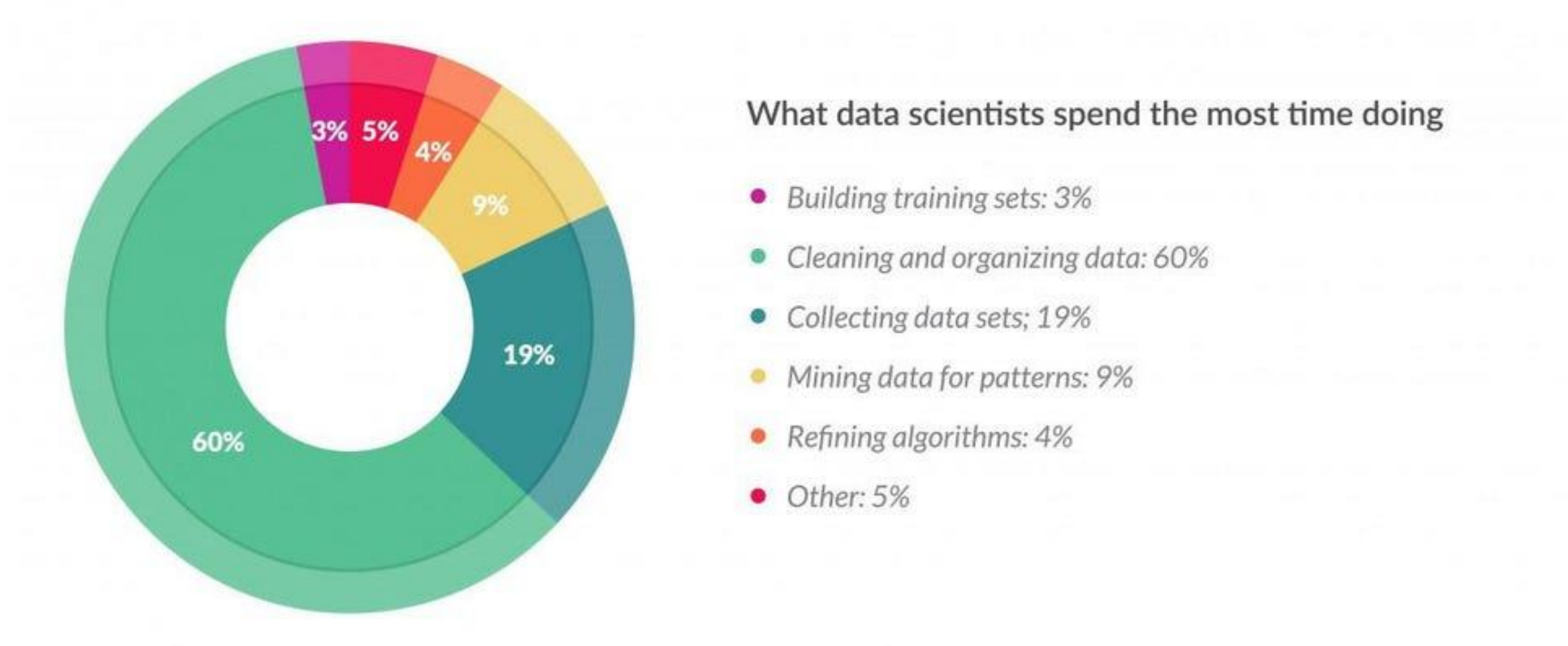
Word Embedding

Machine Learning Study
JinHo Kim

Contents

1. Introduction

Intro.



https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=5f1eb7286f63

Word Representation

1. Discrete Representation (Local Representation)

   1) One - hot Vector

      - One - hot Vector

   2) Count Based

      - Bag of Words (BoW)
      - Document-Term Matrix (DTM) or TDM
      - TF - IDF
      - N-gram

2. Continuous Representation

   1) Prediction Based (Distributed Representation)

      - Neural Network Language Model (NNLM)
      - Word2Vec
      - FastText
      - Embedding from Language Model (ELMo)

   2) CountBased (Full Document)

      - Latent Semantic Analysis (LSA))

   3) Prediction Based and CountBased (Windows)

      - GloVe

Word Representation

1. Discrete Representation (Local Representation)
   1) One - hot Vector
      - One - hot Vector
   2) Count Based
      - Bag of Words (BoW)
      - Document-Term Matrix (DTM) or TDM
      - TF - IDF
      - N-gram

2. Continuous Representation
   1) Prediction Based (Distributed Representation)
      - Neural Network Language Model
      - Word2Vec
      - FastText
      - Embedding from Language Model (ELMo)
   2) CountBased (Full Document)
      - Latent Semantic Analysis (LSA))
   3) Prediction Based and CountBased (Windows)
      - GloVe

One Hot Encoding

“ I ate an apple and played the piano”

	1	2	3	4	5	6	7	8
I	1	0	0	0	0	0	0	0
ate	0	1	0	0	0	0	0	0
an	0	0	1	0	0	0	0	0
apple	0	0	0	1	0	0	0	0
and	0	0	0	0	1	0	0	0
played	0	0	0	0	0	1	0	0
the	0	0	0	0	0	0	1	0
piano	0	0	0	0	0	0	0	1

https://towardsdatascience.com/word-embedding-in-nlp-one-hot-encoding-and-skip-gram-neural-network-81b424da58f2

One Hot Encoding Problem

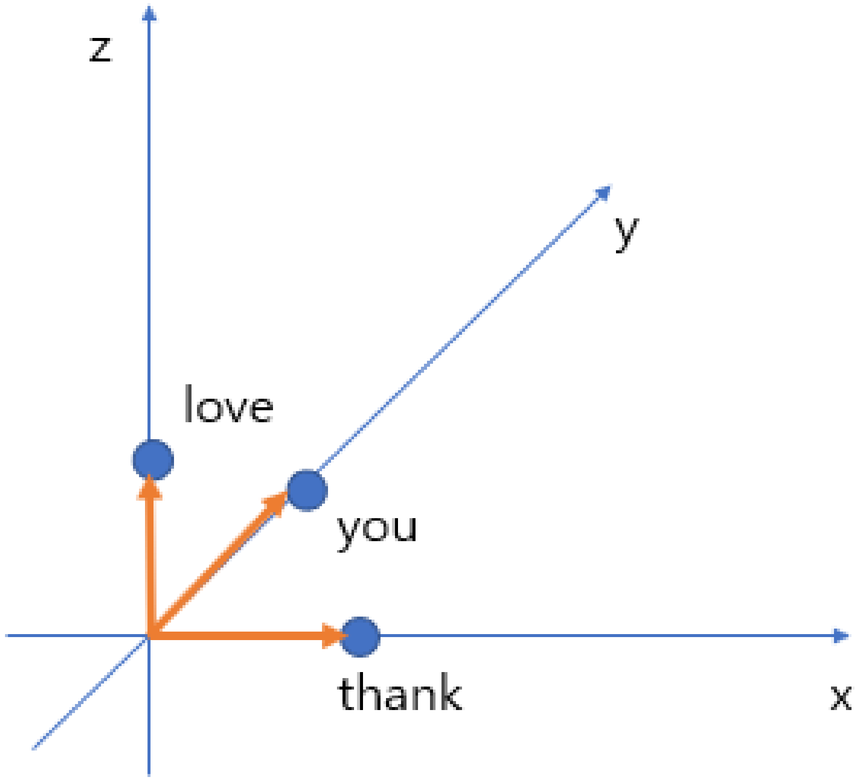
1. It is inefficient to use too many memory spae

2. Every distance is same to each other

3. Cosine similarity also 0 since angle is 90 degree

https://www.youtube.com/watch?v=s2KePv-OxZM&list=PLetSlH8YjIfVzHuSXtG4jAC2zbEAErXWm&index=10

One Hot Encoding Problem



http://ai4school.org/?page_id=2407

One-hot VS Embedding

-	One-hot Vector	Embedding Vector
차원	고차원	저차원
표현	희소 벡터	밀집 벡터
학습 방법	수동	훈련 데이터로 학습
타입	binary(1, 0)	실수

One-hot vector

[0, 0, 0, …, 1, …, 0, 0, 0]

'Sparse'

Embedding vector

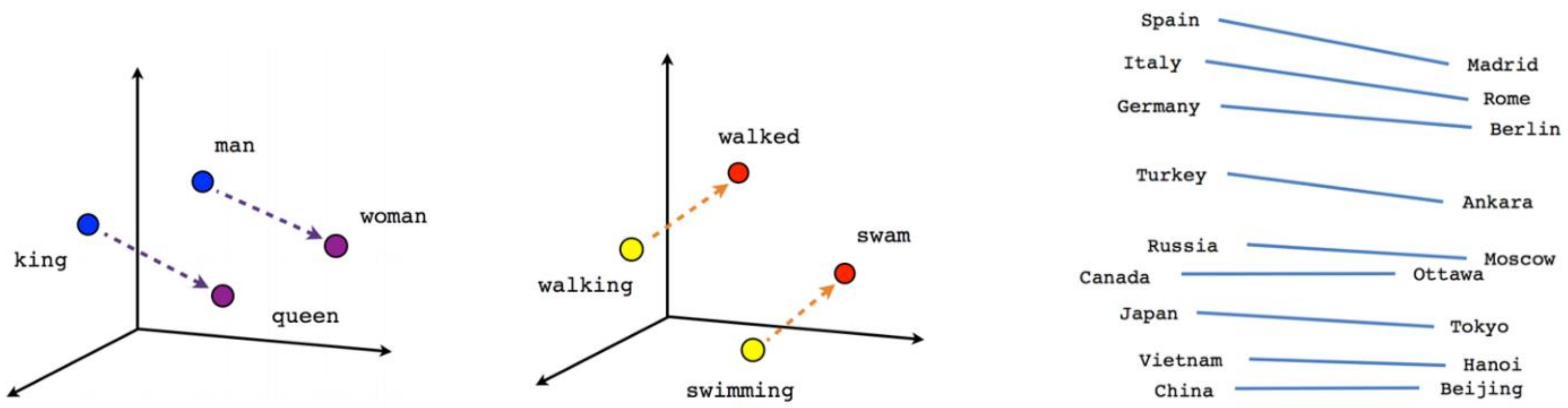
[0.55, 0.6, 0.7,…, 0.21]

'Dense or Distributed'

https://wikidocs.net/33520

2. Word2Vec

Distributed Representation

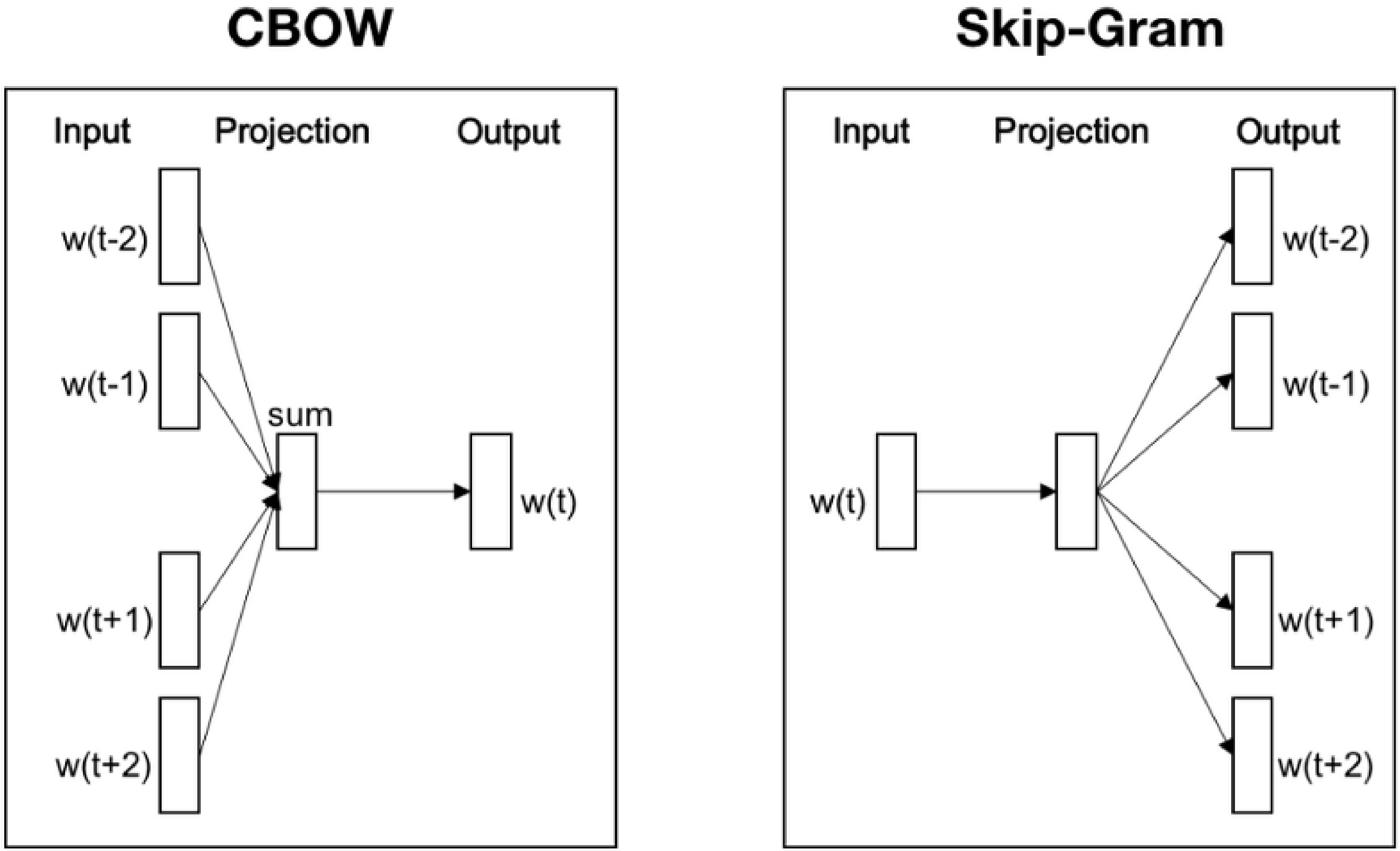


Male-Female   Verb tense   Country-Capital

단어 간 similarity 계산 가능!

https://towardsdatascience.com/creating-word-embeddings-coding-the-word2vec-algorithm-in-python-using-deep-learning-b337d0ba17a8

CBOW VS Skip-Gram



https://wooono.tistory.com/244

CBOW (window = 2)

중심 단어
주변 단어

The fat cat sat on the mat

The fat cat sat on the mat

The fat cat sat on the mat

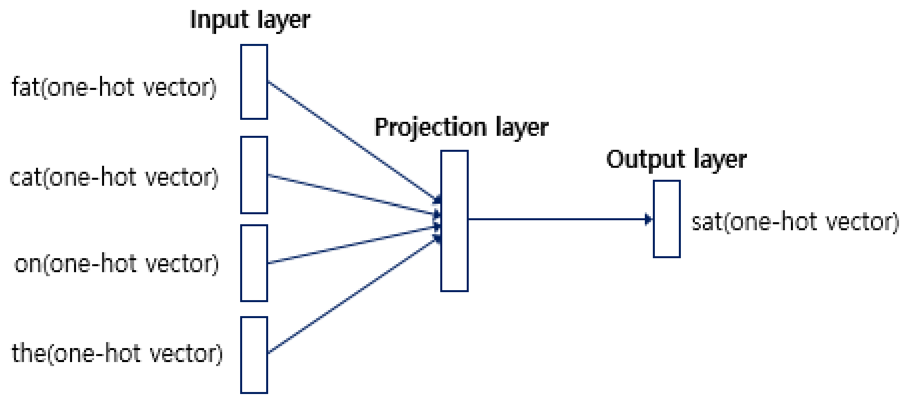
The fat cat sat on the mat

The fat cat sat on the mat

The fat cat sat on the mat

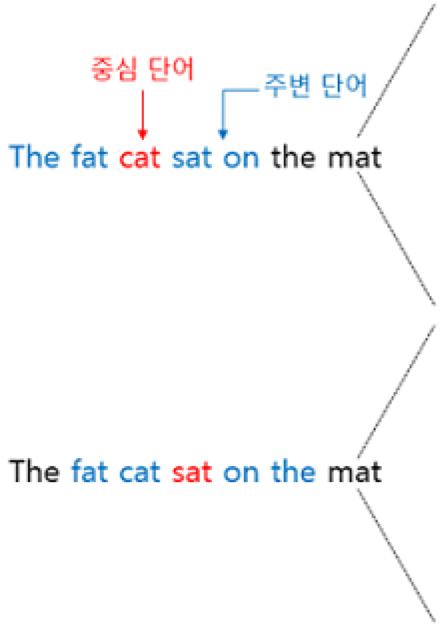
The fat cat sat on the mat

중심 단어	주변 단어
[1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 1, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 1, 0, 0]	[0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0]

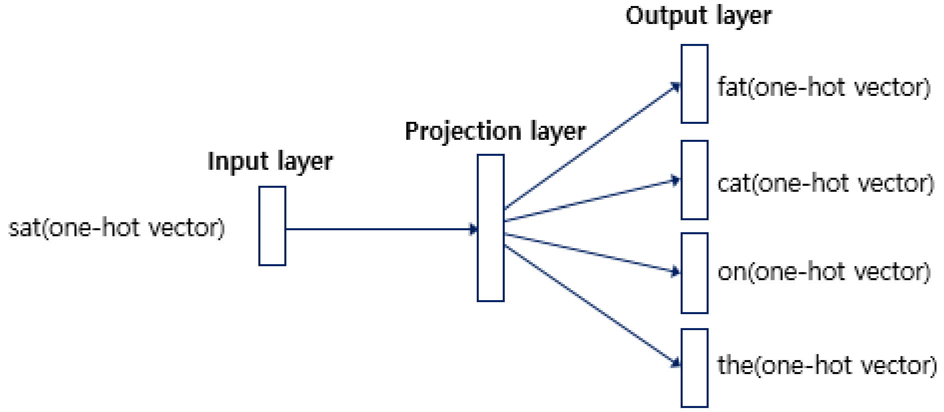


https://wikidocs.net/22660

Skip-gram (window = 2)

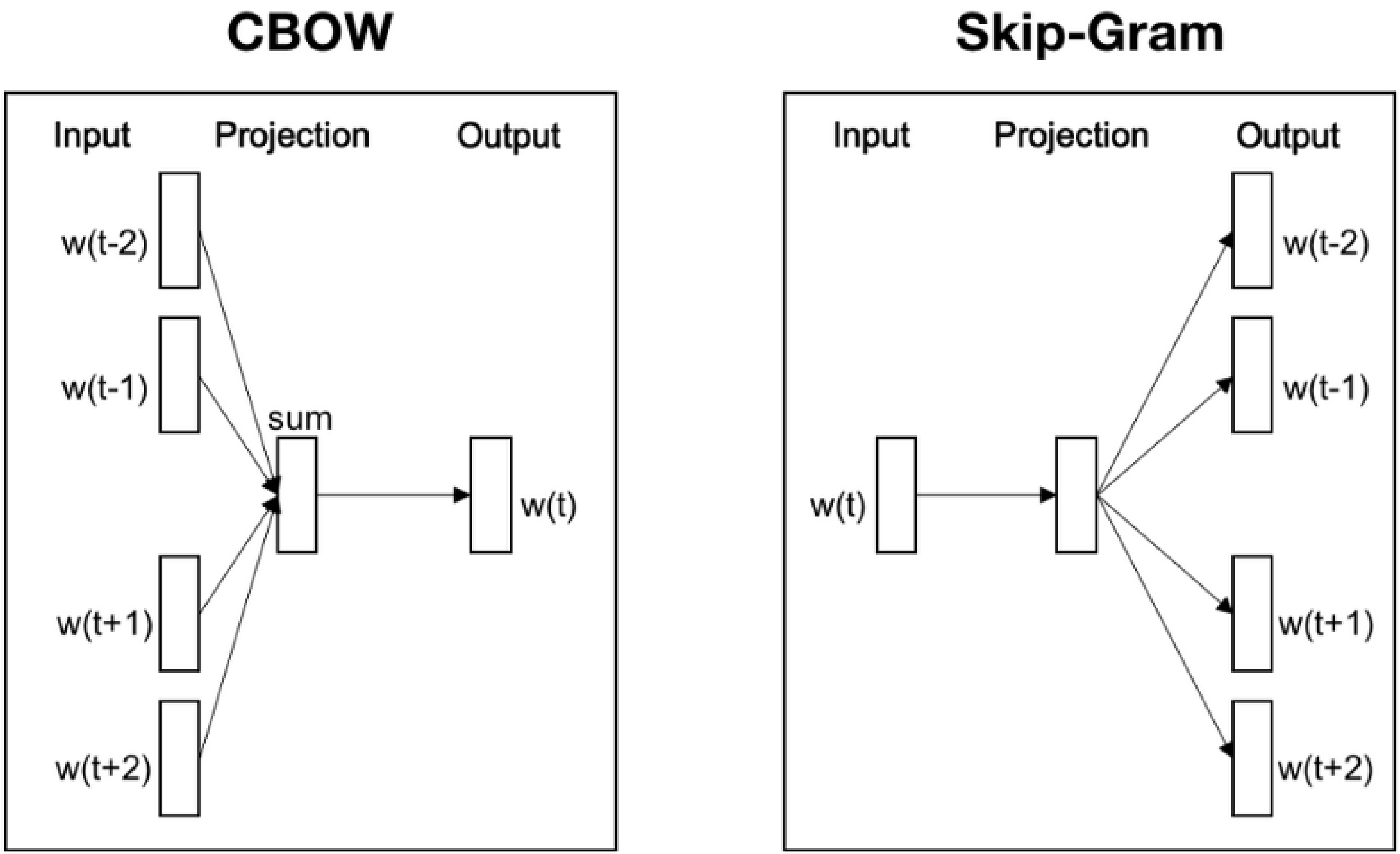


중심 단어	주변 단어
cat	The
cat	Fat
cat	sat
cat	on
sat	fat
sat	cat
sat	on
sat	the



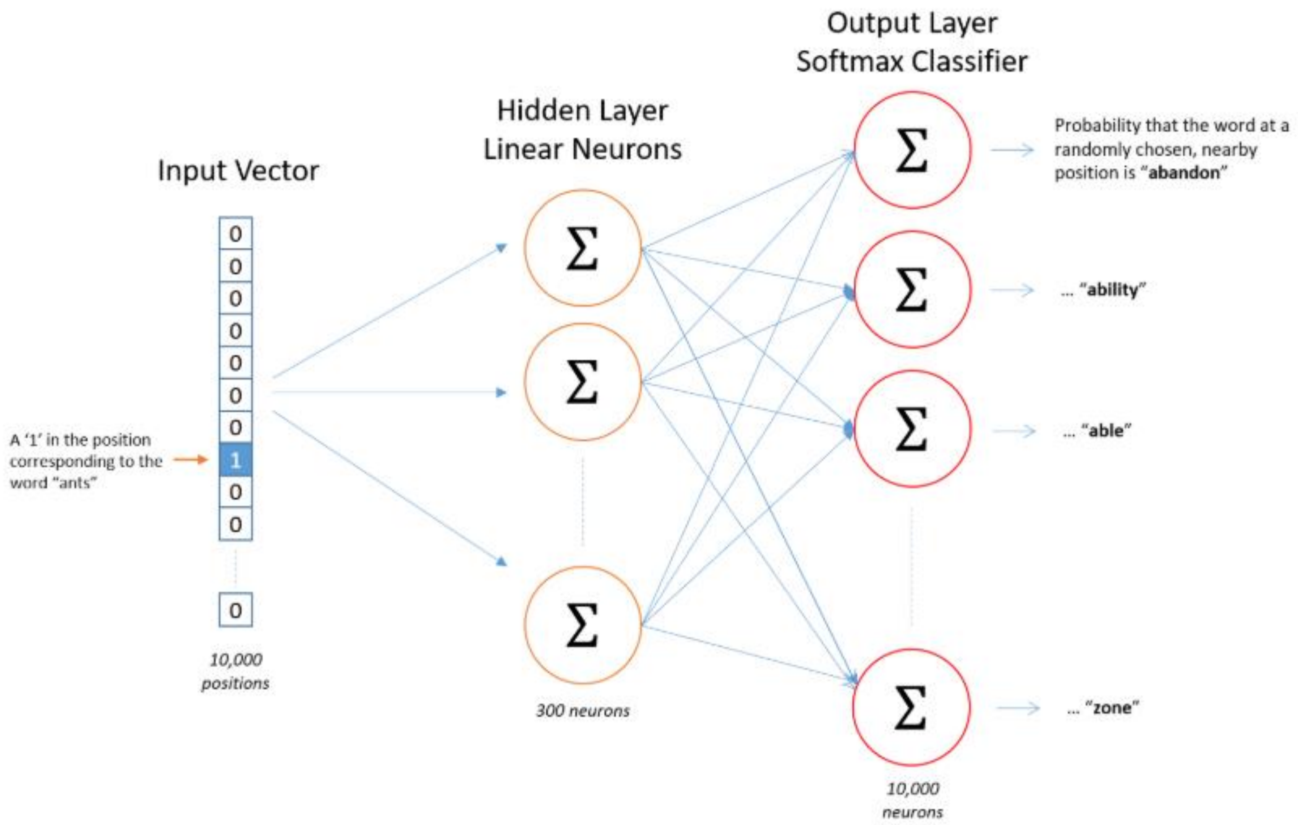
https://wikidocs.net/22660

CBOW VS Skip-Gram



https://wooono.tistory.com/244

Negatvie Sampling



ex) 'The fat cat sat on the mat' 기준 단어와 문맥 단어와 전혀 상관없는 모든 사전 크기 만큼 다시 학습

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Negatvie Sampling

ex) 'The fat cat sat on the mat' 기준 단어와 문맥 단어와 전혀 상관없는 모든 사전 크기 만큼 다시 학습

Embedding 크기를 전체 단어 집합이 아닌, **일부 단어집합**으로 조정

기준 단어 주변에 등장한 단어

일부 단어 집합 = Positive sample + Negative sample

기준 단어 주변에 등장하지 않은 단어

$$P(\omega_i)_n = \left(\frac{f(\omega_i)}{\sum_{j=1}^{n} f(\omega_j)} \right)^{3/4}$$

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

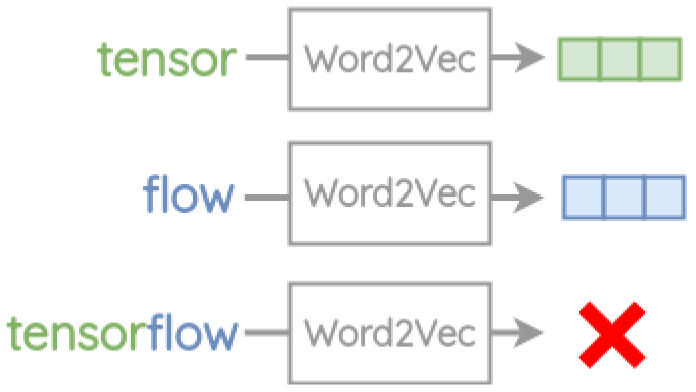
3. FastText

기존 Word2Vec Problem

1. 모든 단어를 각각의 vector로, 즉 1:1로 representation 하는 것의 한계가 있음

→ 특히, training data로 등장하지 않은 **rare word**의 경우 정확한 vector embedding이 어려움

→ **OOV(Out of Vocabulary)** 문제가 있음



https://amitness.com/2020/06/fasttext-embeddings/

기존 Word2Vec Problem

2. 단어 자체의 **내부적 구조**를 무시

   → Morphological(형태학적인) language 언어를 표현하는 것의 한계

Shared radical

eat eats eaten eater eating

https://amitness.com/2020/06/fasttext-embeddings/

FastText 제안

어떠한 corpus 학습에서도
효율적이면서, 형태론적으로 의미 있는 representation 방법을 제안

- Extension of the continuous skip-gram model
- Character level information by character n-gram

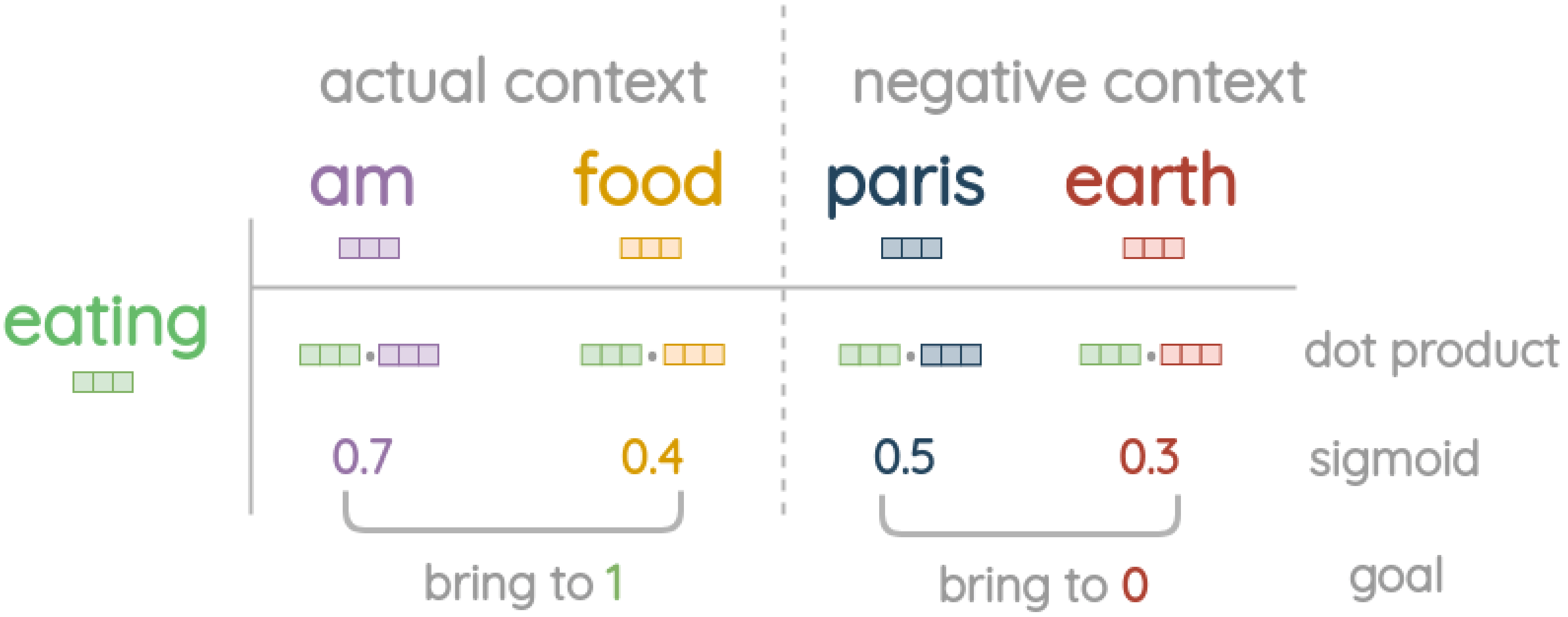
https://www.youtube.com/watch?v=7UA21vg4kKE

FastText: Subword model

Character n-gram

https://amitness.com/2020/06/fasttext-embeddings/

FastText 제안



https://amitness.com/2020/06/fasttext-embeddings/

3. Code Practice

Reference Paper.

1. Tomas Miklov, et al, “Efficient Estimation of Word Representations in Vector Space”, In ICLR, 2013.

2. Tomas Miklov, et al, “Distributed Representations of Words and Phrases and their Compositionality”, In NIPS, 2013.

3. Jeffrey Pennington, et al, “GloVe: Global Vectors for Word Representation”, In EMNLP, 2014.

QnA?

Thank You!