



12

[AI활용텍스트분석]

Word2vec과 다양한 신경망 기법



학습 내용

1. Word2vec
2. FastText
3. GloVe(Global Vectors for Word Representation)
4. Swivel(Submatrix-Wise Vector Embedding Learner)



학습 목표

- Word2vec을 설명할 수 있다.
- 신경망 언어 모델들의 학습 과정을 설명할 수 있다.
- 파이썬 Word2vec 실습을 통해 구체적인 사례를 설명할 수 있다.



오늘의사전 학습

이번 강의부터 필요한 것

- 헌법 텍스트 전처리 파일
- Colab 실습 환경
- 마이크로소프트 엑셀



오늘의사전 학습

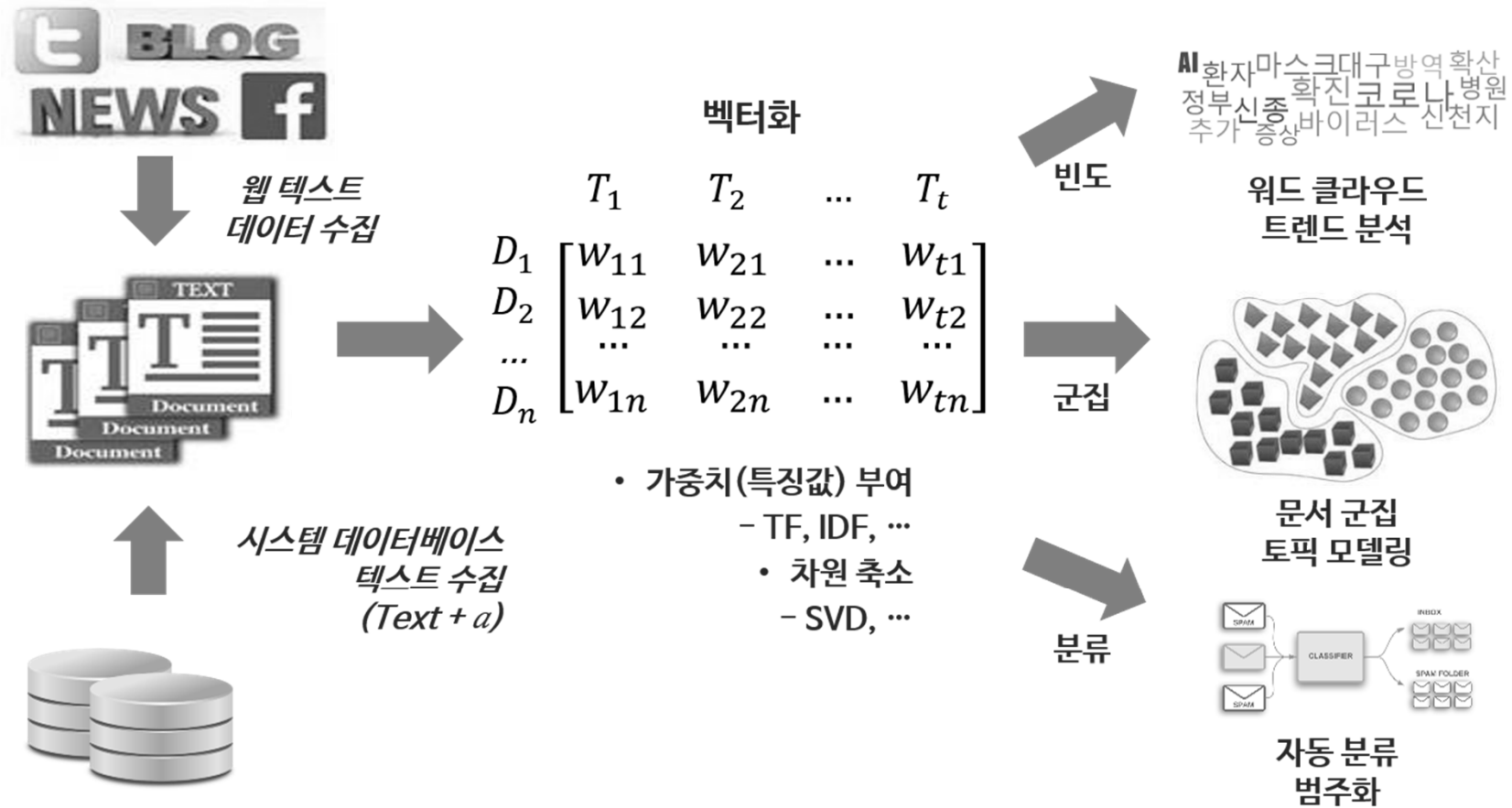
이번 강의에서 얻을 수 있는 것

- Word2vec을 이해함
- 신경망 언어 모델들의 학습 과정을 이해함
- 파이썬 Word2vec 실습을 통해 구체적인 사례를 이해함



복습 하기

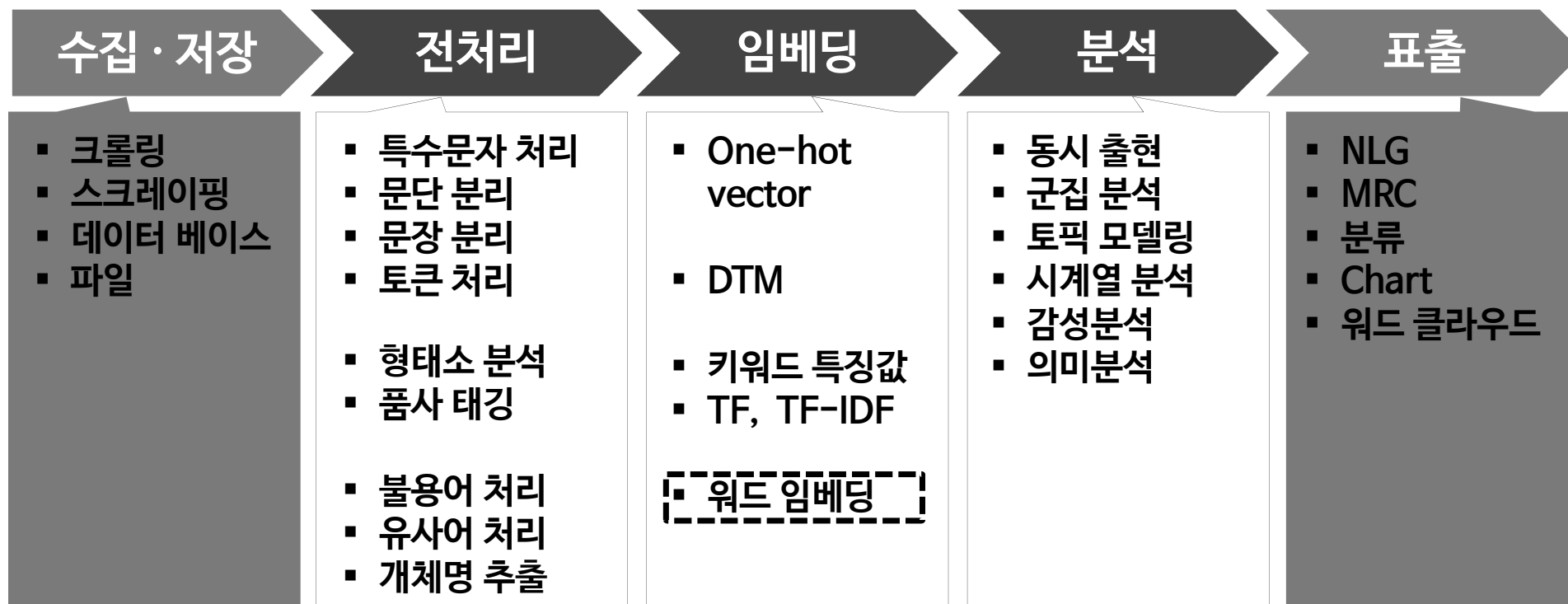
| 텍스트 분석의 대략적인 절차





복습 하기

| 텍스트 분석 절차



비즈니스 목적 / 입증하고자 하는 가설
인문, 사회, 경영, 교육, 보건, ...



복습 하기

| NPLM - Neural Probabilistic Language Model

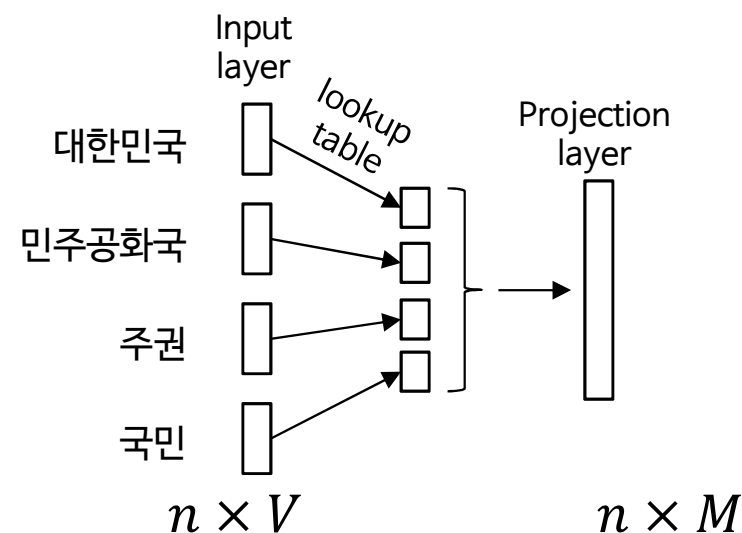
차원 $M = 4$, 단어 벡터 V 의 크기 = 5

$$x_t \times W_{V \times M} = e_t$$

0	0	1	0	0
---	---	---	---	---

0.56	0.55	0.673	0.603
0.087	0.806	0.584	0.913
0.076	0.897	0.778	0.217
0.572	0.335	0.409	0.821
0.235	0.239	0.069	0.999

0.076	0.897	0.778	0.217
-------	-------	-------	-------

 lookup table

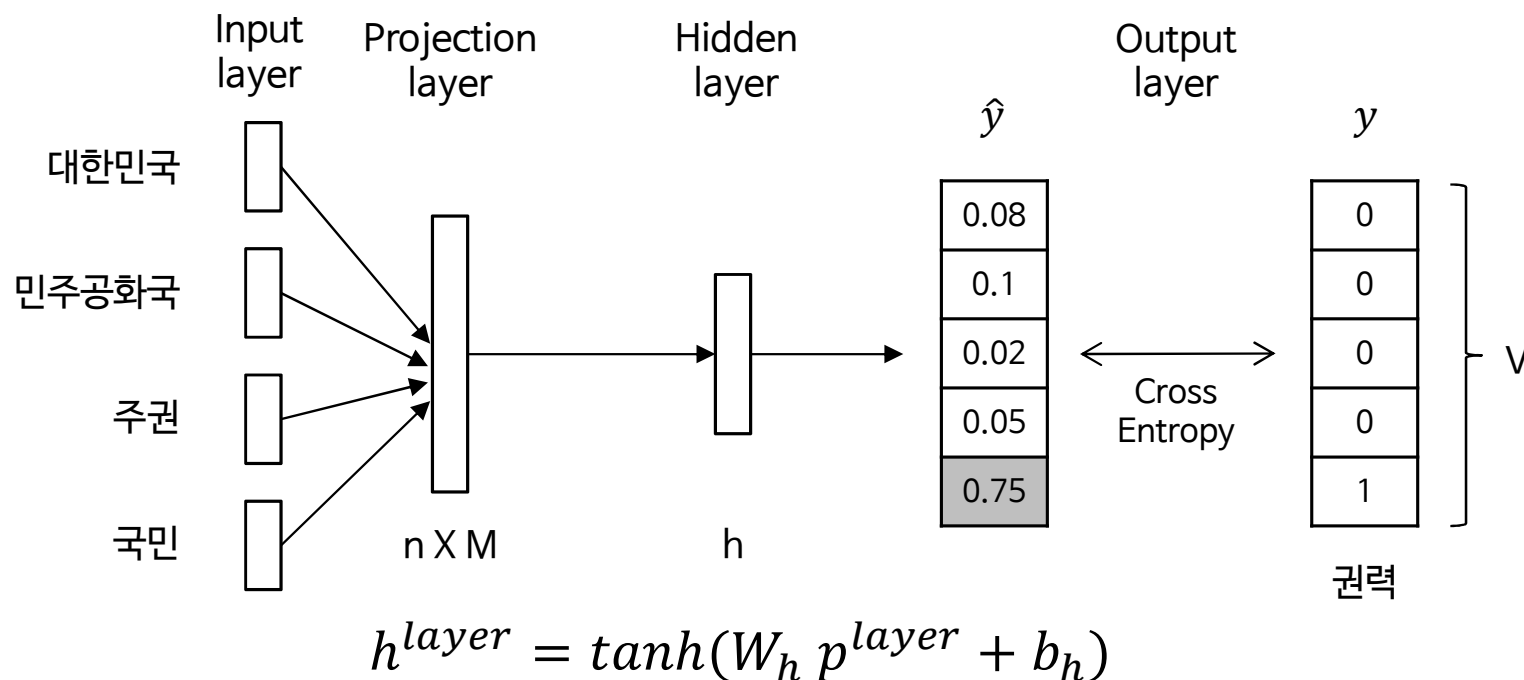


복습 하기

| NPLM - Neural Probabilistic Language Model

차원 $M = 4$, 단어 벡터 V 의 크기 = 5, 윈도우 크기 = 4

$$\hat{y} = \text{softmax}(W_y h^{\text{layer}} + b_y)$$





보충 하기

| 활성화 함수(Activation Function)

활성화 함수

- 어떠한 신호를 입력 받아 이를 적절한 처리를 하여 출력해주는 함수
- 활성화 함수에는 여러가지가 있음
 - 시그모이드, ReLU, tanh, 등등

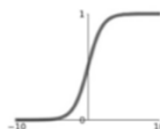


보충 하기

| 활성화 함수(Activation Function)

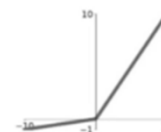
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



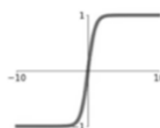
Leaky ReLU

$$\max(0.1x, x)$$



tanh

$$\tanh(x)$$

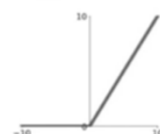


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

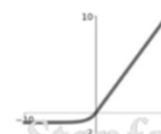
ReLU

$$\max(0, x)$$



ELU

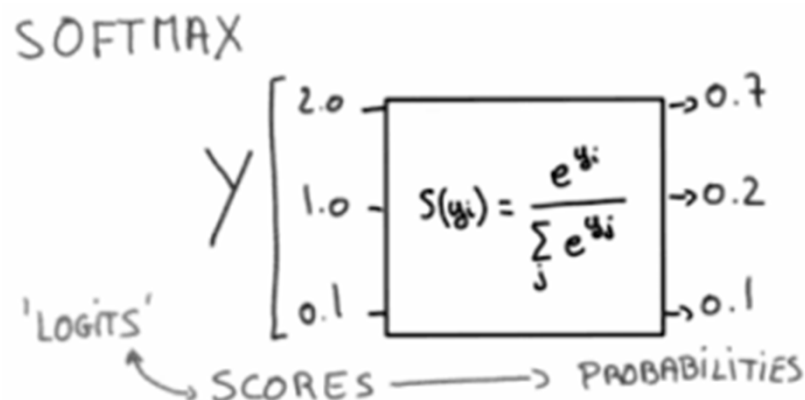
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



- ➡ 시그모이드, ReLU, tanh 등의 함수는 입력된 데이터에 대해서 0과 1사이의 값을 출력하여 해당 값이 둘 중 하나에 속할 확률로 해석

보충 하기

| 활성화 함수(Activation Function)



- ➡ 소프트맥스 함수는 분류 해야 하는 정답(클래스)의 총 개수를 k라고 할 때, k차원의 벡터를 입력 받아 각 클래스에 대한 확률을 추정
- ➡ 주로 모델의 Output Node에서 분류(Classification)을 목적으로 사용



보충 하기

| 손실함수(비용함수) - Cross Entropy

손실함수(비용함수)

- 엔트로피는 불확실성의 척도
- 정보이론에서의 엔트로피는 불확실성을 나타내며, 엔트로피가 높다는 것은 정보가 많고, 확률이 낮다는 것을 의미
- 불확실성이라는 것은 어떤 데이터가 나올지 예측하기 어려운 경우를 의미

$$H(x) = - \sum_{i=1}^n p(x_i) \log q(x_i)$$



보충 하기

| 손실함수(비용함수) - Cross Entropy

크로스 엔트로피

- 실제 분포 q 에 대하여 알지 못하는 상태에서, 모델링을 통하여 구한 분포인 p 를 통하여 q 를 예측 (q 와 p 가 모두 들어가서 크로스 엔트로피라고 함)



보충 하기

| 손실함수(비용함수) - Cross Entropy

머신러닝 /
딥러닝

- 대부분 정답 분포를 모사 하는 것이 목표가 되고, 얼마나 근접했는지를 수치화해서 피드백을 주는 방식으로 구성
- 어떤 문제를 풀고 싶을 때 그 문제엔 (정확히는 알 수 없으나) 어떤 미지의 실제 분포라는 것이 존재하고, 우리가 만드는 머신러닝 / 딥러닝 모델은 그 실제 분포를 모사해주는 확률 모델



보충 하기

| 손실함수(비용함수) - Cross Entropy

Cross entropy

틀릴 수 있는 정보 양을 고려한 최적으로
인코딩할 수 있게 해주는 정보량

- 주로 Regression에서는 최소자승법(MSE)이 비용함수로 쓰이고 Classification문제에서는 Cross entropy가 비용함수로 사용
- 가중치 업데이트 시에 MSE를 비용함수로 쓰니까 업데이트 되는 양이 너무 적어, 이를 보완하기 위한 방법으로 Cross Entropy 사용
- Cross entropy 값은 실제값과 예측값이 맞았을 경우에는 0으로 수렴하고, 값이 틀릴 경우 무한대로 발산하기 때문에 예측값과 실제값이 같도록 Cross entropy를 손실함수로 사용하여 값이 최소화 시키는 방향으로 학습



보충 하기

| 이번 수업에서 소개하는 기법들 : 소개기법과 특징

임베딩 기법	특징
Word2vec Sent2vec Doc2vec	<ul style="list-style-type: none">• 신경망 기법을 응용• 단어의 문맥 상에서의 의미를 내포하도록 학습
FastText	<ul style="list-style-type: none">• Word2vec과 거의 동일• 학습 어휘가 아닌 어휘에 대한 추정 가능
GloVe	<ul style="list-style-type: none">• LSA와 word2vec의 한계 극복을 위해 제안• 단어 문맥 행렬의 분해를 통한 학습• 중심 단어와 주변 단어 벡터의 내적이 전체 코퍼스에서의 동시 등장 확률이 되도록 학습
Swivel	<ul style="list-style-type: none">• PMI 행렬을 분해하여 학습• PMI 행렬의 단점을 극복하는 방향으로 학습



1

word2vec

- 1) 개요
- 2) 활용
- 3) 응용 - sent2vec
- 4) 응용 - doc2vec

1) 개요

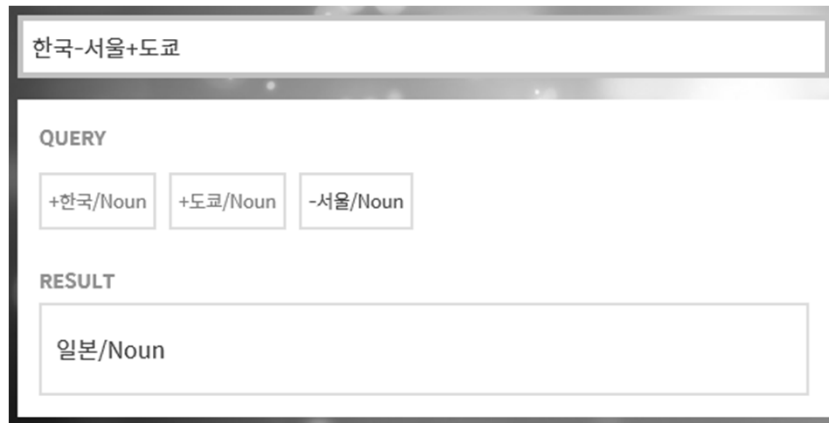


세종사이버대학교



활용 예시

- 고양이 + 애교 = 강아지
- 한국 - 서울 + 도쿄 = 일본
- 박찬호 - 야구 + 축구 = 호나우두



한국-서울+도쿄

QUERY

+한국/Noun +도쿄/Noun -서울/Noun

RESULT

일본/Noun

- <http://w.elnn.kr/search/>
- 약 45만 종류, 4.2억 개의 단어

1) 개요



세종사이버대학교

window와 target word

window target word

↑ ↑

Colorless green ideas sleep furiously

window target word

↑ ↑

Colorless green ideas sleep furiously

Colorless green ideas sleep furiously

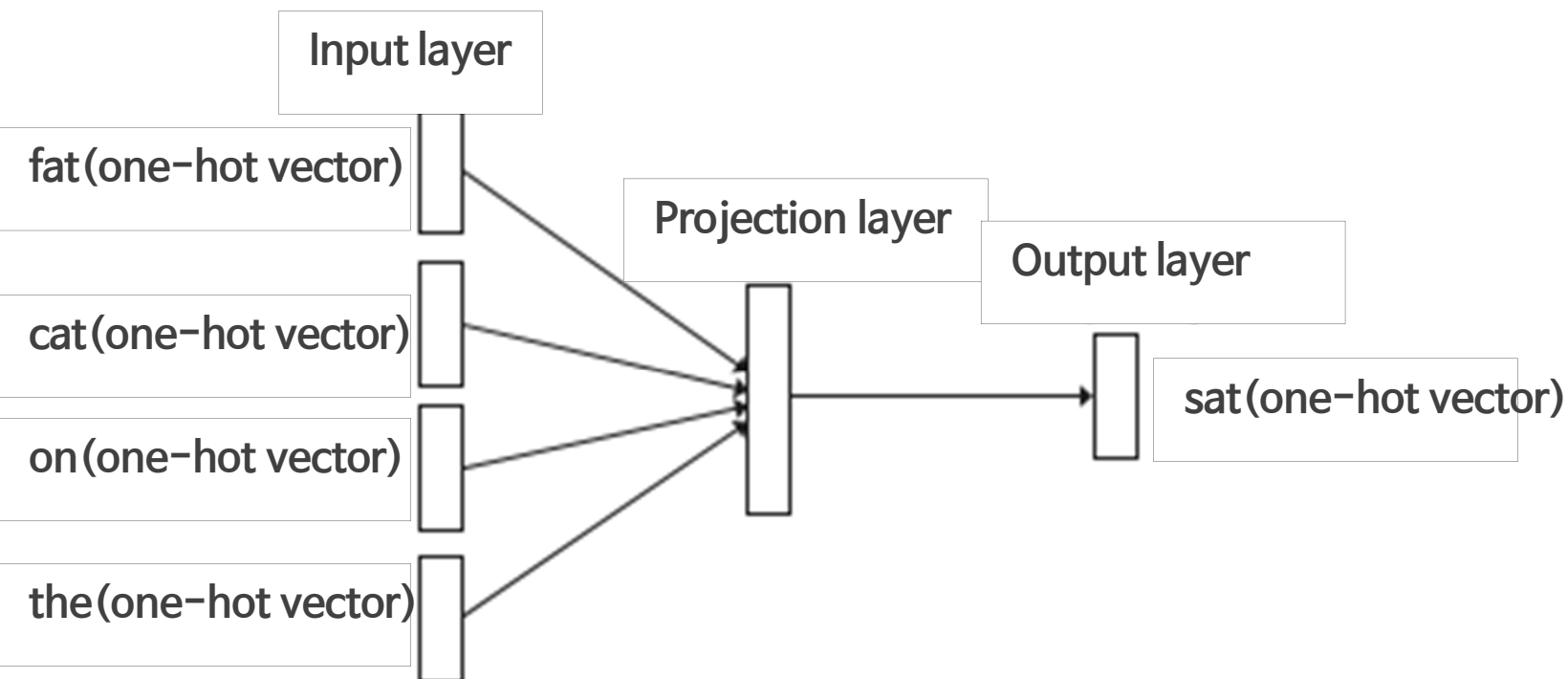
Colorless green ideas sleep furiously

1) 개요



CBOW와 Skip-gram

▶ CBOW (Continuous BOW)



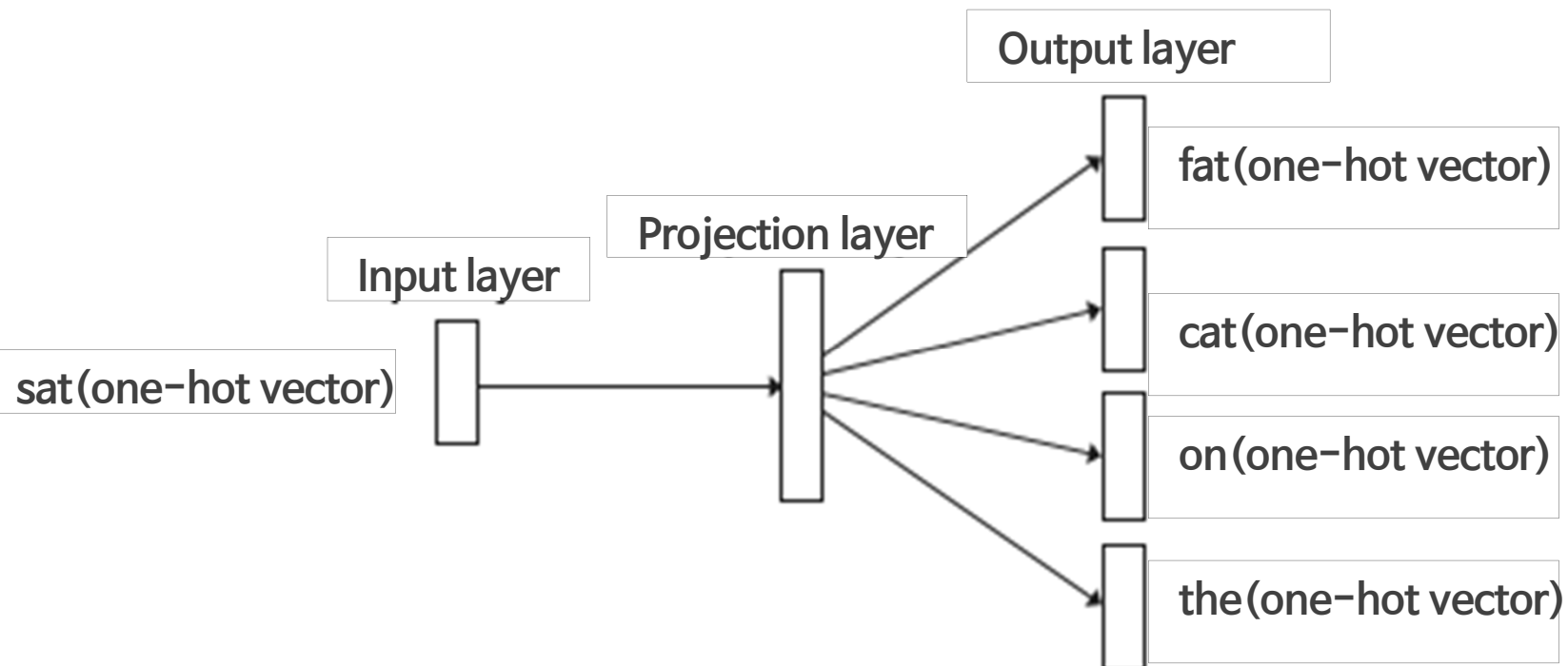
1) 개요



세종사이버대학교

CBOW와 Skip-gram

▶ Skip-gram



1) 개요



입력값을 위한 one-hot 인코딩 예시

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]

1) 개요



CBOW (Continuous BOW) 연산 방식

$$X_{cat} \times W_{V \times M} = V_{cat}$$

0	0	1	0	0	0	0
---	---	---	---	---	---	---

 \times

0.5	2.1	1.9	1.5	0.8
0.8	1.2	2.8	1.8	2.1
2.1	1.8	1.5	1.7	2.7

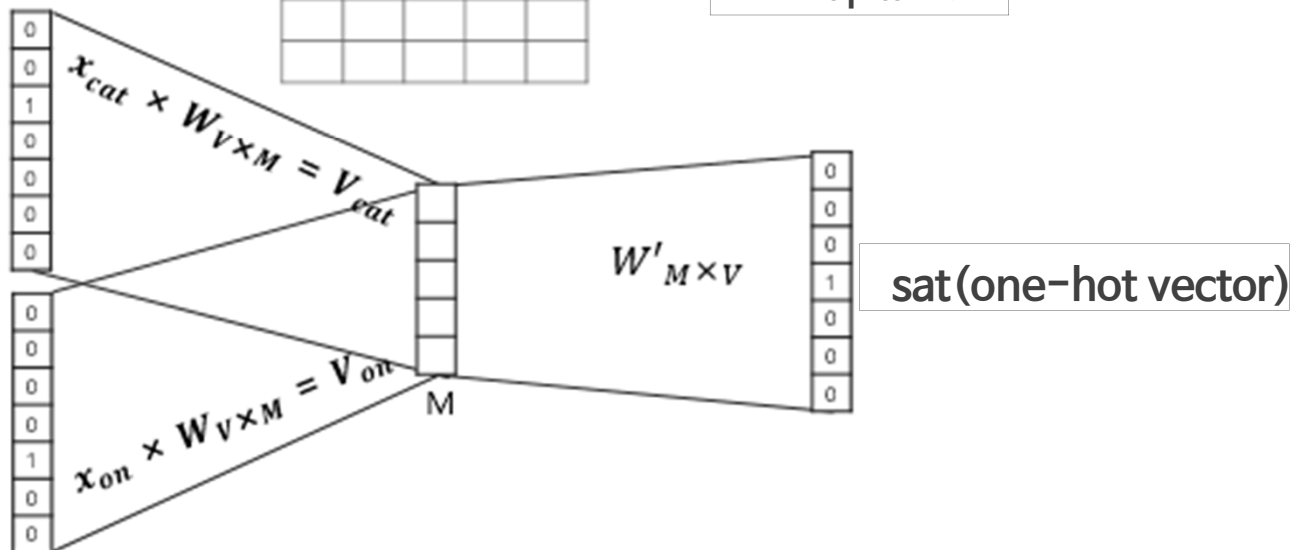
 $=$

2.1	1.8	1.5	1.7	2.7
-----	-----	-----	-----	-----

Lookup table

cat(one-hot vector)

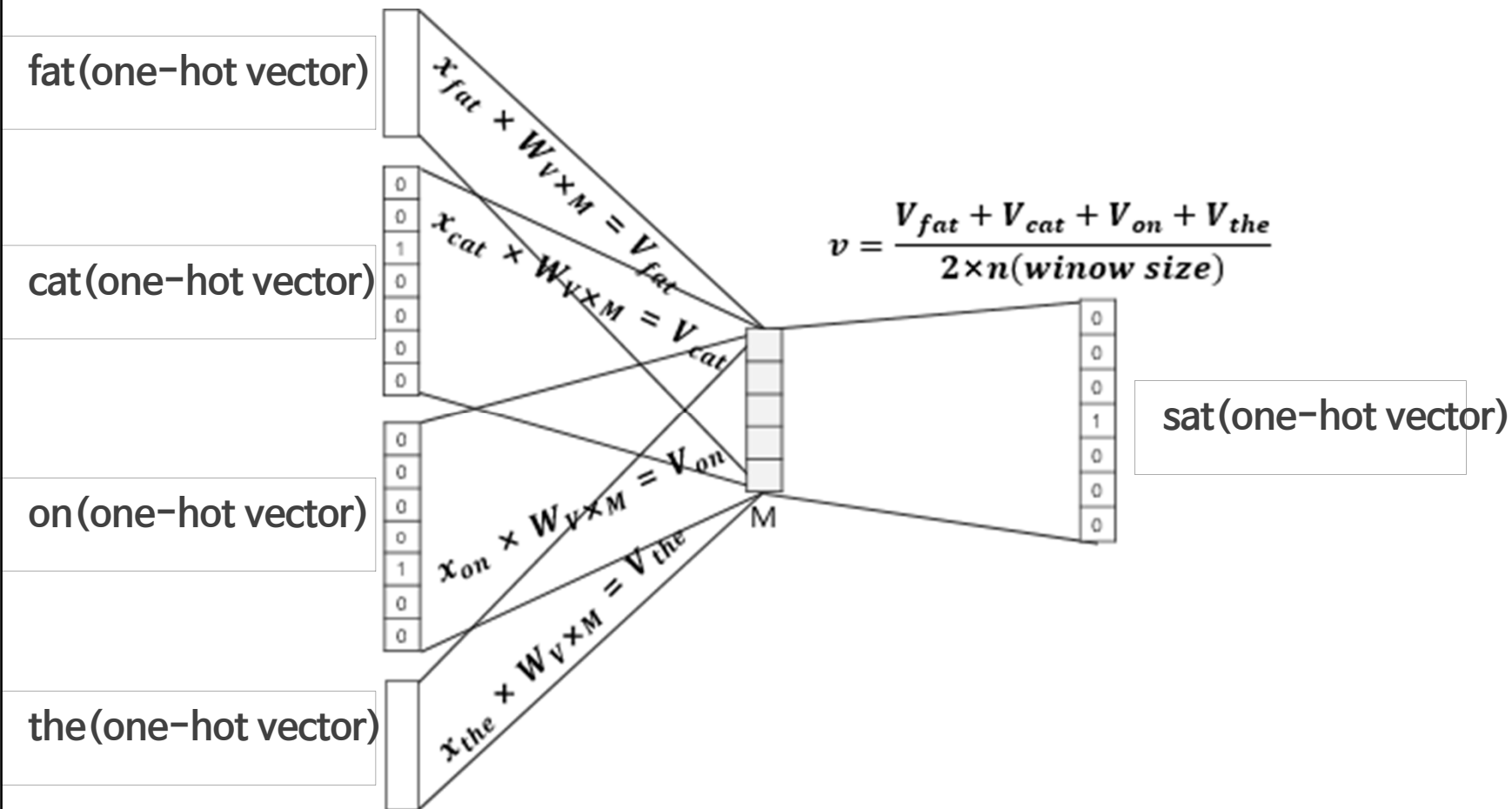
on(one-hot vector)



1) 개요



CBOW (Continuous BOW) 연산 방식

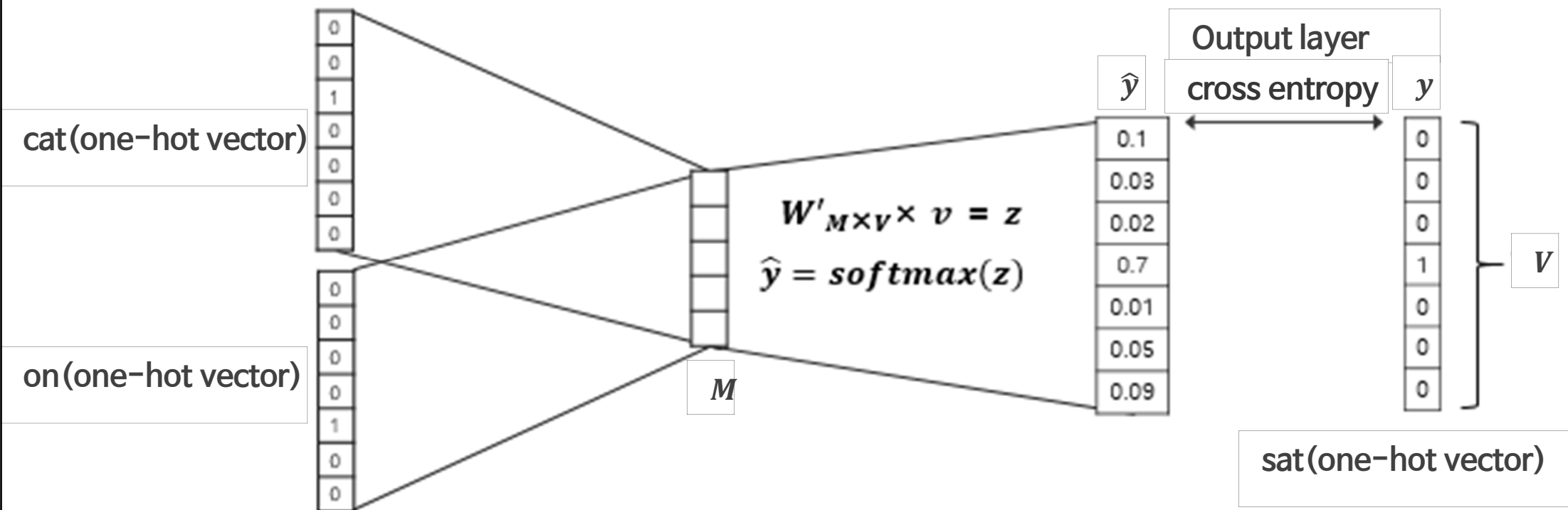


1) 개요



세종사이버대학교

CBOW (Continuous BOW) 연산 방식

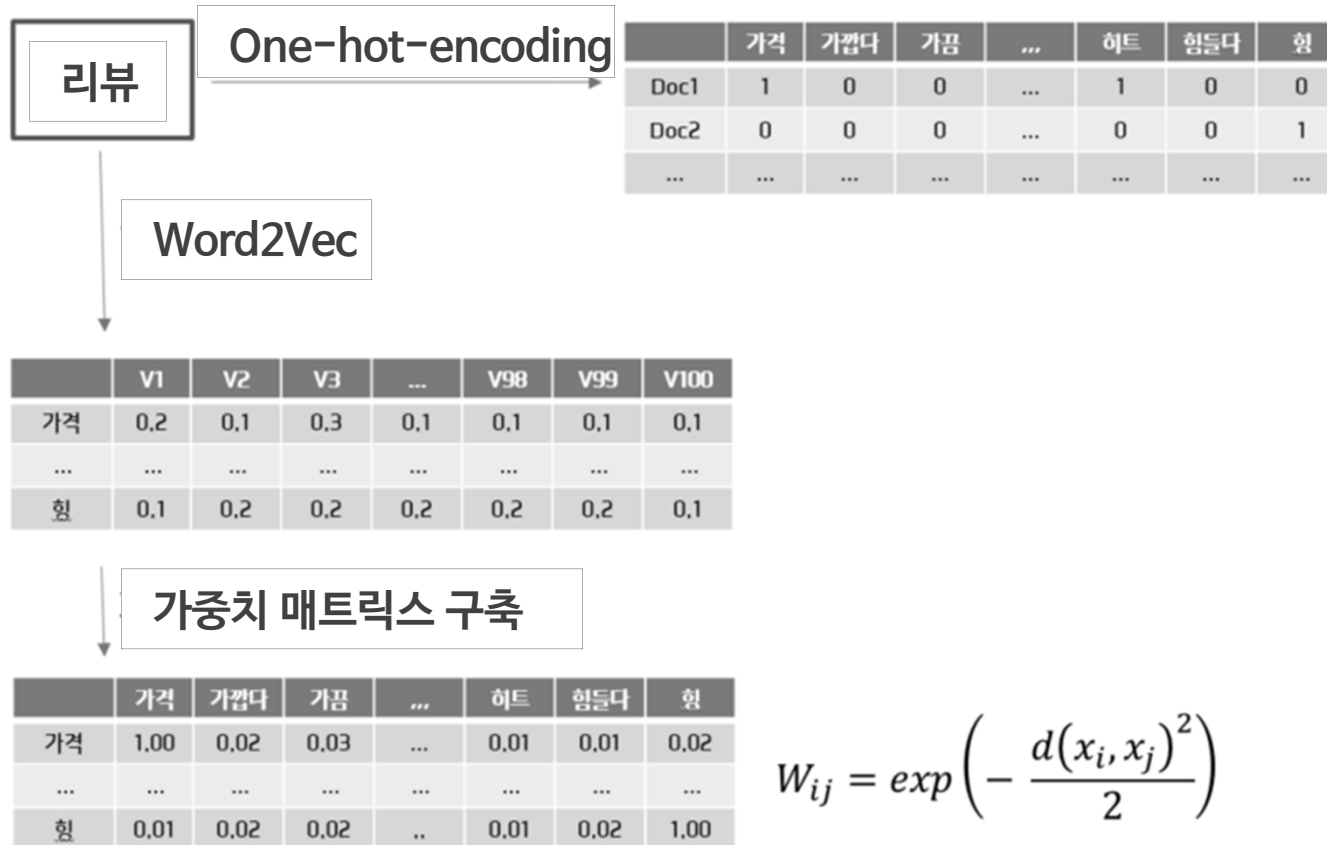


2) 활용



📖 Word2vec을 어떻게 활용할까?

▶ ratsgo's blog : word2vec으로 문장 분류하기



2) 활용



Word2vec을 어떻게 활용할까?

- ▶ ratsgo's blog : word2vec으로 문장 분류하기

Document Term Matrix

	가격	가깝다	가끔	...	히트	힘들다	형
Doc1	1	0	0	...	1	0	0
Doc2	0	0	0	...	0	0	1
...

×

7개 기능만 추출한 가중치 매트릭스

	가격	가깝다	가끔	...	히트	힘들다	형
배터리	1.00	0.02	0.03	...	0.01	0.01	0.02
디자인	0.01	0.02	0.02	..	0.01	0.02	0.01
...

기능별 Score

=

	배터리	디자인	화면	촬영	스펙	사운드	운영체제
Doc1	0.04	0.03	0.01	...	0.02	0.01	0.01
Doc2	0.01	0.01	0.02	...	0.03	0.01	0.01
...

3) 응용 - sent2vec



세종사이버대학교

개요

Word2vec의 CBOW를 응용

단어 대신 문장에 대한 벡터를 생성

유사 문장 검색 등의 영역에서 사용

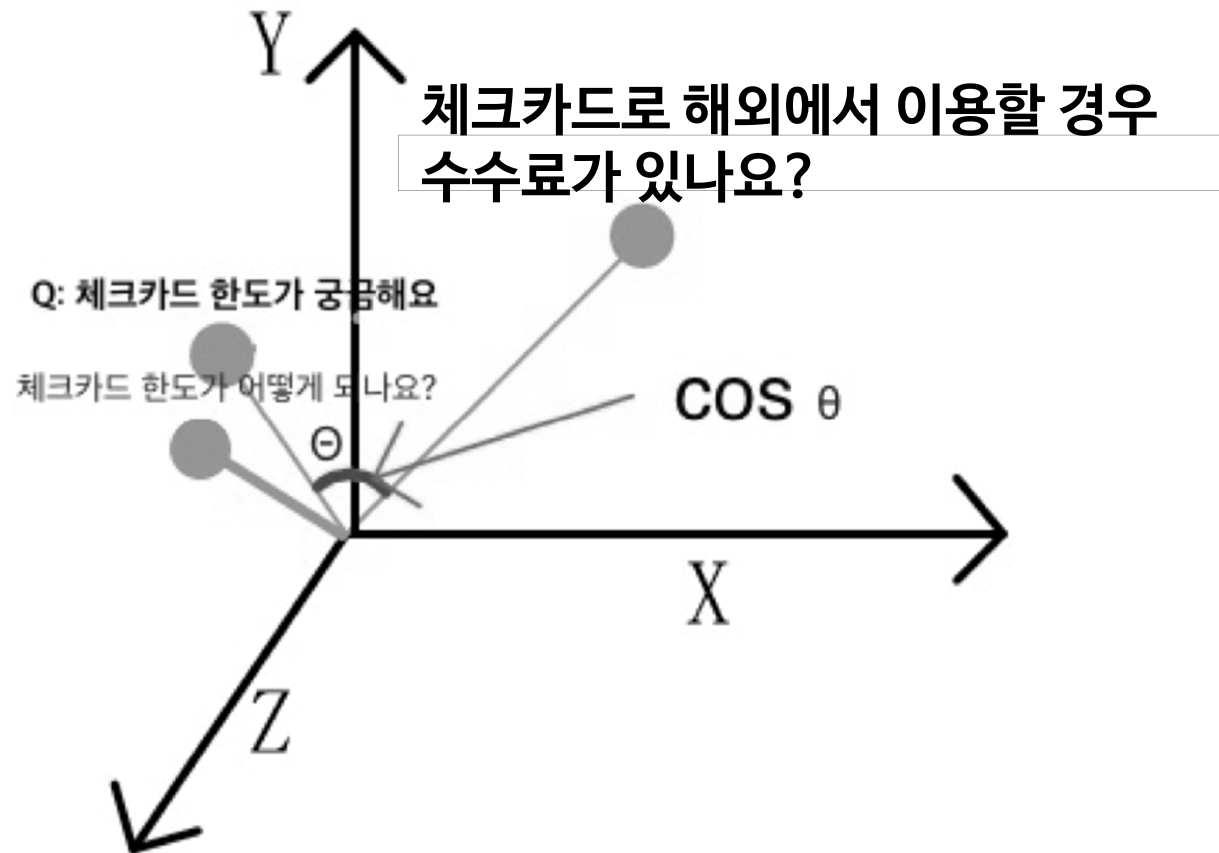
문장에 대한 N-gram 으로 학습

3) 응용 - sent2vec



세종사이버대학교

개요



3) 응용 - sent2vec



세종사이버대학교

문장에 대한 N-gram 생성

문장 = (A, B, C, D, E) 로 구성되어 있을 때

- 단어 N그램 $\rightarrow N = 3$: (A), (A,B), (A,C)
- 단어 N그램 $\rightarrow N = 4$: (A), (A,B), (A,C), (A,D)

N그램 확장은 뒤로 하되 이전 단어는 포함하지 않음

3) 응용 - sent2vec



단어 N그램 N=3, 단어 ID 조합과 새로운 단어 ID

◀ 각각의 N그램에 대한 단어 ID 조합의 새로운
키 값을 생성하여 새로운 단어 벡터를 생성 ▶

타입	단어	단어 ID
단어	카드결제	14
	알림	17
	서비스	26
	계좌변경은	1538
	어떻게	2
	하나요	6
단어 n-그램=3	단어 ID 조합	새로운 단어 ID
	14 17	692956
	14 26	1780052
	17 26	841078
	17 1538	711288
	26 1538	1285391
	26 2	888413
	1538 2	1747
	1538 6	100493
	2 6	758302

3) 응용 - sent2vec



세종사이버대학교



예시 화면



문장 입력 마감

입력된 문장

4,930개의 입력된 문장이 있습니다.

전체 문장 조회

입력된 문장 조회

	번호	정답
신용대출 기한연장이 가능한가요?	150	신용대출도 기한연장이 가능한가요?
전화 잃어버렸다	446	휴대폰을 분실했는데 어떡하죠?
카드 해지 방법	195	카드를 해지하려면 어떻게 해야하나요?
카드 분실 신고를 했는데 재발급을 어디서 받죠?	192	카드 분실 신고를 했습니다. 재발급 받으려면 어떻게 해야 하나요?
카드를 일시적으로 정지하고 싶습니다.	191	카드를 일시적으로 정지할수 있나요?
일시적으로 카드를 정지하고 싶습니다.	191	카드를 일시적으로 정지할수 있나요?
일시적으로 카드를 정지할수 있나요?	191	카드를 일시적으로 정지할수 있나요?
이용내역서를 받고 싶지 않습니다.	190	이용내역서를 받고 싶지 않는데 어떻게 해야 하나요?

4) 응용 - doc2vec



세종사이버대학교

개요

2014년 구글 연구팀이 발표한 문서 임베딩 모델

단어와 문서를 같은 임베딩 공간의 벡터로 표현하는 방법

Word2Vec 이 확장된 임베딩 방법

Document id 를 모든 문맥에 등장하는 단어로 취급

타겟 단어와 이전 단어 k 개가 주어졌을 때, 이전 단어들 + 해당 문서의 아이디로 타겟 단어를 예측

4) 응용 - doc2vec



세종사이버대학교

개요

두 문서에 등장한 단어가 다르더라도 단어의 벡터들이 비슷하다면 두 문서의 벡터는 서로 비슷함

PV-DM 과 PV-DBOW 두 개의 방식 존재

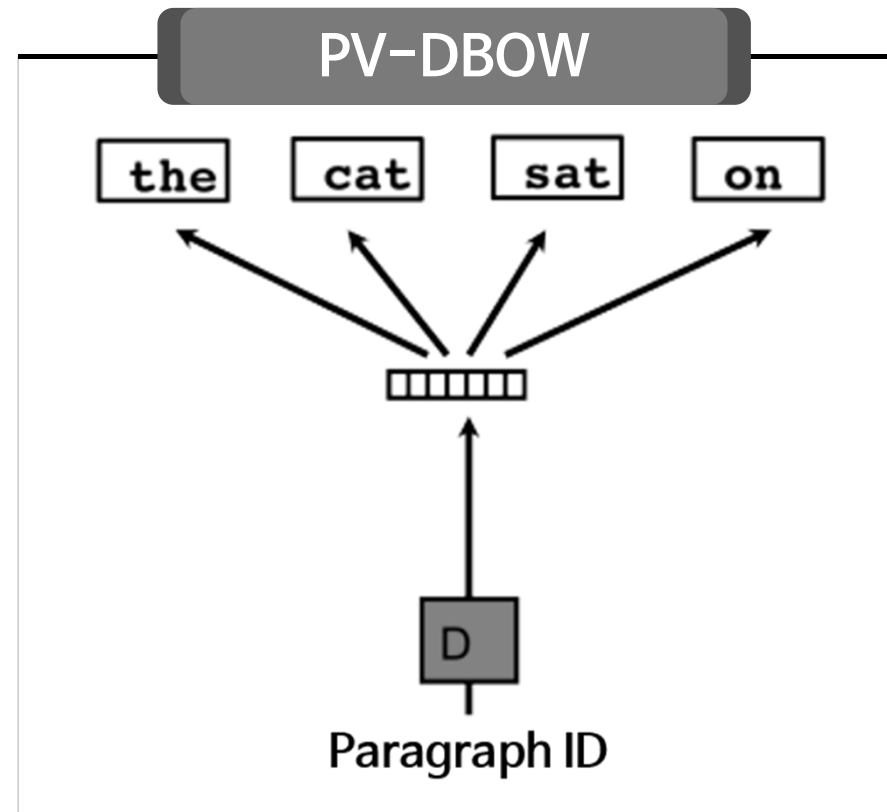
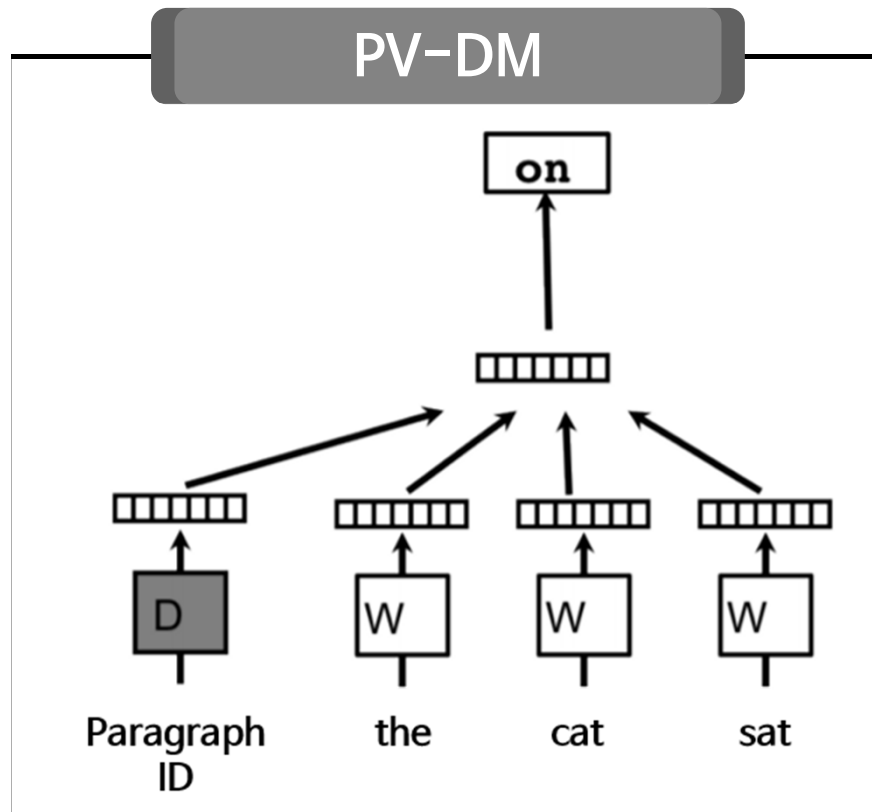
리뷰들을 기반으로 영화 벡터를 학습하고 싶다면 각 리뷰마다 해당하는 영화의 아이디를 Document Id 로 정의 가능

4) 응용 - doc2vec



세종사이버대학교

PV-DM과 PV-DBOW



4) 응용 - doc2vec



세종사이버대학교

학습 데이터 예시

paragraph_1이라는 문서에서 “the cat sat on the mat”
라는 문장이 있을 때의 학습데이터 구성

- 윈도우 사이즈 : $k = 3$
- [paragraph_1, the, cat, sat] - on
- [paragraph_1, cat, sat, on] - the
- [paragraph_1, sat, on, the] - mat

paragraph에서 단어를 예측하는 과정에서 학습

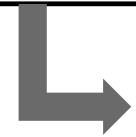
4) 응용 - doc2vec



세종사이버대학교

학습 데이터 예시

각 문서 paragraph는 별도의 문서의 수 \times d 차원 행렬에 저장



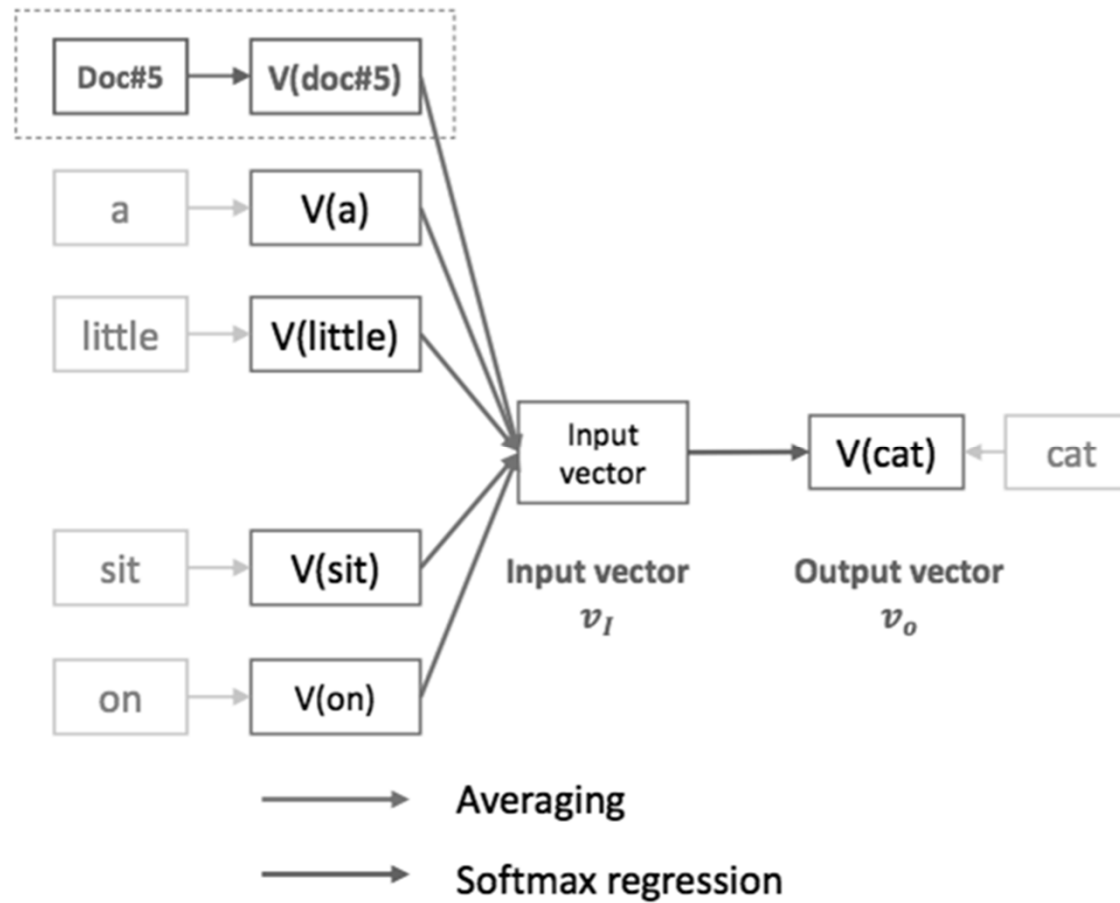
학습이 완료된 후 이 행렬을 이용하여 paragraph의 임베딩 된 벡터를 사용함

4) 응용 - doc2vec



세종사이버대학교

학습 방법





2

FastText

- 1) 특징, 개요
- 2) 학습 팁
- 3) 활용

1) 특징, 개요



세종사이버대학교

페이스북이 2016년에 발표

단어 대신 부분 단어(Subword)를 사용

이외의 부분은 word2vec과 거의 동일

Word2Vec 등장 빈도 수가 적은 단어(Rare word)에 대해서는 임베딩의 정확도가 높지 않다는 단점

FastText는 등장 빈도 수가 적은 단어라 하더라도, N-gram으로 임베딩

- 참고할 수 있는 경우의 수가 증가
- Word2Vec와 비교하여 정확도가 높은 경향이 있음
- 미학습 단어(새로운 단어, 오타 등)에 대해서도 유사도 추정 가능

2) 학습 팁



세종사이버대학교

부분 단어(subword)로 학습 : 음절, 자소 단위로 학습

음절 단위(N=3 일 때)

- “텍스트분석”=[‘텍스’, ‘텍스트’, ‘스트분’, ...]

자소 단위(N=3 일 때)

- [‘ㅌ ㅊ’, ‘ㅌ ㅊ ㄱ’, ‘ㅊ ㄱ ㅅ’, ... ‘ㅅ ㅡ ㅌ’, ‘ㅌ ㅡ ㅂ’, ‘ㅡ ㅂ ㅅ’, ...]

실제 내부적으로는...

- 〈‘텍스’, ‘텍스트’, ‘스트분’, ‘트분석’, ‘분석’〉 〈‘텍스트분석’〉

3) 활용



세종사이버대학교

사전 학습된 모델 활용 가능

페이스북에서는 294개 언어에 대하여
위키피디아로 학습한 사전 훈련된 벡터들을 제공





3

GloVe

(Global Vectors for Word Representation)

- 1) 특징, 개요
- 2) GloVe의 학습 모델 생성 절차
- 3) 윈도우 기반 동시 등장 행렬
(Window based Co-occurrence Matrix)
- 4) 동시 등장 확률 (Co-occurrence Probability)
- 5) GloVe의 목적 함수

1) 개요, 특징



2014년에 미국 스탠포드 대학에서 개발

기존의 카운트 기반의 LSA(Latent Semantic Analysis)와
예측 기반의 Word2Vec의 단점을 지적하며 이를 보완

LSA는 카운트 기반으로 코퍼스의 전체적인 통계 정보를 고려

- 의미를 유추하는데 한계

Word2vec은 주변 단어의 출현 순서에 따른 예측 기반으로
윈도우 크기 내에서만 주변 단어를 고려

- 코퍼스의 전체적인 통계 정보 반영에 한계

2) GloVe의 학습 모델 생성 절차



1 주어진 코퍼스와 지정한 윈도우 사이즈를 통해 동시 등장행렬 (Co-occurrence Matrix) Matrix X을 생성

2 행렬 분해

3 Word2vec과 유사한 방법으로 학습 대상 단어들을 윈도우 사이즈 내에서 선택

4 선택된 단어와 Matrix X를 기반으로 목적함수 (손실함수)를 이용해 학습 진행



주요 아이디어

임베딩 된 중심 단어와 주변 단어 벡터의 내적이 전체 코퍼스에서의 동시 등장 확률이 되도록 만드는 것

3) 윈도우 기반 동시 등장 행렬 (Window based Co-occurrence Matrix) 세종사이버대학교



Ex

- I like deep learning
- I like NLP
- I enjoy flying

카운트	I	like	enjoy	deep	learning	NLP	flying
I	0	2	1	0	0	0	0
like	2	0	0	1	0	1	0
enjoy	1	0	0	0	0	0	1
deep	0	1	0	0	1	0	0
learning	0	0	0	1	0	0	0
NLP	0	1	0	0	0	0	0
flying	0	0	1	0	0	0	0

4) 동시 등장 확률 (Co-occurrence Probability)



동시 등장 확률과 크기 관계 (ratio)	k=solid	k=gas	k=water	k=fasion
$P(k \text{ice})$	0.00019	0.000066	0.003	0.000017
$P(k \text{steam})$	0.000022	0.00078	0.0022	0.000018
$P(k \text{ice}) / P(k \text{steam})$	8.9	0.085	1.36	0.96

임베딩 된 중심 단어와 주변 단어 벡터의 내적이 전체 코퍼스에서의 동시 등장 확률이 되도록 만드는 것

→ 이를 만족하도록 임베딩 벡터를 만드는 것

$$\text{dot product}(w_i, \tilde{w}_i) \approx P(k | i) = P_{ik}$$

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

5) GloVe의 목적 함수



세종사이버대학교

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

$$\text{Loss function} = \sum_{m,n=1}^V f(X_{mn})(w_m^T \tilde{w}_n + b_m + \tilde{b}_n - \log X_{mn})^2$$

$$\begin{matrix} \text{타겟단어} & & \text{문맥단어} \\ |V| & & |V| \\ \log(& \text{Word context Matrix } X &) \approx \begin{matrix} \text{타겟단어} \\ |V| \end{matrix} \begin{matrix} d \\ U \end{matrix} \times \begin{matrix} \text{문맥단어} \\ |V| \end{matrix} \begin{matrix} d \\ V \end{matrix} \end{matrix}$$

U와 V를 랜덤으로 초기화 한 후 목적함수를 최소화 하는 방향으로 학습손실이 더 이상 줄어 들지 않을 때 까지 U, V를 업데이트 후 학습 종료

→ U를 단어 임베딩으로 사용



4

Swivel

(Global Vectors for Word Representation)

- 1) PMI(Point-wise Mutual Information)
- 2) 특징, 개요
- 3) (예습)언어 임베딩 분야의 흐름

1) PMI(Point-wise Mutual Information)



PMI

- 두 단어의등장이 독립일 때 대비 얼마나 자주 같이 등장했는가를 수치화
- 분포 가정에 따르는 단어 가중치 할당 방법
- PMI 행렬은 행 벡터 자체를 해당 단어의 임베딩으로 사용 가능

$$PMI(A, B) = \log \frac{P(A, B)}{P(A) \times P(B)}$$

1) PMI(Point-wise Mutual Information)



세종사이버대학교

개울가, 에서, 속옷, 빨래, 를, 하는, 남녀

문맥 단어 \	개울가	에서	속옷	빨래	를	하는	남녀	Total
개울가 : 빨래 :		+1	+1		+1	+1		20
total			15					1000

※ 원도 사이즈 = 2

$$PMI(\text{빨래}, \text{속옷}) = \log \frac{P(\text{빨래}, \text{속옷})}{P(\text{빨래}) \times P(\text{속옷})}$$

$$= \log \frac{P(\frac{10}{1000})}{P(\frac{20}{1000}) \times P(\frac{15}{1000})}$$

2) 개요, 특징



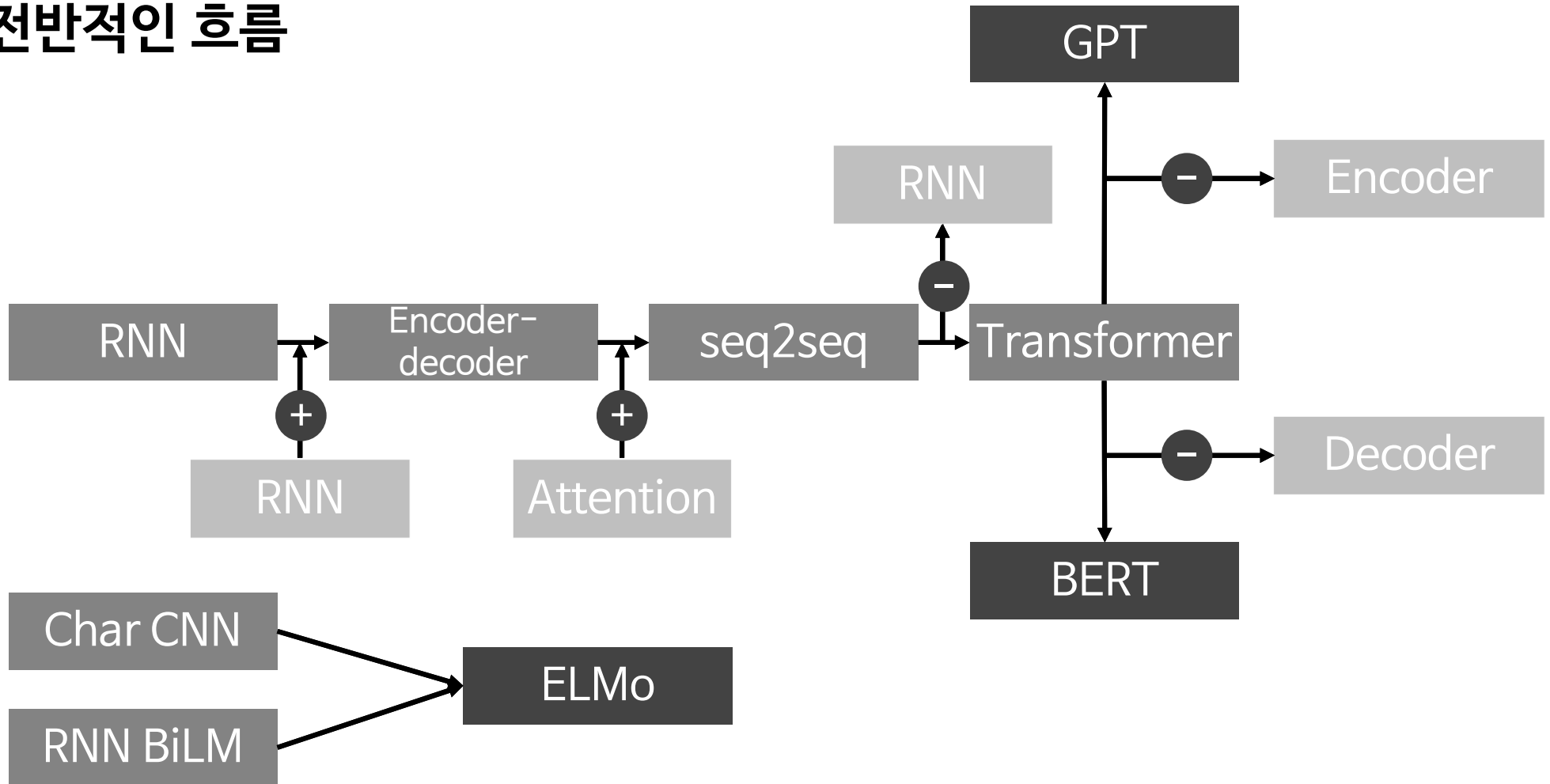
Swivel

- 구글 연구팀 (Shazeer et al. 2016)이 발표한 행렬분해 기반 단어 임베딩 기법
- PMI(Point-wise Mutual Information) 행렬을 분해하여 학습
- PMI 행렬의 단점을 극복할 수 있도록 목적함수를 설계

$$\begin{array}{c} \text{문맥단어} \\ |V| \\ \text{PMI Matrix } X \\ \text{타겟단어} \\ |V| \end{array} \approx \begin{array}{c} d \\ \text{타겟단어} \\ |V| \\ U \end{array} \times \begin{array}{c} \text{문맥단어} \\ |V| \\ d \\ V \end{array}$$

3) (예습)언어 임베딩 분야의 흐름

전반적인 흐름





학습 정리

■ word2vec과 x2vec

- word2vec은 신경망 언어 모델의 아이디어를 이용한 임베딩 모델임
- 주변의 단어들로 중심의 단어를 예측하는 과정, 혹은 중심의 단어로 주변의 단어들을 예측하는 과정에서 학습이 됨
- 문장의 구조를 통해 학습하므로 정답 라벨링이 필요하지 않음
- sent2vec과 doc2vec은 각각 문장과 문서를 임베딩하는 방법이며 word2vec의 아이디어를 응용하였음



학습 정리

■ 그 외의 임베딩 기법들

- Fasttext는 음절이나 자소 단위로 전처리한 텍스트를 word2vec의 방식으로 학습함
- GloVe는 동시 등장 행렬을 학습에 응용했으며, 카운트 기반의 LSA(Latent Semantic Analysis)와 예측 기반의 Word2Vec의 단점을 보완하고자 한 기법임
- Swivel은 GloVe 기법을 보완하고자 PMI 행렬을 분해하여 학습하는 기법임