Efficient Estimation of Word Representations in Vector Space

Hwang Hyeon Tae

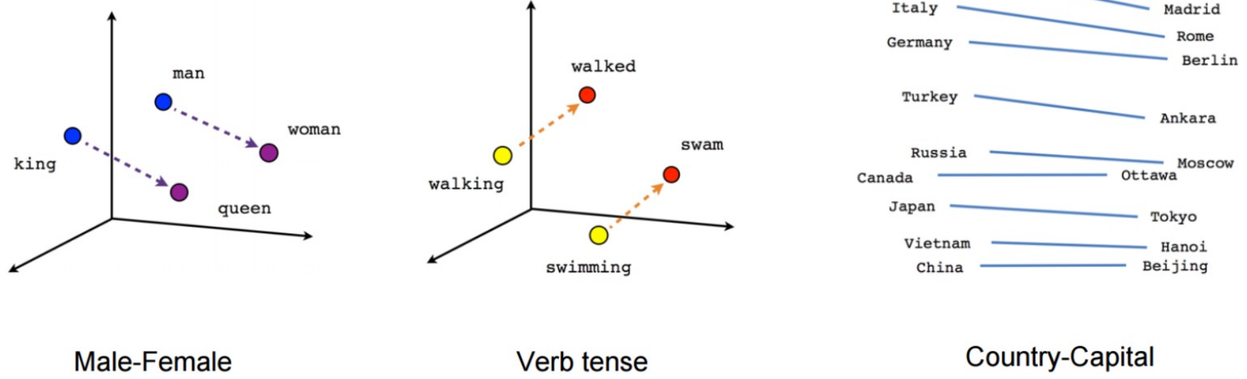
01

Introduction

Goals of the Paper

- 주 목표는 거대한 데이터셋에서 고품질 단어 벡터를 학습하는 데 사용할 수 있는 기술을 소개하는 것
  - 분산 표현된 벡터들은 단어 간 similarity 측정 가능
  - 단어 표현의 similarity가 단순한 구문 규칙성을 넘어서는 것을 확인

Example



Male-Female　　Verb tense　　Country-Capital

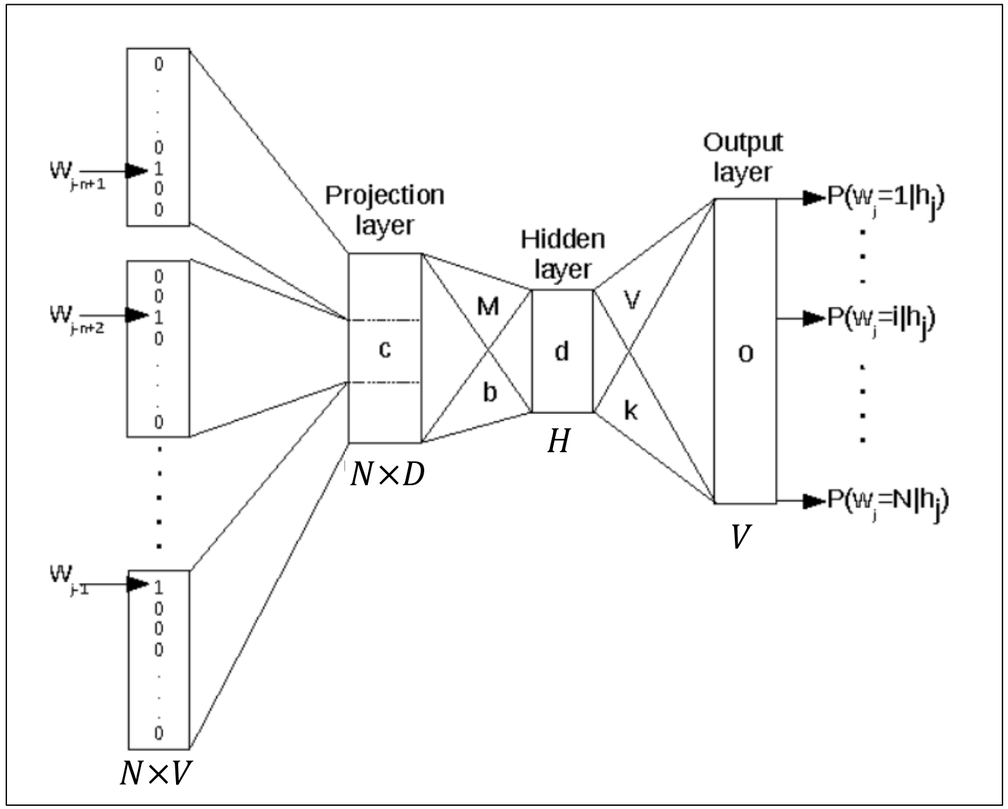
Vector("Queen") = Vector("King") - Vector("Man") + Vector("Woman")

02

Previous work

Feed-forward NNLM

- Structure



Notation

N : Input Layer로 들어가는 이전 단어의 개수
D : Projection 이후의 차원
H : Hidden layer Size
V : Vocabulary Size

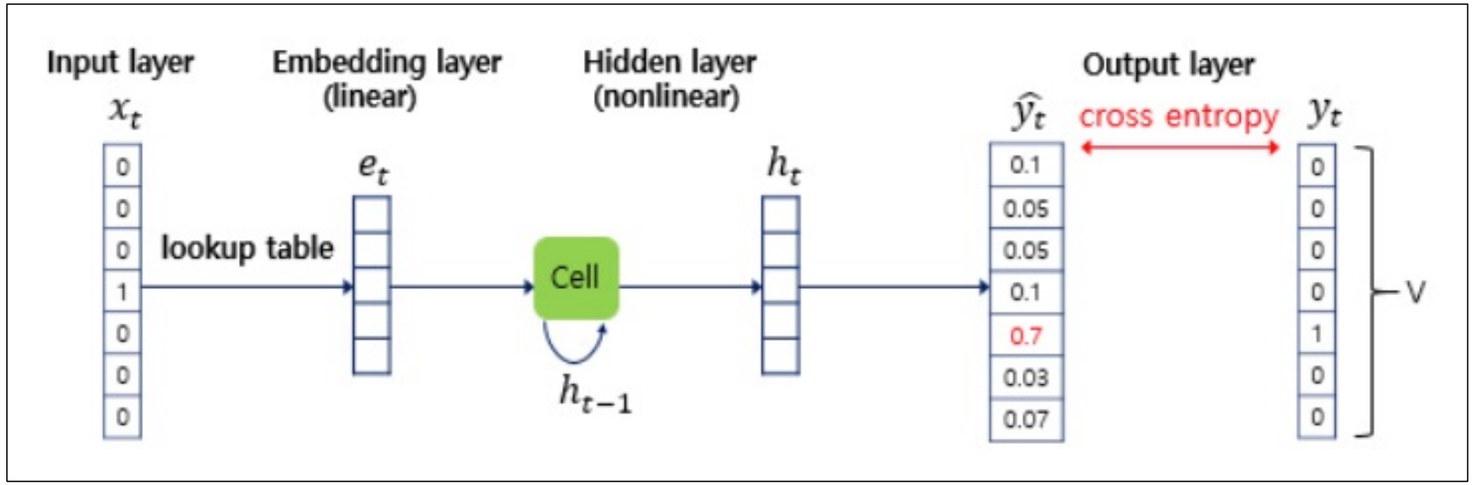
Feed-forward NNLM

- 한계점

- History로 사용할 단어의 개수를 고정해 주어야 한다.
- History만을 보고 예측하기 때문에 미래 시점의 단어들을 고려하지 않는다.
- Computational Complexity : $Q = N \times D + N \times D \times H + H \times V$

RNN

- Structure



Notation

D : The word representations
H : Hidden layer Size
V : Vocabulary Size

RNN

- 특징

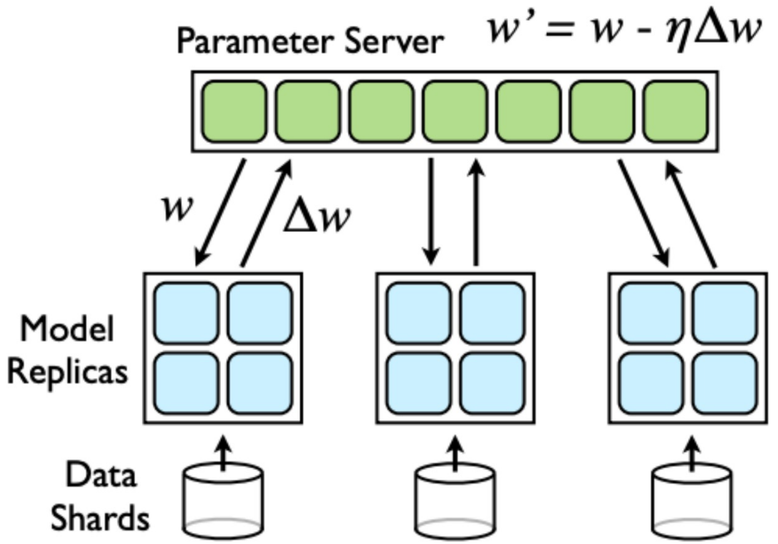
- Context 길이 제한이 있는 Feed-forward NNLM의 한계 극복
- Time delay를 사용해 Hidden layer에 연결함으로써 일종의 단기 메모리 형성
- Computational Complexity : $Q = H \times H + H \times V$

03

Word2Vec

Parallel Training of Neural Networks

- DistBelief 분산 프레임워크



동일한 모델의 여러 복제본을 병렬로 실행함에 따라 multi-core 연산 수행

2 Model Architectures

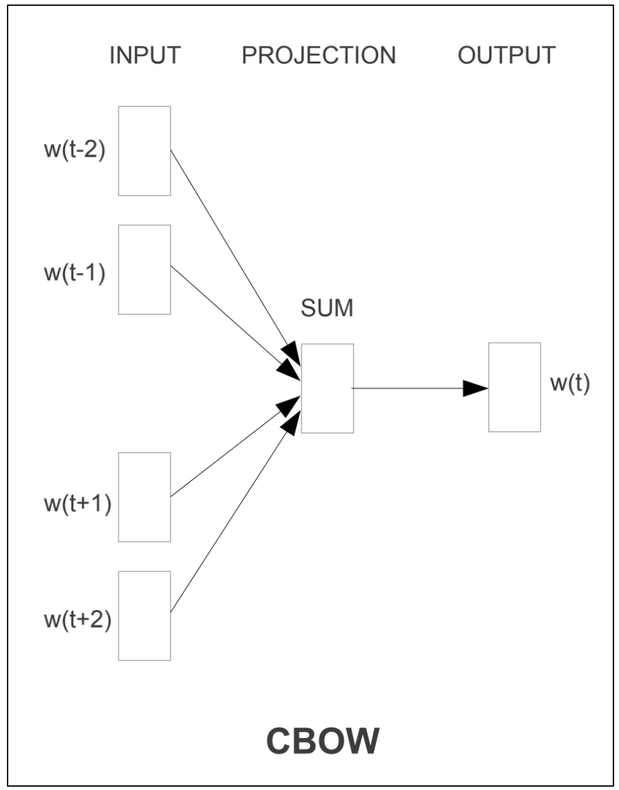
- 이전 세션의 Computational Complexity의 주 원인은 Nonlinear Hidden Layer
- New Model
  - Distributed representation of words를 기반으로 학습
  - Feed-forward NNLM의 Structure를 따름
  - 모든 단어는 같은 Position에 Projection 됨
  - New Model의 경우 Hierarchical Softmax를 적용함에 따라 Output unit 수를 V -> $\log_2 V$ 로 감소
- Structure
  - Continuous Bag-of-Words Model(CBOW)
  - Continuous Skip-gram Model(Skip-gram)

Continuous Bag-of-Words Model(CBOW)

- Structure

Example

The fat cat sat on the mat

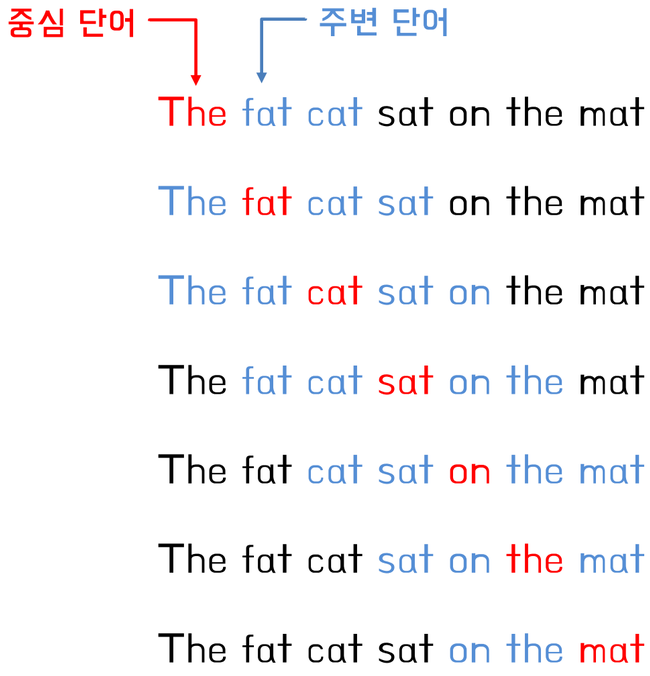


Notation

N : Input Layer로 들어가는 이전 단어의 개수
D : Projection 이후의 차원
V : Vocabulary Size

Continuous Bag-of-Words Model(CBOW)

- Structure



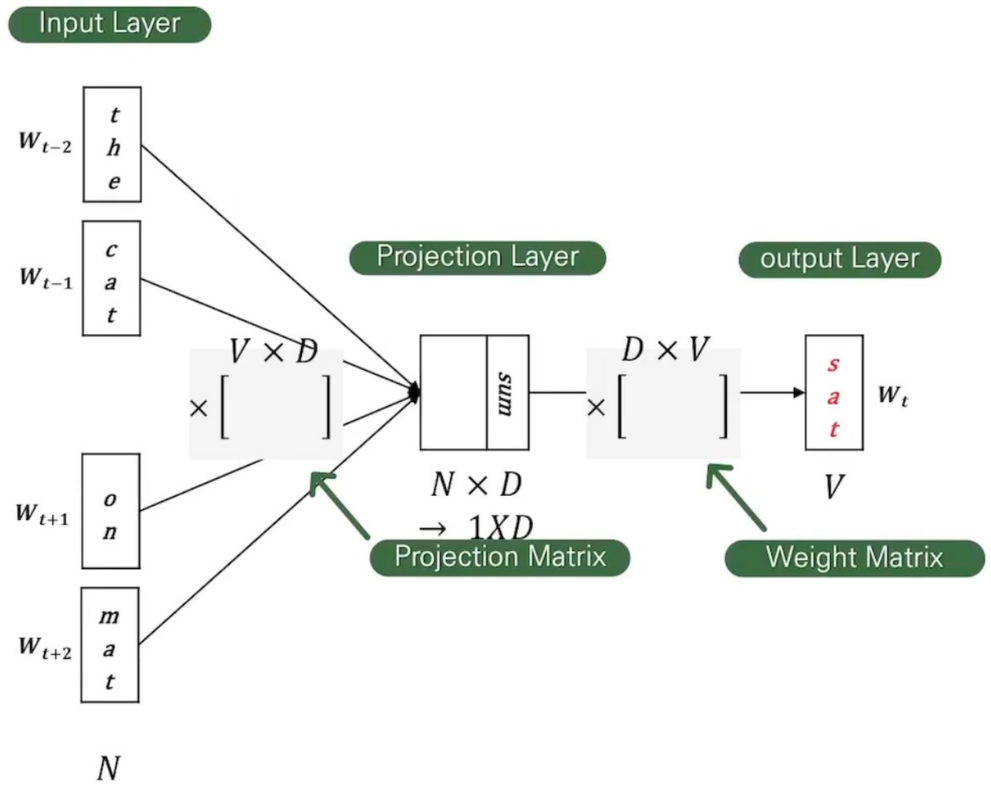
중심 단어	주변 단어
[1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0]
[0, , 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 1, 0, 0, 0]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

Continuous Bag-of-Words Model(CBOW)

- Structure

Example

The fat cat sat on the mat



Continuous Bag-of-Words Model(CBOW)

■ 목표

- 과거 단어와 미래 단어를 사용해 현재의 단어를 올바르게 추측하는 것

■ 특징

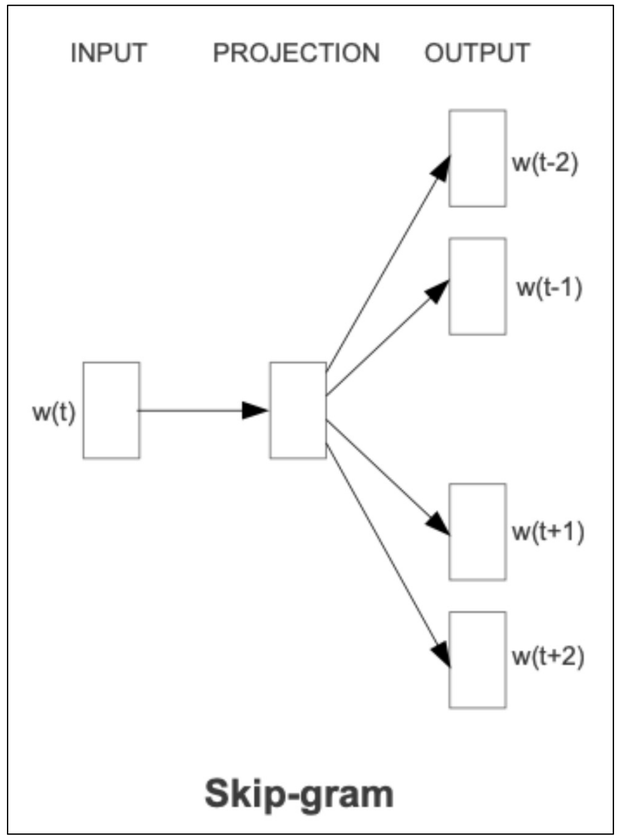
- Feed-forward NNLM와는 달리 미래 단어도 함께 사용
- Input으로 4개의 미래 단어와 4개의 단어 과거를 사용 시 최고의 결과를 얻음
- Computational Complexity : $Q = N \times D + D \times \log_2 V$

Continuous Skip-gram Model(Skip-gram)

- Structure

Example

The fat cat sat on the mat

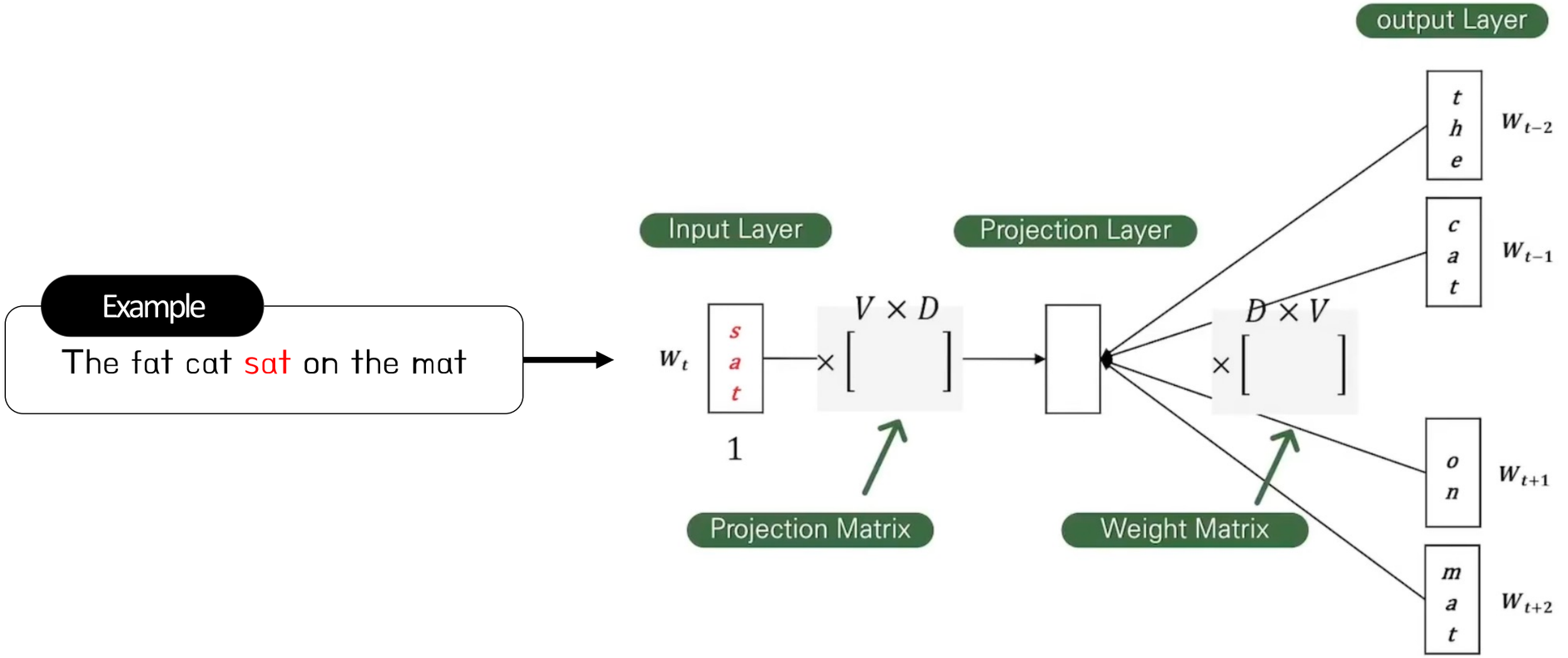


Notation

C : 단어의 최대 거리
D : Projection 이후의 차원
V : Vocabulary Size

Continuous Skip-gram Model(Skip-gram)

- Structure



Continuous Skip-gram Model(Skip-gram)

■ 목표

- 현재 단어를 사용해 동일한 문장의 다른 단어를 예측하는 것

■ 특징

- C(단어 사이의 최대 거리)를 선택하게 되면 각 훈련 단어에 대한 범위 <1, C> 내의 숫자 R을 무작위로 선택

  -> 이후 과거의 R 단어와 미래의 R 단어를 레이블로 선택

- 단어 사이의 거리가 멀어질수록 적은 가중치 부여

- Computational Complexity : $Q = C \times (D + D \times \log_2 V)$

04

Evaluation

Test set

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Maximization of Accuracy(CBOW)

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

어느 시점이 지나면 더 많은 차원을 추가하거나 더 많은 훈련 데이터를 추가하면 효과가 줄어드는 것을 확인

-> **벡터 차원과 훈련 데이터의 양을 함께 늘려야 한다**

Comparison of Model Architectures

Table 3: *Comparison of architectures using models trained on the same data, with 640-dimensional word vectors. The accuracies are reported on our Semantic-Syntactic Word Relationship test set, and on the syntactic relationship test set of [20]*

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Comparison of Model Architectures

Table 4: *Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.*

Model	Vector Dimensionality	Training words	Accuracy [%]		
			Semantic	Syntactic	Total
Collobert-Weston NNLM	50	660M	9.3	12.3	11.0
Turian NNLM	50	37M	1.4	2.6	2.1
Turian NNLM	200	37M	1.4	2.2	1.8
Mnih NNLM	50	37M	1.8	9.1	5.8
Mnih NNLM	100	37M	3.3	13.2	8.8
Mikolov RNNLM	80	320M	4.9	18.4	12.7
Mikolov RNNLM	640	320M	8.6	36.5	24.6
Huang NNLM	50	990M	13.3	11.6	12.3
Our NNLM	20	6B	12.9	26.4	20.3
Our NNLM	50	6B	27.9	55.8	43.2
Our NNLM	100	6B	34.2	**64.5**	50.8
CBOW	300	783M	15.5	53.1	36.1
Skip-gram	300	783M	**50.0**	55.9	**53.3**

Thank you

N. Hwang Hyeon Tae | E. gusxo3975@naver.com | L..linktr.ee/oneul_