Efficient Estimation Of Word Representations In Vector Space (Word2Vec)

2022.03.27.

발표자 : 김성윤

목차

- Introduction
  - Word Representation
  - Word Embedding
- Model Architecture (NNLM, RNNLM)
- Word2Vec
  - CBOW : Continuous Bag-Of-Words
  - Skip-gram : Continuous Skip-gram
- Tasks & Results

Introduction

Introduction

목적

- Continuous vector representation model “Word2Vec” 을 소개한다.
- 단어 유사성 task 로 기존 model architecture과 비교하여 성능을 확인한다.

필요성

- 연산량 대비 성능이 좋은 word embedding 모델
  - 계산 복잡도(computational complexity)가 크면 부담스럽다..
  - 훈련데이터를 무작정 많이 넣는 방법이 과연 최선일까?
  - 많은 데이터를 학습시켜서 High-quality word vector를 만들자.
- 인공신경망으로 구문과 의미(syntax & semantics)에 따른 유사성을 측정

Introduction

가정

- 유사한 단어는 서로 가까이에 존재할 것이다.
- 그리고 유사성에 계층이 여럿 존재할 것이다.

=> multiple degrees of similarity

단어 벡터 간 연산을 통해 예측이 가능하다!

vector("King") - vector("Man") + vector("Woman") ≈ vector("Queen")

따라서, 벡터 연산을 통해 정확도를 최대화하는 model architecture를 개발하였다.

들어가기 전에...

Syntatics (Syntax)

- 언어의 문법과 구조
- 언어 문장 내에 있는 구성요소의 순서

  "타당한 문장을 구성했는가?"

Semantics

- 문장 내 구성요소들의 의미

  "문장이 타당한가?"

Study of the quality of vector representations of word derived by various models
on a collection of *syntactic and semantic language tasks*.

Word Representation

- Sparse(희소), Local Representation
  - 더 높은 차원(higher dimension)

- 각 차원에 정보가 독립적으로 표현되는 방법
  - words as atomic units

- One-Hot Encoding

Word Embedding

- Dense, Distributed(분산) Representation
  - 더 적은 차원 (lower dimension)
  - 더 큰 일반화 능력(generalization power)

- 단어의 의미를 여러 차원에다가 분산하여 표현

- 모든 단어들을 이용하여 유사성(관계)를 알아낼 수 있는 특징적인 표현을 학습

Model Architecture

Model Architecture

- Continuous representation of words – LSA, LDA
- Distributed representations of words learned by neural networks

- Comparing model architectures

  첫째, 모델의 계산복잡도는 '모델을 충분히 훈련시키기 위해 필요한 파라미터의 수'로 정의한다.

  둘째, 모델의 계산복잡도를 최소화하면서 정확도를 최대화하는 것을 목표로 한다.

- Training Complexity

$$O = E \times T \times Q$$

E : training epoch의 수
T : 전체 training set의 단어 수
Q : model architecture 별로 상이하게 정의

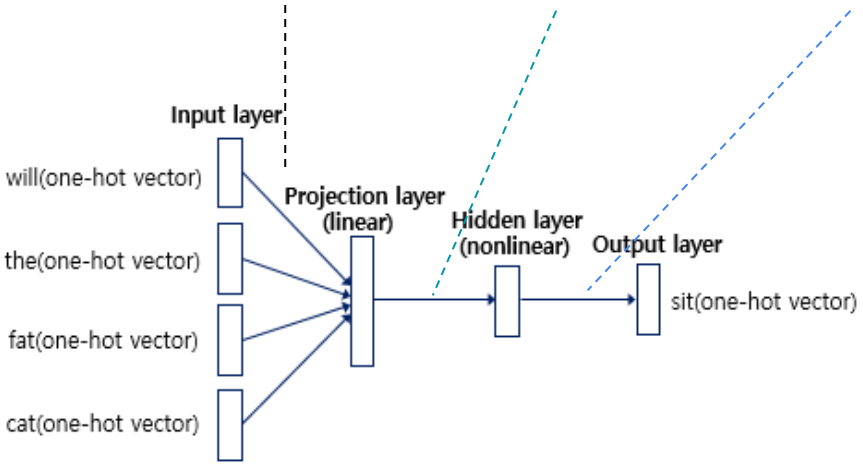
Feedforward Neural Net Language Model (NNLM) =NPLM

- single-word context

- Computational complexity

$$Q = N \times D + N \times D \times H + H \times V$$

N : input 단어 개수
D : dimension
H : Hidden layer 크기
V : vocabulary 크기



- Dominating term “H x V”
  - 해결방법 : 연산량 감소
    - hierarchical softmax (Huffman binary tree)
    - 훈련 시 정규화하지 않는 모델

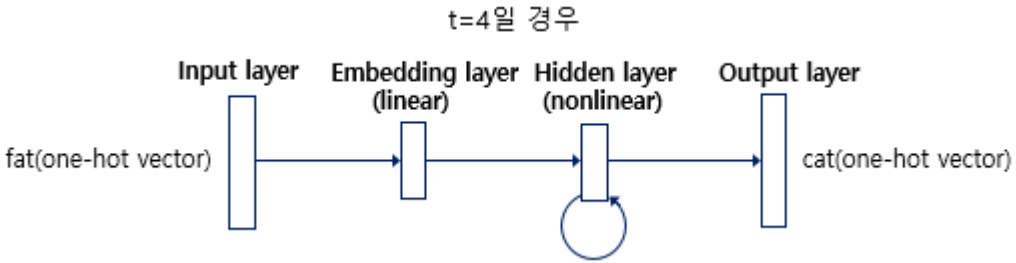
Recurrent Neural Net Language Model (RNNLM)

- NNLM의 한계점
  Specify context length(order of the model N)

- NNLM에서 projection layer를 제거
- Recurrent matrix - Short term memory

- Computational complexity

$$Q = H \times H + H \times V$$



H : Hidden layer 크기
V : vocabulary 크기

(단, word representation D는 hidden layer H와 차원의 크기가 같다)

Word2Vec

계산 비용 최소화
distributed word vector 학습

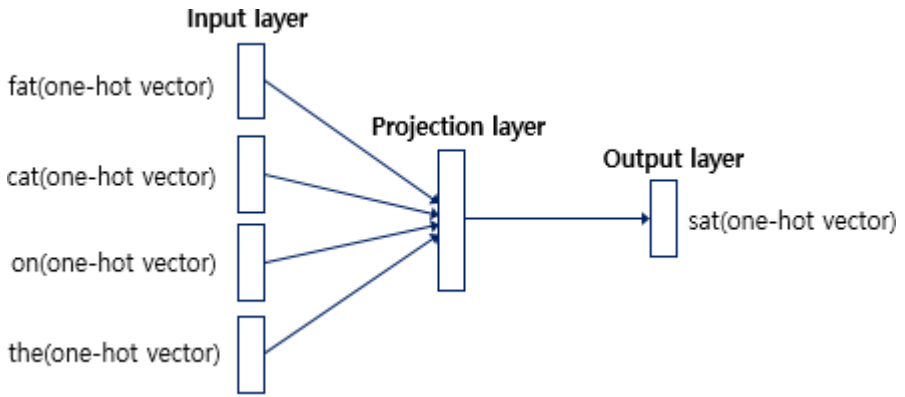
★ Continuous Bag-of-Words Model (CBOW)

- multi-word context
- NNLM에서 non-linear hidden layer를 제거, projection layer를 공유

- 주변 단어들을 확인하여 중심의 target 단어를 예측한다.
- 각 word vector들을 평균내서 사용하기 때문에 등장 빈도가 낮은 word들은 제대로 된 학습을 기대하기 힘들다.
  (단어의 순서가 projection에 영향을 주지 않는 이유)

- Training complexity

$$Q = N \times D + D \times log_2(V)$$



★ Continuous Skip-gram Model (Skip-gram)

- 중심 단어에서 주변 target 단어를 예측한다.
  - 여러 word에 대해 예측을 수행해서 연산량은 CBOW보다 많지만, 그만큼 중심 단어의 학습 기회가 많아서 학습 효과가 우수하다.

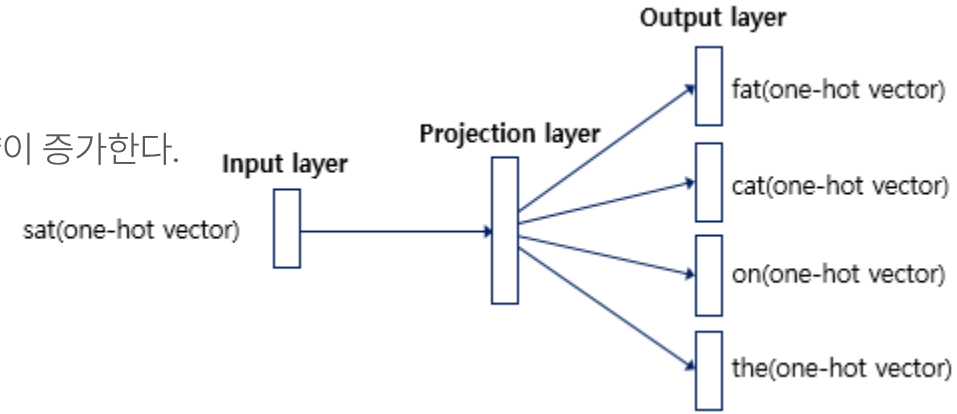
- Training complexity

$$Q = C \times (D + D \times log_2(V))$$

C : predict할 word와 현재 word의 최대 거리

- Distance가 멀수록 weight가 작다

- 범위 R을 증가시키면 성능이 향상되지만 연산량이 증가한다.
  - <1:C> 사이의 무작위 숫자 R를 선정
  - Current word 에 대한 과거와 미래의 R를 '정답'
  - R x 2 → word classification
  - Current word → input
  - R+R words → output



$$Q = N \times D + D \times log_2(V)$$

$$Q = C \times (D + D \times log_2(V))$$

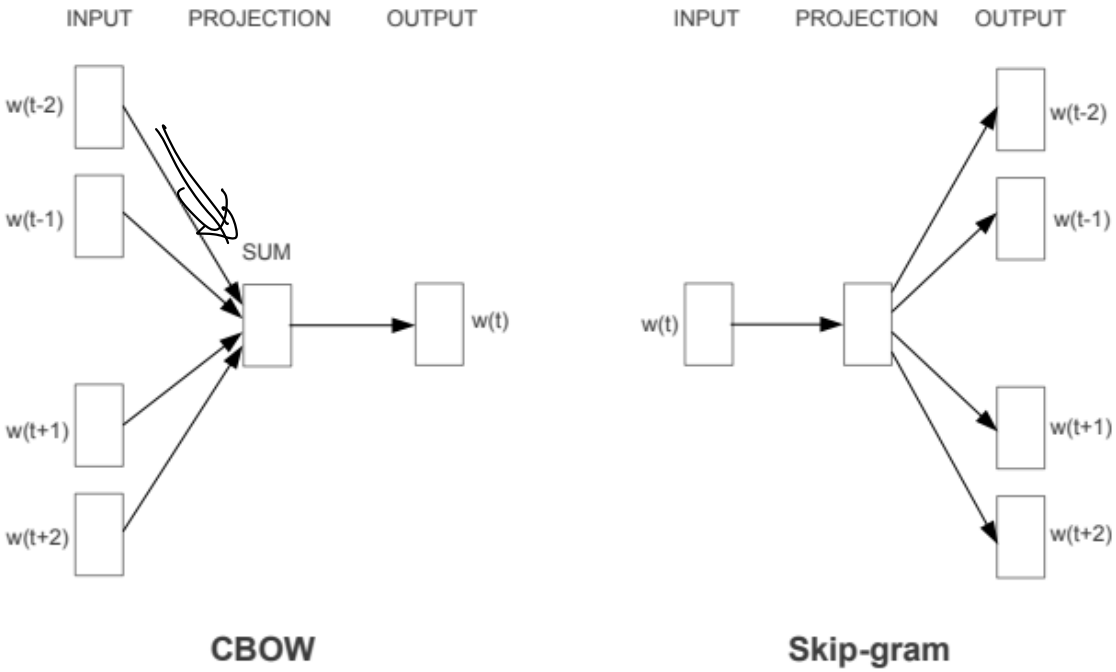
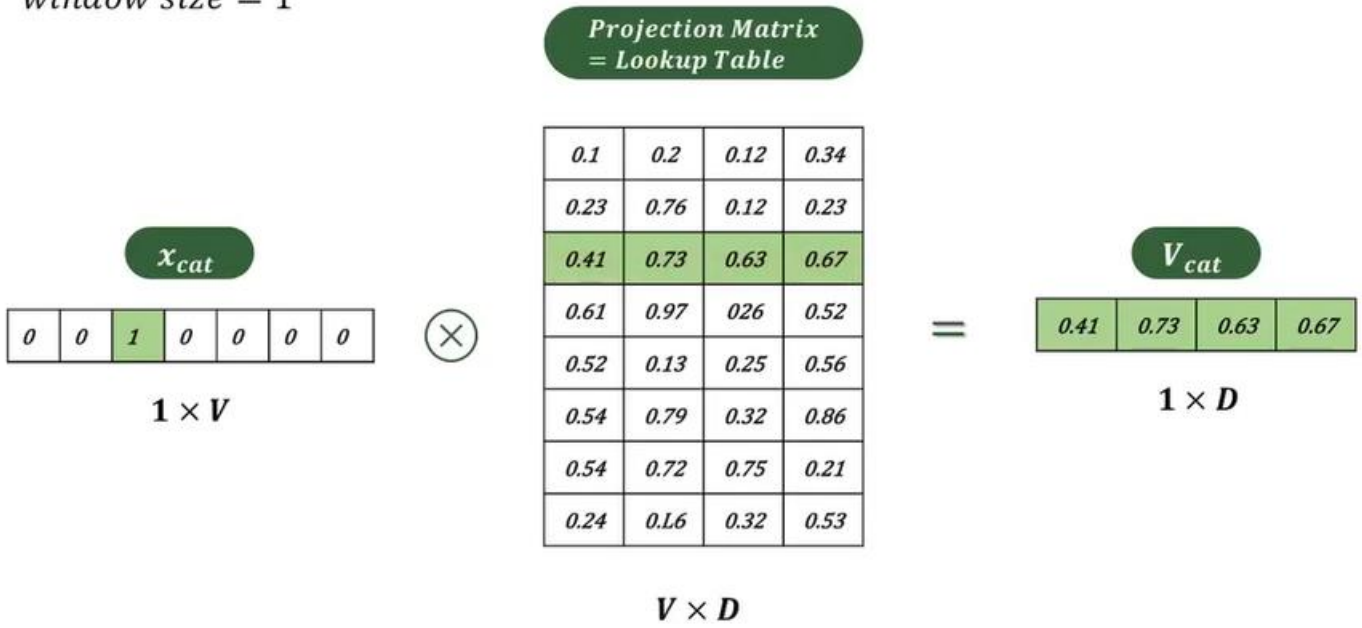
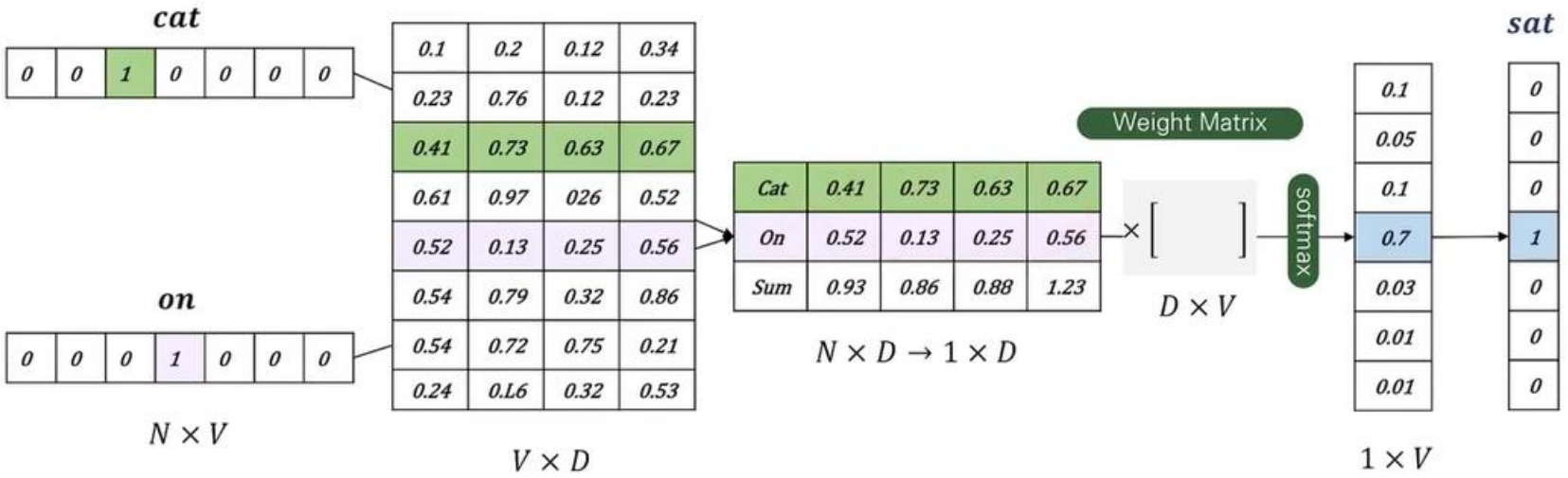


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

- $window\ size = 1$



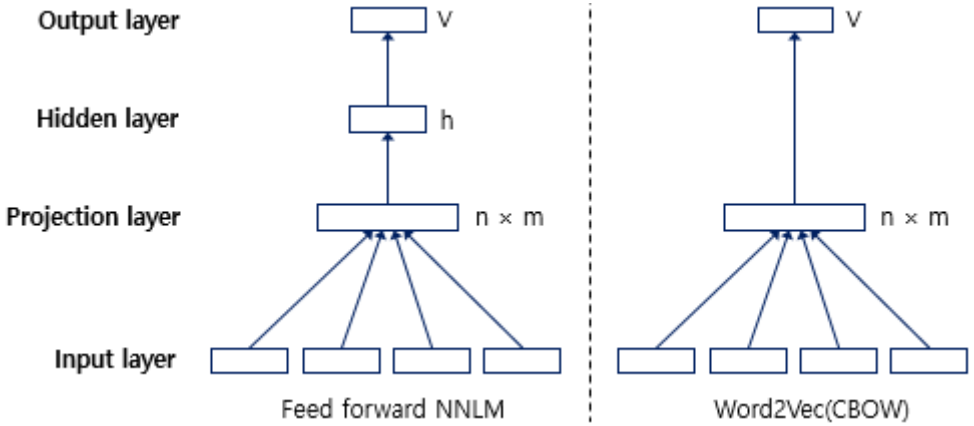
- $window\ size = 1$



NNLM vs CBOW

- Input ~ Projection layer : weight matrix가 동일하다

- Hidden layer의 필요성
  - NNLM - 여러 word embedding vector를 하나의 vector로 압축
  - CBOW - word embedding vector를 non-linear layer를 거치지 않고 단순 평균



Tasks & Results

4.1 Task Description

- Test Set
  Semantic – Syntactic
  Word Relationship

- Accuracy Evaluation
  - 모든 질문 유형에 대해
  - 각각의 질문 유형에 대해
  - 완벽하게 일치할 때만 정답처리

** 단어 형태론(단어구조와 구성요소, 형성 규칙)에 대한 정보는 누락됨

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

	Type of relationship	Word Pair 1		Word Pair 2	
Semantic	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
Syntactic	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

http://www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt

4.1 Task Description

```
// Copyright 2013 Google Inc. All Rights Reserved.
: capital-common-countries
Athens Greece Baghdad Iraq
Athens Greece Bangkok Thailand
Athens Greece Beijing China
Athens Greece Berlin Germany
Athens Greece Bern Switzerland
Athens Greece Cairo Egypt
Athens Greece Canberra Australia
Athens Greece Hanoi Vietnam
Athens Greece Havana Cuba
Athens Greece Helsinki Finland
Athens Greece Islamabad Pakistan
Athens Greece Kabul Afghanistan
Athens Greece London England
Athens Greece Madrid Spain
Athens Greece Moscow Russia
Athens Greece Oslo Norway
Athens Greece Ottawa Canada
Athens Greece Paris France
Athens Greece Rome Italy
Athens Greece Stockholm Sweden
Athens Greece Tehran Iran
Athens Greece Tokyo Japan
```

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship 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

http://www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt

4.2 Maximization of Accuracy

Table 2: *Accuracy on subset of the Semantic-Syntactic Word Relationship test set, using word vectors from the CBOW architecture with limited vocabulary. Only questions containing words from the most frequent 30k words are used.*

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

벡터 차원과 훈련 데이터의 양을 동시에 증가시켜야 한다.

We have to increase both **vector dimensionality** and the **amount of the training data** together.

- 중요한 시사점인 이유?

4.3 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

(1)

(2)

(3)

4.3 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**

4.3 Comparison of Model Architectures

Table 5: *Comparison of models trained for three epochs on the same data and models trained for one epoch. Accuracy is reported on the full Semantic-Syntactic data set.*

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days]
			Semantic	Syntactic	Total	
3 epoch CBOW	300	783M	15.5	53.1	36.1	1
3 epoch Skip-gram	300	783M	50.0	55.9	53.3	3
1 epoch CBOW	300	783M	13.8	49.9	33.6	0.3
1 epoch CBOW	300	1.6B	16.1	52.6	36.1	0.6
1 epoch CBOW	600	783M	15.4	53.3	36.2	0.7
1 epoch Skip-gram	300	783M	45.6	52.2	49.2	1
1 epoch Skip-gram	300	1.6B	52.2	55.1	53.8	2
1 epoch Skip-gram	600	783M	56.7	54.5	55.5	2.5

- Data size 2배 > epoch 횟수 증가
- Train data size 2배 < vector 차원 2배

4.5 Microsoft Research Sentence Completion Challenge

- Semantic Data Model
- Main task : 빈 칸에 알맞은 단어 찾기
  - 1040개의 문장 / 한 문장 당 한 단어 / 5지선다
- Performance of Skip-gram architecture [Table 7, pg 9]

Was she his [client | musings | discomfiture | choice | opportunity] , his friend , or his mistress?

All red-headed men who are above the age of [800 | seven | twenty-one | 1,200 | 60,000] years , are eligible.

That is his [generous | mother's | successful | favorite | main] fault , but on the whole he's a good worker.

https://www.microsoft.com/en-us/research/project/msr-sentence-completion-challenge/

Examples of the Learned Relationships

- Improve performance
  - Word vectors trained on even larger data sets with larger dimensionality
  - Provide more than one example of relationship

- Apply vector operations to solve different tasks
  - Selecting out-of-the-list words

Conclusion

Study of the quality of vector representations of words derived by various models on a collection of syntactic and semantic language tasks.

- Applications
  - NLP tasks – sentiment analysis(감성분석), paraphrase detection(의역예측)
  - Automatic extension of facts in Knowledge Bases
  - Verification of correctness of existing facts
  - Machine translation(기계번역)

Related Papers

GloVe

GloVe: Global Vectors for Word Representation (2014)

FastText

Enriching Word Vectors with Subword Information (2017)