Efficient Estimation of Word Representations in Vector Space

https://arxiv.org/abs/1301.3781

0. Introduction

- 자연어 처리(NLP)에서 단어 의미를 효율적으로 표현하는 방법의 필요성이 커지고 있음
- 기존의 단어 표현 방식(One-hot, LSA)은 계산 비용이 크고 의미 표현이 제한적임
- 단어를 저차원 벡터로 임베딩하는 효율적인 방법을 제안
- Word2Vec의 Continuous Bag-of-Words (CBOW) 와 Skip-Gram 모델 제안

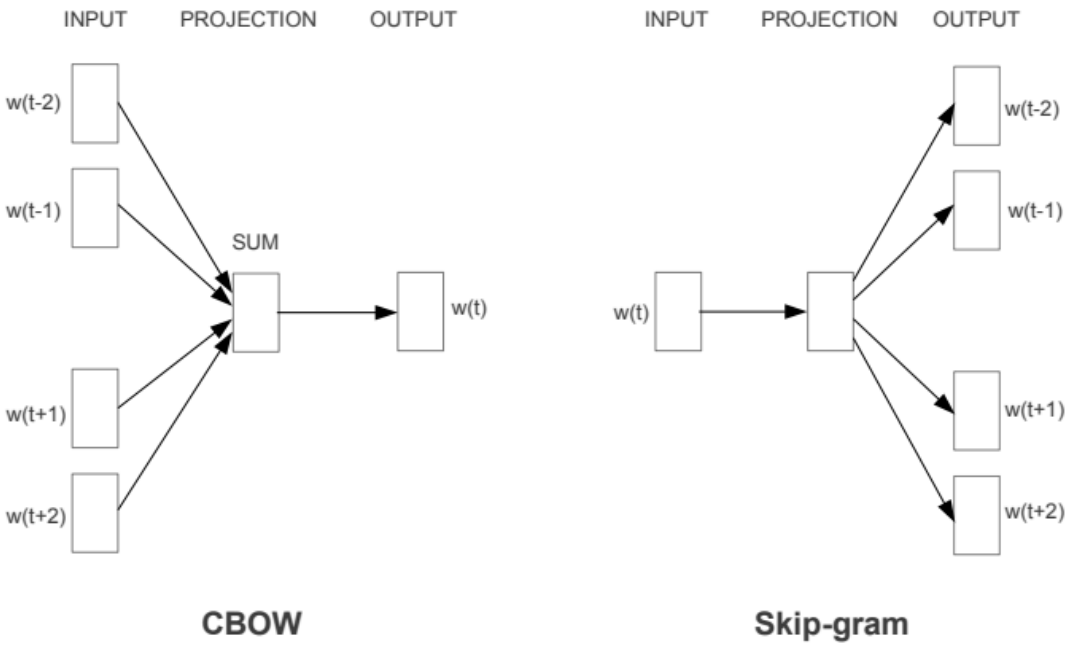
1. Overview

- 목표 : 대규모 말뭉치(corpus)에서 단어 벡터를 효율적으로 학습
- Word2Vec : 단어의 분포 정보를 벡터 공간에 보존하는 방식
- 특징 : 빠른 학습 속도, 낮은 계산 비용, 뛰어난 의미 표현 능력
- 주요 응용 : 단어 유사도 측정, 문서 분류, 기계 번역 등

2. Challenges

- 대규모 말뭉치 처리 시 계산 비용이 매우 높음
- 기존 방법(LSA, NNLM)의 학습 속도가 느림
- 단어 의미를 잘 보존하면서 효율성을 유지하는 것이 어려움

3. Method



- Continuous Bag-of-Words (CBOW) : 주변 단어(context)를 기반으로 중심 단어(target) 예측
- Skip-Gram : 중심 단어(target)로 주변 단어(context) 예측
- Optimization : Hierarchical Softmax, Negative Sampling 도입 → 계산 효율 향상
- 구현 : 간단한 신경망 구조로 빠른 학습 가능

4. Experiments

- 데이터셋 : Google News Corpus (약 100B 단어)
- 비교 baseline : LSA, NNLM 등
- 평가 지표 : 단어 유사도, 벡터 연산(예: king − man + woman ≈ queen)
- 다양한 벡터 차원(dimension), window size, negative samples 실험

5. Results

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

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

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**

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

- Word2Vec은 기존 방법 대비 계산 속도가 수십 배 빠름

- 의미 관계 보존 능력이 뛰어나며, 단어 간 유사도를 잘 반영
- 벡터 연산을 통한 의미 추론 가능성 입증
- CBOW는 학습 속도가 빠르고, Skip-Gram은 희소 단어에 강함

6. Insight

- Word2Vec은 단어 의미 표현 방식의 혁신적 전환점
- NLP에서 단어 임베딩 표준으로 자리 잡음
- 이후 GloVe, FastText, Transformer 기반 임베딩 등 다양한 발전의 기반
- 핵심은 효율성과 의미 보존의 균형