AI ROBOTICS KR : NLP STUDY

박병준

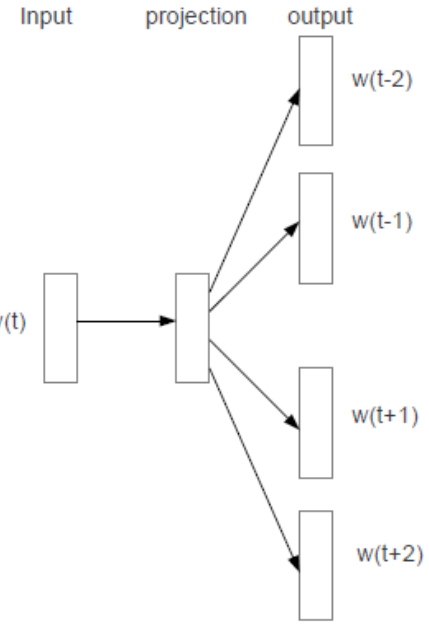
Distributed Representations of Words and Phrases and their Compositionality

목차

요약

Skip-gram 모델: 의미적 단어 관계를 포착하는 고품질 분산 벡터 표현을 학습하기 위한 효율적인 방법

- 대부분의 뉴럴 네트워크 아키텍처와 달리 dense matrix 곱셈이 필요하지 않다.
- 매우 효율적



벡터의 품질과 훈련 속도를 향상시키는 몇 가지 모델의 확장

- 서브 샘플링: 상당한 속도를 얻고 또한 더 많은 정규 단어 표현을 학습
- 네거티브 샘플링: 계층적 소프트맥스 대안을 설명

단어표현의 본질적인 제한을 극복

- 단어순서 표현
- 관용구를 표현

Skip-gram 모델

Skip-gram 모델의 훈련 목적은 문장이나 문서에서 주변 단어를 예측하는 데 유용한 단어 찾는 것

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

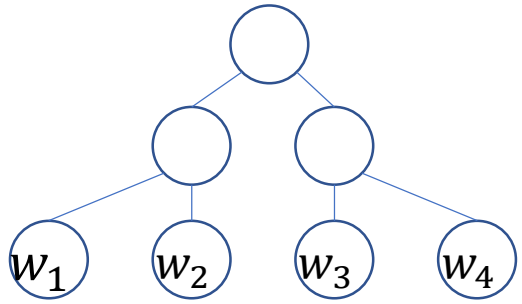
- T: 학습시킨 단어 수
- t: 학습중인 단어
- c: 훈련 문맥 크기(클 수록 정확도 향상, 훈련 시간 증가)

$$p(w_O|w_I)=\frac{\exp\left({v'_{w_O}}^{\top}v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\top}v_{w_I}\right)}$$

- W: vocabulary의 단어 수

하지만 이 식은 보통 W가 10^5~10^7정도이므로 비현실적이다.

Skip-gram 모델: Hierarchical Softmax



계층적 소프트맥스는 전체 소프트맥스의 근사치
장점은 확률 분포를 얻기 위해 $\log_2(W)$ 노드에 대해서만 평가

계층적 소프트맥스는 W 워드를 잎으로 하는 출력 계층의 이진 트리 표현
하위 노드의 상대적 확률을 명시적으로 나타냄
단어에 확률을 무작위로 할당하여 정의

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma\left([\![n(w,j+1) = \mathrm{ch}(n(w,j))]\!] \cdot {v'_{n(w,j)}}^{\top} v_{w_I}\right)$$

더 정확히 말하면, 각 단어 w는 나무의 뿌리에서 적절한 경로로 도달할 수 있다.

- n(w, j)가 루트부터 w까지 경로의 j번째 노드, n(w, 1) = root
- L(w)가 이 경로의 길이, n(w,L(w)) = w
- 내부 노드 n의 경우 ch(n)가 n의 임의 고정 자식
- [x]가 참이면 1, 거짓이면 -1

Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252, 2005.

Skip-gram 모델: Negative Sampling

Noise Contrastive Estimation(NCE)은 계층적 소프트맥스의 대안
좋은 모델은 Logistic regression의 평균에 의한 잡음이 데이터를 구별할 수 있음을 가정
hinge loss와 비슷, 위 잡음으로 데이터 순위를 매기고 모델을 학습

NCE는 소프트맥스의 로그 확률을 대략적으로 최대화하는 반면
Skip-gram 모델은 고품질 벡터 표현을 배우는 것에 집중
→벡터 표현이 품질을 유지하는 한 NCE를 자유롭게 단순가능.
→ → Skip-gram의 모든 로그 P(wO|wI) 를 대체하는 데 음성 샘플링(NEG)을 정의

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

각 데이터 샘플에 대해 k개의 네거티브 샘플에서 로지스틱 회귀 분석을 사용하여 대상 단어 wO와 노이즈 분포 Pn(w)를 구별하는 것이 과제다.
5-20 범위의 k 값은 소규모 훈련 데이터 집합에 유용, 대규모 데이터 집합의 경우 k는 2-5만큼 작을 수 있음

NCE가 샘플과 소음 분포의 수치적 확률을 모두 필요로 하는 반면 음의 샘플링은 샘플만 사용
NCE는 소프트맥스의 로그 확률을 대략 최대화하지만, 이 특성은 여기 적용에 중요하지 않다.

Noise Contrastive Estimation (NCE): Michael U Gutmann and Aapo Hyv¨arinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics.

Skip-gram 모델: Subsampling of Frequent Words

큰 코퍼스에서 빈번한 단어가 보통 희귀한 단어들보다 정보적 가치를 적게 제공
희귀한 단어와 빈번한 단어의 불균형을 해결하기 위해 서브 샘플링 사용

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

- P(wi)는 버려질 단어 wi의 확률
- t는 상수(보통 10^-5)
- f(wi)는 단어 wi의 빈도

빈도의 순위를 유지하면서 주파수가 t보다 큰 단어를 공격적으로 서브 샘플링
서브샘플링 공식은 휴리스틱하게 선택
학습을 가속화하고 희귀한 단어의 학습된 벡터의 정확도를 상당히 향상

결과

Hierarchical Softmax, Noise Contrastive Estimation, Negative Sampling, subsampling 실험
평가방법: 유추(아래 논문 참고)

Germany" : "Berlin" :: "France" : ?

vec("Berlin") - vec("Germany") + vec("France")

구문론적 유추 "quick" : "quickly" :: "slow" : "slowly"
의미론적 유추, such as the country to capital city relationship

데이터 전처리

- 뉴스 기사(10억 단어의 내부 구글 데이터셋)
- 5번 미만으로 발생한 모든 단어 지움
- 데이터 전처리 후 크기: 692,000개 어휘

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

결과

네거티브 샘플링이 계층적 소프트맥스를 능가하며, 노이즈 대비 추정치보다 성능이 더 우수
서브샘플링은 훈련 속도를 여러 번 향상시키고 단어 표현을 훨씬 더 정확하게 만듬

Method	Time [min]	Syntactic [%]	Semantic [%]	Total accuracy [%]
NEG-5	38	63	54	59
NEG-15	97	63	58	**61**
HS-Huffman	41	53	40	47
NCE-5	38	60	45	53
The following results use 10^{-5} subsampling				
NEG-5	14	61	58	60
NEG-15	36	61	61	**61**
HS-Huffman	21	52	59	55

표준 시그모이드 RNN으로 학습한 벡터가 훈련 데이터 양이 증가함에 따라 유의미하게 개선됨을 보여줌
으로 비선형 모델도 단어 표현의 선형 구조에 대한 선호를 가지고 있음을 보여줌

관용구 학습

많은 구들은 그 각 단어의 의미의 합이 아님
- Korean Air → 한국의 공기(X) 대한한공(O)

구문에 대한 벡터 표현을 배우려면 여러 문맥을 보며 함께 자주 or 함께 안 나타나는 단어를 찾음
- This is New York Times.
- This is Toronto Maple Leaf.

이런 방식으로 어휘의 크기를 크게 늘리지 않고도 합리적인 구절을 많이 생산

n-gram을 사용하여 Skip-gram 모델을 훈련시킬 수 있지만, 너무 메모리 집약적
So, 데이터 중심 접근법을 사용하기로 결정, (유니그램과 바이그램 카운트를 기반)

관용구 학습: Phrase Skip-Gram Results

동일한 데이터로 먼저 구를 기반으로 훈련 말뭉치를 제작
여러 개의 Skip-gram 모델을 훈련
전과 같이 벡터 300차원과 컨텍스트 크기 5를 사용

Method	Dimensionality	No subsampling [%]	10^{-5} subsampling [%]
NEG-5	300	24	27
NEG-15	300	27	42
HS-Huffman	300	19	**47**

- 약 330억 단어의 데이터 세트, 계층적 소프트맥스, 벡터 1000차원, 컨텍스트로 문장 전체 이용했을 때 72%의 정확도
- 데이터셋을 60억단어로 줄였을 때 정확도가 66%로 감소

→훈련 데이터의 양이 매우 중요

	NEG-15 with 10^{-5} subsampling	HS with 10^{-5} subsampling
Vasco de Gama	Lingsugur	Italian explorer
Lake Baikal	Great Rift Valley	Aral Sea
Alan Bean	Rebbeca Naomi	moonwalker
Ionian Sea	Ruegen	Ionian Islands
chess master	chess grandmaster	Garry Kasparov

기존 단어 표현과 비교

Collobert and Weston , Turian et al. ,Mnih and Hinton 과 비교

Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohona karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating