집 현 전 - 송 성 희

Distributed Representations of Sentences and Documents

01 Introduction

논문 선정 이유

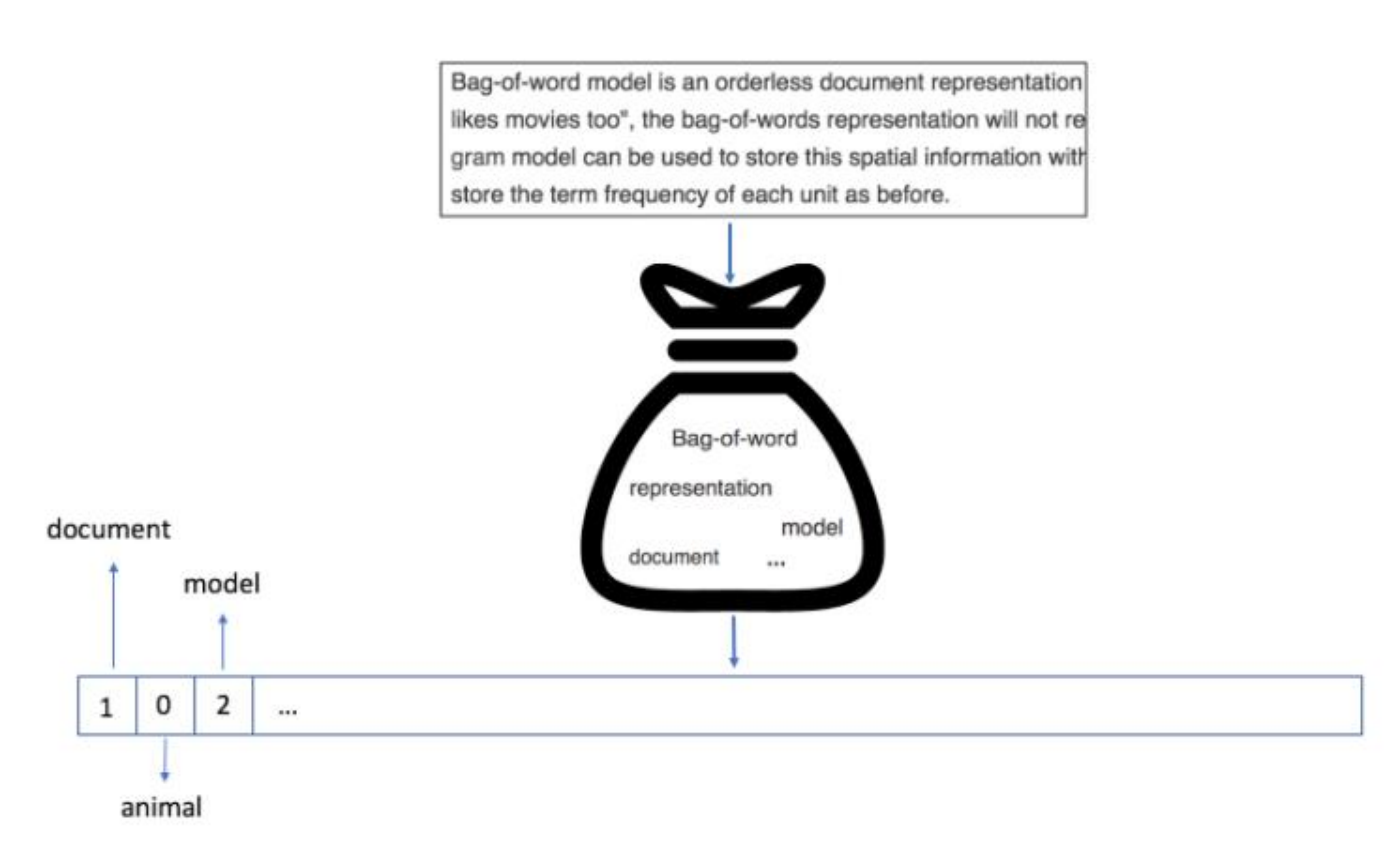
Sentiment Analysis에 적용할 때 낮은 error rate

Classification, 문서 유사도, 정보 검색 등에서 유용하게 사용

LDA2Vec, Live2Vec등 다양한 응용 가능

01 Introduction

Bag-of-words (BOW) : 단어의 순서를 고려하지 않음



01 Introduction

Bag-of-grams : 순서를 고려하지만 고차원일수록 이해도가 복잡

bag of words : [hello, world] , [world, hello]

bi-gram : [hello world], [world hello]

Tri-gram : [hello world python]

01 Introduction

Bag-Of-Words와 Bag-Of-Grams의 단점 :

“powerful,” “strong” and “Paris” are equally distant despite the fact that semantically, “powerful” should be closer to “strong” than “Paris.”

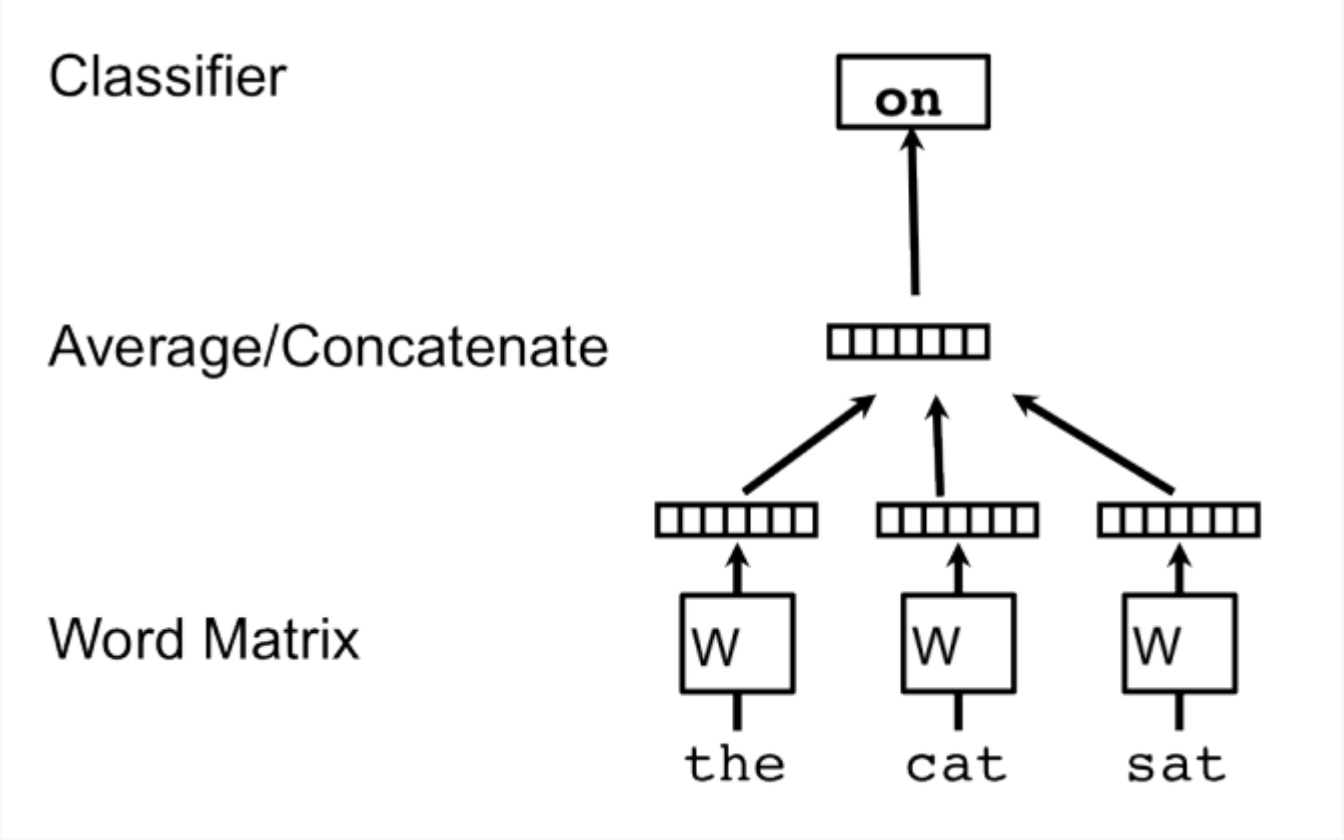
Paragraph Vector – BOW와 Bag-of-grams의 단점 보완을 위해 만들어진 기법

Paragraph를 Vector로 변환하는 방법

단어들의 순서가 저장되어 있는 Memory의 역할을 함

02 Algorithm

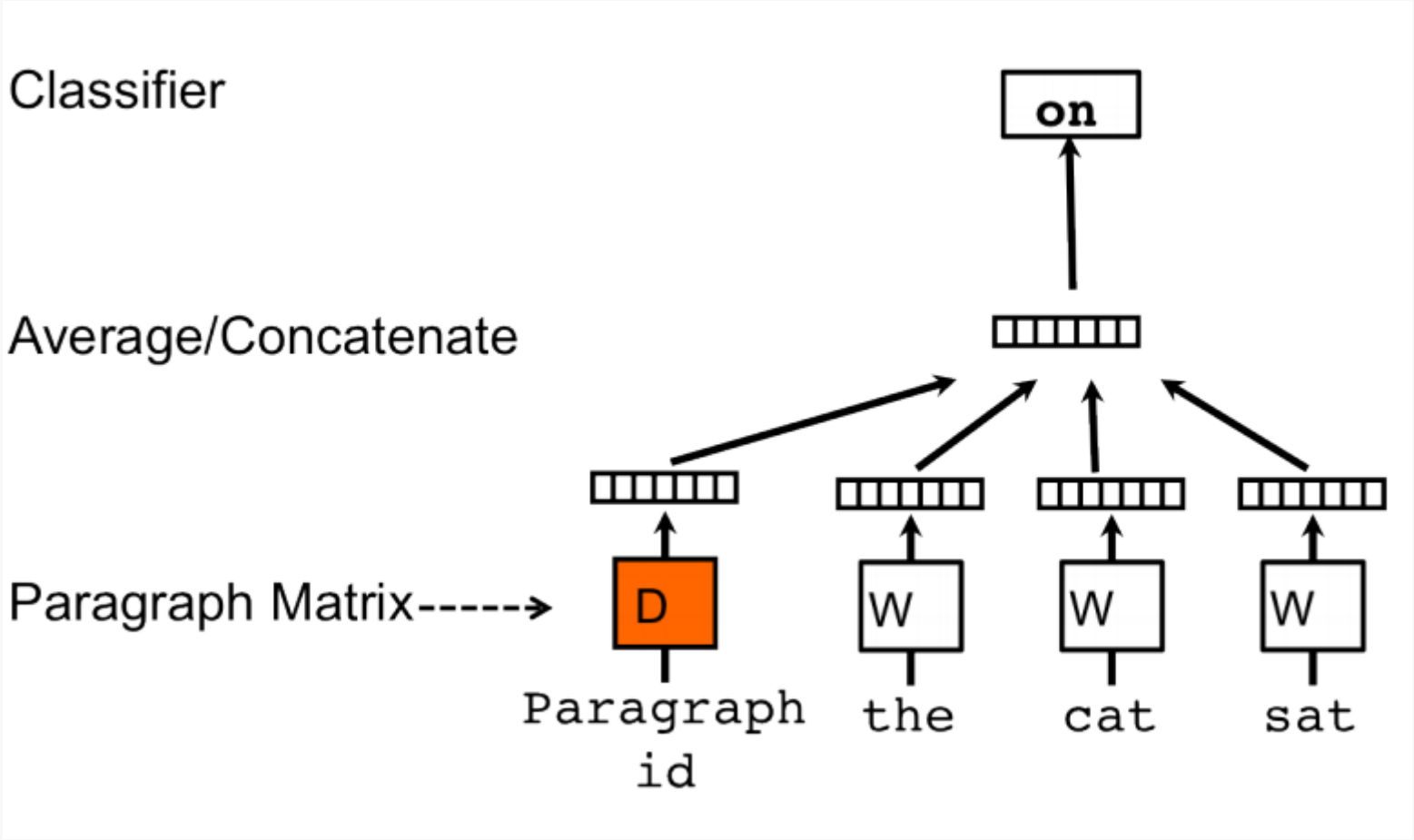
"The cat sat on the table"



W1 : the, W2: cat, W3: sat -> on

02 Algorithm

ParagraphVector-DistributedMemory (PV-DM)



02 Algorithm

"동구 밖 과수원길 아카시아 꽃이 활짝 폈네 "

Paragraph Matrix

ID	Paragraph
1	동구 밖 과수원길 아카시아 꽃이 활짝 폈네

Word Matrix

ID	Word
1	동구
2	밖
3	과수원길
4	아카시아
5	꽃이
6	활짝
7	폈네

02 Algorithm

EX) Window가 3일 경우

Paragraph Matrix

ID	Paragraph
1	동구 밖 과수원길 아카시아 꽃이 활짝 폈네

Word Matrix

ID	Word
1	동구
2	밖
3	과수원길

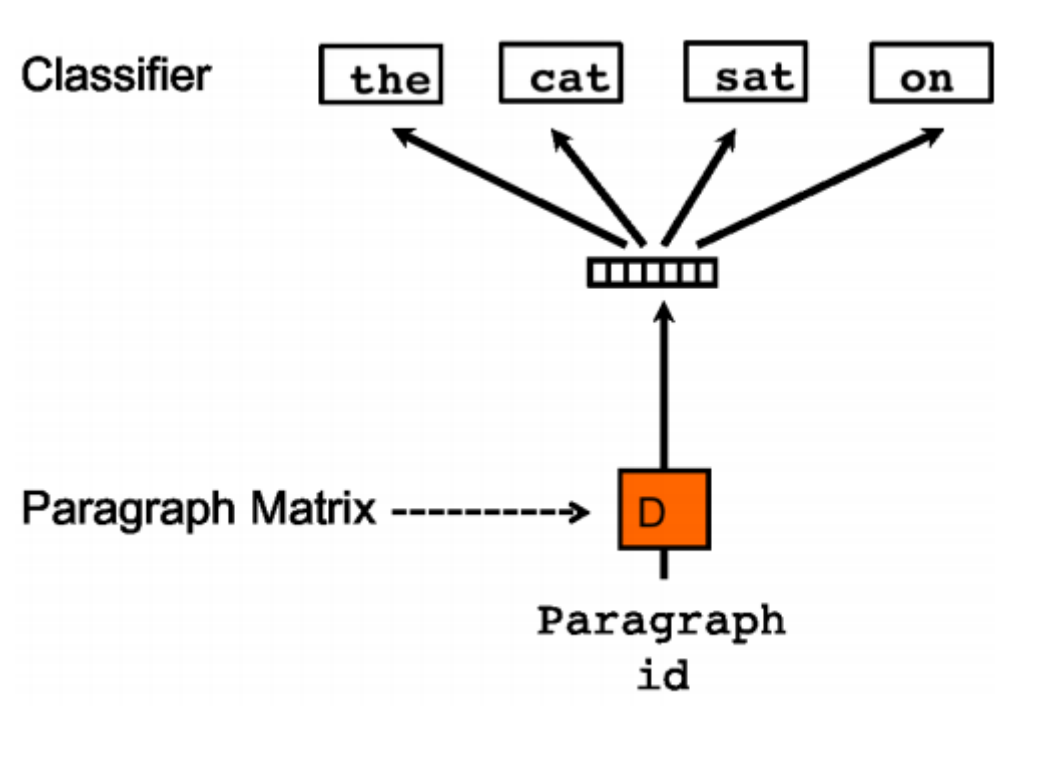
Softmax 함수

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

4	**아카시아**

02 Algorithm

ParagraphVector-DistributedBagOfWords(PV-DBOW) : PV-DM보다 속도가 빠름



워드 벡터를 필요하지 않음

03 Result

Dataset: Stanford Sentiment Treebank Dataset

the movie review site **Rotten Tomatoes** : 11855 문장

Training set : 8544 문장

Validation set : 1101 문장

Test set : 2210 문장

From very negative to very positive in the scale from 0.0 to 1.0.

The labels are generated by human annotators using Amazon Mechanical Turk.

Table 1. The performance of our method compared to other approaches on the Stanford Sentiment Treebank dataset. The error rates of other methods are reported in (Socher et al., 2013b).

Model	Error rate (Positive/ Negative)	Error rate (Fine-grained)
Naïve Bayes (Socher et al., 2013b)	18.2 %	59.0%
SVMs (Socher et al., 2013b)	20.6%	59.3%
Bigram Naïve Bayes (Socher et al., 2013b)	16.9%	58.1%
Word Vector Averaging (Socher et al., 2013b)	19.9%	67.3%
Recursive Neural Network (Socher et al., 2013b)	17.6%	56.8%
Matrix Vector-RNN (Socher et al., 2013b)	17.1%	55.6%
Recursive Neural Tensor Network (Socher et al., 2013b)	14.6%	54.3%
Paragraph Vector	**12.2%**	**51.3%**

03 Result

IMDB dataset

The dataset consists of 100,000 movie reviews taken from IMDB.

데이터 구성

25,000 labeled training instances,
25,000 labeled test instances
50,000 unlabeled training instances.

two types of labels: Positive and Negative

Table 2. The performance of Paragraph Vector compared to other approaches on the IMDB dataset. The error rates of other methods are reported in (Wang & Manning, 2012).

Model	Error rate
BoW (bnc) (Maas et al., 2011)	12.20 %
BoW (bΔt'c) (Maas et al., 2011)	11.77%
LDA (Maas et al., 2011)	32.58%
Full+BoW (Maas et al., 2011)	11.67%
Full+Unlabeled+BoW (Maas et al., 2011)	11.11%
WRRBM (Dahl et al., 2012)	12.58%
WRRBM + BoW (bnc) (Dahl et al., 2012)	10.77%
MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector	**7.42%**

03 Result

Paragraph 1: calls from (000) 000 - 0000 . 3913 calls reported from this number . according to 4 reports the identity of this caller is american airlines .

Paragraph 2: do you want to find out who called you from +1 000 - 000 - 0000 , +1 0000000000 or (000) 000 - 0000 ? see reports and share information you have about this caller

Paragraph 3: allina health clinic patients for your convenience , you can pay your allina health clinic bill online . pay your clinic bill now , question and answers...

Table 3. The performance of Paragraph Vector and bag-of-words models on the information retrieval task. "Weighted Bag-of-bigrams" is the method where we learn a linear matrix W on TF-IDF bigram features that maximizes the distance between the first and the third paragraph and minimizes the distance between the first and the second paragraph.

Model	Error rate
Vector Averaging	10.25%
Bag-of-words	8.10 %
Bag-of-bigrams	7.28 %
Weighted Bag-of-bigrams	5.67%
Paragraph Vector	**3.82%**

04 Discussion

PV-DM이 PV-DBOW보다 일관되게 우수한 결과를 나타냄

PV-DM에서 concatenation을 사용하는 것이 sum을 사용하는 것보다 우수함

Window size는 5에서 12 사이에서 좋은 성능을 나타냄

일반적으로 PV-DBOW와 PV-DM을 혼합해서 사용

THANK YOU -

감사합니다.