Distributed Representations of Sentences and Documents

QUOC LE , TOMAS MIKOLOV

PRESENTERS:

AMIN and ALI

YORK UNIVERSITÉ UNIVERSITY

Outline

- Introduction
- Algorithm
  - Learning Vector Representation of Words
  - Paragraph Vector: A distributed memory model
  - Paragraph Vector without word ordering: Distributed bag of words
- Experiments
- Conclusion
- Demo

2

YORK UNIVERSITÉ UNIVERSITY

Introduction

- Many machine learning algorithms require the input to be represented as a fixed-length feature vector.
- When it comes to texts, one of the most common fixed-length features is bag-of-words.

YORK UNIVERSITÉ UNIVERSITY

3

Bag of Words

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List
for
is
of
the
to

YORK UNIVERSITÉ UNIVERSITY

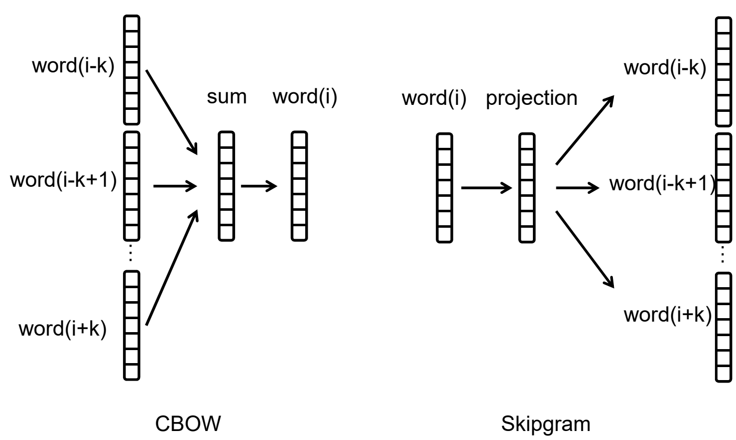
4

Bag of Words Disadvantages

- The word order is lost, and thus different sentences can have exactly the same representation, as long as the same words are used.
- Even though bag-of-n-grams considers the word order in short context, it suffers from data sparsity and high dimensionality.
- Bag-of-words and bag-of-n-grams have very little sense about the semantics of the words or more formally the distances between the words. (powerful, Paris, strong)

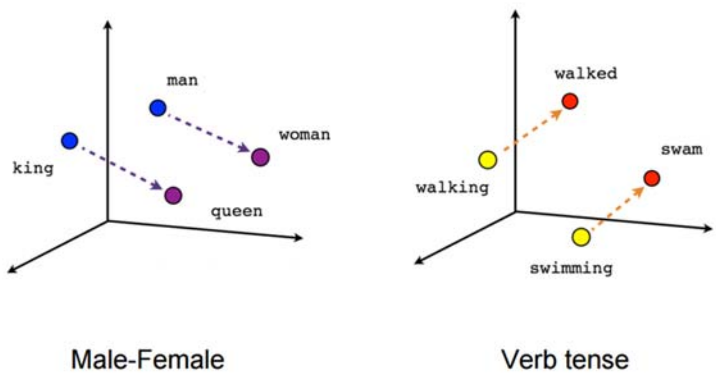
5

Word Embedding



6

Word Embedding



7

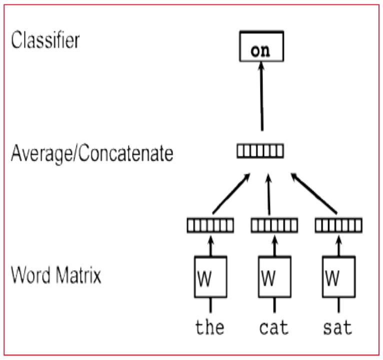
Proposed Method

- Distributed Representations of Sentences and Documents model was proposed.
- Paragraph Vector, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts.
- Proposed algorithm represents each document by a dense vector which is trained to predict words in the document.

8

Learning Vector Representation of Words

- The task is to predict a word given the other words in a context.



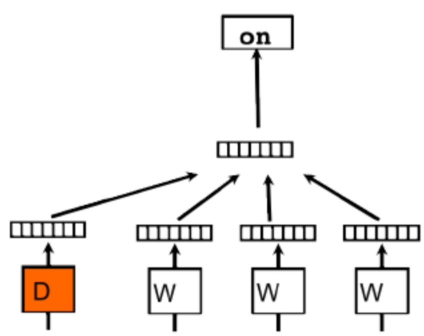
$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t|w_{t-k},\ldots,w_{t+k})$$

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

9

Paragraph Vector: A distributed memory model (**PV-DM**)

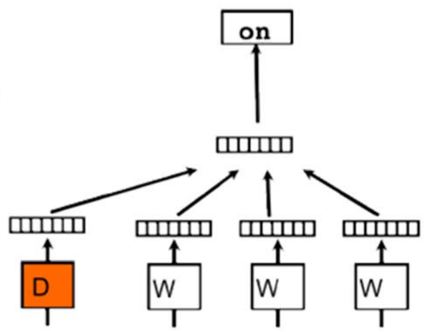
- Paragraph vectors are used for prediction
- Every paragraph is mapped to a unique vector.
- Every word is also mapped to a unique vector



10

Paragraph Vector: A distributed memory model (**PV-DM**)

- The contexts are sampled from a sliding window over paragraph
- Paragraph vector is shared across all contexts from the same paragraph.
- Word vectors are shared across paragraphs



YORK U UNIVERSITÉ UNIVERSITY

11

Advantages over BOW

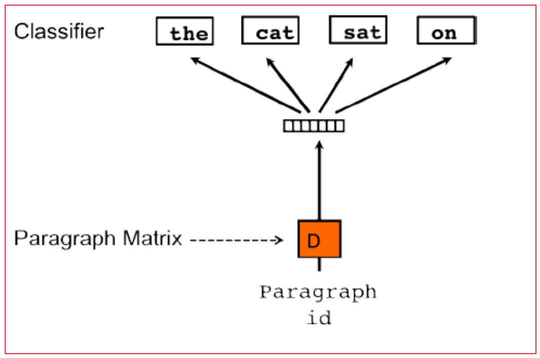
- Semantics of the words. In this space, "powerful" is closer to "strong" than to "Paris"
- Take into consideration the word order.

YORK U UNIVERSITÉ UNIVERSITY

12

Paragraph Vector Distributed Bag of Words (**PV-DBOW**)

- In this version, the paragraph vector is trained to predict the words in a small window.



YORK U UNIVERSITÉ UNIVERSITY

13

Experiment

- Each paragraph vector is a combination of two vectors: one learned by PV-DM and one learned by PV-DBOW.
- Sentiment Analysis.
  - Stanford sentiment treebank
    - 11855 sentences
  - IMDB
    - 100000 movie reviews
- Information Retrieval

14

Stanford sentiment treebank

- Learn the representations for all the sentences
- The paragraph vector is the concatenation of two vectors from PV-DBOW and PV-DM
- Logistic Regression was used for prediction
- Every sentence has label which goes from 0.0 to 1.0

15

Stanford sentiment treebank

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%**

16

IMDB

- Using Neural Networks and Logistic Regression for prediction
- The paragraph vector is the concatenation of two vectors from PV-DBOW and PV-DM

YORK UNIVERSITÉ UNIVERSITY

17

IMDB

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%**

YORK UNIVERSITÉ UNIVERSITY

18

Information Retrieval

- calls from (000) 000 - 0000 . 3913 calls reported from this number . according to 4 reports the identity of this caller is american airlines .
- 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
- allina health clinic patients for your convenience , you can pay your allina health clinic bill online . pay your clinic bill now , question and answers...

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%**

YORK UNIVERSITÉ UNIVERSITY

Observations

- PV-DM is consistently better than PV-DBOW
- PV-DM alone can achieve good results
- The combination of PV-DM and PV-DOW can gain best results.
- A good guess for window size is between 5 and 12.
- The proposed method must be run in parallel.

YORK UNIVERSITÉ UNIVERSITY

Advantages and Disadvantages

- The proposed method is competitive with state-of-the-art methods.
- The good performance demonstrates the merits of Paragraph vector in capturing the semantics of paragraphs.
- It is scalable (sentences, paragraphs, and documents).
- Paragraph vectors have the potential to overcome many weaknesses of bag-of-words (word orders, word meaning, …)

- Paragraph vector can be expensive.
- Too many parameters.
- If the input corpus is one with lots of misspellings like tweets, this algorithm may not be a good choice

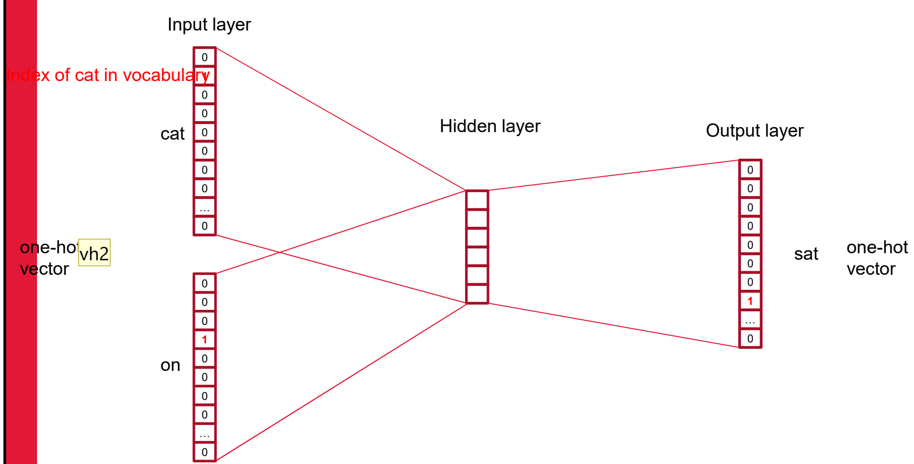
YORK UNIVERSITÉ UNIVERSITY

21

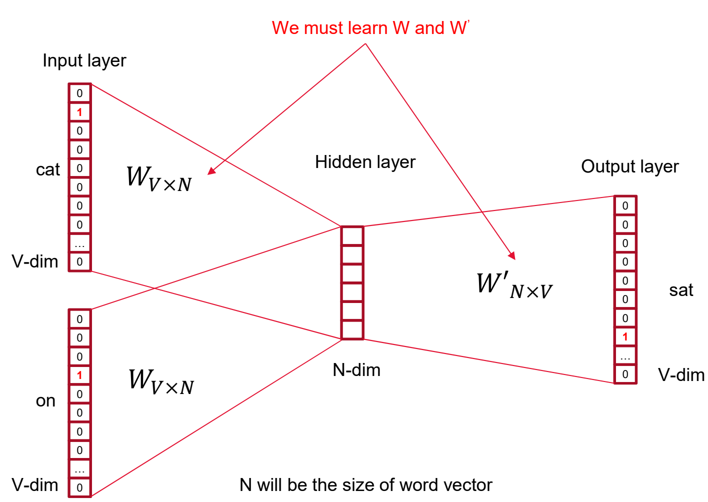
Demo

YORK UNIVERSITÉ UNIVERSITY

22


Input layer
Index of cat in vocabulary
cat
one-ho
vh2
vector
on
Hidden layer
Output layer
sat
one-hot
vector

23

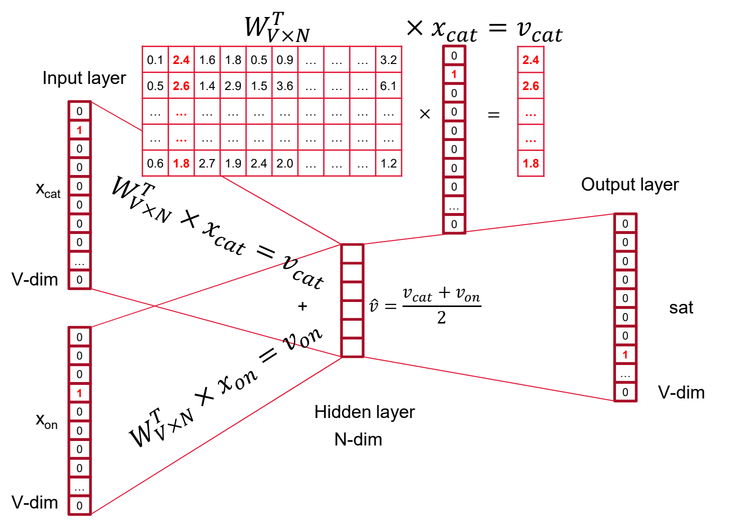

We must learn W and W'
Input layer
cat
$W_{V\times N}$
V-dim
on
$W_{V\times N}$
V-dim
Hidden layer
N-dim
$W'_{N\times V}$
Output layer
sat
V-dim
N will be the size of word vector

24

Slide 23

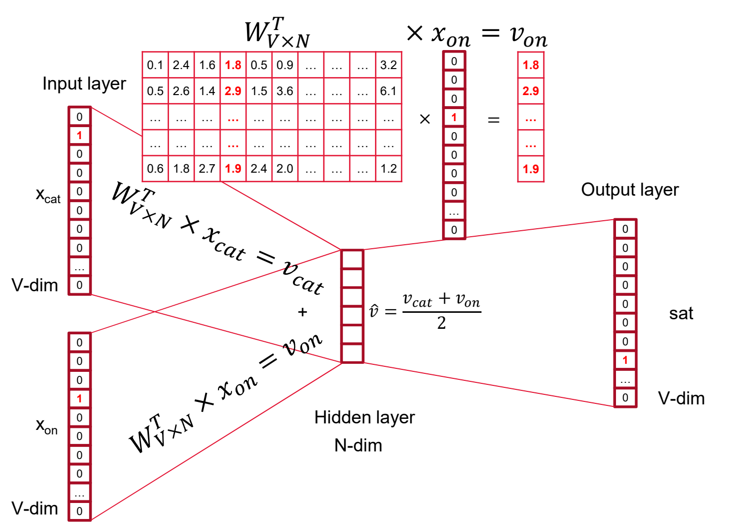
vh2 One hot encoding technique is used to encode categorical integer features using a one-hot aka one-of-K scheme.

Suppose you have 'color' feature which can take values 'green', 'red', and 'blue'. One hot encoding will convert this 'color' feature to three features, namely, 'is_green', 'is_red', and 'is_blue' which all are binary.
vagelis hristidis, 2016-11-06



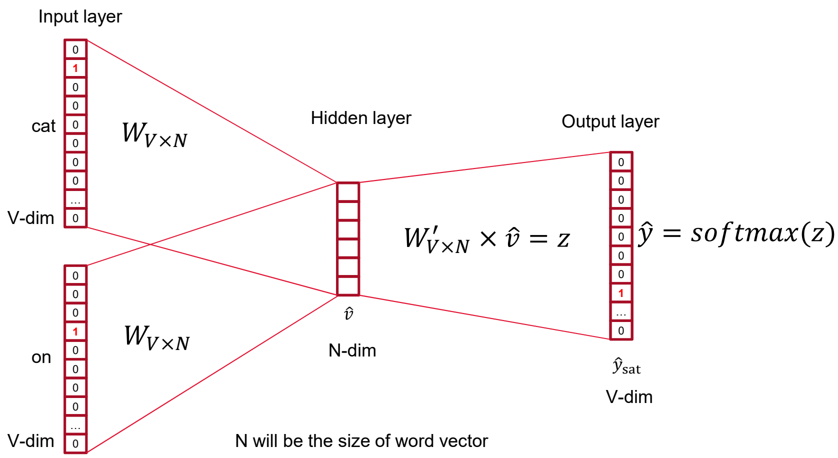
$W^T_{V\times N} \times x_{cat} = v_{cat}$
Input layer
x_{cat}
V-dim
x_{on}
V-dim
$W^T_{V\times N} \times x_{cat} = v_{cat}$
+
$W^T_{V\times N} \times x_{on} = v_{on}$
$\hat{v} = \frac{v_{cat} + v_{on}}{2}$
Hidden layer
N-dim
Output layer
sat
V-dim

25



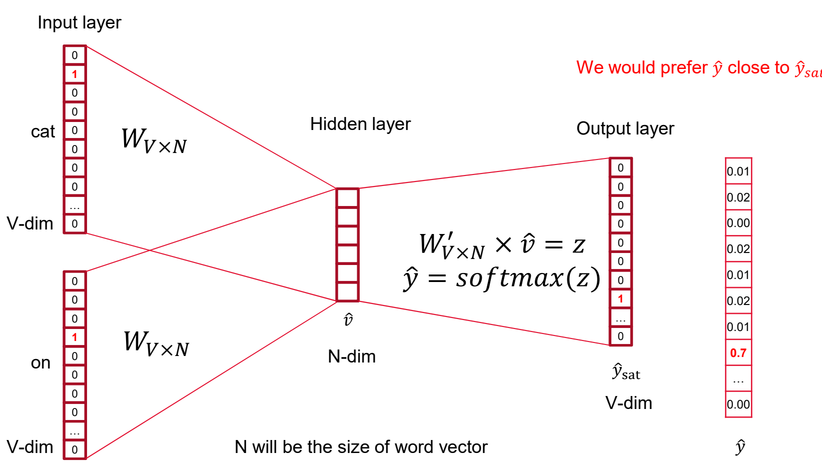
$W^T_{V\times N} \times x_{on} = v_{on}$
Input layer
x_{cat}
V-dim
x_{on}
V-dim
$W^T_{V\times N} \times x_{cat} = v_{cat}$
+
$W^T_{V\times N} \times x_{on} = v_{on}$
$\hat{v} = \frac{v_{cat} + v_{on}}{2}$
Hidden layer
N-dim
Output layer
sat
V-dim

26



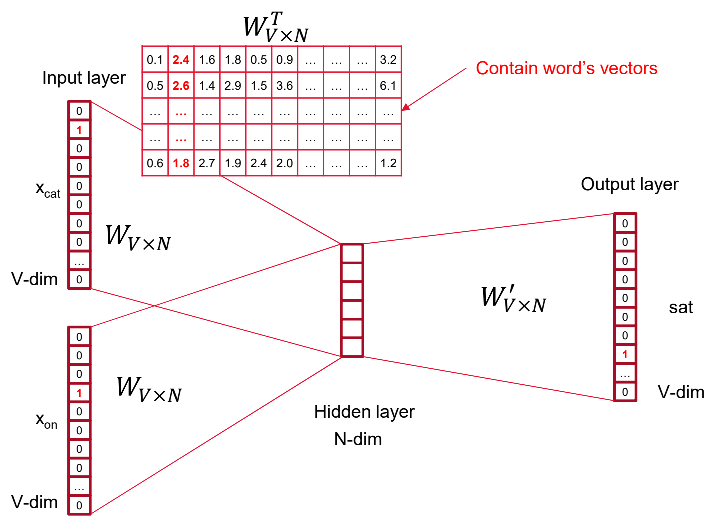
N will be the size of word vector

27



N will be the size of word vector

28



We can consider either W or W' as the word's representation. Or even take the average.