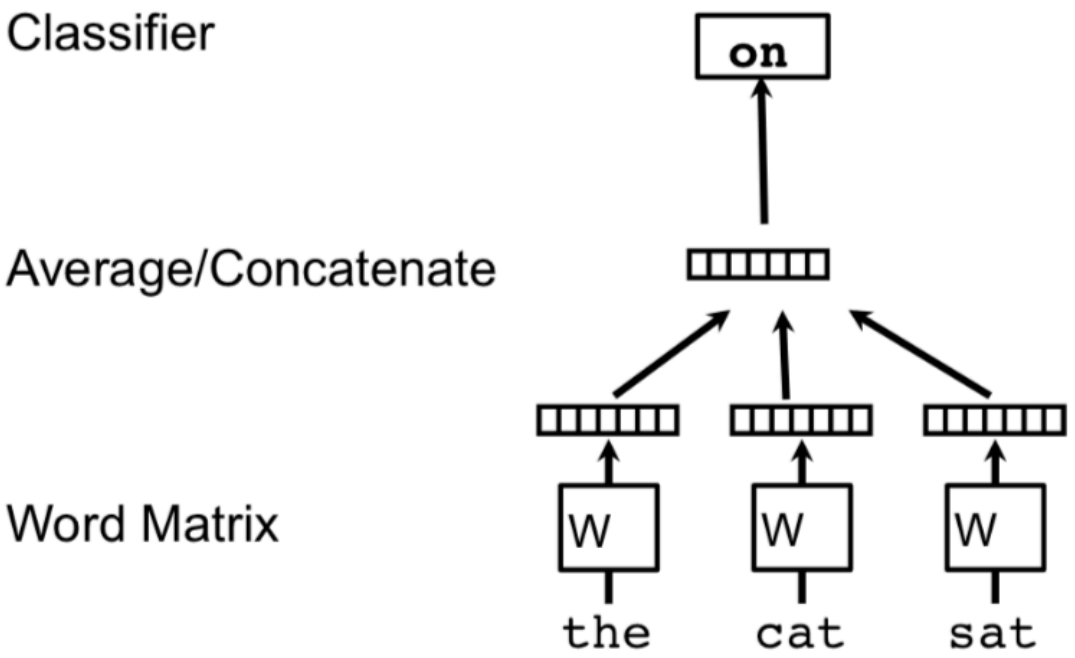
Distributed Representations of Sentences and Documents

- Presenter: Bonggun Shin
- PT date: Sep 24th 2018
- Authors: Quoc Le and Tomas Mikolov
- Date: 2014
- Venue: ICML

Motivation

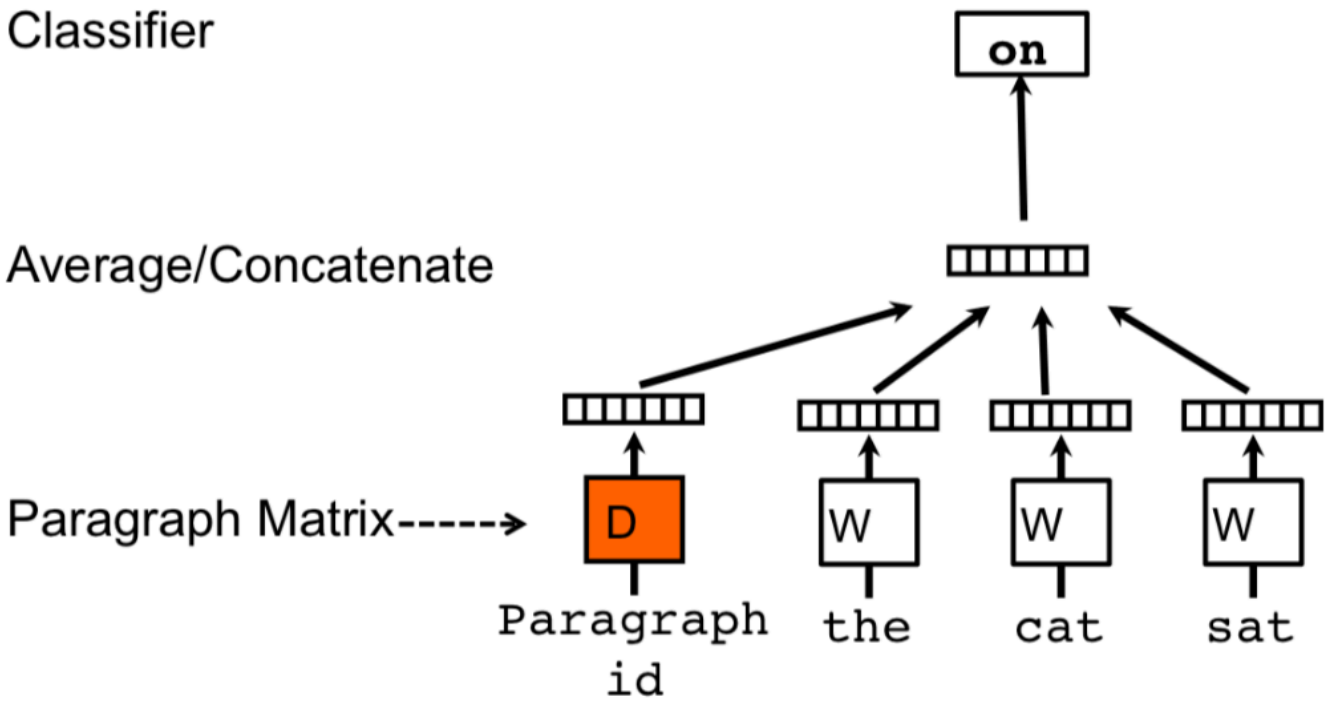
- Text processing requires fixed length vector
- Text has variable length
- Traditional method
  - Bag-of-words (BOW)
    - loses the ordering of the words
    - Ignore semantics of the words
  - Word2Vec
    - Partial context awared model - but still BOW
    - Not applicable to model a lengthy paragraph
  - Doc2Vec!

Model



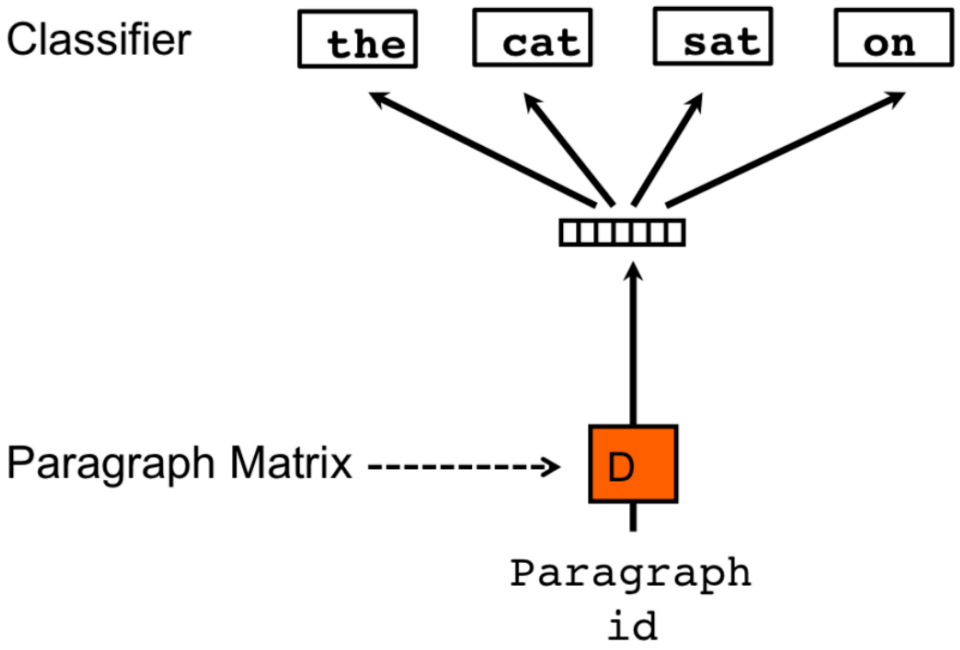
- Word2Vec - Input(Context words), Output(the last word)

Model



- Distributed Memory Model of Paragraph Vectors (PV-DM)
  - Input(Context words, documentID), Output(the last word)

Model



- Distributed Bag of Words of Paragraph Vector (PV-DBOW)
  - Input(documentID), Output(Context words)

Experiment

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%**

- Dataset: SST-1 and SST-2
- Vector representations have 400 dimensions for both words and paragraphs

Experiment

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%**

- Dataset: 3 paragraphs for a given keyword (2 paragraphs are the correct result while the other one is random paragraph)
- Metric: The number of times when each method produces smaller distance for the first two paragraphs than the first and the third paragraph

Discussion

- PV-DM is consistently better than PV-DBOW
- PV-DM+ PV-DBOW is better than single model
- Using concatenation in PV-DM is often better than sum
- Recommended window size: [5,12]
  - Cross-validation will give the best size