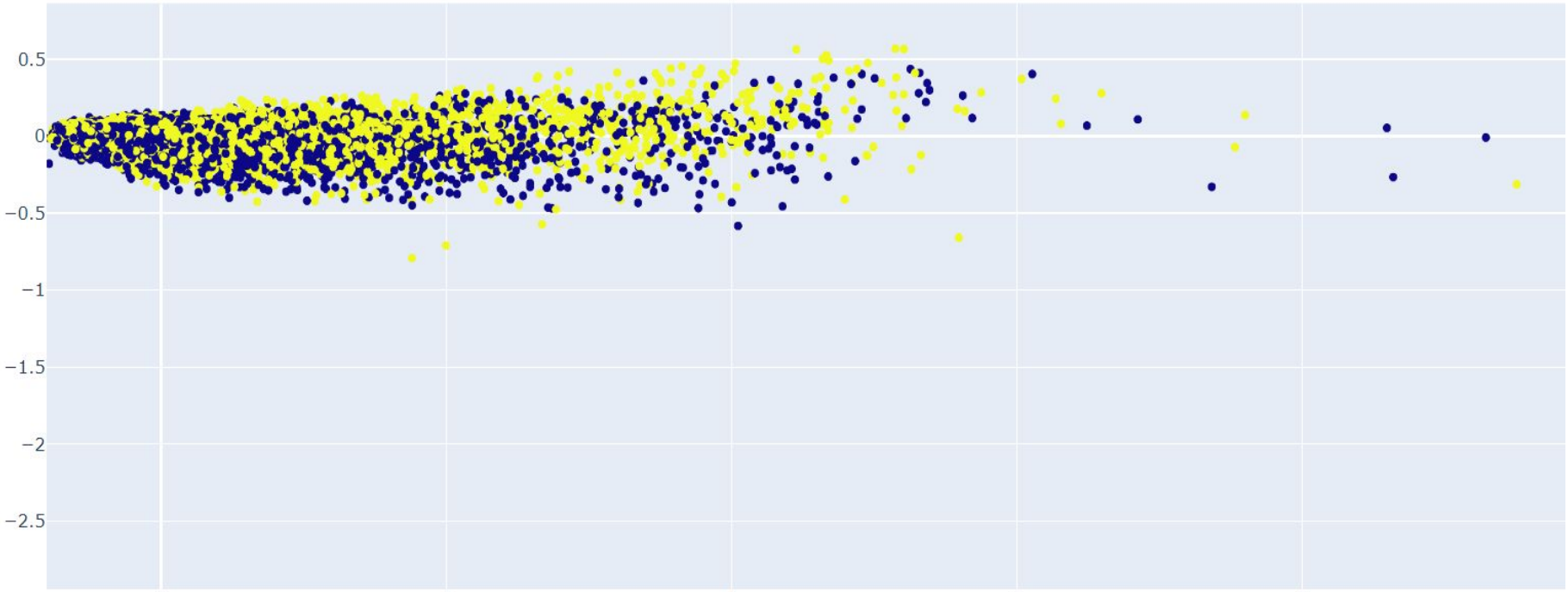
Paragraph Vector

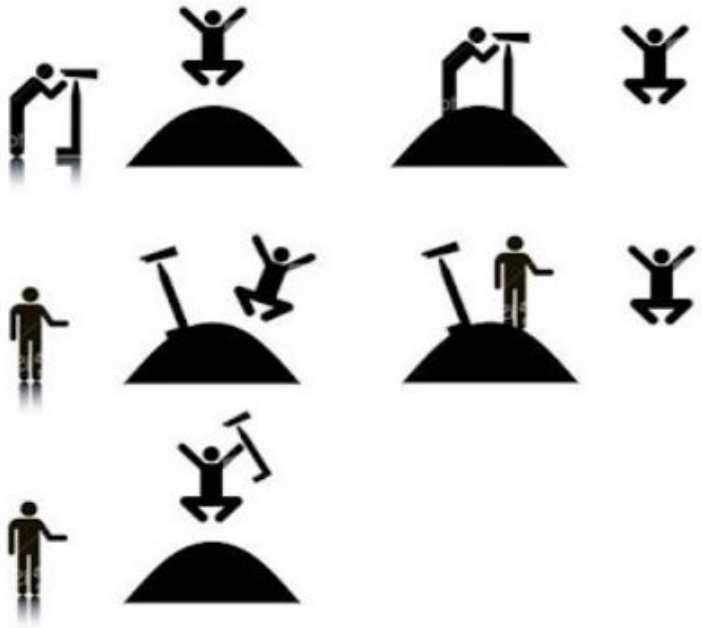
Tianyi Sun, Qingyuan Xue, Eric Darve



Clustering using PCA on labelled sentiment moview review data（learned paragraph vectors embeddings)

Understanding of language

I saw a man on a hill with a telescope



Ambiguity is Explosive and Ambiguity is Ubiquitous

How can machine understands language

The: [0100000]

cat: [0010000]

sat: [0001000]

on: [0000100]

the: [0000010]

mat: [0000001]

a	are	been	day	have	how	nice	see	to	you
1	1	1	1	2	2	2	1	1	3

One-hot Encoding

represents word as one-hot vectors

Drawbacks: inefficiency and no similarity representation

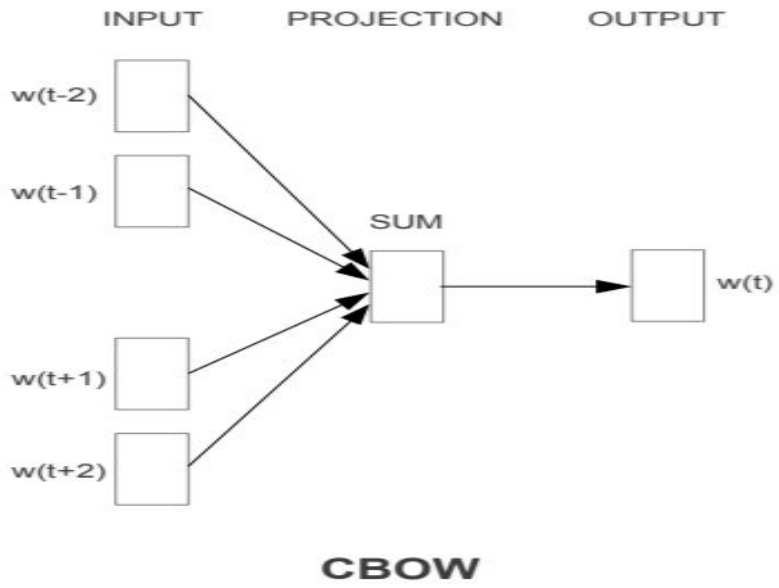
Bag of Words

describe the occurrence of words

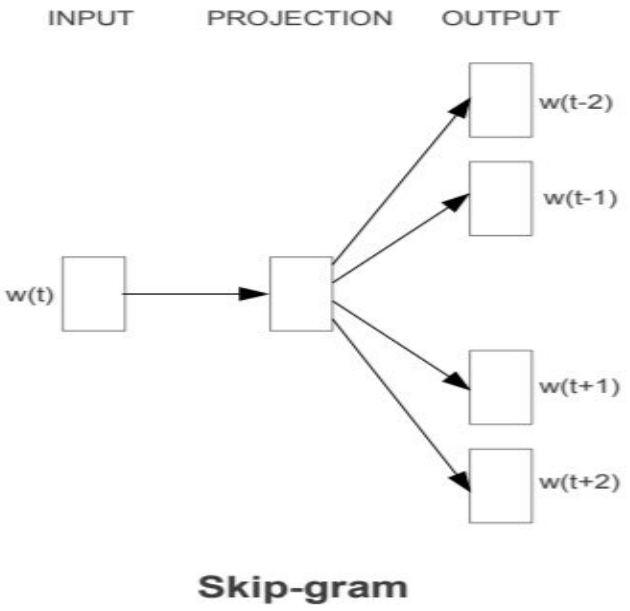
Drawbacks: orders and structures are ignored

The breakthrough: training through tasks

The efficient algorithms of continuous **Bag-of-Words model** and continuous **Skip-gram model** compute the word representations



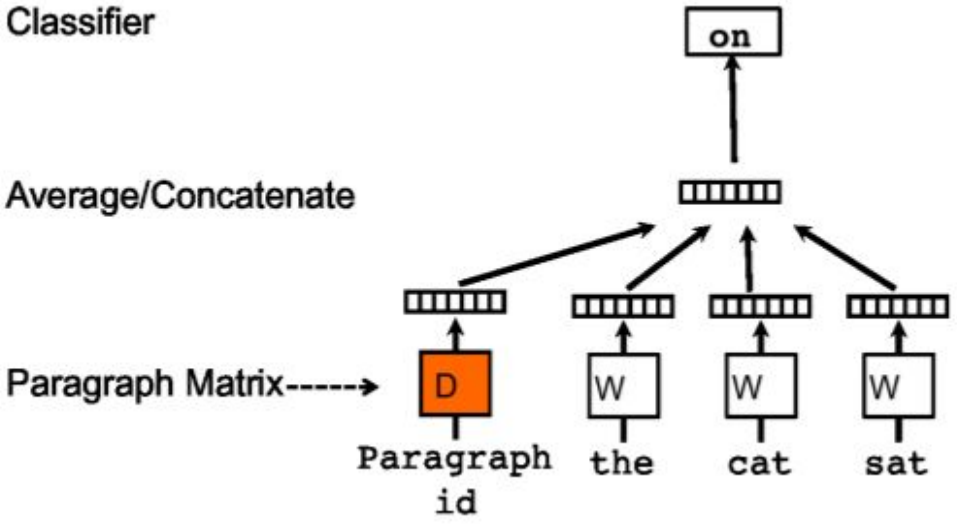
Using the context words to predict the centre word

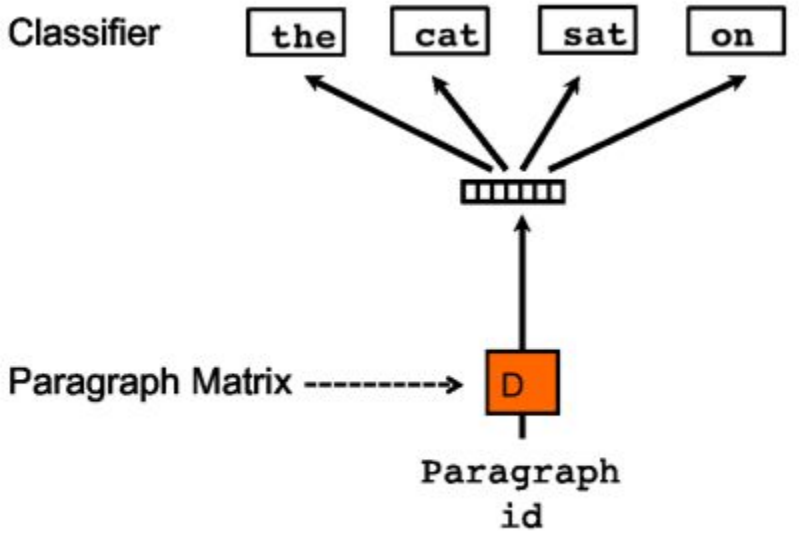


Using the current word to predict the context words

What about sentences and paragraphs?
Paragraph Vector

- Training: words prediction using paragraph vector and words vectors
- Prediction: compute paragraph vector with the new paragraph input





The algorithm and python implementations

- Objective function:

$$\max_{w,b} \quad 1/T \sum_{t=k}^{T-K} \log \mathrm{P}(\mathrm{Wt} \mid W^{t-k}, \ldots W^{t+k}, D)$$

- Training and implementations

Training algorithm	dm_concat	Dimension Size	Negative	Initial learning rate	Window_size
distributed memory	1（Concatenate the context vectors）	400	2(noisy words)	0.065	10 (maximum distance)
distributed bag of words	1（Concatenate the context vectors）	400	2(noisy words)	0.065	10 (maximum distance)

Evaluations and Results

- **Linguistic acceptability Judgments** (10,657 labeled english sentences)
  (two examples from the data set)

The professor talked us	**0**
Anson became a muscle bound	1

- **Classification Prediction**
  - **Logistic regression model accuracy: 0.69** (versus the chance rate: 0.69)

samples	True labels	Predictions
who does john visit sally because he likes ?	0	1
the more does bill smoke , the more Susan hates him	0	1
the bookcase ran	0	1

Evaluations and Results

- **IMDB Movie Reviews**
  - Doc2Vec: 25,000 labelled training samples, 50,000 unlabelled samples
  - Sentiment classification task: 25,000 labelled training samples, 25,000 labelled testing samples.

- **Sentiment classification task**

Models	Accuracy
LDA (Latent Dirichlet Allocation)	67.42%
Chance Rate (baseline)	50%
Paragraph Vector	75%

References

- *Course Information & Introduction to the NLP*, Univeristy of Sydney: Lecture notes https://canvas.sydney.edu.au/courses/30912/pages/course-contents

- Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). PMLR.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, *7*, 625-641.
- Liu, F., Jiao, Y., Massiah, J., Yilmaz, E., & Havrylov, S. (2021). Trans-Encoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. *arXiv preprint arXiv:2109.13059*.