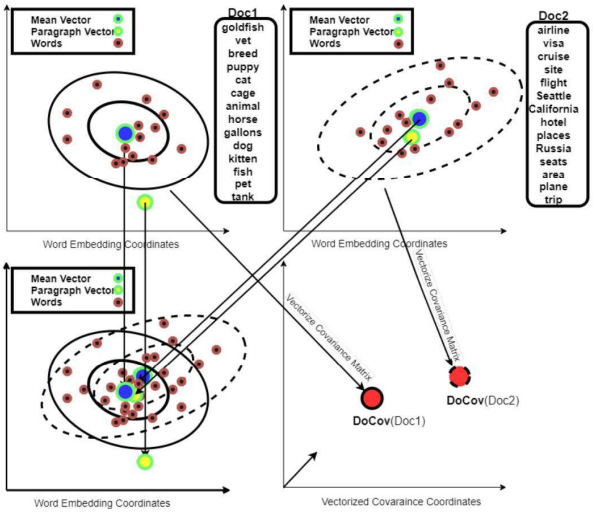
A Document Descriptor using Covariance of Word Vectors

Marwan Torki, Alexandria University
mtorki@alexu.edu.eg

ACL 2018

Abstract - In this paper, we address the problem of finding a novel document descriptor based on the covariance matrix of the word vectors of a document. Our descriptor has a fixed length, which makes it easy to use in many supervised and unsupervised applications. We tested our novel descriptor in different tasks including supervised and unsupervised settings. Our evaluation shows that our document covariance descriptor fits different tasks with competitive performance against state-of-the-art methods.

Motivation



Two Documents are represented using DoCov. Doc1 is about "pets" and Doc2 is about "travel".

Top: The first two dimensions of a word embedding for each document.

Bottom Left: The embedding of the words of the two documents. The Mean vectors and the paragraph vectors are shown. Covariance matrices are shown via the confidence ellipses.

Bottom Right: Corresponding covariance matrices are represented as points in a new space

Approach

Given a d-dimensional word embedding model and an n-terms document. We apply the steps to get the document vector ***v***

1. Compute Observation Matrix ***O***

$$O = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

2. Compute Mean Vector $\bar{x}$

$$\bar{x} = \begin{bmatrix} \bar{x_1} & \bar{x_2} & \cdots & \bar{x_d} \end{bmatrix}^T \in R^d$$

3. Compute Covariance Matrix C

$$C = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_d} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_1X_d} & \sigma_{X_2X_d} & \cdots & \sigma^2_{X_d} \end{pmatrix}$$

4. Compute vector version ***v***

$$v = \mathbf{vect}(C) = \begin{cases} \sqrt{2}C_{p,q} & \text{if } p < q \\ C_{p,q} & \text{if } p = q \end{cases}$$

Results

1. IMDB Movie Review Dataset (Choice of Embedding) Error-Rate performance

Model/Dim					
BOW Gensim	9.66%				
	Mean	DoCoV	DoCoV +Mean	DoCoV Bow	DoCoV +Mean+Bow
d=100	14.13%	11.64%	11.16%	9.39%	9.44%
d=200	12.86%	11.08%	10.80%	9.39%	9.58%
d=300	12.83%	11.08%	10.85%	9.41%	9.47%
Glove	Mean	DoCoV	DoCoV +Mean	DoCoV Bow	DoCoV +Mean+Bow
d=100	20%	13.07%	12.88%	9.63%	9.62%
d=200	16.95%	12.36%	12.22%	9.64%	9.65%
d=300	16.29%	12.00%	11.91%	9.63%	9.66%
d=300,*Lrg*	14.94%	11.70%	11.56%	9.5%	9.6%
Gnews	Mean	DoCoV	DoCoV +Mean	DoCoV Bow	DoCoV +Mean+Bow
d=300	14.03%	11.11%	10.75%	**9.32%**	9.6%

2. Classification Benchmark

Representation \Dataset	MR	CR	Trec	Subj	Overall
Mean	77.4	79.2	80	91.3	81.98
BOW +tf-idf weights	77.1	78.5	89.3	89.3	83.55
P2vec (Le and Mikolov, 2014)	74.8	78.1	91.8	90.5	83.8
Skip-uni (Kiros et al., 2015)	75.5	79.3	91.4	92.1	84.58
bi-skip (Kiros et al., 2015)	73.9	77.9	89.4	92.5	84.43
comb skip (Kiros et al., 2015)	**76.5**	**80.1**	**92.2**	**93.6**	**85.6**
FastSent (Hill et al., 2016)	70.8	78.4	76.8	88.7	78.68
FastSentAE (Hill et al., 2016)	71.8	76.7	80.4	88.8	79.43
SAE (Hill et al., 2016)	62.6	68	80.2	86.1	74.23
SAE+embs (Hill et al., 2016)	73.2	75.3	80.4	89.8	79.68
SDAE (Hill et al., 2016)	67.6	74	77.6	89.3	77.13
SDAE+embs (Hill et al., 2016)	74.6	78	78.4	90.8	80.45
COV	79.7	79.4	89.5	92.8	85.35
COV+Mean	80.2	80.1	90.3	93.1	85.93
COV+Bow	80.7	80.5	**91.8**	**93.3**	86.58
COV+Mean+BOW	**81.1**	**81.5**	91.6	93.2	***86.85***

3. Spearman/Pearson correlations on unsupervised (relatedness) evaluations.

Model	STS 2014							SICK
	News	Forums	Wordnet	Twitter	Images	Headlines	All	
P2vec (Le and Mikolov, 2014)	0.42/0.46	0.33/0.34	0.51/0.48	0.54/0.57	0.32/0.30	0.46/0.47	0.44/0.44	0.44/0.46
FastSent (Hill et al., 2016)	0.44/0.45	0.14/0.15	0.30/0.34	0.42/0.43	0.55/0.60	0.43/0.44	0.27/0.29	0.57/0.60
FastSent+AE (Hill et al., 2016)	**0.58/0.59**	**0.41**/0.36	**0.74/0.70**	0.63/0.66	**0.74/0.78**	0.57/0.59	**0.63/0.64**	**0.61/0.72**
Skip-Thought (Kiros et al., 2015)	0.56/**0.59**	**0.41/0.40**	0.69/0.64	**0.70/0.74**	0.63/0.65	**0.58/0.60**	0.62/0.62	0.60/0.65
SAE (Hill et al., 2016)	0.17/0.16	0.12/0.12	0.30/0.23	0.28/0.22	0.49/0.46	0.13/0.11	0.12/0.13	0.32/0.31
SAE+embs (Hill et al., 2016)	0.52/0.51	0.22/0.23	0.60/0.55	0.60/0.60	0.64/0.64	0.11/0.11	0.42/0.43	0.47/0.40
SDAE (Hill et al., 2016)	.07/0.04	0.11/0.13	0.33/0.24	0.44/0.42	0.44/0.38	0.36/0.36	0.17/0.15	0.46/0.46
SDAE+embs (Hill et al., 2016)	0.51/0.54	0.29/0.29	0.56/0.50	0.57/0.58	0.59/0.59	0.43/0.44	0.37/0.38	**0.46/0.46**
Mean	**0.65**/0.68	0.46/0.45	0.75/0.78	**0.71**/0.75	0.76/0.78	0.59/0.64	0.64/0.66	**0.63/0.73**
DoCoV	0.62/0.68	0.50/**0.51**	0.77/**0.79**	0.69/0.75	**0.78**/0.80	0.60/0.63	**0.67/0.70**	0.61/0.69
DoCoV+Mean	0.64/**0.70**	**0.51/0.51**	**0.79**/0.78	**0.71/0.76**	**0.78/0.81**	**0.61/0.65**	**0.67/0.70**	0.62/0.71

Conclusion

- We presented our novel descriptor to represent text on any level such as sentences, paragraphs or documents.
- Our representation is generic which makes it useful for different supervised and unsupervised tasks.
- It has fixed-length property which makes it useful for different learning algorithms.
- our descriptor requires minimal training.
- We do not require a encoder-decoder model or a gradient descent iterations to be computed.
- We showed competitive performance against other state-of-the-art methods in supervised and unsupervised settings.

Thanks to Alexandria University for covering publication costs and travel grant to ACL 2018.
Thanks to former collaborators from QU, RMIT, UMD for fruitful discussions.
Thanks to QU and QNRF for funding the author when the work started. Special thanks to Tamer Elsayed at QU.