Word embedding composition via tensors

Borzdov Bogdan, Dziubenko Ivan, Tatarnikova Anna

December 20, 2018

Outline

- Introduction
- Rand-Walk
- Syntactic Rand-Walk
- Results
- Conclusion

Background

How can words be matched to each other in a semantic sense?

Example:

“king”, “queen”, “prince”

In order to be able to represent semantic proximity, it was proposed to use a comparison of the word vector, which reflects its value in the “space of meanings”.

Known approaches

- TF-IDF
- Word2Vec:
  - Skip-Gtam
  - CBOW
- GloVe
- RAND-WALK

Problem Formulation

Challenge:

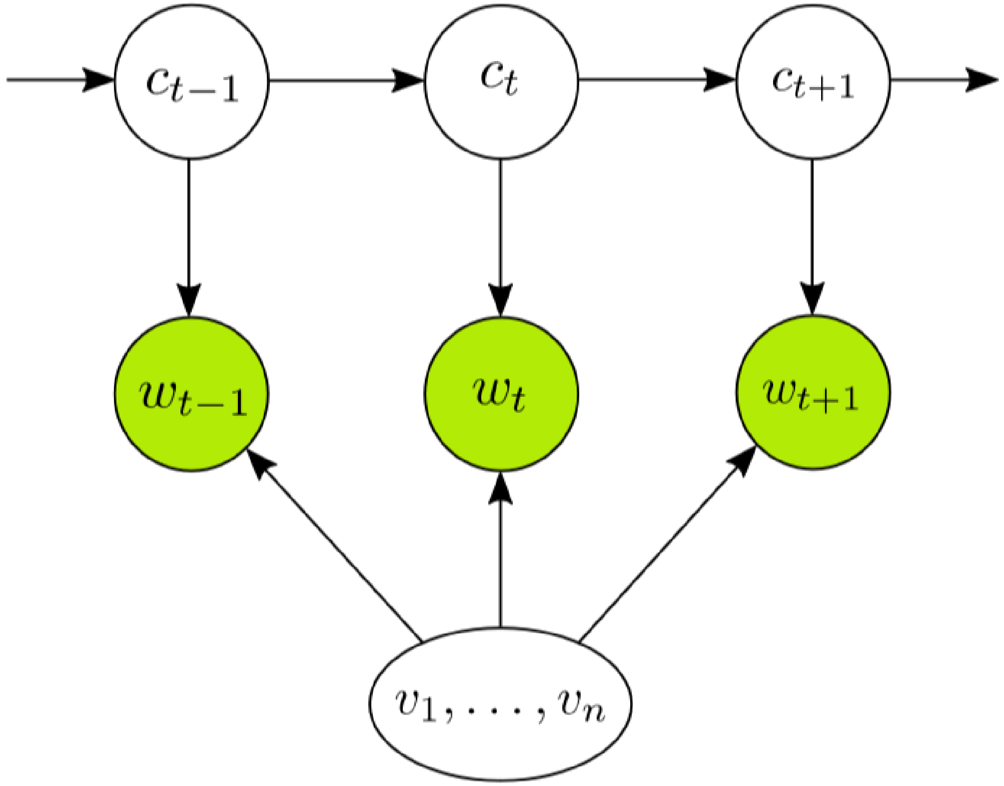
Method that is capable of capturing specific meaning of syntactic relations (e.g., adjective-noun or verb-object pairs) in a way that is impossible by simply “adding” the meaning of the individual words.

Our Approach

- Each word w in vocabulary has a corresponding embedding $v_w \in \Re^d$. The process of corpus generation is driven by the random walk of a discourse vectors $c_t \in \Re^d$ (Arora et al, 2015).
- We use a core tensor $T \in \Re^{d \times d \times d}$ to capture the relations between a pair of words and its context. The process of tensor generation is driven by the syntactic random walk of a discourse vectors $c_t \in \Re^d$.

RAND-WALK

- A corpus of text: a sequence of random variables $w_1, w_2, w_3, \ldots$, where w_t takes values in a vocabulary V of n words. Each word $w \in V$ has a word embedding $v_w \in \Re^d$.
- The process of word embedding generation is driven by the random walk of a discourse vector $c_t \in \Re^d$. Its coordinates represent what is being talked about.



RAND-WALK

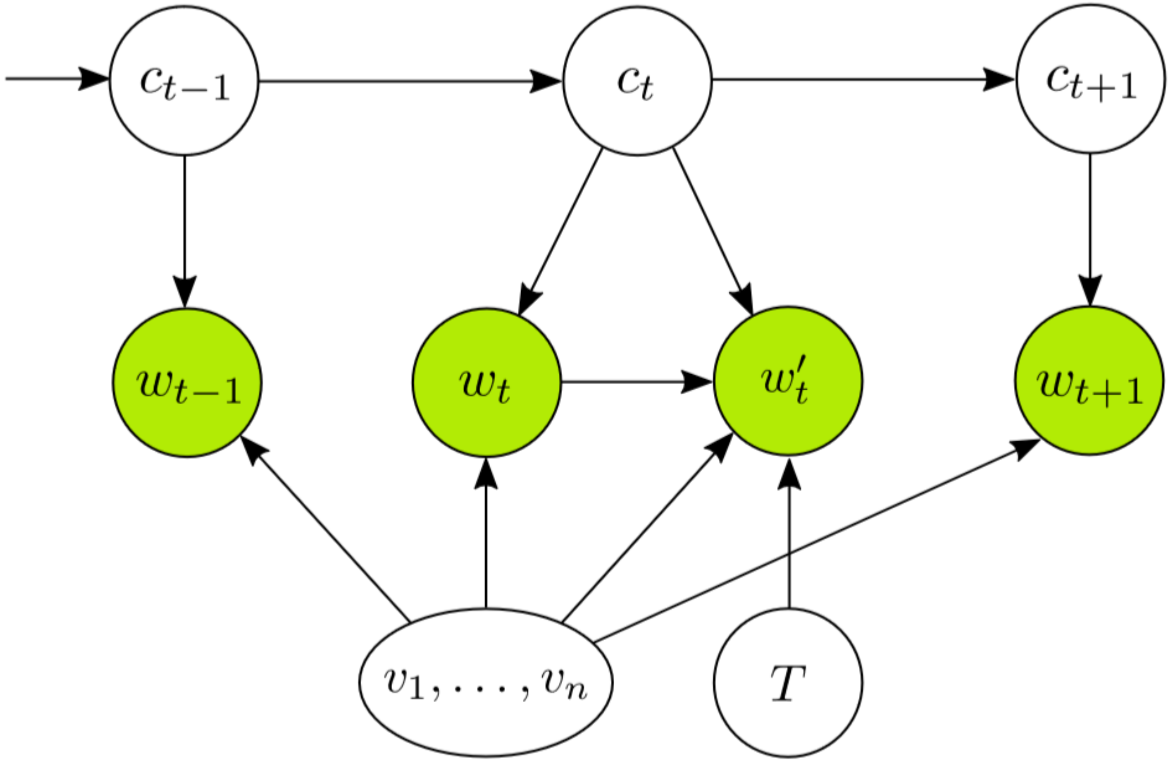
- The discourse vector c_t does a slow random walk, so that nearby words are generated under similar discourses: $\|c_{t+1} - c_t\|$ is small.
- Let $X_{w,w'}$ be the number of times words w and w' co-occur within the same window.
- The maximum likelihood values for the word vectors correspond to the following optimization

$$\min_{\{v_w\},C} \sum_{w,w'} X_{w,w'} (\log(X_{w,w'}) - \|v_w + v_{w'}\|_2^2)^2$$

Syntactic RAND-WALK

We use a core tensor $T \in \Re^{d \times d \times d}$ (composition tensor) to capture the relations between a pair of words and its context, where c is a discourse vector, v and v' are word embeddings of the two relevant words.

$$T(v, v', c) = \sum_{i,j,k=1}^{d} T_{i,j,k} v(i) v'(j) c(k).$$



Syntactic RAND-WALK

- $X_{(a,b),w}$ denotes the number of co-occurrences of word *w* with the syntactic word pair (*a*,*b*).
- The maximum likelihood values for the word vectors correspond to the following optimization

$$\min_{T,\{C_w\},C} \sum_{(a,b),w} f(X_{(a,b),w})\left(\log(X_{(a,b),w}) - \left\| v_w + v_a + v_b + T(v_a, v_b, \cdot) \right\|^2\right)^2,$$

where $f(x) = \min(x, 100)$.

Implementation

- Python with Tensorflow and Spacy
- Only adjective-noun syntactic pairs considered
- Embedding vectors are trained by optimizing the objective using AdaGrad (Duchi et al., 2011) with initial learning rate of 0.05 and 5 epoch.

Training method

- We first train the word embeddings according to the RAND-WALK model, following Arora et al. (2015).

$$\min_{\{v_w\},C} \sum_{w,w'} f(X_{w,w'})\left(\log(X_{w,w'}) - \left\|v_w + v_{w'}\right\|_2^2\right)^2$$

- Using the learned word embeddings, we next train the composition tensor T via the following optimization problem

$$\min_{T,\{C_w\},C} \sum_{(a,b),w} f(X_{(a,b),w})\left(\log(X_{(a,b),w}) - \left\|v_w + v_a + v_b + T(v_a,v_b,\cdot)\right\|^2\right)^2$$

$$\left[T(x,\cdot,\cdot)\right]^T y = T(x,y,\cdot)$$

Composite embedding

- Classic: $v_a + v_b = v_c$
- In SRW: $v_a + v_b + T(v_a, v_b, \bullet) = v_c$

Where $T(x, y, \bullet)_k = \sum_{i,j=1}^{d} T_{i,j,k}\, x(i)\, y(j)$

Dataset

- We train our model using enwik8 compressed Wikipedia articles.
- The text is pre-processed to remove non-textual elements, stop words, and rare words (words that appear less than 300 within the corpus).
- We generate a matrix of word-word co-occurrence counts using a window size of 5. Triple co-occurrence tensor for SRW is generated only on found adjective-noun pairs.

Results

Civil war

Additive of RAND-WALK	Additive of Glove	Tensor
War	Civil	Self
Rights	Order	Rule
Public	Data	Force
Minister	States	Area
Battle	Languages	Spanish

Results

United states

Additive of RAND-WALK	Additive of Glove	Tensor
States	States	Young
Left	Apollo	Southern
Game	President	Capital
President	Center	Famous
Society	Famous	Industry

Results

European union

Additive of RAND-WALK	Additive of Glove	Tensor
European	European	Subject
Important	Low	Forces
Official	Self	Video
Austrian	British	Image
Means	Soviet	Energy

Results

Soviet union

Additive of RAND-WALK	Additive of Glove	Tensor
Union	Union	World
Mission	Social	Prize
Japanese	Force	Austria
Eastern	Battle	Class
Mother	America	Union

Conclusion

- Method is highly dependent on amount of epoch and iterations, can produce completely different results on similar numbers of iterations.
- It is necessary to remove conjunctions, articles, etc, as they will otherwise spoil the resulting similar words
- While optimization part of SRW is not much more difficult than the other method’s, co-occurrence tensor generation takes a lot of time: Total learning time: Time(Glove): 40,9760s; Time(Rand-walk):33,6288s; Time(SRW): 1887,7156s
- Semantic results are hard to estimate. Needs additional testing on phrase similarity datasets.

References

- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Rand-walk: A latent variable model approach to word embeddings, 2015.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 2011.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. Proceedings of the Empiricial Methods in Natural Language Processing, 2014.
- Understanding composition of word embeddings via tensor decomposition, 2018.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations, 2013a.