NLP Assignment 2 Report

Doc2Vec for Sentiment Detection of Reviews

Zébulon Goriely, Queens', zg258

Friday 22nd November, 2019

Word Count: 983[1]

1 Introduction

Le and Mikolov [2014] introduced `doc2vec` for learning embeddings for sequences of words. Here, I investigate the use of `doc2vec` document vectors in the task of sentiment detection, using a Support Vector Machine (SVM) classifier.

I compare the performance of the classifier trained with `doc2vec` vectors with the same classifier trained with simple word-frequency and presence-based vectors. Finally, I examine means of qualitatively analysing the hypothesis that `doc2vec` encodes semantics, in particular how sentiment-coding adjectives are represented.

2 Background

To classify reviews, I use an SVM classifier trained and tested on document vectors. I was given a dataset of 1000 positive and 1000 negative movie reviews in the framework of a course in NLP.

2.1 Support Vector Machine

Pang et al. [2002] introduced the use of SVM classifiers for sentiment analysis, operating on document vectors.

A frequency-based document vectors is defined $\vec{d} := (n_1(d), n_2(d), \ldots, n_m(d))$ where $n_i(d)$ is the occurrences of feature f_i in document d.

A presence-based document vector is defined as $\vec{d} := (s_1(d), s_2(d), \ldots, s_m(d))$ where

$$s_i = \begin{cases} 1 & \text{if } d \text{ contains } f_i \\ 0 & \text{otherwise} \end{cases}$$

[1] *texcount docs/assignment2/report.tex*

Using supervised learning, the procedure produces a hyperplane represented by the vector $\vec{w}$ which divides the vectors into two classes. This hyperplane exists in the n-dimensional vector space of the document vectors.

This allows for classification to proceed by determining which side of the hyperplane each d_i falls on.

2.2 Doc2Vec

Mikolov et al. [2013] introduced the ideas of skipgrams and `word2vec` to create compact vector-space representations of words. Unlike the frequency-based document vectors and presence-based vectors, the dimensions of these vectors are not directly interpretable.

Le and Mikolov [2014] extended this idea to `doc2vec` which learns embeddings of *sequences* of words. A key feature is that it is agnostic to granularity, generating fixed-length vectors from variable-length pieces of text. In this case, I train `doc2vec` to output document embeddings, a new type of vector that Le and Mikolov claim is effective for sentiment analysis.

The two architectures in `doc2vec` are the distributed memory and distributed bag of words model which are closely correlated to the `word2vec` and skip-gram models respectively.

3 Method

Selecting unigrams as a feature, I compare the performance of training and testing an SVM classifier on:

- frequency-based document vectors

Hyperparameter	Description	Best Value
Vector Size	Dimension of word vectors	124
Window Size	Left/right context window size	6
Min Count	Minimum frequency threshold for word types	20
Negative Sample	No. of negative word samples	0
Epoch	Number of training epochs	5

Table 1: A description of `doc2vec` hyperparameters and the best values found for this task.

	Unigram-frequency	Unigram-presence	doc2vec
Unigram-frequency	1	2.00×10^{-4}	2.00×10^{-4}
Unigram-presence	2.00×10^{-4}	1	1.60×10^{-3}
doc2vec	2.00×10^{-4}	1.60×10^{-3}	1

Table 2: The significance of difference between systems from Table 3.

- presence-based document vectors
- vectors inferred from a trained `doc2vec` model

I implemented[2] the SVM classifier using Joachim's (1999) SVM^{light} package[3], using default parameters. For the `doc2vec` implementation, I used the gensim library[4].

I trained the `doc2vec` model on a large external corpora of 100,000 movie reviews provided by the Stanford Large Movie Review Dataset [Maas et al., 2011]. In a 2014 evaluation of `doc2vec` , Lau and Baldwin [2016] suggest relevant parameters to use for training, as described in Table 1.

To tune my parameters, I set aside a validation corpus comprising of 10% of our 2000 movie review dataset. For each set of parameters chosen, I trained `doc2vec` on the 100,000 unlabelled files, used this model to generate vectors to train SVM on 90% of our 2000 movie review dataset and tested on this validation corpus. The validation corpus was never used for training and once suitable parameters were found, it was not used for testing.

Parameters were initially chosen using Lau and Baldwin's suggestions. I then tested a range of values for each parameter, in order to increase the accuracy when testing on the validation corpus.

During this process I also found that the distributed bag of words (`dbow`) model was more effective than the distributed memory (`dm`) model and that using hierarchical softmax also improved performance. The best parameters found from this process are shown in Table 1, giving an accuracy of 89% on the validation corpus.

4 Results

	Vector Type	Accuracy
(1)	Unigram-frequency	75.9
(2)	Unigram-presence	86.7
(3)	doc2vec	89.3

Table 3: Average ten-fold cross-validation accuracies, in percent. Dataset: 1800 movie reviews.

I ran ten-fold cross-validation for my three experiments, as seen in Table 3, using our 1800 movie reviews.

The result of 89.3% achieved using `doc2vec` is significantly better than the 75.9 and 86.7 achieved by the frequency-based vectors and presence-based vectors respectively. The significance was calculated using a Monte-Carlo permutation test, with p-values recorded in Table 2.

[2]https://github.com/ZGoriely/cambridge-nlp
[3]http://svmlight.joachims.org
[4]https://radimrehurek.com/gensim

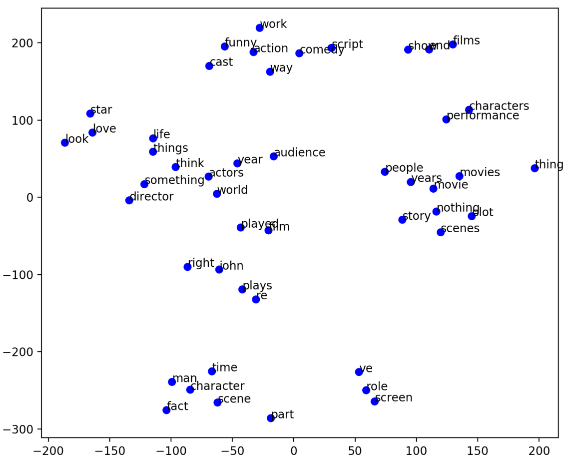

Figure 1: Two-dimensional t-SNE projection of doc2vec embeddings of the 50 most frequent nouns

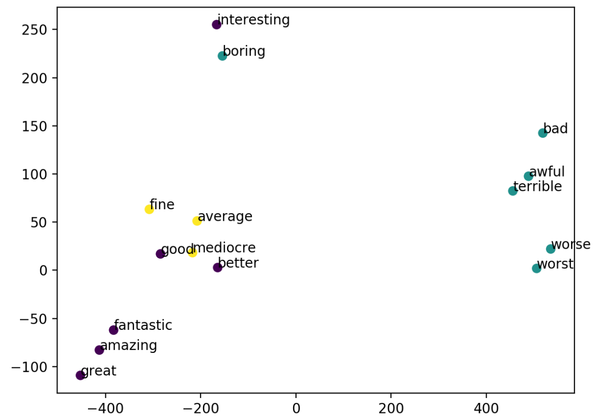

Figure 2: Two-dimensional t-SNE projection of doc2vec adjective embeddings

5 Discussion

My results show that the more complicated `doc2vec` model, considering the *context* of words, performs better at sentiment analysis than the simple bag-of-word models for generating document vectors. Using T-sne and heatmap graphs, I can qualitatively examine what the model is doing.

Plotting the 50 most frequent nouns in our 2000 files using t-SNE reveals how `doc2vec` learns contex, seen in Figure 1. Nouns used in the same context such as *plot, story* and *man, character* are grouped together and are thus represented by similar vectors in the `doc2vec` embedded vector space.

I hypothesise that `doc2vec` learns sentiments of common adjectives from their usage in the positive and negative files that `doc2vec` is trained on. I plotted these adjectives using t-SNE in Figure 2.

Word	Positive	Negative
amazing	2004	516
good	15025	14728
bad	3747	14726
average	595	831

Table 4: Occurrences of words in positive and negative training files

I observe that the model seems to learn the sentiment of these adjectives. *Fantastic, amazing, great* are grouped together and *bad, awful, terrible, worse, worst* are also grouped. Interestingly, the neutral adjectives *fine, average, mediocre* seem to be grouped with the positive adjectives *good, better*. This may be due to the comparative occurrence of these adjectives in positive and negative files. As seen in Table 4, the words *amazing* and *bad* dominate in positive and negative files respectively whereas the words *good* and *average* are distributed more evenly between positive and negative files.

Li et al. [2015] proposed a method to visualise indicator words by plotting the variance of word vectors in a sentence against each other. Figure 4 shows a variance visualisation of the sentence *"this has to be one of the greatest movies of all time"*; each grid corresponds to $||e_{i,j} - \frac{1}{N_s}\sum_{i' \in N_s} e_{i',j}||^2$ where $e_{i,j}$ denotes the value for jth dimension of word i and N denotes the number of token within the sentences. This gives the word *greatest* as the strongest indicator for this sentence.

A heat-map visualisation of three short reviews, shown in Figure 4, reveals another analysis that `doc2vec` learns sentiment; the neutral review seems to lie between the heat-maps of the positive and negative reviews.

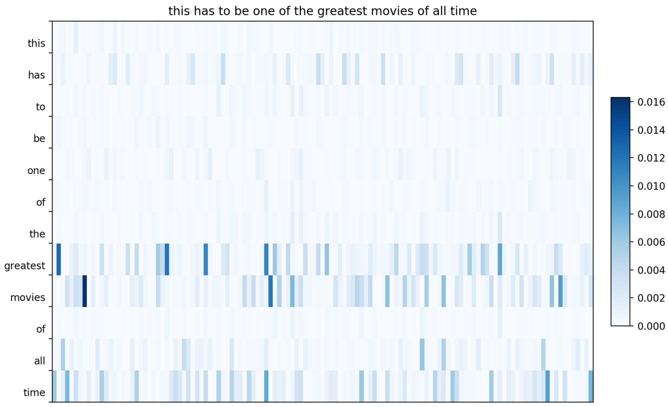

Figure 3: Variance visualisation of a single-sentence review

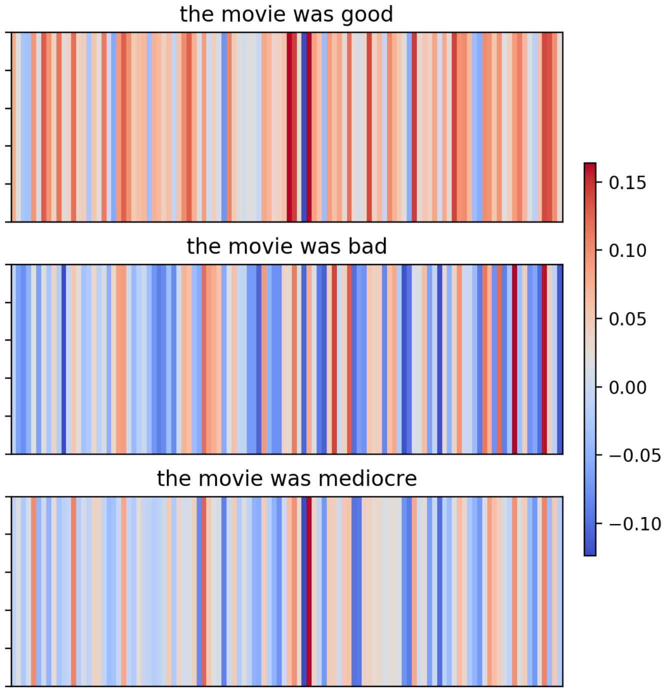

Figure 4: Heat-map visualisation of a good, bad and neutral sentiments

Conclusion

I have shown how using `doc2vec` vectors in an SVM classifier outperforms naive frequency-based vectors and presence-based vectors. I further qualitatively analysed the behaviour of `doc2vec`, finding that the model learns sentiment through occurrence of adjectives in positive and negative documents.

References

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.