ENSAE
ParisTech

École nationale
de la statistique
et de l'administration
économique

Compressed Sensing Project

A compressed sensing view of unsupervised text embeddings

Salomé Do, Lucas Zanini

March, 22nd 2019

Introduction

Text documents have been, until recently, an underexploited source of information due to their complex, non-numerical structure. Low level natural language processing (NLP) tasks as text classification, sequence tagging (POS-tagging, named entity recognition, ...) and parsing have long been tackled through bag-of-words methods and hidden markov models. The application of deep learning to NLP tasks (started around 2010s) has opened new, complementary and effective ways to solve these problems, eventually leading to core innovations in high-level tasks such as question answering or language generation. Using feed-forward networks to generate low-dimensionnal representation of sparse, high dimensionnal word vectors; recurrent neural networks as LSTMs to fit the sequential nature of sentences; and later on attention mechanims, deep learning techniques have set new baselines on traditional benchmarks. However, the efficiency of these results is totally empirical and is not backed by theoretical results, in addition to having high computationnal costs. The article studied in this project [1] aims at giving some results on unsupervised text embeddings.

To do so, it focusses on a compressed sensing view of the text embedding problem, which aims at provinding low-dimensionnal representation for texts. As words can be viewed as very sparse vectors in the vocabulary space, providing a low-dimensionnal representation of a text can be seen as measuring words signals in a compressed way. The article explores this idea, and links compressed sensing to some LSTMs-learned text representations. In order to evaluate such representations (through their performances on text classification), authors adapts a result in [2] on the quality of low-dimensionnal compressed sensing representations as inputs for a classification problem.

Contents

1 Text embeddings

1.1 Word embeddings

1.2 From word embeddings to text embeddings

1.2.1 LSTMs

1.2.2 Bag of n-Grams

1.2.3 DisC embeddings

1.3 Why text embeddings?

2 Text embedding as a Compressed Sensing problem

2.1 Results on compressed sensing for classification

2.2 Links with LSTMs

2.3 Sparse recovery with pretrained embeddings

3 Experiments and discussion

3.1 Results reproducing

3.2 Testing DisC embeddings on real life data

3.3 Discussion

References

[1] S. Arora, M. Khodak, N. Saunshi, and K. Vodrahalli, "A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and lstms," in Proceedings of the 6th International Conference on Learning Representations (ICLR), 2018.

[2] R. Calderbank, "Compressed learning : Universal sparse dimensionality reduction and learning in the measurement domain," 2009.