Deep Learning for Document Classification using Skip-Thoughts

Project Proposal

Amlan Kar amlan@iitk.ac.in

Sanket Jantre jsanket@iitk.ac.in

Abstract

Document Classification problems have been applied to to various tasks, such as automatic tag suggestion, document indexing, sentiment analysis etc. Traditionally, most of these methods involve processes that do not utilize information such as text order, such as BoW models or Tf-Idf techniques to create document vectors. Later, powerful semantic word embeddings emerged, including word2vec and GloVe that have been shown to work well for benchmark sentence classification tasks[1]. Recently, a new semantic sentence embedding, dubbed Skip-Thoughts[2] has emerged which models sentences as vectors. We intend to explore how a Convolutional Neural Network(CNN) can work with these skip-thought embeddings to model data for various Document Classification tasks.

1 Introduction

In the recent past, deep learning methods have consistently set new benchmarks for a variety of NLP Tasks, such as part-of-speech tagging [3], sentiment classification [4], neural language models [5] and machine translation. These models have been heavily influenced in the recent past by the availability of robust embeddings[6] that have boosted their results. Recently, a sentence embedding model, dubbed Skip-Thoughts[2] has emerged, which employs a Gated Recurrent Neural Network based encoder-decoder model to learn generic unsupervised sentence encodings. We attempt to learn a deep framework (specifically, a convolutional neural network), which given skip-thought representations of sentences in a document, learns to to perform various Document classification tasks. We will try to focus on the task of Multi-Label document classification, which has not been explored a lot using deep learning methods in published literature[7].

2 Related Work

While not a lot of work has been done using the sentence domain (to the best of the authors' knowledge), similar work has been done using the word vector domain[1]. The work done by Kalchbrenner et al.[8] is the only material we found that is roughly similar to our idea, where the sentence representations are being learned within their DCNN(Dynamic Convolutional Neural Network) structure. We intend to utilize some aspects of his DCNN structure in our model.

3 Methodology

3.1 Flowchart

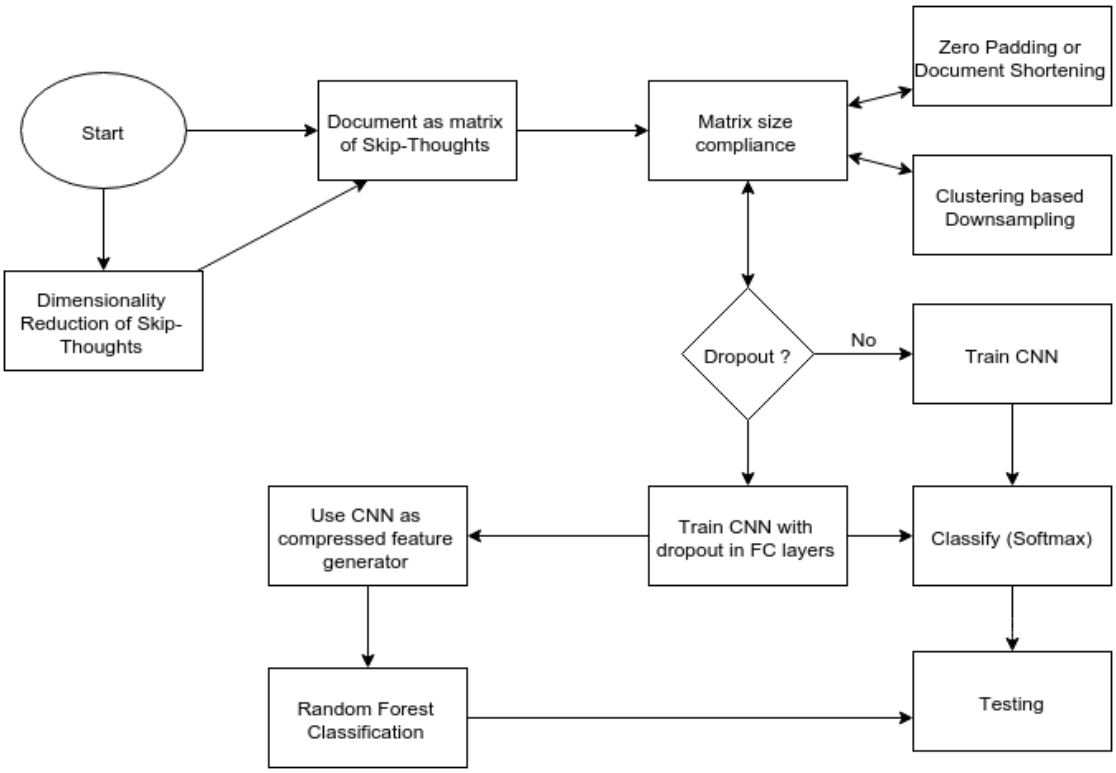

Figure 1: Basic Flowchart

3.2 Datasets

- $NLM - 500$: PubMed documents with MeSH terms
- $20 - Newsgroups$: 18821 docs with train/test splits on 20 news classes
- $CiteULike - 180$: 180 documents with high quality human tags from the CiteULike database

References

[1] Yoon Kim. Convolutional neural networks for sentence classification. *EMNLP 2014*, 2014.

[2] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011.

[4] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[5] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048, 2010.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] Mark J. Berger. Large scale multi-label text classification with semantic word vectors. 2014.

[8] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.