Document Classification for Movie Review

Project Proposal

Patel Jay
Roll No:31603118

Priyanka Jariha
Roll No:31603203

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 embed- ding, 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. We try to classify review of the movie blogs.

I. Introduction

In the recent past, a variety of NLP Tasks, such as part-of-speech tagging [4], sentiment classification [5], neural language models [6] and machine translation have consistently set new benchmarks. 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 unsuper- vised sentence encodings. We attempt to train a convolutional neural network, which given a representation of a document, learns to to perform various Document classification tasks on it. We present the network doing a binary sentiment classification task, but show how other tasks can be easily performed by slightly modifying the networks structure. We consider a document/sentence as a 2-D matrix consisting of concatenated vectors of its sen- tences/words. The size of the input to the convNet is calculated according to the dataset. The height of the 2-D input is set to the maximum length of a document(in sentences)/sentence(in words). Short documents are zero-padded and fed to the Convolutional Neural Network. The ConvNet is trained with 0.5 dropout in the fully connected layers. Effectively, we would be using the CNN as a feature generator then.

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, L eon 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, An- drew Y Ng, and Christopher Potts Potts. *Recursive deep models for semantic compositionality over a sentiment treebank.*In EMNLP, 2013.

[5] R Tomas Mikolov, Martin Karafi at, Lukas Burget, Jan Cernocky‘, 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 10451048, 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 31113119, 2013.