iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** X **Month of publication:** October 2017

DOI: http://doi.org/10.22214/ijraset.2017.10263

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue X, October 2017- Available at www.ijraset.com

Text Categorization with Dynamic Input Length Using Deep Learning – A Review

Pradnya Saval[1], Sheetal Rathi[2]
[1, 2] *Computer Engineering Department, Mumbai University*

Abstract: ***The increase in the development of internet and social networks brings a lot of user created texts on the internet such as product review, ratings, comments, etc. To classify the semantics of such texts is of great research and practical value. The categorization of these texts is difficult because of the less contextual information they provide. The existing research on text categorization basically includes feature extraction and sentiment analysis. With the development of deep learning various deep learning models have achieved remarkable results. The author explains and compares the various deep learning models to give a basic guidance for deep neural network selection.***
Keywords: Deep Learning, Recurrent neural network, Convolutional neural network, Text categorization, Neural Networks

I. INTRODUCTION

Text categorization is the task of assigning predefined categories to text documents. It provides conceptual views of documents and is used in various applications in the real world. For example, cooking recipes are typically organized by type of ingredients used, belonging to various geographical regions, recipes given by respective chefs, academics can be categorized to different domains, medical reports can be categorized to different tests and so on. Spam filtering is another application for text categorization where we filter or categorize emails which are not to our relevance.

Text categorization is based on words in which simple statistics of the combinations of words are performed till date. On the other hand, various researchers have found deep learning useful in separating information from raw data ranging from computer vision to speech recognition to text categorization. Natural language processing (NLP) has helped a lot from the rejuvenation of deep neural networks (DNN). There are various types of DNN architectures namely Recursive Neural Network (RecursiveNN), Recurrent Neural Network (RecurrentNN) and Convolutional Neural Network (CNN). To overcome the limitations of RNN algorithm two new types of algorithms have been generated namely Long short-term memory (LSTM) and Gated Recurrent Unit (GRU)[1].

A. Dep Neural Network Models

1) Recursive Neural Network (RecursiveNN): RecursiveNN has proved to be efficient in constructing sentences. The RecursiveNN seize the explanation of a sentence using a tree structure. The performance of this model wholely depends on the representation of the tree structure giving the time complexity of O(n2) where n is the length of the text. Using this model could be time consuming when the model works on long sentences and even relationship betweentexts is hard to represent using a tree structure. Therefore it is stated that modelling long sentences by recursive neural network is inadequate [2]. The main advantage of recursive neural network is that it is efficient in constructing sentences however it is inefficient to model long sentences and documents. The performance of this model depends on the textual tree structure and therefore the performance degrades as the tree structure increases depending upon the text.

2) Recurrent Neural Network (RecurrentNN): RecurrentNN is a model that provides better conceptual information and is sequential architectures. It analyzes a word and stores the information of all the preceding words in a fixed sized hidden layer. This model is stated as a biased model where all the next words in a sentence are effective than the previous one due to which the information is available at the end of the document. But this can be a problem as the effective information will not always be present at the end of the document and the model uses its internal memory to process arbitrary sequences of inputs. The time complexity of RecurrentNN is O(n). RNN has the capture the contextual information and can capture semantics of long texts as it can handle arbitrary input and output lengths. The application of RNN is mostly used for text and speech analysis [3].

3) Convolutional Neural Network (CNN): CNN reports higher performance compared to RecurrentNN. CNN computes the most informative n-grams of all the words in a sentence and extracts the most valuable information. CNN can be used of sentiment categorization as the sentiments are determined by key phrases. It is an unbiased model and uses feed-forward ANN to fairly discriminate phrases using a max-pooling layer. Therefore CNN can capture semantic information better than recurrent and recursive neural network. The time complexity of CNN is also O(n). However, the previous studies tend that CNN use simple

convolutional kernels due to which it is difficult to determine the size of the window. . The application of CNN is mostly used for image and video processing [4].

4) Recurennt Convolutional Neural Network (RCNN): To overcome the demerits of the above model RCNN was introduced for text categorization. First, the bi-directional recurrent structure is applied which introduce less noise compared to other traditional window based neural network and hence it provides better contextual information. Second, we introduce a max-pooling layer which fairly determines which feature plays an important role compared to other words in the sentence. Hence, the RCNN is a combination of recurrent neural network and convolutional neural network with a time complexity of O(n) [4].

II. RELATED WORK

In the field of neural network, researchers have worked upon few challenges on the various models of deep learning. In [1], the CNN is combined with LSTM without the method of word segmentation. The methodology used was Char-CNN based on Zhang's model where k-max pooling is implemented instead of max-pooling. Max-pooling is a method for down sampling where a sliding window is used on a row and selects a cell with maximum value and then the window is passed to next layer. K-max pooling on the other hand doesn't have a window instead it performs selecting operation for the entire row. The top k values with maximum value is selected and passed to the next layer. Therefore, the authors have used Char-CNN method by using k-max pooling which has the capability to accept any length of input before a fully-connected layer and gives better accuracy than other traditional word-level methods. This method can be applied to other natural languages using word segmentation for future research.

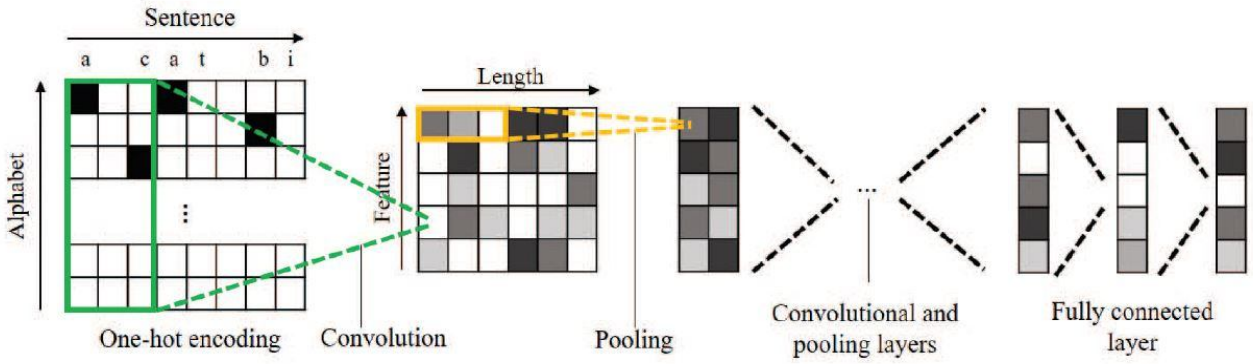


Fig. 1 Char-CNN proposed by Zhang [1]

In [2], the author has applied RCNN a recurrent structure to capture the contextual information using a bi-directional recurrent structure along with the max-pooling layer which helps to extract the most relevant feature and key components in the text or document. Using this biased model the later words are more dominant than the previous one. However, the effectiveness is reduced as the key components will not always be present at the end of the document as the simple convolutional kernels are difficult to determine the window size. Hence, it is important to determine the window size for a simple convolutional kernel to determine the features of a document.In [3], the authors have given a systematic comparison of the deep learning models such as CNN, LSTM and GRU used in natural language processing and aiming to select the most appropriate deep neural network models. The methods implemented contains the training data using basic setup without complex tricks and have used for searching parameters for each model separately. The comparison mentions that the CNN model uses convolutional layer and GRU models the input from left to right and use the last layer as the result of the input. Due to this the variation in the batch size and hidden size causes oscillation which can be overcome if it can be optimized to increase the performance of the models. The usage of CNN model helps to extract lexical and sentence level features. The model takes all the word tokens and transforms to vectors using word embeddings. The lexical features are then extracted according to the nouns present in the sentence. After the extraction both the methods are combined to produce the final feature vector. To specify which pair of words needs to be labels to the respective word. The author has used SemEval-2010 Task 8 dataset which is freely available and contains 10,717 samples along with 8,000 and 2,717 training and test instances respectively. [4].In [5], the author has used various large-scale dataset such as AG's news with 4 classes, 120,000 and 7,600 train and test samples respectively along with other datasets and implemented in the convolutional neural network. The comparison of these dataset is among the various other traditional models such as bag-of-words, n-grams and other deep learning models.

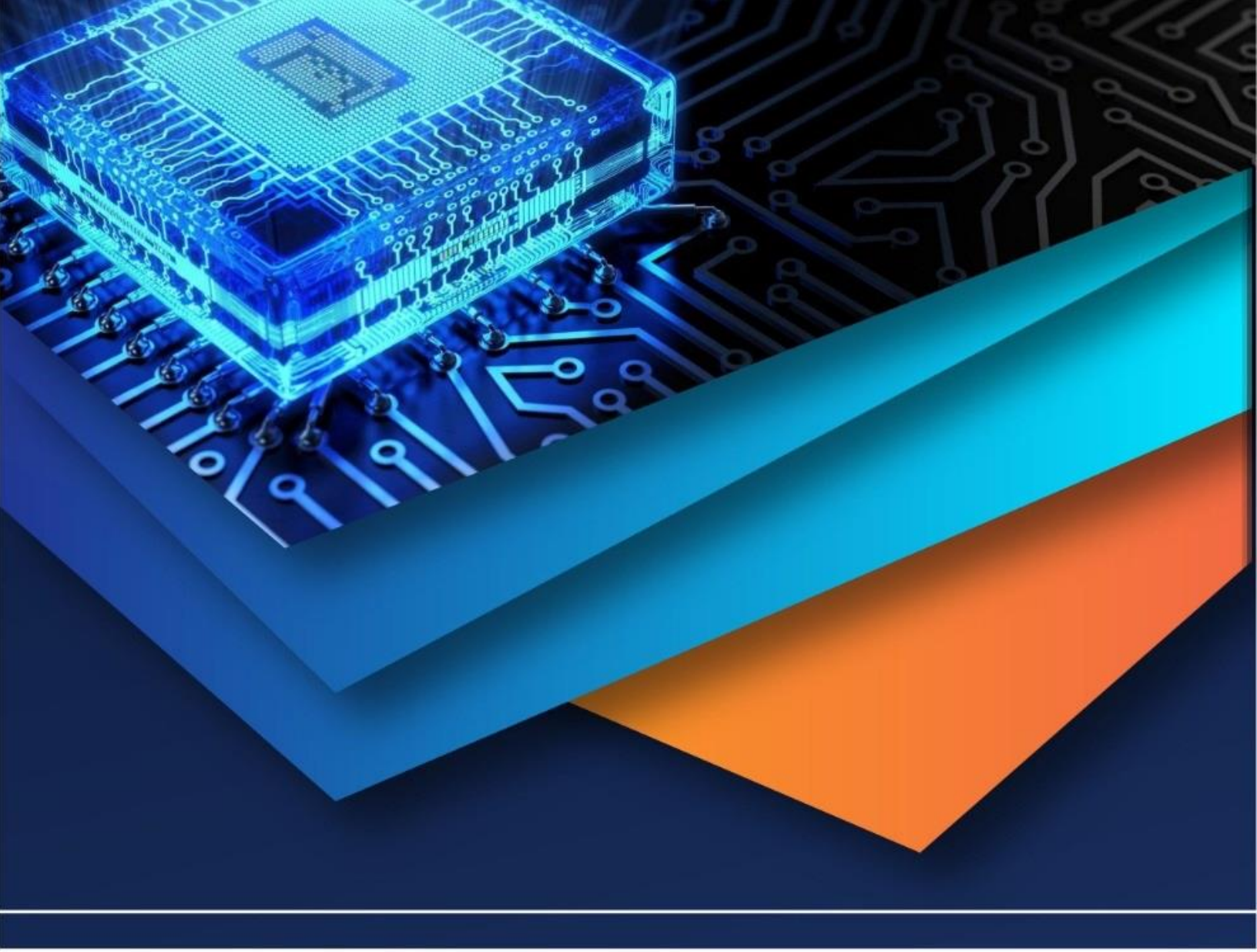
There are two types of CNN namely, straightforward adaptation of CNN from image to text and second is simple CNN using bag-of-words. There various methods used to categorize using CNN are CNN for image, CNN for text, seq-CNN for text and bow-CNN for text. The comparison of error rates (%) shown in the paper, shows that seq-CNN outperforms other methods of categorization using three datasets namely Movie Reviews (IMIDB) with 8.74, Electronics Product Reviews (Elec) with 7.78 and News Articles (RCV1) with 9.96 [6].

III.CONCLUSIONS

This work compared the most widely used deep neural networks namely RecurrentNN, RNN and CNN by discussing various challenges faced by other researchers and further scope of the research. The author has described all the models with respect to its description, the methodology used by other researchers in the respective field along with the solutions and future scope of the model. The author even mentioned the demerits of the models which need to be taken care of before implementing any of the mentioned models.

REFERENCES

[1] Thanabhat Koomsubha,"A Character-level Convolutional Neural Network with Dynamic Input Length for Thai Text Categorization", 978-1-4673-9077-4/17/$31.00 ©2017 IEEE.
[2] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao,"Recurrent Convolutional Neural Networks for Text Classification", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[3] Wenpeng Yin, Katharina Kann, Mo Yu, Hinrich Schutze,"Comparative Study of CNN and RNN for Natural Language Processing", Computation and Language, February 2017.
[4] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao,"Relation Classification via Convolutional Deep Neural Network",Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2335–2344, Dublin, Ireland, August 23-29 2014.
[5] Xiang Zhang, Junbo Zhao, Yann LeCum,"Character-level Convolutional Networks for Text Classification", Computation and Language, December 2015.
[6] Rie Johnson, Tong Zhang,"Effective Use ofWord Order for Text Categorization with Convolutional Neural Networks", Computation and Language, December 2014.
[7] Tharani.S, Dr. C. Yamini," Classification using Convolutional Neural Network for Heart and Diabetics Datasets", International Journal of Advanced Research in Computer and Communication Engineering, ISO 3297:2007 Certified Vol. 5, Issue 12, December 2016.
[8] Shiyao Wang, Zhidong Deng,"Tightly-coupled convolutional neural network with spatial-temporal memory for text classification", International Joint Conference on Neural Networks (IJCNN), 2017.
[9] Rita Georgina Guimarães, Renata L. Rosa, Denise De Gaetano,"Age Groups Classification in Social Network Using Deep Learning", IEEE Access, Vol.5,2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY