ISSN: 2057-5688

Deep Learning-Based Text Classification Using TCNN-DAM Model

[1]GODUGU ARAVIND BABU, [2]M. NARESH

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract*: In recent years, deep learning-based models have significantly improved the Natural Language Processing (NLP) tasks. Specifically, the Convolutional Neural Network (CNN), initially used for computer vision, has shown remarkable performance for text data in various NLP problems. Most of the existing CNN-based models use 1-dimensional convolving filters (n-gram detectors), where each filter specialises in extracting ngrams features of a particular input word embedding. Deep learning technology develops rapidly. Convolutional Neural Network (CNN), as a key technology in deep learning, has been favored and concerned by many scholars and widely applied in information retrieval, classification, data management, mining and other fields. In order to fully obtain the local features and key words of text, a TCNN-DAM model is proposed based on the study of TCNN model, which aims at maximizing the representation of text features, improving the text classification effect, and promoting the model to better classify in sogou news corpus. Tests show that the improved model has outstanding classification effects, which can effectively improve the accuracy, precision, and F1-score.*

Keywords*: Deep learning, text classification, convolutional neural network, TCNN model.*

I. INTRODUCTION

Deep learning models have achieved remarkable results in computer vision and speech recognition in recent years. Within natural language processing, much of the work with deep learning methods has involved learning word vector representations through neural language models and performing composition over the learned word vectors for classification. Word vectors, wherein words are projected from a sparse, 1-of-V encoding (here V is the vocabulary size) onto a lower dimensional vector space via a hidden layer, are essentially feature extractors that encode semantic features of words in their dimensions.

Text classification also known as text categorization, is a classic problem in natural language processing (NLP), which aims to assign tags or labels to textual devices such as sentences, questions, paragraphs, and files. It has a wide range of programs including question answering, spam detection, sentiment analysis, news classification, user categories, content moderation, etc. Text data can come from unique sources such as web statistics, emails, chats, social media, tickets, insurance claims, customer reviews, and buyer submitted questions and answers, to name a few. Text is a very rich source of data. But extracting information from textual content can be difficult and time-consuming due to its unstructured nature. Text class guide can be done by both annotation and automatic tagging. With the increasing scale of textual information in industrial applications, the type of computerized text content is becoming more and more important. Approaches to the computerized text category can be grouped into categories:

- Principle-based strategy
- Machine domain-based methods (facts driven)

Rule-based methods classify text into different categories using a set of pre-defined rules, and require a deep domain knowledge. On the other hand, machine learning based approaches learn to classify text based on observations of data. Using pre-labeled examples as training data, a machine learning algorithm learns inherent associations between texts and their labels. Machine learning models have drawn lots of attention in recent years. Most classical machine learning based models follow the two-step procedure. In the first step, some hand-crafted features are extracted from the documents (or any other textual unit). In the second step, those features are fed to a classifier to make a prediction. Popular hand-crafted features include bag of words (BoW) and their extensions. Popular choices of classification algorithms include Naïve Bayes, support vector machines (SVM), hidden Markov model (HMM), gradient boosting trees, and random forests. The two-step approach has several limitations. For example, reliance on the handcrafted features requires tedious feature engineering and analysis to obtain good performance. In addition, the strong dependence on domain knowledge for designing features makes the method difficult to generalize to new tasks. Finally, these models cannot take full advantage of large amounts of training data because the features (or feature templates) are pre-defined. Neural approaches have been explored to address the limitations due to the use of hand-craft features. The core component of these approaches is a machine-learned embedding model that maps text into a low-dimensional continuous feature vector, thus no hand-crafted features is needed.

One of earliest embedding models is latent semantic analysis (LSA) developed by Dumais et al. [1] in 1989. LSA is a linear model with less than 1 million parameters, trained on 200K words. In 2001, Bengio et al. [2] propose the first neural language model based on a feed-forward neural network trained on 14 million words. However, these early embedding models underperform classical models using hand-crafted features, and thus are not widely adopted. A paradigm shift starts when much larger embedding models are developed using much larger amounts of training data. In 2013, Google develops a series of word2vec models [3] that are trained on 6 billion words and immediately become popular for many NLP tasks. In 2017, the teams from AI2 and University of Washington develops a contextual embedding model based on a 3-layer bidirectional LSTM with 93M parameters trained on 1B words. The model, called ELMo, works much better than word2vec because they capture contextual information. .

II. LITERATURE SURVEY

Sanjay K Dwivedi et al. [4] provided an overview of the approaches of the question-answer system. Question answer system (QA) have three stages, i.e., Question analysis: parsing, question classification, and query reformulation; Document analysis: extract possible documents and recognize the answer, and Answer analysis: extract possible answer and grade the best one. They discussed classification for characterizing question answer approaches which includes.

Linguistic Approach: depending on the structure and knowledge of the language, fetched same meaning in the different expression. Linguistic techniques such as tokenization, POS tagging, and parsing are executed to user's question and extract the particular response from the structured database [5].

Statistical Approach: is used for a large amount of data having sufficient amount of data for statistical comparison to be considered important. This approach is made to provide the enough amount of learning data during training statistical models, and the system could possibly deliver the positive response of even complex questions if the statistical model has been properly trained [6].

Pattern Based Approach: is an astonishing efficient technique for utilizing the web as a source data. Instead of linguistic analysis, pattern-based approach utilizes eloquence of text pattern, pattern matching not only decreasing the linguistic computations but also help in automatic wrapper generation for controlling the divergent web data [7].

Some of the previous literature review such as survey of text question answering techniques explained the types of Question Answering system which includes: Web based Question Answering system (using search engines to get back hyperlink that containing answers to the questions), IR/IE Question Answering system (returning a group of top ranked documents as responses to the query), Restricted Domain Question Answering system (needed a linguistic support to identify the natural language text to get the answer of the questions correctly), Rule Based Question Answering system (extended form for IR based question answering system) and Classification of Questioners level (the questions are classified into different levels according to its context) in the paper.

Xiang Zhang, et al. provided an overview on character-level Convolutional Networks for text classification. They offer an actual method on the use of character-level convolutional networks for text classification. Applying convolutional network to text classification shown that convolutional network can be directly applied to distinct set of words without any information on the syntactic or semantic structures of the languages. They also constructed several large-scale datasets to show that character-level convolutional network could achieve competitive results [8].

Siwei Lai et al. [9] provided an overview for text classification through recurrent convolutional neural network without human desired features. In this paper, they provide the model of deep neural network to learn the semantics of the text and using the four datasets such as Fudan set (Chinese document classification that consist of twenty categories includes energy, education and art), 20Newsgroups (Choose four major headings like comp, politics, rec, religion from twenty newspaper), Stanford Sentiment Treebank (dataset which contains information about movie reviews and classified in different ways like Very Negative, Negative, Neutral, Positive, Very Positive), ACL Anthology Network (dataset holds the scientific documents) which are initially trained through different learning methods such as word representation learning (by combination of a word and its context we get a more precise word meaning) and text representation learning (represents the text data). They also use the max-pooling layer in text representation learning for down sampling the data that automatically judges the most important semantic factors in the dataset and determine which word play key role in text classification to get the key components in texts.

III. PROPOSED SYSTEM

The implemented Convolution Neural Network model has three layers. Basic functionality of convolution neural network resembles to the visual cortex of the animal brain. In text classification task convolution neural network gives promising result. Criteria for text classification is similar to image classification only difference is that instead of pixel values we have matrix of word vectors. Proposed model is implemented in python using tensorflow library.

SYSTEM ARCHITECTURE

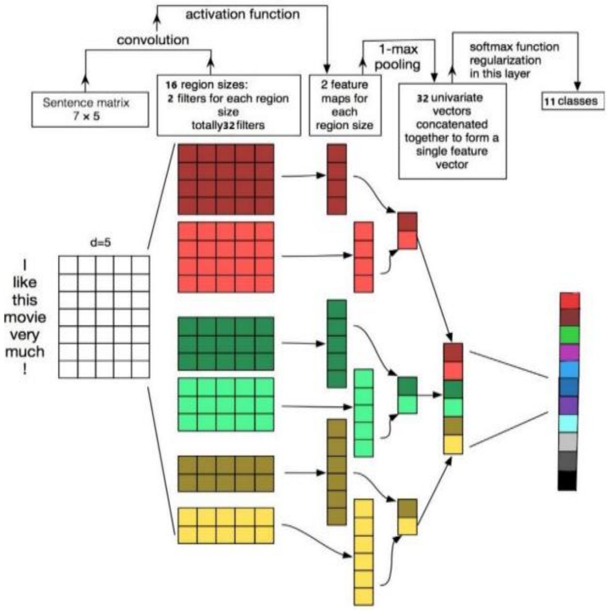

activation function
convolution
Sentence matrix 7 × 5
16 region sizes: 2 filters for each region size totally32 filters
2 feature maps for each region size
1-max pooling
32 univariate vectors concatenated together to form a single feature vector
softmax function regularization in this layer
11 classes
d=5
I
like
this
movie
very
much
!

Fig.1 System architecture

The model implemented includes three layers. First is the Embedding layer which convert the words to its respective embedding vectors, second is the convolution layer in which main processing of the model takes place. Predefined filters roll over the sentence matrix and reduce it into low dimensional matrix. Third layer is softmax layer which is a down sampling layer capable of reducing sentence matrix and calculating loss function. Embedding lookup function is used to get the word embedding of the sentence. The matrix generated as a result of embedding layer is than padded to equalize all the sentences. The defined filters will than start reducing the matrix and generate convolved features. These convolved features are than further reduced. The output generated as a result of convolved features is than spread over the max pooling layer for further down sampling of output. Filters of different sizes and shapes are defined. The shape of the filters use in the proposed model is (3, 4, 5). Total number of filter is 33. The filters will roll over the original sentence matrix thus reducing it to low dimensional matrix. Instead of training our own embedding we use embedding lookup function of tensorflow. Embedded sentences are than padded to make all the sentence matrices of same size and shape.

Improvement of the TCNN model

The emergence of the TCNN model has prompted the application of convolutional neural networks in text classification to take a step forward. The unique feature of the TCNN model is to obtain high-quality local text information in the feature extraction process through a variety of convolution operations. The disadvantage lies in the inability to maximize local key features. In order to solve the shortcomings of the TCNN model, an improved deep learning model text classification algorithm TCNN-DAM is proposed. The network structure of the two models is shown in Figure 1 and Figure

Comparing the TCNN-DAM model and the TCNN model, the advantage is that it can express the local features of the text to the maximum, enhance the keyword features, and maximize the classification effect. The specific content of the TCNN- DAM model is:

The text set is stored in the form of rows. When processing the text set, select convolution kernels with different sizes. The intercept size set for each row is 200, so the convolution kernel size Set to 3 to 5 is the best

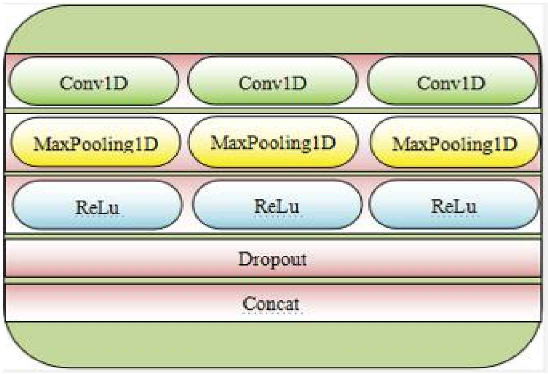


Fig.2 The TCNN model network structure

(2) The DAM dual-channel attention mechanism is added to the TCNN model to assign weights to the extracted text feature information, aiming to strengthen the keyword features and maximize the representation of text information.

(3) The model output adopts the gradient descent method Admax and combines with Softmax for normalization processing, aiming to reduce the upper layer output change.

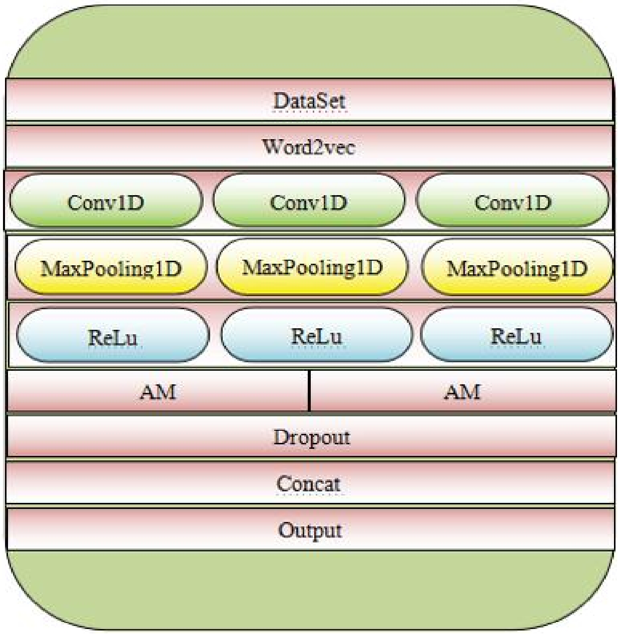


Fig.3 The TCNN-DAM model network structure

IV. EXPERIMENTAL EVALUATION

We conduct series of experiment on convolution neural network using dataset described as follow:

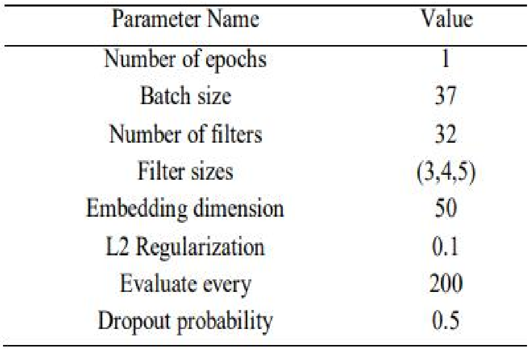
Dataset

The dataset used is of Consumer Complaints. There is total 11 categories in which data is classified. Total number of rows is 555958. Dataset is in csv MS Excel format. Dataset is split into train, test and validation set. We used predefined function of split dataset.

Learning

We experimented with model to check the flexibility of it. List of parameters is as follow

Table 1. Default Parameters.



Parameter Name	Value
Number of epochs	1
Batch size	37
Number of filters	32
Filter sizes	(3,4,5)
Embedding dimension	50
L2 Regularization	0.1
Evaluate every	200
Dropout probability	0.5

We experimented with parameters by increasing number of epochs we found significant change in the results. Batch size and number of filters are defined according to the dataset. These are the default for this dataset. Embedding dimensions is defined according to the maximum sentence length. Evaluation of training takes place every 200 steps on the validation set. Dropout probability is used to down sample the output by 0.5 %

Evaluation Criteria

An evaluation criterion is based on the following metrics:

Accuracy

Accuracy is a measure of percentage value for correct predictions of data. It is calculated by the following formula

$$Accuracy = \frac{correct\ predictions}{all\ predictions}$$

Precision

Precision is a measure for which model is correctly predicting any particular category. It is calculated by following formula

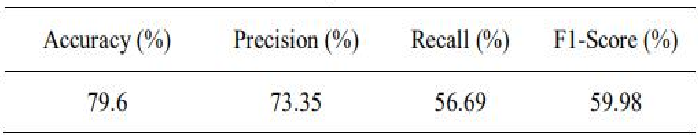
$$Precision = \frac{particular\ category\ predicted\ correctly}{all\ category\ predictions}$$

Recall

Recall is calculated by following formula:

$$Recall = \frac{correctly\ predicted\ category}{all\ real\ categories}$$

Table.2



Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
79.6	73.35	56.69	59.98

V. CONCLUSION

This paper summarizes the previous experience and theory, focusing on the use of deep learning method for news text classification. In view of the TCNN model can not fully obtain the information of text parts and keywords, the TCNN-DAM model is proposed, so that its accuracy, accuracy, recall rate and value are improved compared with the previous models. Of course, the data set adopted in this paper is relatively single, and a variety of data sets will be used for training in the later stage. Meanwhile, the research based on deep learning text classification will be strengthened.

REFERENCES

[1]. V. G. Poonam Gupta, "A Survey of Text Question Answering Techniques," International Journal of Computer Applications (0975 – 8887) , Vols. Volume 53– No.4,, September 2012 [2]. Moreda, Paloma., Llorens Hector., Saquete, Estela. and Palomar, Manuel. "Combining semantic information in question answering systems" Journal of Information Processing and Management 47, 2011. 870- 885. DOI: 10.1016/j.ipm.2010.03.008. Elsevier.

[3]. C. dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 69–78, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[4]. R. Johnson and T. Zhang. Effective use of word order for text categorization with convolutional neural networks. CoRR, abs/1412.1058, 2014.

[5]. Prasadu Peddi (2021), "Deeper Image Segmentation using Lloyd's Algorithm", ZKGINTERNATIONAL, vol 5, issue 2, pp: 1-7.

[6]. Li, DU. Jia. and Fang, YU .Ping. 2010. Towards natural language processing: A well-formed substring table approach to understanding garden path sentence. 978-14244-6977-2/10, IEEE.

[7]. Suarez, O. S., Riudavets, F. J. C., Figueroa, Z. H., and Cabrera, A. C. G.

"Integration of an XML electronic dictionary with linguistic tools for natural language processing" Journal of Information Processing & Management, vol. 43, 2007, 946-95

[8] Kim Y. Convolutional neural networks for sentence classification [C]//Empirical Methods in Natural Language Processing. 2014:1746- 1751.

[9] Prasadu Peddi (2018), Data sharing Privacy in Mobile cloud using AES, ISSN 2319-1953, volume 7, issue 4.