Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/149

TEXT CLASSFICATION USING CNN AND CNN-LSTM

Noor Mohd[1], Harsiddhi Singhdev[1], Deepak Upadhyay[2]

[1]Department of Computer Science & Engineering Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
[2]Department of Computer Science & Engineering, Graphic Era Hill University, Dehradun, Uttarakhand India, 248002

ABSTRACT

Text classification, language modelling, and machine translation are some of the applications. CNNs excel at retrieving local and position-invariant characteristics, but RNNs excel at classification based on a long-range semantic dependency instead of local key-value pairs. It shows out that CNNs perform admirably when used to NLP problems. The basic (Bow) model is a clear oversimplified based on faulty beliefs, but it has been the usual technique for years and has produced reasonable accuracy. The speed of CNNs is a major selling point. In this paper, Text classification is carried out by using a deep learning model that is CNN and a hybrid model using CNN-LSTM and compare the performance of two models.

Keywords: CNN, LSTM, NLP and Machine learning

INTRODUCTION

Many applications, such as internet searches, information filtering, and text categorization, rely on text classification. As a result, it has attracted the attention of a large number of researchers. Feature representation, which is often based on the bag-of-words method, is a major issue in text classification. (BoW) model, in which unigrams, bigrams, n-grams, or some combination of the three are used. Typically, features are extracted from exquisitely designed patterns. Classical action recognition methods, on the other hand, frequently overlook contextual details or sentence structurein texts are not captured, and hence the semantics of the words are not captured. The continued advancement of skilled word embedding and deep neural networks has recently sparked new interest in a variety of Nlp applications. Traditional text classification can be divided into two categories: (1) Text classification based on expert systems and knowledge engineering. (2) Machine learningbased text classification. The former has numerous flaws in terms of flexibility, extensibility, and classification effect. K-Nearest Neighbor (KNN),Support Vector Machine (SVM), [4], Maximum Entropy(ME) and other algorithms are commonly used in text classification based on machine learning. When attempting to deal with large amounts of data, such algorithms' text representation is usually a high dimensional-degree sparse vector, which has a weak feature expression capability and necessitates manual feature engineering. As a result of the loss of word order data, the BOW model's text features do not capture the semantics of texts. High-dimensionality and data sparsity also issue. Word embedding as distributed representations of words is proposed as a solution to these issues. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are the most common deep learning models used in text classification

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/149

(RNNs). An even more highly regarded architecture for interacting with NLP tasks is the Convolution Neural Network (CNN). Detecting local associations in a equivalent manner with convolutional filters, CNN excels at extracting ngram features. The weighted sum of hidden states is more accurately reflects the original sentence. In Text classification, sentence modelling [17], and semantic parsing [18], CNN has been recommended for dealing with NLP tasks, and it has produced impressive results. However, the width of convolutional filters affects their ability to extract n-gram features. Given the importance and use of news articles, several studies have looked into the Deep learning capacity of the word embedding tool word2vec and associated CNNs. LSTM mainly used for processing long sentences or text. It can memorize the relationship of long-distance dependence and keep the main semantic information. In this paper, Text classification is carried out by using a deep learning model that is CNN and a hybrid model using CNN & LSTM and compare the performance of two models.

In[1] Classical text classification research focuses on three main areas: feature engineering, attributes selection, and the application of various ML algorithms. In feature engineering, the bag-of-words (BOW) feature is by far the most popular. Furthermore, some more advanced functions have been added to the system. Tree kernels noun phrases +, and POS tags are some of the things that have been designed. In [2] Feature selection tries to improve the performance of the classifier by removing noisy features. The most typical way for selecting features is to remove stop words (for example, "the"). To choose useful characteristics, advanced approaches use information gain, mutual information, or L1 regularization. Classifiers like support vector machine (SVM) , naïve Bayes (NB), and logistic regression (LR), are frequently used in machine learning methods . These approaches, however, have a data sparsity issue. In[3] Deep neural networks and feature learning have recently sparked new concepts for addressing data sparsity, and various word representations have been learned with the help of neural models. Word embedding is a real-valued vector that represents a neural representation of a word. We can make use of the separation in between two embedding vectors to estimate using word embedding to determine word interdependence. In[4] Neural networks accomplish admirably in a wide range of NLP jobs when using skilled word embeddings. To forecast the sentiment of a text, utilise semi-supervised recursive auto encoders. Using an RNN, suggested a technique for paraphrase identification. To analyses the sentiment of words and sentences, a recursive neural tensor network was introduced. To create language models, a recurrent neural network is used. For dialogue act classification, a novel recurrent network was presented. Introduce a semantic role labelling convolutional neural network. In [13, 14] there are several ways talked about CNN modeling.

In [5-6] a character-level CNN classification approach was proposed. Six convolutional layers with convolution kernel sizes of 3 and 7 are used in their method, followed by three fully linked layers. It's the first programme to analyses sentiment at the character level. In [8-9] developed a hierarchical attention network that uses a "word-sentence-article" hierarchical system to incorporate attention processes at the word and sentence levels, allowing words and phrases of differing significance in the text to have different "attention" capacities. In [10] A fastText model with only three layers was proposed: an input layer, a hidden layer, and an output layer, which is convenient and effective. To obtain the vector of the document, the model employs the hierarchical softmax in response to shorten the model's training time by adding the character level n-gram feature of the word as an advanced features to the word vector of the full document as input. On some data sets, it achieves

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/149

higher accuracy than the above approaches with a faster training pace. In [11] developed three distinct methods for text categorization exchanging information with recurrent neural networks. The approach seeks to enhance categorization by learning tasks in parallel and makes use of the correlation between related tasks. On four benchmark text categorization tasks, it outperforms the competition.

proposed model

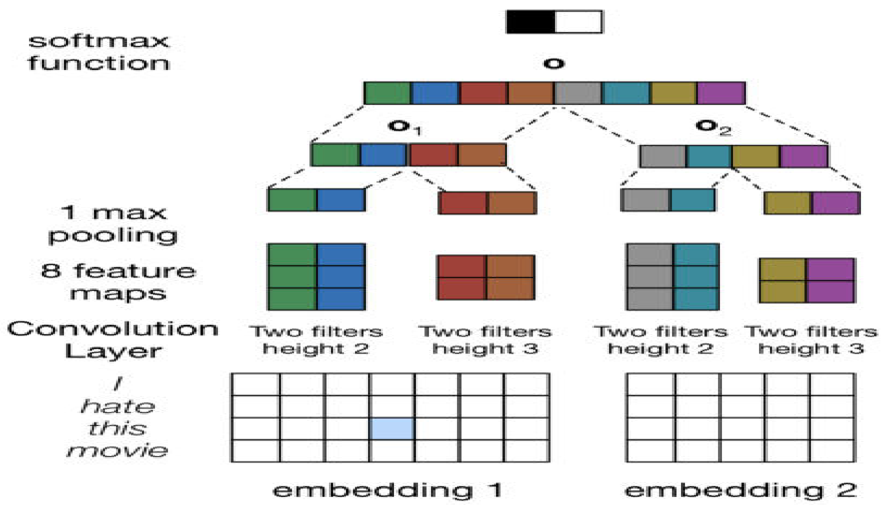

Figure 1. Schematic of CNN Structure

use of neural networks can help to decrease reliance on feature engineering. By using pooling, the text classification approach based on CNN (Convolutional neural network) may acquire significant text properties, but contextual information is difficult to acquire. Text classification procedure based on CNN (Convolutional neural network) can acquire prominent attributes of text through pooling, but it is hard to access significant information. RNN (Recurrent neural networks) can gather context information, however the order in which words are said will bias the results; Using pooling, the text assessment based on CNN (Convolutional neural network) may extract significant text properties, however it is difficult to access contextual features. The primary goal of text classification is to detect whether the thoughts represented in a document, sentence, or entity feature/aspect are positive, negative, or neutral, and at the document, phrase, or attribute/aspect level, identify the polarity of that text.

working flow of cnn

A Convolutional Neural Network, in general, contains three layers, which are as follows:

- **Input:** A picture's 32 widths, 32 heights, and three R, G, and B channels serve as the raw pixel ([32x32x3]) values.
- **Convolution:** Each neuron calculates a dot product between weights and a small region in the input volume to which they are truly linked in order to determine the outcome of those neurons that are associated to the local regions of the input. For instance, if we choose to utilise 12 filters, the volume will be [32x32x12].
- **ReLU Layer:** Element-wise activation functions like max (0, x) thresholding at zero are built using it. It produces ([32x32x12]), which is equivalent to the same-sized volume.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/149

- **Pooling:** This tier is used to perform a down-sampling technique with (width, height) in the spatial dimensions, producing a volume of [16x16x12].
- **Locally Connected: -** It can be characterized as a typical NN (Neural Network) layer that receives input from the layer above, analyses class scores, and generates a 1-dimensional array with the same number of dimensions as classes.

Cnn-lstm

The architectural design of the CNN-LSTM approach is shown in Figure 1. For text classification, It has a convolutional and recurrent layer approach. To discover long-term dependencies, In this investigation, the Maxpool layer is replaced by a recurrent layer. Instead of employing the maxpool layer, we're using the CNN-LSTM architecture.

Text Evaluations- Movie critiques are used as system input. Only reviews written in English are included, and punctuation such as periods, brackets, and commas are separated by white space.

Pre-processing- Pre-processing activities are required to increase the performance of our model. Before incorporating the sample into the deep learning model, we first performed data preprocessing processes on it. We used RE to remove punctuation, halt words, bracket integers, and conduct stemming. After cleaning the dataset, use the tokenize API to conduct tokenization. The technique of tokenization involves transforming words into a single number. Then we built a lexicon to translate these words into numbers.

Embedding words- The embedding level is responsible for a succession of inputs and implants each word into a relevant size feature vector space.

Convolution Layer- Neural networks can extract information horizontally from many words. These features are critical for categorization jobs. Word embedding will be used as input to a convolutional layer. A convolutional operation is applied to the input by several convolutional layers, with the outcome passed on to the next layer.

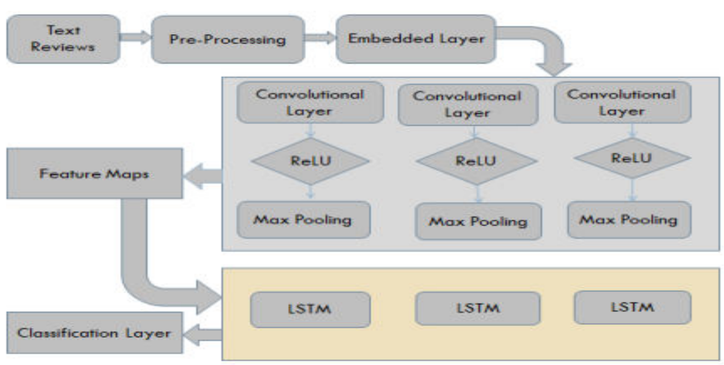

Figure 2. CNN-LSTM Model

The maximum pooling layer- The sub sampling layer, also known as the pooling layer, is a layer that collects data from multiple sourcess, can reduce the number of features and the system's processing complexity. The most important data is kept in the subsampling stage.

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/149

LSTM Layer- The LSTM layer is used to classify the extracted characteristics. It's a special kind of RNN that can adjust to long-term connections.. The viability of LSTM is based on its ability to identify shifting opinions in a twitter post. That seems to have a recall, which "remembers" previous input data and takes actions as a result of it. Because of the layer's stability, past data can be used to influence future input.

Classification Layer- A classification layer is formed by a fully connected layer. After that, the results of the LSTM are loaded into a Keras dense layer-based Fully Connected Layer (FCL). This layer is followed by a basic sigmoid activation function that confirms the output is between 0 and 1.

Result

The proposed method makes use of a data collection of 1000 good and 1000 negative movie reviews from IMDB. For training and validation, we randomly divide the complete set of training samples. Training and validation datasets are separated from the input dataset. Keras provides two methods for evaluating deep learning models. The first is dataset verification that is carried out automatically, while the second is manual dataset verification. The performance of our model is assessed on the validation dataset at each epoch using Keras, which separates a portion of the training data into validation data. The experiment used the following settings, which were adjusted manually during experimentation.

Embedding Dimension	100
Epoch	10
Filters	32
Kernel Size	8
Pool Size	2
Dropout	0.5

Table 1. Hyper-Type Parameters

Accuracy- In ten epochs, the proposed CNN-LSTM network achieved 79 percent accuracy. The system's accuracy is proportional to the number of epochs.

Table 2. Comparison Table

Model	Epoch 10 (Avg Accuracy)	Epoch 20 (Avg Accuracy)	Epoch 30 (Avg Accuracy)
CNN	72%	73%	73.8%
CNN-LSTM	78%	79%	79.5%

Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/149

In ten epochs, the proposed CNN-LSTM network achieved 79 percent accuracy. The system's accuracy is proportional to the number of epochs.

Conclusion

The amount and pace of text data are growing on a daily basis. Sentiment analysis is a key part of our decision-making process on a daily basis. Various approaches are used to conduct sentiment analysis. Deep learning approaches have improved the accuracy of sentiment classifications in recent years when compared to previous methods. The Convolutional Neural Network is an excellent feature learner, extracting information horizontally from a variety of text. For categorization jobs, these properties are critical. In extended words, the LSTM Network keeps track of long-term dependencies. The viability of LSTM is based on its ability to detect shifting mood in a twitter post. That seems to have a recollection that "remembers" past information and makes judgments depending on what it has learned. In our technique, the benefits of both LSTM and CNN are integrated. In this study, we investigated two various deep learning models for evaluating IMDB movie reviews. A CNN model makes up the first, while a CNN-LSTM hybrid model makes up the second. On the IMDB movie review dataset, our technique performed admirably and exceeded other conventional models in terms of accuracy. Future work includes experimenting with various LSTM types, including CNNs and Bi-LSTMs (for example multilayer CNN). In the future, this technology will be applied to very large datasets.

References

1. Aggarwal, C. C., and Zhai, C. 2012. A survey of text classification algorithms. In mining text data. Springer. 163–222.
2. Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL, 238–247.
3. Li, W.; Sun, L.; and Zhang, D. 2008. Text classification based on labeled-lda model. Chinese Journal of Computers 31(4):620–627.
4. Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In NAACL-HLT, 746–751
5. Post, M., and Bergsma, S. 2013. Explicit and implicit syntactic features for text classification. In ACL, 866–872.
6. A.M.Abirami and V.Gayathri "A Survey on Sentiment Analysis Methods and Approach" 2016 IEEE Eighth International Conference on Advanced Computing (ICoAC).
7. Qiongxia Huang, Xianghan Zheng, Riqing Chen, Riqing Chen "Deep Sentiment Representation Based on CNN and LSTM" International Conference on Green Informatics 2017.
8. S. M. Vohra, J. B. Teraiya "A Comparative Study of Sentiment Analysis Techniques" Journal of Information, Knowledge and Research in Computer Engineering, 2016.
9. Yaya Heryadi, Harco Leslie Hendric Spits Warnars "Learning Temporal Representation of Transaction Amount for Fraudulent Transaction Recognition using CNN, Stacked LSTM, and CNN-LSTM" Journal of Electronics Vol.25, No.4, July 2016.
10. Alec Yenter , Abhishek Verma "Deep CNN-LSTM with Combined Kernels from Multiple Branches for IMDb Review Sentiment Analysis" IEEE International Conference on Computer Communication and the Internet 2017.

11. Zhou, Yu, Ruifeng Xu, Lin Gui "A sequence level latent topic modeling method for sentiment analysis via CNN based diversified restrict boltzmann machine" Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. Vol. 1. IEEE, 2016.
12. Xiaobo Zhang, Qingsong Yu ” Hotel Reviews Sentiment Analysis Based on Word Vector Clustering” 2nd IEEE International Conference on Computational Intelligence and Applications 2017.
13. Tiwari, P., Upadhyay, D., Pant, B., & Mohd, N. (2022). Multiclass Classification of Disease Using CNN and SVM of Medical Imaging. In International Conference on Advances in Computing and Data Sciences (pp. 88-99). Springer, Cham.
14. Tiwari, P., Pant, B., Elarabawy, M. M., Abd-Elnaby, M., Mohd, N., Dhiman, G., & Sharma, S. (2022). CNN Based Multiclass Brain Tumor Detection Using Medical Imaging. *Computational Intelligence and Neuroscience*, *2022*.