Deep Learning in Chinese Text Recognition Program: Exploring the Influence of Program Parameters on Recognition Accuracy

Zhengyu Peng

HenanVictoria University, Kaifeng, Henan, China, 475001
Zhengyu.peng@live.vu.edu

Abstract

Deep learning (DL), a class of pattern analysis methods for learning the intrinsic patterns and levels of representation of sample data, has gained more and more attention in recent years. However, deep learning still faces many problems. The difference of various parameters when deep learning is performed for a specific kind of data will have a large impact on the results, while the adjustment of parameters lacks a systematic logical support. Therefore, this paper, through literature research methods, empirical research methods and other relevant methods, will summarize the influencing factors of Chinese text recognition procedures, including the effect of each parameter in the procedure on the results and the differences between different algorithms, and suggest possible directions for future improvements. The paper finds that the choice of algorithm has a large impact on the accuracy of deep learning, and programmers should give priority to choosing the best algorithm rather than obsessing over fine-tuning parameters. The future direction of deep learning should focus on optimizing and improving old algorithms or even creating new, more targeted algorithms that are more efficient for some problems.

Keywords: *deep learning, algorithm, reinforcement learning, text recognition, recurrent neural network*

1 INTRODUCTION

Deep learning, as a learning method derived from artificial neural networks, is efficient and accurate in modeling and classifying complex data such as speech, images, and text. The large amount of data available for training is the basis for the success of this learning method, and the program can obtain a good ability to classify and recognize data by learning the key parameters that can be used for recognition of this type of data [4]. It can be understood as a process of learning the functions and nonlinear characteristics of complex computational models with multiple levels of abstraction and multiple processing layers [9]. In image recognition, deep learning algorithms can be even more accurate than real people if the right parameters and learning patterns are set [5]. However, in the field of text recognition, Chinese text recognition is still a challenge because the language logic of Chinese is different from that of English [1]. In this paper, we designed and implemented a Chinese text recognition system using a deep learning procedure with two different learning algorithms to evaluate the impact of each parameter on the recognition quasi-accuracy. In order to study the problem, literature research methods, empirical research methods, and other methods were used in this process. This paper firstly introduces the basic concepts of two deep learning algorithms, recurrent neural networks and convolutional neural networks. Then the algorithms of deep learning will be used for Chinese text recognition and the effect of each parameter in the procedure on the results will be examined, as well as the differences between different algorithms, and it discusses the possible reasons for the influence of each parameter. Finally, the paper will present ideas for future improvements of the algorithm. It hopes to provide insightful suggestions for the research in this field.

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 724-729, 2023.
https://doi.org/10.2991/978-94-6463-040-4_110

2 BACKGROUND

2.1 Definitions

The name deep learning is meant to distinguish it from so-called "shallow" machine learning [11]. Deep learning has an input layer and an output layer, which can be transformed prior to training by artificial feature engineering on the input. One or more hidden layers exist between the input layer and the output layer. A non-linear transformation, activation function, etc. is applied to the input of a unit to obtain a new representation of the input [6]. Besides, after the computation flows from the input to the output, the error derivative can be computed backwards on the output layer and each hidden layer and the gradient propagated backwards to the input layer so that the weights can be updated to optimize some of the loss functions.

Convolutional neural networks are classified as a class of supervised deep feature learning models. Convolutional networks are made up of many layers and connections designed to learn hierarchical feature representation. Three strategies are mainly adopted: local acceptance field, shared weight and spatial or temporal subsampling [2]. However, in order to make CNN suitable for practical applications, especially in the application of high-dimensional input data such as image and speech processing, and to achieve reasonable performance compared with shallow learning methods, this deep network requires a large amount of data.

Another deep supervised feature learning (and unsupervised) algorithm is recurrent neural networks (RNN), which have feedback connections that allow them to have internal states. This means that they have a memory that retains previous input information, making them advantageous for applications with chronological and sequential data. Previous RNNs architectures to learn long-term dependencies in continuous data would cause gradient disappearance or gradient explosion, but the current improved RNNs can effectively deal with these problems [6].

2.2 Related Work

The scope of the related work in this paper is limited to text character recognition meaning to whole sentence meaning recognition of Chinese text sentences. In 2014, Yoon Kim used pre-trained word vectors to train a series of convolutional neural networks for sentence level classification tasks. By fine-tuning the learning of specific task vectors, it is equivalent to teaching the computer what to focus on, which can further improve the performance. At the same time, they modified the architecture to use both task-specific and static vectors and finally improved their technology, especially for sentiment analysis and problem classification [10].

Similar to this study, Xiang Zhang et al. proposed a character-level convolutional network (ConvNets) for text classification. They constructed several large-scale data sets to show that character-level convolutional networks can achieve state-of-the-art or competitive results and compared them with traditional models such as word packets, n-grams and their TFIDF variants, as well as deep learning models such as word-based ConvNets and recurrent neural networks [7].

In contrast, Ren et al. focus on the improvements needed to recognize Chinese text in a CNN architecture compared to English. They use a progressive training approach in order to ensure that the small-sized natural character data and the large-sized artificial character dataset are comparable in terms of training and end up finding more recognition accuracy compared to the baseline approach [8].

3 RESEARCH ENVIRONMENT

In order to produce a result that is close to the real situation, it is necessary to use a Chinese text dataset with a large amount of data and real information collected from the Internet. In this paper, the Chinese text classification dataset from THUCNews was chosen [3]. It contains filtered historical data from the Sina News RSS feeds between 2005 and 2011, containing 740,000 news documents. The benefits of using this dataset are obvious, as Sina, a well-known Chinese news website, has news messages of significant authenticity and complexity compared to using manually generated data, with many rare occurrences.

As the focus of this research is to explore the influencing factors related to Chinese text, this paper decided to build a deep learning program using a framework such as tensorflow in a python environment. This eliminates the need for this study to build the entire model from scratch and allows it to focus on the problem of program parameters. At the same time, TensorFlow does not take up compilation time, which allows research to quickly verify relevant conjectures while eliminating the need for dedicated waiting time.

4 PROCEDURAL FRAMEWORK

This section outlines the procedural structure of the study. The program is divided into three main parts: data preprocessing, configuration of convolutional and recurrent neural networks, as well as the design and debugging of various parameters, and training and validation.

During preprocessing, the program first reads the Chinese text from the dataset. In this program, in order to prevent the use of special parameters from affecting the comparison of accuracy of training models, the operation parameters of the data set are set relatively

conventionally, the number of 5000 features is reserved, and the whole data set is divided into a training set, a verification set, and a test set, the ratio of which is 7:1:2. To separate data sets, the program consolidates multiple files into three separate files.

This part opens the training set file using the open function, then splits the data into labels and contents using the strip and split functions, and splits it into two lists using the try and append functions. After that, use the counter class function and the list function to convert the book to a list, then use the pad function to make sure all the text is the same length, and finally create a file and store it. Next, the program needs to convert the representation of the glossary to become {category:id}. To do this, the program needs to convert the glossary to a {word:id} representation first, then let the category list be fixed before finally converting it to a {category:id} representation. More importantly, in the final stage, the program reconverts the data represented by the ids into text and continues to convert it into a fixed-length id sequence representation, so that the Chinese text data in the dataset is converted into training data for deep learning, which only needs to undergo a shuffle process to have some randomness before it can be used for formal training.

This study considers different learning methods of deep learning as the most influential factor to be explored for Chinese text recognition and classification. Therefore, this study designed the program with both learning methods using convolutional neural networks and recurrent neural networks at the same time. Their structure is shown in Figures 1 and 2. Text classification usually include feature selection, feature dimension reduction, classification model study three steps, and how to select the appropriate text in Chinese text classification feature is especially important, this thesis mainly based on the perspective to solve the problem of the parameter Settings, in order to prevent cannot learn Chinese text characteristics lead to owe fitting end to the premise of the comparative accuracy. In the algorithm program of CNN, the program needs to use the time function to obtain the use time, use the method based on Tensorflow to evaluate the accuracy and loss of data, load the training set and verification set, and use process_file, session.run and other methods in the process. In the training process, the add_summary method should be used to determine the writing of training results every few rounds. This program is configured as 20 rounds. It is worth noting that if the accuracy of the verification set does not improve for a long time, the training can be ended in advance with the break statement to save resource consumption. This program is set to end the training if the 1000 rounds do not improve as well as saves the best results and displays the data. When using CNN algorithms, as Chinese character text recognition is prone to overfitting problems, Just as CNN usually uses three convolution kernels to extract the features of three color channels respectively for superposition when extracting RGB image features, multiple convolution kernels can better identify and retain data with more features, while word vector dimension has an impact on the text classification model's random initialization of vectors of different words and other steps. This study considers that the influencing factors that should be focused on are the dropout retention ratio, the size and number of convolutional kernels, and the word vector dimension. In the algorithm program using RNN, the program also needs to use the time function to obtain the current time and use various methods of Tensorflow. The difference lies in the study method and the relevant configuration parameters. The program uses the GRU as the solution to solve the problem of short-term memory, LSTM as another kind of solution, although it is due to the use of more parameters and a more complex logic expression in the large data set performance is better, but this article is not only for training model for use after is compared under the condition of various training model accuracy, less so GRU helped parameter easier convergence characteristics, this article finally decided to use it. At the same time, the program selects the relatively regular and reasonable configuration parameters to contrast the choice of algorithm for accuracy and try to avoid the influence of fine-tuning that may be the result of the impact of the parameters, such as setting the dropout of reserve ratio of 0.8. In RNN programs, the meaning of the hidden layer is the input data, the characteristics of the abstract to another dimension of space, to show the more abstract characteristics, these characteristics can better linear classification, based on priority logic, to solve the problem of the appropriate feature extracting experiment will give priority attention to hidden layer parameters such as the impact on the program. In addition, for RNN algorithm, in order to prevent the occurrence of fitting problems, the program pays attention to dropout retention ratio and total number of iterations to avoid overtraining.

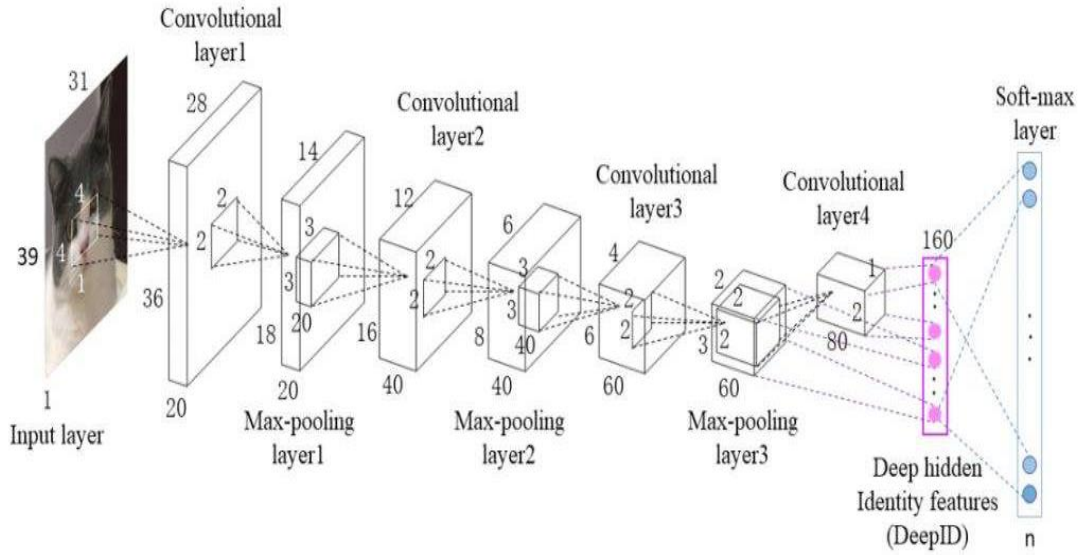

Figure 1: Convolutional neural network structure

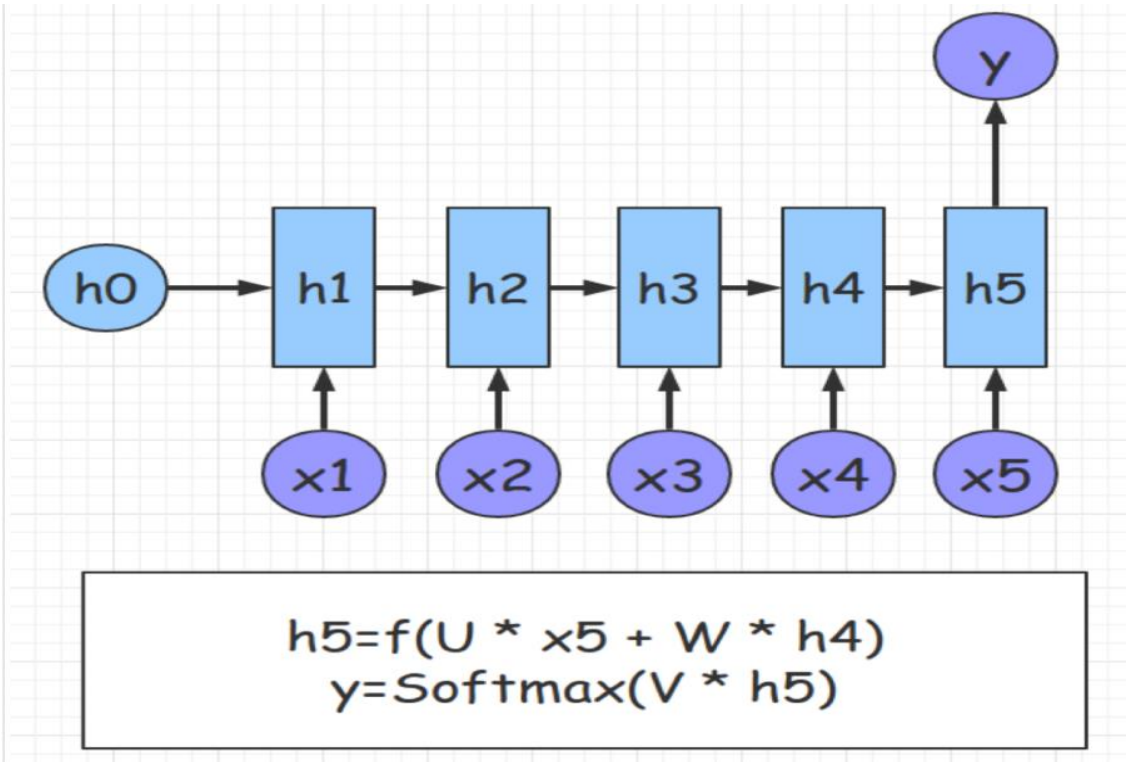

Figure 2: Recurrent neural network structure

Finally, during training and testing, this study argues that attention needs to be paid to the randomness and amount of data allocated during pre-processing to prevent model jitter during training and to train only the sequential features of the learned Chinese text, and to facilitate the robustness of the model.

5 RESULTS AND DISCUSSION

After running the program, it was found that the accuracy of the CNN algorithm program (96.41%) was higher than that of the RNN algorithm (94.66%) for similar parameters, and the precision, recall and f1-score for all categories exceeded 0.9, indicating excellent classification results, compared to the poor performance of the RNN algorithm for some classifications. In addition, the CNN algorithm stopped after only three iterations, while the RNN algorithm required eight iterations, which is much slower than the CNN algorithm.

Discarding the type of algorithm, the important influences on the accuracy of a deep learning procedure are the word vector dimension and the number and size of convolutional kernels. Compared to the accuracy of

the standard algorithm, adjusting the word vector dimension to 64, the number of convolutional kernels to 128 and the size to 5 will increase the accuracy by about four percent and reduces the time and number of rounds required for iterations. In this paper, it is considered that the word vector dimension needs to be more than 8 times logN (N is the size of the word list). Too small parameters will lead to the lack of sufficient capacity of the model to accommodate these words, but the larger the parameter is, the risk of overfitting will increase. In addition, when the same receptive field is achieved, the smaller the convolution kernel is, the smaller the required parameters and computation are. At the same time complex small convolution kernel can also have more nonlinear transformation, enhance the ability to learn features.

For the pre-processed data, it was found that randomness became the most important factor after a certain number of data sets, and it was found that low randomness leading to overfitting increased the accuracy to around 98%. But obviously, this kind of training model is not practical in real environment.

This paper argues that in the future if you need to improve on the depth of the text classifier learning program, to the improvement of the existing programs, or create a new learning algorithm can be effectively enhance learning accuracy and reduce the consumption of time, for example, this paper RNN algorithm has been used by a program using the improved scheme that can solve the problem of short-term memory. Otherwise, the accuracy of the RNN algorithm program trained will be lower than that using the CNN algorithm. The selection of data sets can be improved on the premise that the former cannot be optimized, but the upper limit of their improvement is also determined by the algorithm.

6 CONCLUSION

This paper argues that, based on the above conclusions, the most important direction for future improvement of the Chinese text recognition algorithm program is to innovate the algorithm method. The accuracy difference between different algorithms can reach several percentage points, and the consumption time and number of iteration rounds also vary greatly. The second is the selection and processing of data sets. After pretreatment of huge and real data sets, excellent and non-over-fitting results can be achieved in algorithm training. In addition to these influences, configuration parameters can have a non-negligible effect on training results, but when they are in a generally reasonable range, fine-tuning can have a negligible effect on training results.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my most sincere gratitude to my supervisor, Professor Pietro Lio', for the long hours of work that have gone into the completion of my paper, the conceptualisation, the experiments and the writing of the paper, a process that has not been easy. The completion of this paper would not have been possible without the careful guidance and attention given to me by the professor, whose advice was indispensable from the selection of the topic to the conception of the paper, to its writing and revision. I have benefited greatly from the professor's rigorous teaching attitude and profound academic knowledge.

REFERENCES

[1] Changxu C., Wuheng X., Xiang B. Bin F. & Wenyu L., (2020). Maximum Entropy Regularization and Chinese Text Recognition. DAS, 3-17.

[2] LeCun Y., Bengio Y. et. al., (1995). Convolutional networks for images speech and time series - The handbook of brain theory and neural networks. 10, 3361.

[3] Maosong S., Jingyang L., Zhipeng G., Yu Z., Yabin Z., Xiance S., Zhiyuan L., (2016). THUCTC: An Efficient Chinese Text Classifier.

[4] Reza, S. & Vitaly S., (2015). Privacy-Preserving Deep Learning ACM Conference on Computer and Communications Security. 909-910.

[5] Ren S. & Sun J., (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. https://arxivorg/abs/150201852.

[6] Seyed S., Mousavia M. S. & Enda H., (2018). Deep Reinforcement Learning: An Overview Lecture Notes in Networks and Systems. 426-440.

[7] Xiang Z., Junbo Z., Yann L., (2015). Character-level Convolutional Networks for Text Classification. https://arxivorg/abs/150901626.

[8] Xiaohang R., Kai C., Jun S., (2016). Character-level Convolutional Networks for Text Classification. https://arxivorg/abs/160401891.

[9] Yann L., Yoshua B. & Geoffrey H., (2015). Deep Learning Nature No. 7553, 436-444.

[10] Yoon K., (2014). Convolutional Neural Networks for Sentence Classification. https://arxivorg/abs/14085882.

[11] Yuxi L., (2017). Deep Reinforcement Learning: An Overview arXiv:learning arXiv:181006339.