

Assignment_Chen_2304602

by Yinzihang Chen

Submission date: 18-Apr-2024 02:39PM (UTC+0100)

Submission ID: 229840957

File name: 51826_Yinzihang_Chen_Assignment_Chen_2304602_571450_1228980566.docx (35.39K)

Word count: 3054

Character count: 17032

Ensemble model on IMDB dataset

Name: Yinzihang Chen

Student Number: 2304602

Contents

Abstract.....	3
1. Introduction	4
2. Related work	5
2.1 Models.....	5
2.2 Dataset	6
2.3 Ensemble technique	6
3. Methodology.....	7
4. Results and discussion.....	9
4.1 CNN	9
4.2 LSTM.....	9
4.3 GRU	10
4.4 Bert.....	11
4.5 Ensemble Model	11
4.6 Discussion.....	12
5. Conclusion	13
Appendix.....	14
References	15

Abstract

With the development of the times, artificial intelligence is gradually becoming popular. More and more fields have seen the emergence of artificial intelligence, while also giving rise to many different types of models.

This study aims to compare the performance of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), Bidirectional Encoder Representations from Transformers (BERT) and their ensemble model on the IMDB. We first preprocess the IMDB dataset, then apply the above four models for training and testing, and finally evaluate the models using relevant parameters.

1.Introduction

AI, or artificial intelligence, is a branch of computer science that aims to create machines or software that can perform tasks that typically require human intelligence. These tasks include but are not limited to visual perception, language recognition, decision-making, language and speech between translations, etc. Artificial intelligence can be divided into two main types:

- a)  1. **Narrow AI:** It refers to the design and training of artificial intelligence systems used to perform specific tasks without direct human intervention. Narrow AI systems perform well in their specialized fields, but they lack self-awareness or a true understanding of the tasks they perform. Most existing artificial intelligence applications, such as image recognition software, search engines, and personal assistants (such as Siri and Alexa), are examples of narrow AI.
- b) **Strong AI:** It is a theoretical form of artificial intelligence that can understand, learn, and apply knowledge, just like humans, capable of executing intelligent behaviors in a broad and abstract manner. Universal AI will be able to understand and learn any intelligent task that humans can perform, and possess emotions, consciousness, and self-awareness. At present, universal AI is still the theme of science fiction and futuristic research and has not yet been realized.

The development of artificial intelligence involves various technologies and disciplines, including machine learning (ML), deep learning, neural networks, natural language processing (NLP), and so on. Machine learning is a core component of AI that allows computers to improve their task execution capabilities by learning and recognizing patterns, rather than through explicit programming instructions. Deep learning is a special machine learning method that uses large neural networks to simulate the way the human brain processes and analyzes data. The advancement of these technologies has driven the application of AI in healthcare, finance, autonomous vehicles, customer service, and many other fields.

This paper compares the results of five different models (CNN[1], LSTM[2][3], GRU[4], BERT[6] and their ensemble model) on IMDB to measure their performance and differences in processing text data, and then analyzes and compares them to draw conclusions.

2.Related work

As mentioned earlier, this paper analyzes five models by comparing them on IMDB. Next, we will introduce the four base models and datasets.

2.1 Models

- a) **Convolutional Neural Networks (CNN):** CNN is a deep learning algorithm that performs well in the fields of image processing and visual recognition. CNN understands images by simulating human visual perception mechanisms, and its main advantage lies in its ability to automatically and effectively extract features from images.
- b) **Long Short-Term Memory (LSTM):** LSTM (Long Short-Term Memory Network) is a special type of Recurrent Neural Network (RNN) that is particularly skilled at processing and predicting important events with long time gaps in sequence data. LSTM is designed to solve the problem of gradient vanishing or exploding that traditional recurrent neural networks encounter when processing long sequences. This makes LSTM perform well in tasks such as language modeling, speech recognition, and machine translation that require understanding long-distance dependencies. LSTM is highly effective in many applications, especially in processing sequential and temporal data, such as Machine translation, speech recognition, sentiment analysis, stock market prediction, weather prediction, as well as behavior recognition and event detection.
- c) **Gated Recurrent Units (GRU):** GRU is a recurrent neural network (RNN) architecture used for processing sequential data, similar to long short-term memory networks (LSTM), but with a more concise structure. It was proposed by Cho et al. in 2014 with the aim of solving the gradient vanishing problem in standard recurrent neural networks while reducing computational complexity[4][5].
- d) **Bidirectional Encoder Representations from Transformers (BERT):** BERT is an advanced natural language processing pre training model developed by the Google AI team. The key innovation of the ERT model lies in the use of a bidirectional Transformer

architecture to handle language modeling tasks in a completely new way. Due to its powerful representation ability, BERT has achieved breakthrough performance in various natural language processing tasks, including but not limited to: Text classification, question answering systems, entity recognition, sentiment analysis, and machine translation. The launch of BERT not only greatly promoted the development of natural language processing technology, but also provided a foundation for many subsequent research and model innovations, such as enhanced or variant models such as RoBERTa and ERNIE.

2.2 Dataset

In this paper, we used the IMDB dataset. Below is an introduction to the IMDB dataset.

The IMDB dataset (Internet Movie Database Dataset) is a standard dataset commonly used for sentiment analysis tasks in the field of natural language processing (NLP). This dataset contains 50000 movie review texts from the Internet Movie Database (IMDB). These reviews were carefully selected and marked as positive or negative, making them suitable for training and testing the emotional analysis model. These 50000 evaluations include 25000 positive evaluations and 25000 negative evaluations, ensuring a balanced classification of the dataset, which helps to avoid bias during model training. At the same time, these evaluations are divided into two parts: one is used for training (usually 25000), and the other is used for testing (also 25000). This distribution ensures that the model can be trained on sufficient data and evaluated fairly on an independent test set. This dataset is widely used in the development and evaluation of sentiment analysis models, especially classification models, which can learn to distinguish whether the emotional tendencies of comments are positive or negative. Due to its rich natural language texts and clear emotional polarity labels, the IMDB dataset is an ideal choice for evaluating the effectiveness of different NLP techniques, such as text preprocessing, feature extraction, model training, and tuning.

2.3 Ensemble technique

Below is a brief introduction to three Ensemble techniques:

- a) Bagging: Bagging, also known as Bootstrap Aggregating, is an ensemble learning

technique used to improve the stability and accuracy of machine learning algorithms. Bagging helps prevent overfitting by reducing the variance of the model, especially suitable for models with high variance and low bias (such as decision trees). It is the foundation of the random forest model, which works by constructing multiple decision trees and aggregating their predictions[7][8].

- b) Stacking: Stacking, also known as stacked generalization, is an ensemble learning technique used to improve the prediction accuracy of various different models. In this method, multiple different models (such as decision trees, neural networks, etc.) are first used to predict the training data, and then a new model (usually called a meta model) is used to integrate the prediction results of these first layer models. This method can effectively reduce model bias and improve prediction accuracy[7][8].
- c) Boosting: Boosting is an ensemble learning technique aimed at improving the accuracy of weak learning models by gradually improving multiple models to form a powerful overall model. The core idea of Boosting is to train continuous learners, with each learner attempting to correct the errors of the previous learner[7][8].

3.Methodology

The purpose of this paper is to use four different artificial intelligence models to test their performance on the IMDB dataset, in order to measure the performance of different models when facing text datasets.

The first thing we need to do is to process the data, and then build four models: CNN, LSTM, GRU, and BERT. We plan to test the performance of the four models separately on IMDB first, and then integrate them to test the performance after integrating the four models. For ensemble learning technology, we chose to use bagging on the CNN, LSTM, and GRU models, and stacking on Bert. On the ensemble model, we are preparing to use bagging too. The final performance evaluation indicators used are accuracy, precision, recall, and F1 score. Accuracy is a fundamental indicator for measuring the accuracy of classification model predictions. It is the ratio of the correctly predicted quantity to the total predicted quantity. Accuracy provides

the model with the ability to make accurate predictions on all classification tasks. Precision measures the proportion of instances correctly predicted as positive by the model to actual positive examples. Recall, also known as true rate, measures the proportion of actual positive classes correctly predicted by the model. F1 score is the harmonic average of precision and recall, used to strike a balance between precision and recall.

4.Results and discussion

4.1 CNN

We first need to import the necessary Python libraries and then process the IMDB dataset, limiting the word index sequence of comments to the most common 10000 words. Meanwhile, use 'pad_sequence' to crop or fill all comments to a uniform length of 500 for subsequent input into the CNN model. Then build a CNN model, including Embedding layer, Conv1D layer, Maxpooling1D layer, GlobalMaxPooling1D layer, and Dense layer. In the model using ensemble technique, we used bagging technique to construct and train 10 CNN models. Finally, the performance of the model is evaluated using precision, accuracy, recall, and F1 score. The following will demonstrate the differences in direct performance of the model.

CNN	Accuracy	Precision	Recall	F1 Score
Bagging	0.856	0.861	0.850	0.856
Without Bagging	0.841	0.846	0.834	0.840

Table 1 Comparison between CNN Model with and without bagging

From the table, we can see that after using Bagging, all four indicators have improved by about 1.5%. It can be clearly seen that Bagging technology has helped to improve the performance of the model. However, compared to other models, CNN's various indicators are relatively low, which we will discuss later.

4.2 LSTM

In the LSTM model, the way data is processed is the same as above, and it will not be further elaborated here. The LSTM model consists of an embedding layer, an LSTM layer (including dropout and recurrent_dropout parameters), and an output layer (a fully connected layer using sigmoid activation function). In the bagging model, we created and trained 5 LSTM models; We make predictions for each independently trained model and average the results, which can reduce variance, improve stability, and accuracy.

LSTM	Accuracy	Precision	Recall	F1 Score
Bagging	0.882	0.879	0.886	0.883
Without Bagging	0.865	0.868	0.861	0.860

Table 2 Comparison between LSTM Model with and without bagging

From the table, we can see that bagging has a relatively significant improvement on LSTM, generally around 2%. Compared to GRU and CNN, LSTM has a much longer training time, probably due to its core being a gating mechanism, including forget gates, input gates, and output gates. These gating mechanisms involve multiple matrix operations and nonlinear activation functions, which leads to an increase in the number of LSTM model parameters and computational complexity, requiring more time to process.

4.3 GRU

In the GRU model, the way data is processed is the same as above, and it will not be further elaborated here. The GRU model includes an embedding layer, a GRU layer, and an output layer. The GRU layer is used to process sequential data, and the 'dropout' and 'recurrent_dropout' parameters help reduce overfitting of the model during training. The output layer uses a fully connected layer with sigmoid activation function. For the GRU model, we also used bagging to compare it with models without bagging.

GRU	Accuracy	Precision	Recall	F1 Score
bagging	0.881	0.867	0.900	0.883
Without bagging	0.871	0.872	0.869	0.870

Table 3 Comparison between GRU Model with and without bagging

After continuous debugging, we found that the results of the model performing best are shown in the table above. It can also be seen that when we use bagging technology, the model performance has improved to a certain extent. In this GRU model, when we increase the epoch to 10, we find that the result is not as good as the output when epoch=3, I think this should be due to overfitting caused by too many epochs, resulting in a decrease in

performance.

4.4 Bert

As is well known, Bert is currently an excellent deep learning model for handling NLP tasks. Before this, I had high expectations for it, and the final result was as expected. In the data processing stage, compared to the first three models, Bert added a data conversion stage to convert integer sequences in the dataset into text. Then load the pre-trained 'TFBertForSequenceClassification' model from the 'transformers' library, and set the optimizer, loss function, and evaluation metrics. For the Bert model, we use stacking for ensemble, and

Specific reasons are described in Chapter 3, which will not be elaborated here.

Bert	Accuracy	Precision	Recall	F1 Score
Stacking	0.912	0.908	0.906	0.915
Without Stacking	0.863	0.855	0.874	0.864

Table 4 Comparison between Bert Model with and without stacking

From the table, it can be seen that even without stacking, Bert's performance is still very outstanding. After ensemble, the performance improvement is even more significant, which is unmatched by the other three networks.

4.5 Ensemble Model

Finally, there is the ensemble model. We combine CNN, LSTM, GRU, and Bert to form a new model. Then use bagging on the new model to enhance the performance of each model, especially the weaker parts of the model. Let's take a look at how this model performs.

Ensemble	Accuracy	Precision	Recall	F1 Score
Boosting	0.938	0.937	0.939	0.938

Table 5 Ensemble model with bagging

From the table, it can be seen that all indicators of the ensemble model are very high and the performance is excellent. Compared to the four individual networks, there is a very significant improvement. After continuous debugging, its performance advantages are further demonstrated.

4.6 Discussion

From the above five models, we can see that CNN performs the worst on the IMDB dataset, while Bert undoubtedly performs the best. The main reason is that CNN performs well in image recognition; CNN is more commonly used for local recognition, such as image edges and shapes, but in sentiment analysis, relying solely on capturing local features is clearly not enough to fully understand the sentiment of the text. Compared with recursive neural networks, RNN performs poorly because it is designed to process sequential data and can maintain memory of previous information.

LSTM and GRU perform similarly on the IMDB dataset because they are both RNN variants used for processing sequence data; They are very similar in structure, but GRU is more concise than LSTM, both of which control information retention and forgetting through the use of gating mechanisms. Both are highly skilled at capturing long-term dependencies, which is crucial for sentiment analysis tasks.

The most important ensemble model integrates four models together, and these networks complement each other, greatly improving the performance of the model. CNN's ability to extract local features can also be utilized, as it can capture key phrases in sentences, which may be crucial for sentiment analysis; LSTM and GRU have similar functions, and they can effectively process and remember information in long sequences, which is crucial for understanding the context of the entire comment. Bert's bidirectional Transformer architecture enables it to understand the relationship between words in text. This deep understanding ability is highly effective for sentiment analysis, and Bert's monomer performance is also excellent, allowing for quick adaptation to NLP tasks. Through integration, their capabilities complement each other and the model performance has been greatly improved.

5. Conclusion

This paper first introduces relevant knowledge in the field of artificial intelligence and introduces the objectives of this paper. This paper analyzes the performance of four base models and the ensemble model composed of four base models on the IMDB dataset, analyzes the advantages and disadvantages of different models in processing text data, and also points out the reasons why some models perform poorly on the IMDB dataset; Finally, through the analysis of the ensemble model, I gained a deeper understanding of each model and ensemble technology; I understand why ensemble technology can make models better.

Ensemble learning is a powerful machine learning paradigm that improves predictive performance by combining multiple models, especially when addressing the potential overfitting or insufficient generalization ability of individual models. The ensemble method can usually provide better robustness and accuracy than any single model, and is an effective means to solve complex problems.

Appendix

<https://youtu.be/HuBFWzA8ALU>

References

- [1]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
- [2] S. F. Pane, J. Ramdan, A. G. Putrada, M. N. Fauzan, R. M. Awangga and N. Alamsyah, "A Hybrid CNN-LSTM Model With Word-Emoji Embedding For Improving The Twitter Sentiment Analysis on Indonesia's PPKM Policy," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022, pp. 51-56, doi: 10.1109/ICITISEE57756.2022.10057720.
- [3] N. Kapali, T. Tuhin, A. Pramanik, M. S. Rahman and S. R. H. Noori, "Sentiment Analysis of Facebook and YouTube Bengali Comments Using LSTM and Bi-LSTM," *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2022, pp. 1-6, doi: 10.1109/ICCCNT54827.2022.9984395.
- [4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
- [5] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
- [6] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [7] Mohammed, Ammar and Rania Kora. "A comprehensive review on ensemble deep learning: Opportunities and challenges." *J. King Saud Univ. Comput. Inf. Sci.* 35 (2023): 757-774.
- [8] Kora R, Mohammed A. An enhanced approach for sentiment analysis based on meta-ensemble deep learning[J]. *Social Network Analysis and Mining*, 2023, 13(1): 38.

FINAL GRADE

GENERAL COMMENTS

57 / 100

PAGE 1

PAGE 2

PAGE 3

PAGE 4



Comment 1

Is this information relevant to the context of this study ?

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11



Comment 2

chapter three ?

PAGE 12

GRADING FORM: ACCA 7026 GRADING

YINZIHANG CHEN

56.50

ENGAGEMENT WITH LIT. SKILLS

(including current research, referencing, academic conventions and academic honesty)



Engagement with a limited range of literature. Contents are largely descriptive in nature, instead of providing a critical evaluation.

7.5

KNOWLEDGE&UNDERSTANDING SKILLS

(A systematic, conceptual understanding of knowledge, and a critical awareness of current problems and/or new insights, much of which is at, or informed by, the forefront of the discipline; a comprehensive understanding of techniques applicable to their own research)



Contents are presented in a descriptive nature. It lacks synthesis of information.

5

COGNITIVE&INTELLECTUAL SKILLS

(Critical evaluation of current research and methodologies and develop critiques of them and, where appropriate, to propose new hypotheses; make sound judgements in the absence of complete data.)



A limited attempt was made to describe the methodology.

11

PRACTICAL APPLICATION SKILLS

(Originality / creativity in the application of knowledge, tools and techniques and in tackling and solving problems in complex and unpredictable professional situations; practical use of established techniques of research and enquiry to create and interpret knowledge in the discipline.)



An appropriate application of methods such as LSTM, GRU , CNN and BERT provided. However the results of bagging and stacking ensemble is provided but it lacks a proper explanation. Discussion of the results provided, with comparing and contrasting with the literature.

28

TRANSF. SKILL & PROF.PRACTICE

(Exercise of self-direction, autonomy and personal responsibility; plan and implement tasks at a professional level; independent learning; use appropriate media to communicate effectively and professionally to a variety of audiences; fluency of expression; systematic approach; clarity and effectiveness in presentation and organisation.)



Need improvement in presentation of the many of the sections such as abstract, introduction,

5

methodology. Demonstrated video does not provide the execution of the code.