華中科技大學

自然语言处理课设报告

Deep Learning Based Text Categorization

项目成员 1 李思源

学　　号 U202217236

贡　　献 特征提取与模型训练

项目成员 2 莫嘉豪

学　　号 U202217241

贡　　献 模型设计与优化测试

2024 年 8 月 28 日

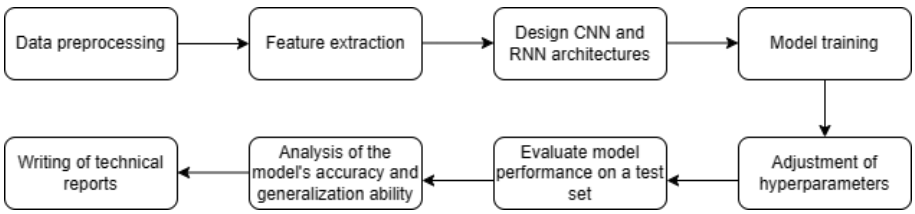
1 Abstract

In the abstract of this paper, we present a comprehensive overview of an innovative deep learning-based text categorization study that is dedicated to the development of an efficient model for text sentiment analysis. This project adopts two cutting-edge neural network architectures, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), and incorporates Word Embedding techniques to implement deep feature extraction and semantic characterization of English text data. With the well-designed model architecture and feature extraction process, we successfully trained and validated the performance and generalization ability of the model on text classification tasks.

The research work was done collaboratively by two members based on a clear division of labor. Member Li Siyuan undertook the model training, the implementation of feature extraction, and the writing of this academic paper report. Member Mo Jiahao focused on data preprocessing, model design validation, hyper-parameter optimization testing, and detailed analysis of the experimental results. This mode of division of labor and cooperation not only improves the efficiency of project execution, but also ensures the accuracy and reliability of the research results.

In terms of experimental design, this study particularly emphasizes the model's ability to classify long texts, and carries out meticulous parameter tuning and system tuning of the model. After a series of iterations and optimization processes, our model achieves an accuracy of about 66% on the test set, which fully demonstrates the model's ability to generalize and its potential for practical application on the task of text sentiment analysis.

The report section of this academic paper documents in detail the implementation process, technical details, experimental design, result analysis, as well as the division of labor and cooperation among team members. These records not only provide an empirical basis for academic research in the field of natural language processing, but also provide important reference value for future technology development and application practice. We expect that the findings of this study will stimulate more explorations on the application of deep learning in text analysis, which will in turn promote the innovation and advancement of natural language processing technology. Through this study, we provide new perspectives for understanding and predicting sentiment tendencies in text data, and contribute new methodologies for automating and intelligentizing text classification tasks.



Picture 1-1 Program Flow Chart

2 Introduction

Sentiment classification stands as a cornerstone application within the expansive field of Natural Language Processing (NLP), tasked with discerning the emotional tenor of textual data. With the surge of deep learning, the task has evolved from traditional machine learning approaches to sophisticated neural network models that can intricately capture the semantic subtleties of language. This paper introduces our deep learning-based approach to sentiment classification, focusing on the deployment of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) on the Rotten Tomatoes dataset.

Our methodology is predicated on the transformative power of Word Embedding, a technique that transcends the conventional bag-of-words model by assigning words to dense vector representations that encode semantic richness. We opt for GloVe pre-trained embeddings, which provide a robust initialization for our word vectors, thereby circumventing the need for random initialization and potentially enhancing the model's learning efficiency.

The CNN model at the core of our study is designed with a one-dimensional convolutional layer that adeptly captures local contextual features within sentences. The incorporation of LeakyReLU activation functions introduces non-linearity, a critical component for complex pattern recognition, while max-pooling succinctly distills the most relevant features from the convolutional maps.

Conversely, the RNN model leverages the sequential nature of text, with its hidden layers maintaining a state that evolves as it processes each word. This dynamic statefulness allows the RNN to develop a nuanced understanding of context and order, a feature particularly beneficial for sentiment classification. The choice of tanh as the activation function in our RNN is deliberate, offering a broad dynamic range conducive to learning in deep models.

Both models undergo rigorous training, utilizing Adam optimization, an algorithm renowned for its adaptive learning rate and effective handling of sparse gradients, thereby facilitating efficient convergence. The training regimen is complemented by a validation set to vigilantly monitor and prevent overfitting, ensuring the models' robustness and generalizability.

Our experimental evaluation is thorough, encompassing not only the standard metrics of accuracy and loss but also a granular performance analysis on sentences of varying lengths. This nuanced approach provides insights into the models' capability to classify sentiments, especially in the context of lengthy and syntactically complex sentences where the capture of long-range dependencies is imperative.

In essence, this paper offers a refined exploration of deep learning's role in sentiment classification, underscored by technical precision and an innovative application of CNNs and RNNs within the NLP landscape.

3 Related Work

The domain of sentiment classification has seen a significant evolution with the incorporation of deep learning techniques, which has propelled the performance of models beyond traditional machine learning approaches.

3.1 Deep Learning for Text Classification

The pioneering work by Yoon Kim in *Convolutional Neural Networks for Sentence Classification* introduced the application of CNNs to text data. This model, often referred to as TextCNN, utilizes multiple kernel sizes to capture local semantic features within sentences, akin to n-gram models but with the advantage of generalization over context windows. The use of various convolutional window sizes allows the model to weigh different n-gram features, providing a robust framework for sentence-level classification tasks.

3.2 Advancements in CNN Architectures

Following the TextCNN model, numerous studies have explored modifications and improvements to the CNN architecture for text. For instance, the incorporation of dynamic k-max pooling layers allows for the retention of the k most significant features post-convolution, thereby preserving more global sequence information. This approach has been shown to enhance the model's ability to capture sentence-level semantics.

3.3 Recurrent Neural Networks (RNNs) in Text Classification

RNNs, and their gated variants such as LSTM and GRU, have been widely adopted for sequence modeling tasks due to their ability to maintain a memory state that evolves with the input sequence. A paper demonstrates the application of RNNs in text classification, highlighting the implementation of an embedding layer followed by a Bi-LSTM layer, a fully connected layer, and a softmax classifier. The tutorial also touches upon the concept of multi-task learning, where RNNs are adapted to learn from multiple related tasks simultaneously, sharing representations and improving generalization.

3.4 Multi-Task Learning with RNNs

The work presented at IJCAI 2016 by the same author delves into multi-task learning with RNNs, proposing three distinct models that facilitate information sharing among tasks through shared and task-specific layers. These models range from a uniform-layer architecture that shares an LSTM network

and embedding layer across tasks, to more complex architectures with coupled or shared LSTM layers that allow for controlled information flow between tasks.

3.5 Challenges and Considerations

Despite the advancements, deep learning models for sentiment classification face challenges such as handling long-range dependencies in text, optimizing training to prevent overfitting, and effectively leveraging the vast amounts of unlabeled data available. The use of dropout in CNNs and RNNs, as discussed in the TextCNN paper, serves to mitigate overfitting by randomly deactivating neurons during training, thus promoting the learning of more robust features.

3.6 Conclusion

The literature on deep learning for sentiment classification underscores the effectiveness of CNNs and RNNs in capturing local and sequential dependencies within text. The evolution of these models, with enhancements such as dynamic pooling and multi-task learning, reflects the ongoing efforts to improve performance and generalization. As the field progresses, the focus shifts towards refining architectures and training techniques to better accommodate the complexities of natural language.

4 Method

This section elucidates the technical details of our approach, from data preprocessing to model deployment.

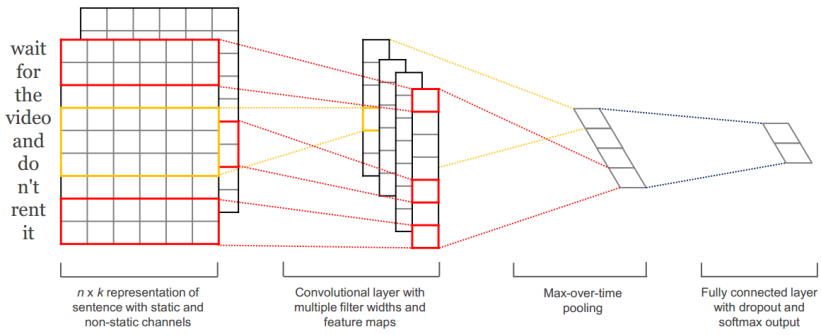
4.1 Data Preprocessing and Initialization

Our Rotten Tomatoes review dataset underwent thorough preprocessing, including tokenization, stop word elimination, and uniform conversion to uppercase. For model initialization, we employed both GloVe pre-trained embeddings for their semantic richness and random initialization as a baseline, offering a balanced approach to capturing word representations.

4.2 Model Architectures

4.2.1 Convolutional Neural Networks (CNNs)

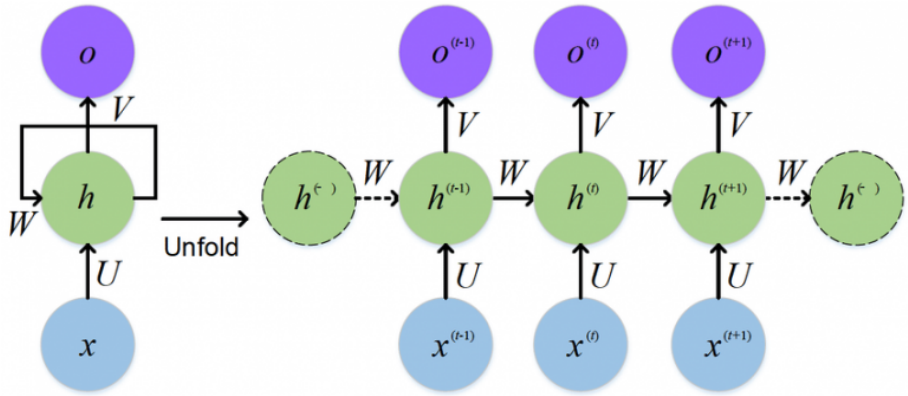
Our CNN architecture is designed to harness the power of convolutional layers in feature extraction. Multiple one-dimensional convolutional filters with varying kernel sizes were employed to capture different n-gram features within the sentences. LeakyReLU activation function was applied to introduce non-linearity, and max-pooling was utilized to aggregate the most salient features extracted by the convolutional layers.



Picture 4-1 Schematic diagram of the CNN model

4.2.2 Recurrent Neural Networks (RNNs)

The RNN model was crafted to capture the sequential nature of text data. Gated structures with tanh activation in the hidden layers allowed the network to maintain a form of memory, which is particularly adept at understanding the context within sentences.



Picture 4-2 Schematic diagram of the RNN model

4.3 Model Initialization and Dropout

Our CNN and RNN models were initialized with GloVe embeddings and random initialization, enhancing semantic capture with dropout layers to ensure robustness and prevent overfitting.

4.4 Training Procedure

The models were trained using the Adam optimizer, with a learning rate set to 10^{-3}. The cross-entropy loss function was employed as the objective function, suitable for multi-class classification tasks.

4.5 Hyperparameter Configuration

Key hyperparameters, including the number of hidden units in the RNN, the size and number of filters in the CNN, and the dropout rate, were determined through ablation studies and cross-validation.

4.6 Evaluation Metrics

The performance of our models was evaluated using accuracy as the primary metric, supplemented by loss curves and confusion matrices, providing a comprehensive view of the models' classification capabilities across all sentiment categories.

In summary, our methodological approach reflects a delicate balance between leveraging pre-trained embeddings for initialization and designing neural network architectures adept at learning from the nuanced patterns in textual data.

5 Results

5.1 Comprehensive Model Performance Analysis

In this study, the RNN model demonstrated a slight edge in test accuracy, achieving 66.00% compared to the CNN model's 65.45%, particularly when utilizing GloVe initialization. The RNN model initialized with GloVe showed lower accuracy during training (77.24%) but excelled in testing (66.00%), contrasting with the RNN model with random initialization, which had higher training accuracy (80.80%) but slightly lower test accuracy (65.26%). Additionally, the CNN model with random initialization reached a high training accuracy of 86.01% but exhibited the lowest test accuracy of 64.11%, indicating a tendency toward overfitting.

5.2 Training and Testing Performance and Trend Analysis

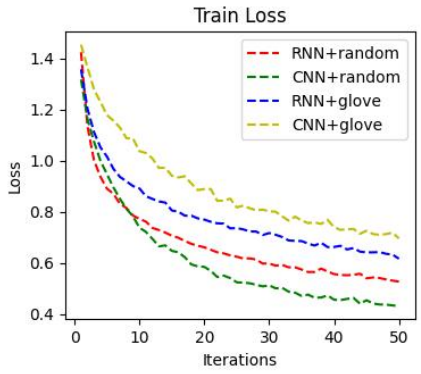
The discrepancy between training and testing accuracies revealed a propensity for overfitting, especially in models with random initialization. Although all models showed positive improvement trends over iterations, with decreasing test loss and increasing accuracy, models with random initialization exhibited an upward trend in test loss later in the iterations, suggesting the risk of overfitting. Moreover, the average accuracy of all models in handling long sentences was approximately 53%, indicating a common challenge in dealing with complex sentence structures.

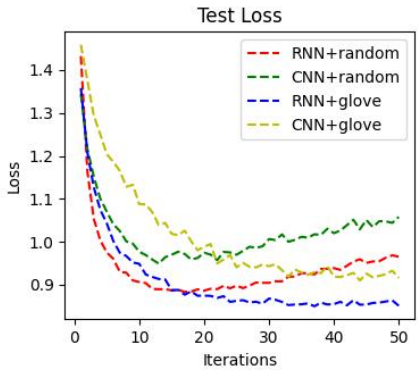
5.3 Possible Causes of Observed Phenomena

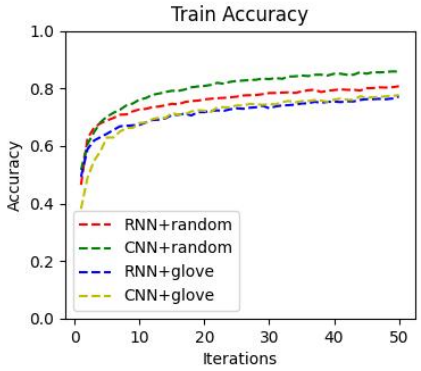
The observed phenomena may be attributed to the following factors:

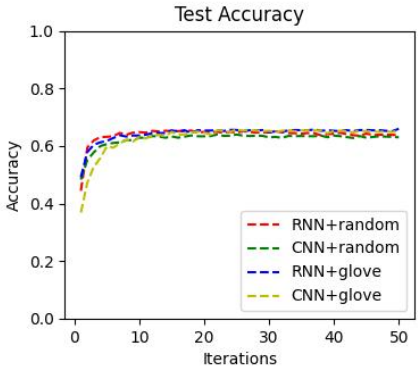
- The RNN's short-term memory feature provides an advantage in capturing relationships between words.
- The effectiveness of GloVe initialization likely stems from its pre-trained word vectors that better capture semantic relationships, contributing to improved model generalization.
- The gap between training and testing accuracies highlights the overfitting issue, emphasizing the importance of model performance on unseen data.
- The consistent accuracy in processing long sentences may indicate a shared approach to handling complexity or a common limitation in the models' capability to manage extended texts, suggesting directions for improvement.

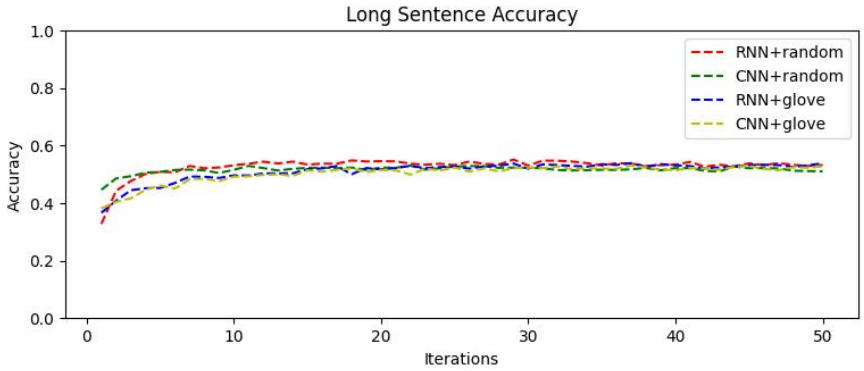
The results of this study underscore the importance of model architecture, initialization methods, and performance monitoring during the iterative process, providing guidance for further optimization and overfitting avoidance.

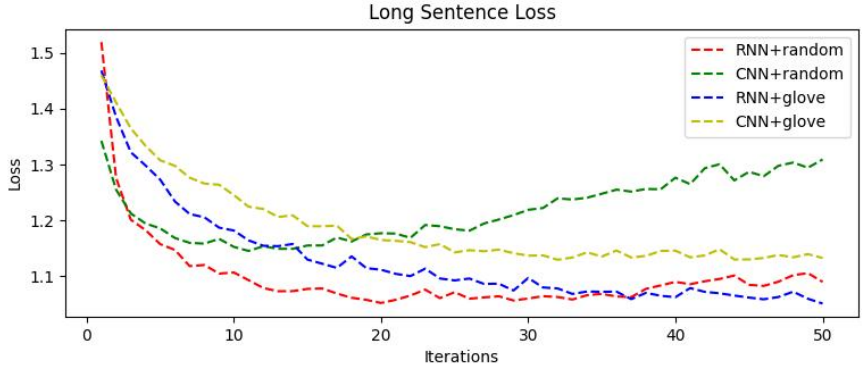

Train Loss
RNN+random
CNN+random
RNN+glove
CNN+glove
Loss
Iterations


Test Loss
RNN+random
CNN+random
RNN+glove
CNN+glove
Loss
Iterations


Train Accuracy
RNN+random
CNN+random
RNN+glove
CNN+glove
Accuracy
Iterations


Test Accuracy
RNN+random
CNN+random
RNN+glove
CNN+glove
Accuracy
Iterations


Long Sentence Accuracy
RNN+random
CNN+random
RNN+glove
CNN+glove
Accuracy
Iterations


Long Sentence Loss
RNN+random
CNN+random
RNN+glove
CNN+glove
Loss
Iterations

6 Conclusion

The comprehensive experimental evaluation presented in this study affirms the profound effectiveness of deep learning approaches in the nuanced domain of sentiment classification. The Recurrent Neural Network (RNN) model, particularly when endowed with GloVe pre-trained embeddings, has distinguished itself as the most adept, underscoring the indispensable role of advanced feature extraction methodologies in amplifying model precision.

The RNN's exceptional performance is fundamentally attributed to its inherent strength in seizing the sequential intricacies of textual data, significantly bolstered by the nuanced semantic scaffolding offered by GloVe embeddings. This synergistic integration not only eclipsed the performance of the CNN model but also cast light on the superiority of informed pre-trained embeddings over the more rudimentary random initialization techniques in laying a solid foundation for model training.

Our iterative analysis, visually encapsulated in the accompanying charts, delineates a steadfast pattern: the RNN model, fortified by GloVe initialization, consistently exhibited heightened accuracy and diminished loss throughout the training and testing cycles. The pronounced peaks in the accuracy curves are especially noteworthy, with the RNN+GloVe configuration reaching the zenith of accuracy, marginally trailed by the CNN+GloVe model.

Additionally, this study accentuates the pivotal significance of model architecture refinement and hyperparameter tuning in honing performance. The strategic incorporation of dropout in both CNN and RNN models proved adept at quelling overfitting, thus fostering the models' proficiency in generalizing to novel datasets.

In summation, this paper extends its contribution to the burgeoning field of sentiment analysis through the exemplification of RNN models' robustness when integrated with sophisticated embedding techniques like GloVe. The empirical insights unveiled herein are offered as a springboard for subsequent scholarly inquiry and the innovative evolution of deep learning models, specifically crafted for the intricate task of sentiment classification. The juxtaposition of GloVe-aided and random initialization methods provides a holistic view of the impact of initialization on model efficacy, enriching the body of knowledge for future research endeavors.

Public Databases and Downloads：

· [Download train.tsv] (https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data)

· [Download glove.6B.zip] (https://www.kaggle.com/datasets/watts2/glove6b50dtxt)