Statistical Modeling and Application Exploration of Unstructured Data Based on Deep Learning

Fei Yin

Zhuhai College of Science and Technology, Zhuhai, 519000, Guangdong, China

986221728@qq.com

Keywords: Deep Learning; Unstructured Data; Statistical Modeling; Medical Imaging Diagnosis; Social Media Public Opinion Analysis

Abstract: This article focuses on the exploration of deep learning (DL) in the field of statistical modeling and application of unstructured data. In view of the complexity and importance of unstructured data processing, the research aims to build an efficient model with the help of DL technology to mine the value of such data. By preprocessing unstructured data such as images and texts, DL model frameworks such as convolutional neural network (CNN) and Long Short-Term Memory (LSTM) are designed and used respectively, and the model is trained by combining random gradient descent algorithm. In the application of medical image diagnosis, the CNN model based on lung X-ray images has a diagnostic accuracy of 85% for pneumonia and 88% for tuberculosis. In the analysis of social media public opinion, the LSTM-based model has an accuracy rate of 80% for positive emotion recognition and 83% for negative emotion recognition. The research results show that DL has achieved remarkable results in statistical modeling of unstructured data, and can effectively handle complex unstructured data. However, there are also some problems such as high data labeling cost and poor model interpretability, and related technologies need to be further optimized and improved in the future.

1. Introduction

In today's digital age, data is growing explosively, in which unstructured data accounts for a considerable proportion. Unstructured data, such as text, images, audio and video, have brought great challenges to traditional data processing and analysis methods because of their diverse formats and lack of clear organizational structure [1-2]. How to effectively model and analyze these unstructured data and mine the valuable information contained in them has become the focus of academic and industrial circles [3].

As an important branch of artificial intelligence, DL has shown great potential in dealing with unstructured data by virtue of its powerful ability of automatic feature extraction and complex pattern recognition [4]. DL model can automatically learn the intrinsic features and patterns of data from massive data, without the need to manually design complex feature extraction algorithms [5]. This provides a new idea and method for the processing of unstructured data.

At present, DL has made remarkable achievements in many fields of unstructured data processing. In the field of image recognition, CNN can classify and detect various images with high precision [6]. In the field of natural language processing, recurrent neural network (RNN) and its variants LSTM, Gated Recurrent Unit (GRU), etc., perform well in text classification, machine translation and other tasks. Although DL has made some progress in unstructured data processing, it still faces many problems to be solved urgently [7]. DL model usually needs a lot of data for training, and the cost of data labeling is high. The interpretability of the model is poor, and it is difficult to understand the logic behind the model decision. There are still technical problems in the fusion of different types of unstructured data. Based on this, this article explores the statistical modeling and application of unstructured data based on DL. The purpose of this study is to build a more effective statistical model of unstructured data through in-depth study of DL technology, break through the bottleneck of existing technology and improve the efficiency and accuracy of

  DOI: 10.25236/icceme.2025.035

unstructured data processing.

2. DL and the present situation of unstructured data processing

DL has developed rapidly in the field of unstructured data processing since its birth. With the help of neural network architecture, it can automatically extract features from unstructured data, thus achieving efficient processing [8]. In the field of image, CNN has become the mainstream method, which is widely used in image classification, object detection and semantic segmentation, such as in medical image analysis, which can help doctors identify lesions. In the field of natural language processing, RNN and its derivative models perform well, which can process text sequence information and realize text generation, sentiment analysis and other functions, such as public opinion monitoring of social media [9]. In audio processing, DL technology is also used in speech recognition and music classification, which improves the human-computer interaction experience.

However, DL still faces challenges in dealing with unstructured data. In data annotation, a large number of unstructured data need to be manually annotated, which is costly and inefficient [10]. As for the interpretability of the model, the complex DL model is like a "black box", so it is difficult to explain its decision-making process, which limits its application in fields that require high security and reliability. The problem of data fusion is also prominent. Different types of unstructured data have different characteristics, and effective fusion is difficult, which affects the comprehensive processing ability of the model for multi-source data.

3. Construction of statistical modeling method for unstructured data based on DL

The construction of statistical modeling method of unstructured data based on DL needs to start from several key links. According to the characteristics of unstructured data, data preprocessing is carried out. There are various forms of unstructured data, such as text noise, different image resolutions, and audio noise. For text data, researchers can perform cleaning operations, remove noise such as HTML tags and special characters, and perform word segmentation to transform the text into units suitable for model processing. Image data should be normalized and resized to a uniform size to enhance image contrast and improve data quality. Audio data needs to be denoised to filter out environmental noise, and at the same time, it needs to be converted into an appropriate format and sampling rate. There are various methods for data denoising. The median denoising formula adopted by the algorithm is:

$$d_i^{'} = \text{median}\left(\left\{d_{i-k}, d_{i-k+1}, \ldots, d_{i+k}\right\}\right) \quad (1)$$

Where: $d_i^{'}$ represents the denoised data point; d_i represents the original data point; k stands for neighborhood size; $\text{median}(\ )$ is a function of calculating the median.

After preprocessing, we should design an appropriate DL model architecture. CNN is a common choice for image-based unstructured data. CNN is constructed by convolution layer, pooling layer and fully connected layer. The convolution layer uses convolution kernel to extract local features of the image, while the pooling layer reduces the dimensions of the features, reducing the amount of calculation. In this article, according to the specific characteristics of the image, such as resolution, color channel, etc., the convolution kernel size, step size and pooling method are reasonably adjusted to better capture the image characteristics. For text data, RNN and its variants, such as LSTM and GRU, perform well. RNN can process sequence data, but it has the problem of gradient disappearance or explosion. By introducing gating mechanism, LSTM and GRU can effectively solve this problem and better capture the long-term dependency in the text. According to the text length, semantic complexity and other factors, we can choose the appropriate RNN variants and adjust the parameters such as the dimension and number of hidden layers. For audio data, a mixed model based on CNN and RNN can be adopted, in which CNN captures the frequency spectrum characteristics of audio and RNN processes the time series information of audio:

$$F(x) = LSTM(RNN(x)) \quad (2)$$

Where $F(x)$ represents the output of the whole model, that is, the evaluation result of the input text. x represents the input text data. $LSTM(RNN(x))$ means that the output of RNN is used as the input of LSTM, which further processes these features and outputs the final evaluation results.

After the model architecture is determined, it is necessary to choose an appropriate algorithm to train the model. Random gradient descent (SGD) and its variants are commonly used optimization algorithms. The formula of SGD is:

$$\theta_{new} = \theta_{old} - \alpha \cdot \nabla_{\theta} J(\theta) \quad (3)$$

Where θ is the model parameter, α is the learning rate, $J(\theta)$ is the loss function, and $\nabla_{\theta} J(\theta)$ is the gradient of the loss function with respect to the parameter. In the training process, the preprocessed unstructured data is divided into training set, verification set and test set. The training set is used to update the model parameters, the verification set is used to adjust the model parameters and prevent over-fitting, and the test set is used to evaluate the final performance of the model. Appropriate learning rate, batch size, and other hyperparameters should be set as needed to balance training speed and model accuracy. In order to improve the generalization ability of the model, regularization technology is used to constrain the model parameters to avoid the model being too complicated. In the training process, the difference between the predicted value and the real value of the model is measured by loss function, the mean square error loss function is used for regression task, and the cross entropy loss function is used for classification task. Researchers can constantly adjust the model parameters to minimize the loss function value, thus optimizing the model performance. After several rounds of training, when the performance of the model is no longer improved on the verification set, it is considered that the model has reached a good state and can be applied to the statistical modeling task of unstructured data to mine the potential laws and characteristics in the data.

4. Application case analysis of statistical modeling of unstructured data

In practical application, the statistical modeling of unstructured data based on DL shows remarkable value, which is analyzed by two typical application examples.

(1) Medical image diagnosis field: In medical image analysis, X-ray image disease diagnosis is an important application scenario. We selected 2000 lung X-ray images from several hospitals as the data set. Among them, 1,500 are used for training statistical models based on CNN, 300 are used for verification, and the remaining 200 are used for testing. When the model is built, the parameters such as convolution kernel size and number of layers are adjusted through many experiments to adapt to the characteristics of lung X-ray images.

After the training is completed, the model is applied to the test set, and the result is shown in Figure 1. It can be intuitively seen from Figure 1 that there are obvious differences between normal lung images and diseased lung images in the features extracted from the model. In the actual diagnosis, the diagnostic accuracy of the model for pneumonia reached 85%, and the diagnostic accuracy of tuberculosis was 88%. Through further analysis of misdiagnosis and missed diagnosis cases, it is found that some misdiagnosis is due to interference factors in the image, such as artifacts caused by metal objects on patients' clothes under X-ray, which affects the model judgment; Missed diagnosis is mostly due to the early characteristics of the disease is not obvious. The performance evaluation of the lung disease diagnosis model is shown in Table 1, and it is considered from multiple dimensions such as recall rate and F1 value.

The results showed that the recall rate of pneumonia was 82%, and the F1 value was 83.5%. The recall rate of pulmonary tuberculosis was 85%, and the F1 value was 86.5%. These data show that the model has high reliability in the diagnosis of lung diseases, but there is still room for improvement.

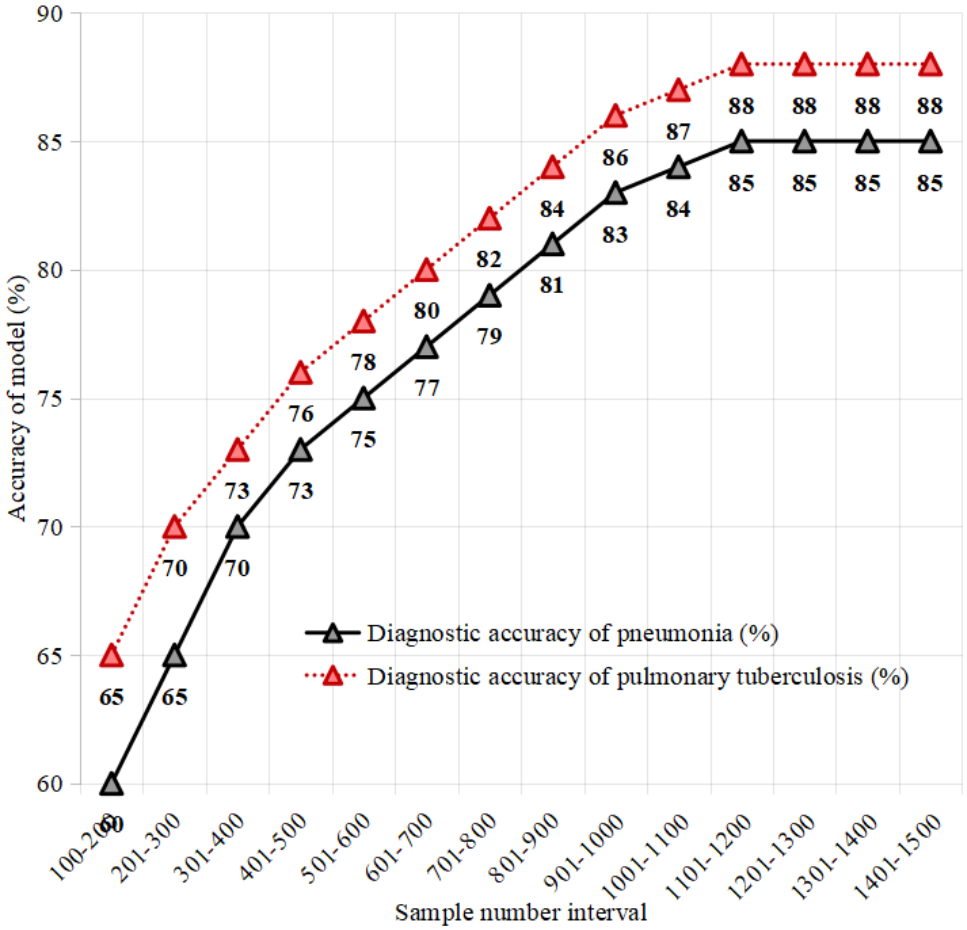

Figure 1 Changes of model accuracy under different sample sizes of lung diseases

Table 1 Performance Evaluation Table for Lung Disease Diagnostic Models

Disease Type	Accuracy	Recall	F1 Score
Pneumonia	85%	82%	83.5%
Tuberculosis	88%	85%	86.5%

(2) The field of social media public opinion analysis: Taking the comment data about electronic products on a popular social media platform as the research object, 5000 comments were collected as a data set. Preprocessing these text data, including removing irrelevant symbols, stop words, etc., and then using LSTM to build a statistical model for emotional analysis, emotional tendencies are divided into three categories: positive, negative and neutral.

Table 2 Proportion of Sentiment Tendencies in Electronic Product Reviews across Different Time Periods

Time Period	Proportion of Positive Reviews (%)	Proportion of Negative Reviews (%)	Proportion of Neutral Reviews (%)
Week 1 after Product Release	70	15	15
Week 2 after Product Release	65	18	17
Week 3 after Product Release	60	22	18
Week 4 after Product Release	55	25	20
Week 5 after Product Release	50	30	20
Week 6 after Product Release	45	32	23
Week 7 after Product Release	40	35	25
Week 8 after Product Release	38	38	24

In the process of model training, word vectors are used to transform the text into numerical

vectors and input them into the model. After the training, it is applied to the test set, and the results are shown in Table 2. As can be seen from Table 2, at the initial stage of product release, the proportion of positive comments is relatively high, and with the passage of time, if the product has problems, the negative comments will gradually increase. The recognition accuracy of the model is 80% for positive emotion, 83% for negative emotion and 78% for neutral emotion. The model has some limitations in dealing with complex semantics and implied emotions. Although some comments are positive on the surface, they actually hide negative attitudes, which leads to misjudgment of the model. On the whole, however, the model can quickly and effectively capture the emotional tendency of comments on electronic products on social media, and provide strong support for enterprises to understand product reputation and adjust market strategies.

5. Conclusions

This article focuses on the statistical modeling and application of unstructured data based on DL, and has achieved a series of valuable results. In the field of medical imaging diagnosis, the model constructed by DL shows high accuracy in the diagnosis of lung diseases, with the accuracy of pneumonia reaching 85% and tuberculosis 88%. This shows that DL model can effectively extract key features from unstructured data such as medical images, assist doctors to judge diseases, and improve the efficiency and accuracy of diagnosis. In the analysis of social media public opinion, the LSTM-based model also has considerable accuracy in identifying the emotional tendency of electronic product reviews, with the accuracy of 80% for positive emotions and 83% for negative emotions, which provides strong support for enterprises to master product reputation and formulate marketing strategies.

However, the research process also exposed some problems. DL model training depends on a large number of labeled data, which is costly and inefficient. The complex structure of the model leads to poor interpretability, just like "black box" operation, which makes it difficult to clarify its decision logic and limits its application in some fields that require high security and reliability. In addition, the fusion processing technology of different types of unstructured data is not mature, which affects the comprehensive processing ability of the model to multi-source data.

Future research can focus on reducing data labeling costs, improving model interpretability and optimizing data fusion methods. To reduce dependence on large amounts of labeled data, semi-supervised learning and active learning methods can be employed. Visualization technology and feature importance analysis can be utilized to improve model interpretability. Furthermore, more effective data fusion algorithms should be explored to further unlock the potential value of unstructured data and promote the sustainable development and widespread application of deep learning in the field of unstructured data processing.

References

[1] Wang Yaqiang, Yang Xiao, Zhu Tao, et al. A Postoperative Risk Prediction Model Enhanced by Unstructured Data Representation[J]. Journal of Chinese Information Processing, 2024, 38(1): 156-165.

[2] Liang Xueqing, Du Shuming. A Distributed Storage Method for Unstructured Data Based on the MapReduce Model[J]. Microcomputer Applications, 2022, 38(07):106-109.

[3] Hou Benzhong, Zhang Yongqiang, Shang Ying, et al. Natural Language-Based Extraction of Unstructured Data in Cloud Databases[J]. Information Technology, 2023, 47(3):57-63.

[4] Xu Weiyou, Wu Zheng, Nong Zhenchang, et al. An Encryption Storage Method for Unstructured Network Data in Cloud Computing Environments[J]. Electronic Design Engineering, 2025, 33(06):159-163.

[5] Wan Lei. A Chunk Storage Method for Unstructured Cloud Data Based on Decision Tree Models[J]. Microcomputer Applications, 2024, 40(9):197-201.

[6] Zhang Xiaorong, Xue Pengcheng, Li Yan. A Fast Mining Algorithm for Low-Dimensional Redundant Unstructured Data Based on Rough Neural Networks[J]. Microcomputer Applications, 2024, 40(12):199-201.

[7] Zhang Mingming, Zha Yiyi, Wang Chong. Research on Unstructured Data Mining for Equipment Using Improved Fuzzy Clustering[J]. Information Technology, 2023, 47(11):168-172.

[8] Luo Jinbin, Zhang Jian, Guo Qidi, et al. Feature Extraction Technology for Unstructured Data Oriented Towards Intelligent Inspection Terminals[J]. Electronic Design Engineering, 2025, 33(1):100-103.

[9] Zhu Juntao, Liu Jiaqi, Yang Lu. Semantic Segmentation of Drivable Areas for Unstructured Roads[J]. Journal of Tianjin University of Technology, 2025, 41(2):105-112.

[10] Hu Tao, Wang Zhongjie, Zhang Lianming, et al. Simulation of Density-Based Clustering for Unstructured Big Data Using Deep Learning[J]. Computer Simulation, 2024, 41(5):501-505.