International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 4094-4105

https://doi.org/10.70517/ijhsa463349

# Research on multi-level information extraction and optimal training based on recurrent neural network

**Xinwen Chen**[1,*]
[1] College of Information Engineering, Ezhou Vocational University, Ezhou, Hubei, 436000, China
Corresponding authors: (e-mail: cxw0012024@163.com).

**Abstract** In this paper, we select the structure of Cell Simultaneous Recurrent Neural Network (CSRN), which is good at dealing with two-dimensional structural data processing, as a prediction method for multilevel information data, and explain its network structure and operation principle. It also describes the structure of four important elements, namely, unit state, forgetting threshold, input threshold and output threshold, in the special long and short-term memory network of recurrent neural networks. The multilevel information is transformed into multimodal information, and the data information of different modalities is analyzed using the projection tracing method. Combine with recursive neural network algorithm to construct a multilevel information extraction model. Comparing the extraction performance of similar modeling algorithms on multimodal information data, the designed multilevel information extraction model performs the best in all indicators on data set-1. The F1 value of 86.34%, the precision rate of 88.84%, recall rate of 88.22% and F1 value of 88.52% in Marco-F1 values show excellent multilevel information extraction performance.

**Index Terms** multilevel information extraction, long and short-term memory networks, projection tracing method, unitary simultaneous recurrent neural network

## I. Introduction

With the popularization and rapid development of the Internet, media data in different modalities such as images, texts and videos grow rapidly. In the face of huge multimedia data, it is of great practical significance to extract the specific modal information you need quickly and accurately [1]. Currently, the processing means for information can be divided into global feature representation and local feature representation. Global feature representation is broader and describes the semantic expression of the whole body of information, while local features are fine-grained features, which focus only on the important regions detected and extract deep features for each region individually [2]-[4]. Since local feature information extraction can accurately extract the significant features contained in the original information, local feature-based information extraction methods are more widely used [5]. However, the use of local feature representation method also brings certain shortcomings. Because the user only extracts features from the detected significant areas, which leads to the background environment information that does not contain global information in the features, and the local features are extracted independently of each other, the user can not mine the high-level relationship between objects among independent objects [6]-[9]. So it is not enough to use only local features in complex scenes.

Based on this, the method of multilevel information extraction and optimized training comes into being. Multi-level information refers to the global and local information, in the specific implementation is not only to extract the overall information features, but also to carry out the detection of important regions and extract the regional features separately [10]. In this way, for the input information, not only get the important target features in the information extraction process but also get the background information of the image, and also combine the local features and global features to dig out the potential relationship between local objects [11]-[13]. With the development of deep learning, various deep neural networks have been utilized to extract deep features of images and texts and use these features for similarity computation, which plays an important role in the information extraction task of complex scenes [14]-[16].

This paper first analyzes in detail the computational methods and training methods of unitary simultaneous recurrent neural networks. Secondly, it discusses the structure of four important elements, corresponding formulas, and variants of the long and short-term memory network, which is better than the general recurrent neural network as a whole. Then, we map the multilevel information into multimodal information, and describe the process and steps of extracting multimodal information using the projection tracing method, so as to form a multilevel information extraction model based on recurrent neural network. Finally, the overall performance of the proposed

model is examined by comparing the performance of similar modeling algorithms on many different multimodal datasets. To address the shortcomings of the proposed model, remote sensing image multi-source datasets are used to optimize the training and improve the generalization ability of the model.

## II. Establishment of multilevel information extraction model

### *II. A.Principle analysis of unitary simultaneous recurrent neural networks*

CSRN, is a special type of neural network, which differs from MLP in that the network structure incorporates a recurrent structure, unlike the recurrent over time recurrent structure of RNN networks, which iterates through a loop with the outputs of the internal hidden layer nodes as inputs again at the same moment of inputs instead of iterating through time. They were first shown to be a powerful tool for solving the maze traversal problem, a task that cannot be solved using typical feedforward neural networks (multilayer perceptrons). Since the initial application of maze traversal, CSRNs have been applied to a variety of problems such as image processing.CSRNs are neural networks consisting of sub-networks arranged in sub-cells to match the topology of the problem domain, typically a two-dimensional grid. For example, in maze traversal, each cell can represent a wall cell or a channel cell. In image processing, each cell may correspond to a pixel of an image to be processed. In this study, each cell may correspond to a measurement of signal strength. The most basic mechanism of this network structure is the Simultaneous Recurrent Neural Network, referred to as SRN structure in this paper, which is shown in Fig. 1.The SRN uses recurrentity to represent a more complex relationship between an input vector $X(t)$ and an output $Y(t)$, regardless of other time $t$.

In Fig. 1, the neural network output is returned again as the same network input, conceptually without time delay, and the input and output should be simultaneous, hence the term simultaneous recurrent neural network CSRN. In practice, there is a physical time delay between the output and the input, but when implemented in fast computers, this delay is negligible relative to the delay between different frames of the input data.

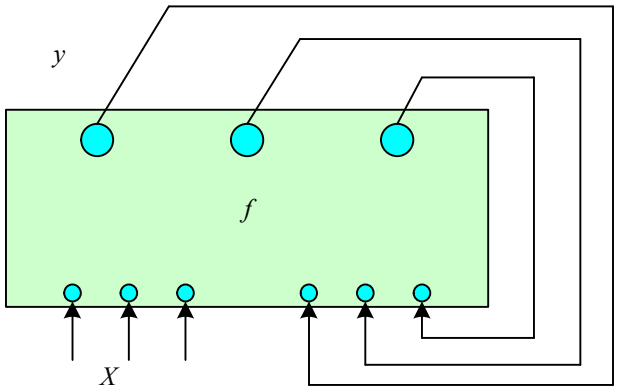



Figure 1: SRN structure

CSRN is defined as mapping as in equation (1):

$$\overline{Y}(t) = F(X(t), W) \tag{1}$$

Calculated by iterating Eq. (2):

$$y^{(n+1)}(t) = f(y^{(n)}(t), X(t), W) \tag{2}$$

where $f$ is some kind of feedforward network or system and $\overline{Y}(t)$ is defined as equation (3):

$$\overline{Y}(t) = \lim_{n \to \infty} y^{(n)}(t) \tag{3}$$

In Fig. 1, $X$ is the input data of the current time frame $t$, and the vector $y$ denotes the intermediate output of the network, which is looped back to the network as an additional input.The core of the SRN is the feed-forward network that realizes the function $f$, and the output $y$ when $n$ is sufficiently large is the final output of the network.

In Eqs. (2) and (3), the integer $t$ denotes the slower time period, e.g., the input data inter-frame delay, which in this paper can denote the next grid point sequence number. The integer $n$ denotes the time with more iterations and faster speed, i.e., the number of forward iterative computations on the same grid point.

CSRN network training methods are varied, this paper introduces the most commonly used BPTT training and uses this method to train the network, the BPTT training process is shown in Figure 2.

Figure 2 shows the application of BPTT in training SRN. The left side (solid arrows) represents the neural network that predicts the desired output $Y$ (in this paper, $Y$ denotes the true value of the signal intensity at all points in the grid), each box denotes the call to the feed-forward system, $X(t)$ is the external input to the system (in this paper, it consists of the two variables indicating the square of containing measurements and the size of the measurements in the grid ), and the initial vector $y(0)$ is set as a constant vector.

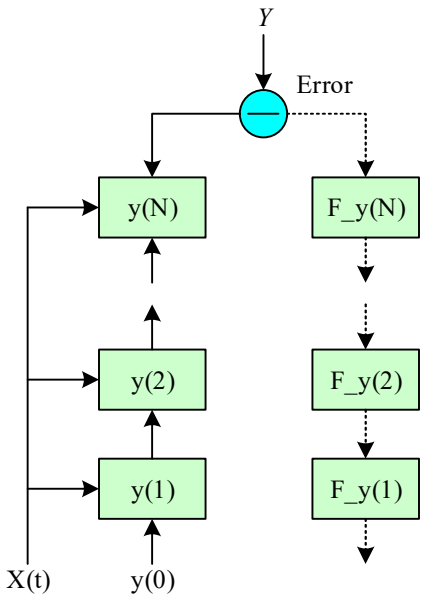



Figure 2: BPTT for SRN

The right side of Fig. 2 illustrates the backpropagation calculation for computing derivatives. For the SRN, the final error depends only on the output of the last iteration, so the last iteration receives feedback only from the final error, and the other iterations receive feedback from subsequent iterations, and each box on the right side represents the backpropagation computation through the left feedforward system.

### *II. B.Long and short-term memory networks*

The percentage of repetitive sections in the long and short-term memory network is shown in Figure 3.

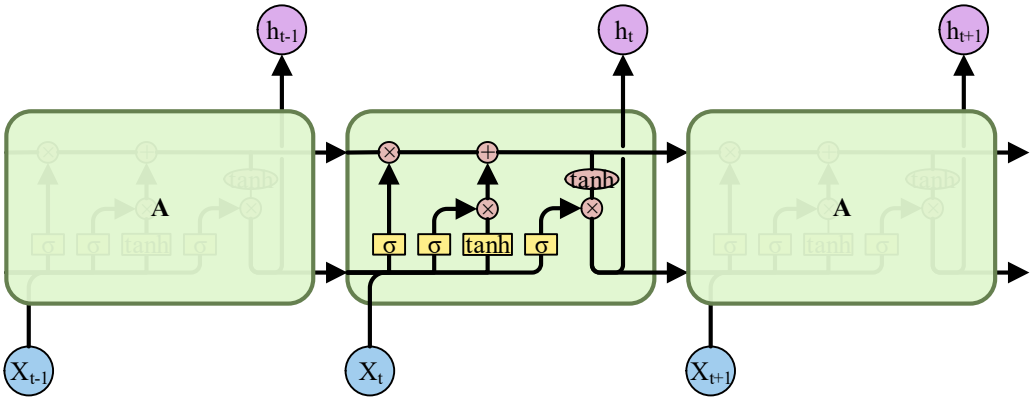



Figure 3: One LSTM memory cell

Long Short-Term Memory (LSTM) networks with input and output thresholds are designed to overcome the phenomenon of gradient vanishing in general recurrent neural networks. The LSTM nodes containing the three thresholds are shown in Fig. 3, where multiple nodes are connected into a chain structure identical to that of a

standard recurrent neural network, but with a different composition of individual nodes. The standard recurrent neural network node consists only of a simple hyperbolic tangent layer, while the long short-term memory network node contains four important elements: unit state, forgetting threshold, input threshold, and output threshold.

Unit state: This is the most important difference between long and short term memory networks and standard recurrent neural networks, the unit state in a long and short term memory network is shown in Fig. 4, the unit state is like a conveyor belt that passes the information on in a chain structure. The information in the cell state can be added or deleted in each node, and these modifications are accomplished through a structure known as a threshold. Thresholds are selective passages of information, consisting of layers of sigmoid neural units and multiplications between elements. The output value of the sigmoid layer is between 0 and 1, which represents how much of the input can pass through that threshold, if the sigmoid output is 0, it means that no information can pass through, on the contrary if the sigmoid output is 1, it means that all the information can pass through. It is also the existence of the cell state that ensures that the gradient in backpropagation can propagate along the time step and there is no gradient vanishing. The new cell state in each node is updated by the cell state in the previous node through a forgetting threshold and an input threshold.

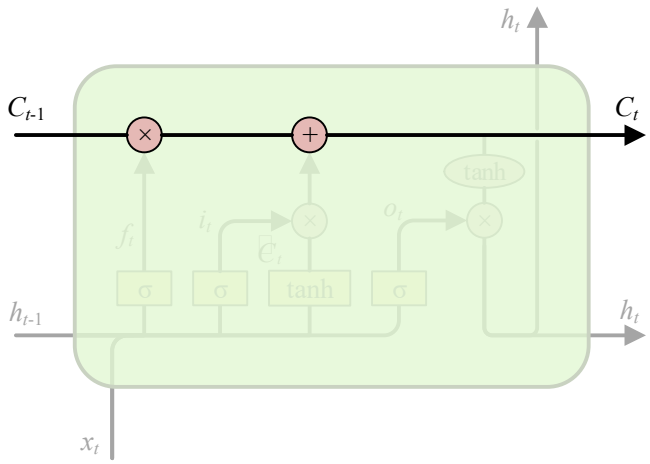



Figure 4: The cell state of LSTM

The forgetting thresholds in the long and short-term memory network are shown in Figure 5.

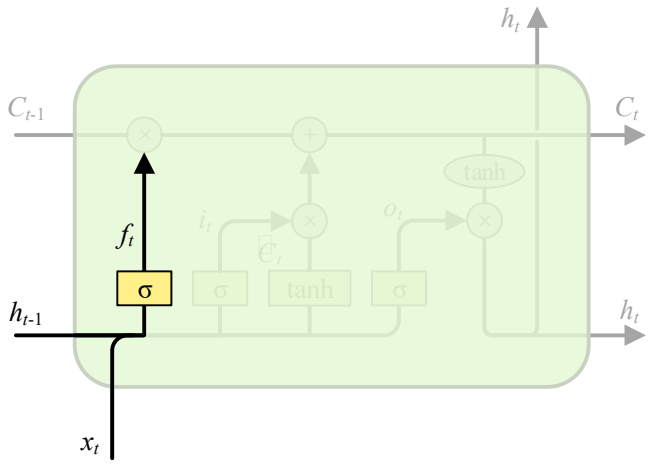



Figure 5: The forget gate of LSTM

As shown in Figure 5, the forgetting threshold determines which information in the unit state should be discarded. The inputs to the forgetting threshold are the hidden state $h_{t-1}$ of the previous time node and the input information $x_t$ of the current node, while the output values, which range from 0 to 1, correspond to each element in the unit state and determine how much of each element in the unit state should be forgotten.

The input thresholds in the long and short-term memory network are shown in Figure 6.

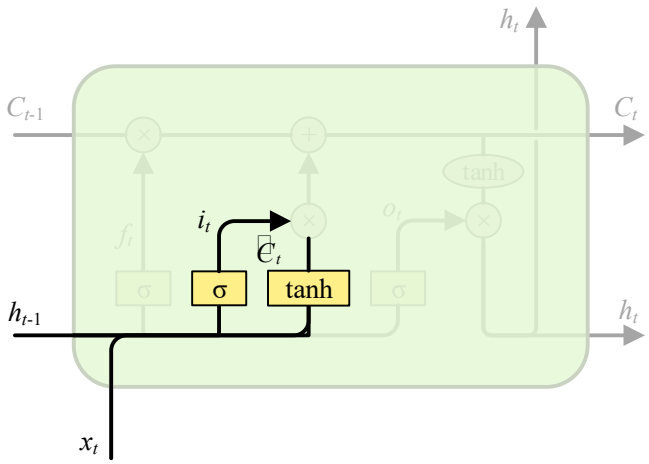



Figure 6: The input gate of LSTM

As shown in Fig. 6, the input threshold determines what new information is added to the cell state and consists of two main parts. One part is a sigmoid layer that uses the hidden state of the previous node and the input from the current node to determine which values will be updated. The other part is a tanh layer that is used to generate new candidate values that can be added to the cell state. The final combination of the above two parts work together to update the cell state.

The output thresholds in the long and short term memory network are shown in Figure 7.

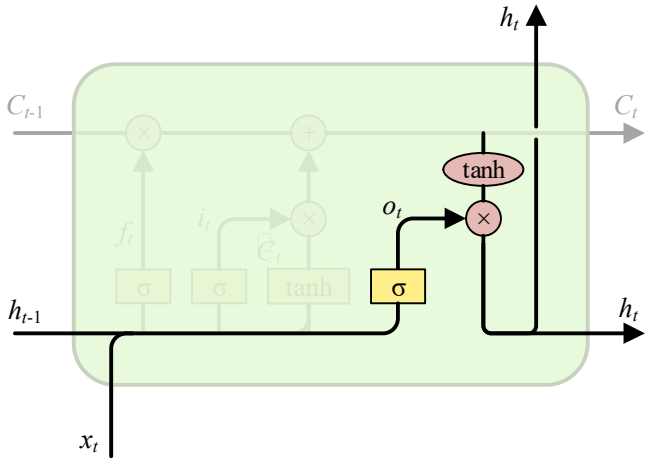



Figure 7: The output gate of LSTM

As in Fig. 7, the output threshold determines the output of the current node. Firstly, the hidden state of the previous node and the input of the current node are utilized through the sigmoid layer to determine which values in the unit state will be output, and in addition, the unit state is processed through the tanh layer and multiplied with the output of the sigmoid layer to obtain the hidden state of the current node.

The long and short-term memory network with forgetting threshold is calculated as Eqs. (4)-(8):

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{6}$$

$$C_t = f_t \square\ C_{t-1} + i_t \square\ \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

$$h_t = o_t \square \tanh(C_t) \tag{8}$$

where $i_t$, $f_t$, and $o_t$ are the input threshold, forgetting threshold, and output threshold of the node at moment $t$, $x_t$ is the node input at moment $t$, $C_t$ is the node unit state at moment $t$, and the vector $h_t$ is the hidden state of the long- and short-term memory network at moment $t$.

The long- and short-term memory network with the unit states introduced into each threshold is shown in Figure 8.

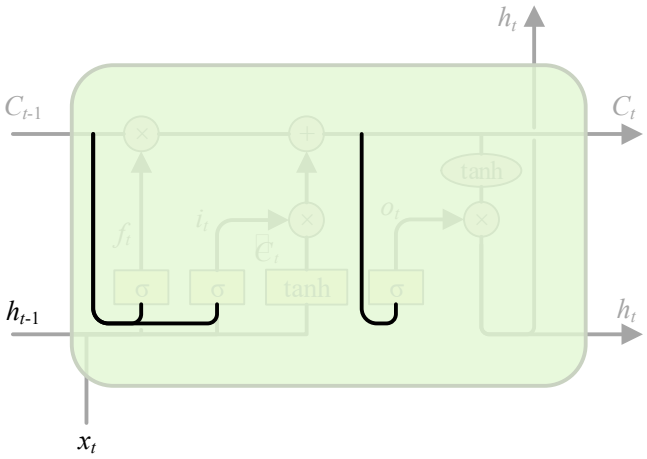



Figure 8: One LSTM variant introducing the cell state into each gate

There are many other variants of long short-term memory networks, one of the more common ones is Figure 8, which was proposed in which the cell states are introduced into the calculation of the thresholds. This variant of the long- and short-term memory network is used in the subsequent experiments, and its computational formulas are given in Eqs. (9)-(13):

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i C_{t-1} + b_i) \tag{9}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f C_{t-1} + b_f) \tag{10}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_{t-1} + b_o) \tag{11}$$

$$C_t = f_t \square C_{t-1} + i_t \square \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{12}$$

$$h_t = o_t \square \tanh(C_t) \tag{13}$$

***II. C.Multimodal information extraction model based on projection tracing method***

In this section, a multimodal information extraction model framework based on projection tracing method is designed, and the specific flow of information extraction is shown in Fig. 9. The model is based on the idea of projection tracing method, using inter-modal isomorphism, modal disassembling of the three modal data in the document, namely, text, pictures, and forms, and then information extraction is carried out respectively on the disassembled three types of single modal, so as to realize the acquisition of text, picture, and form information in the unstructured document.

The disassembly and extraction of pictures, text and tables in unstructured documents mainly includes three parts:

(1) Picture extraction: first based on the html format conversion to obtain the picture in the document, and then based on the regular matching method, combined with the picture name to obtain the picture content with picture tags.

(2) Text extraction: first based on the txt format conversion will be converted to plain text, through regular matching to obtain the text of multi-level title, picture name, table name, multi-level title hierarchical expression, and finally based on the keywords in the multi-level title matching, access to the relevant title and the corresponding text content.

(3) Table Extraction: Firstly, based on table module, obtain the table information in the document. Then, based on the secondary traversal of the table reconstruction method to obtain the table. Finally, based on the reverse order matching method, combined with the table name and table name to get the table content with the table label.

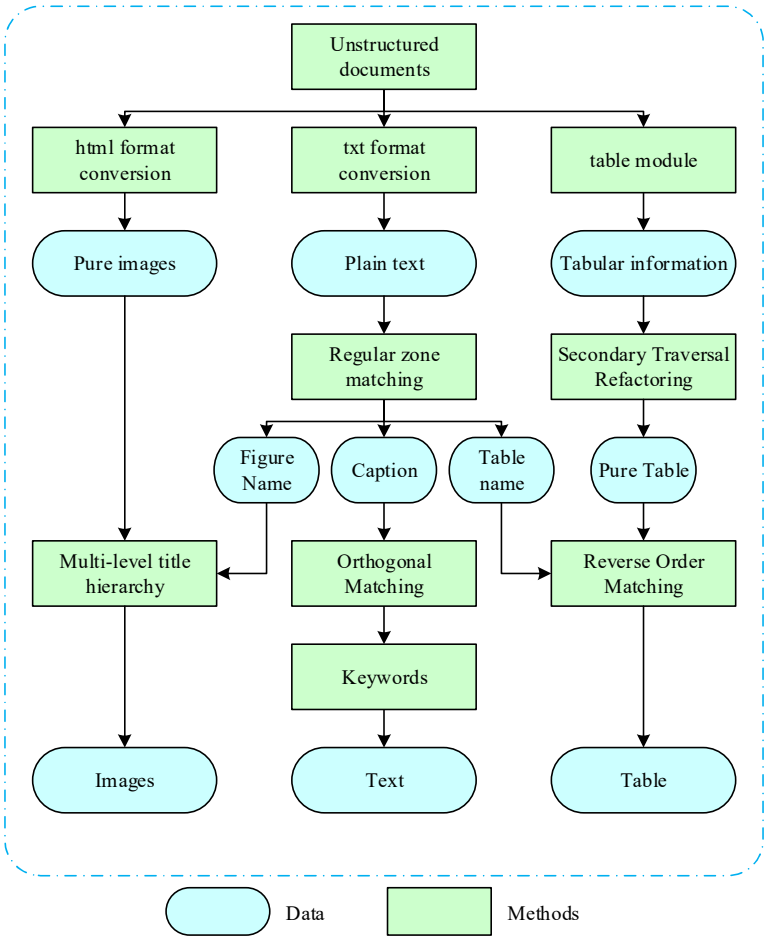



Figure 9: Multimodal information extraction based on projection tracing method

## III. Optimized Training of Multilevel Information Extraction Models

This chapter designs the performance of the multilevel information extraction model in comparison with similar model algorithms, on multimodal information datasets. Combined with the performance results, the model algorithm of this paper is trained using multi-source datasets of remote sensing images to validate and enhance the generalization ability of the model algorithm.

### *III. A. Performance evaluation of the model*

#### III. A. 1) Comparative experiments with similar modeling algorithms

Comparison of the performance of this paper's model with the existing mainstream models on the positive and negative imbalance dataset-1 is shown in Table 1, and the evaluation metric used is the Marco-F1 value, taking into account the positive and negative sample imbalance. The selected mainstream models belong to textual modality, image modality, and multimodal modality, respectively, where there are models under textual modality: (T1) TextCNN, (T2) SIARN, (T3) RMSD, under image modality: (F1) ResNet, (F2) ViT, (F3) ConvNeXt, and under multimodal modality: (M1) HFM, (M2) Res-Bert, (M3) CMGCN.

It can be found that the Marco-F1 values of the 10 models are higher overall, indicating that the positive and negative sample imbalance problem significantly affects the final training results, and the models have better detection performance for negative samples. And the model designed in this paper has the best performance in all the indicators, with an F1 value of 86.34%, a precision rate of 88.84%, a recall rate of 88.22% and an F1 value of 88.52% in the Marco-F1 value.

Table 1: The F1 value results of different models

| Model | F1 Value(%) | Macro-F1 value(%) | | |
|---|---|---|---|---|
| | | Accuracy rate | Recall rate | F1 value |
| T1 | 75.33 | 78.04 | 78.29 | 78.16 |
| T2 | 75.64 | 80.35 | 78.82 | 79.58 |
| T3 | 77.54 | 80.88 | 78.21 | 79.52 |
| F1 | 61.54 | 60.14 | 73.09 | 65.98 |
| F2 | 63.44 | 65.69 | 71.36 | 68.41 |
| F2 | 63.61 | 64.86 | 72.74 | 68.57 |
| M1 | 80.19 | 79.41 | 82.46 | 80.91 |
| M2 | 80.86 | 78.88 | 85.09 | 81.58 |
| M3 | 80.91 | 80.88 | 85.81 | 82.93 |
| Textual | 86.34 | 88.84 | 88.22 | 88.52 |

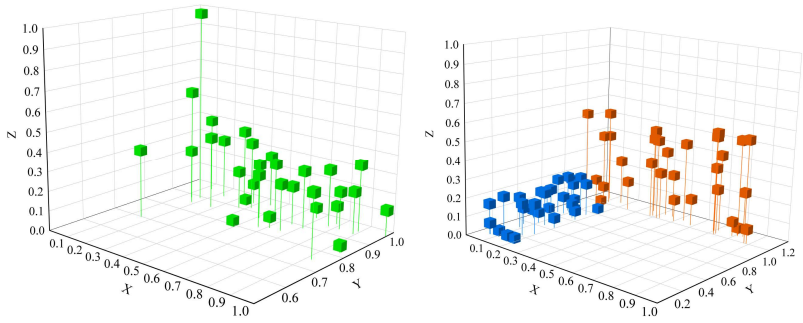
In addition, most of the currently available research works have conducted experiments and comparisons only on dataset-1. In this paper, SarcNet dataset is chosen to test the proposed model.The experimental results of 10 models on SarcNet dataset are shown in Table 2.The model of this paper achieves the best accuracy and F1 value on SarcNet dataset as compared to the other models, and it is the best among the 10 models in terms of accuracy and precision, and F1 value, which are 80.39%, 85.07% and 82.38%. However, it is slightly inferior to 79.86% in recall, which is 0.55% and 0.66% lower than (M2) Res-Bert and (M3) CMGCN models, respectively.

Table 2: The F1 value results of different models on the SarcNet dataset(%)

| Model | Precision rate | Accuracy rate | Recall rate | F1 value |
|---|---|---|---|---|
| T1 | 67.33 | 50.17 | 62.59 | 55.70 |
| T2 | 68.38 | 53.66 | 56.57 | 55.08 |
| T3 | 68.63 | 53.41 | 60.29 | 56.64 |
| F1 | 68.48 | 66.56 | 63.97 | 65.24 |
| F2 | 70.68 | 68.09 | 67.17 | 67.63 |
| F2 | 73.58 | 73.75 | 66.57 | 69.98 |
| M1 | 79.49 | 76.02 | 68.83 | 71.17 |
| M2 | 79.96 | 83.46 | 80.41 | 73.37 |
| M3 | 74.93 | 83.88 | 80.52 | 81.88 |
| Textual | 80.39 | 85.07 | 79.86 | 82.38 |

**III. A. 2) Visualization**

In order to better explain the effectiveness of this paper's model, the loss function optimization results of this paper's model on the positive and negative imbalance dataset are visualized in Fig. 10. Fig. 10(a) shows the optimization results without learning loss function, which has no positive and negative samples. Fig. 10(b) shows the result after optimization with learned loss function, and the positive and negative samples are well optimized and separated in the spatial distribution.




(a) Before optimization (b) After optimization

Figure 10: The optimization result of the triple loss function

### *III. B. Optimization of model training and analysis*

From the evaluation experiments in the previous section, it can be seen that the performance of the recurrent neural network-based multilevel information extraction model on different datasets is not exactly the best among similar model algorithms. In order to improve the operational performance of the designed model, this section takes multi-source remote sensing image data as the experimental object to train and optimize the model's ability to extract different information, and at the same time, to test the model algorithm's migratability on different structured data objects. Multi-source labeled datasets are constructed by combining existing publicly available remote sensing data images:(S1) GID, (S2) LoveDA, (S3) TriCities, and integrating the three datasets to get (S4) CRMS dataset. The accuracy results of the BANet method are also used as a supporting evidence for the analysis of the model in this paper.

#### III. B. 1) Optimized training results of multilevel information extraction models

The results of testing this paper's model on each dataset are shown in Table 3.The selected evaluation categories are (C1) buildings, (C2) roads, and (C3) cropland, and the evaluation metrics are: category accuracy (ACC), overall accuracy (OA), and mean intersection ratio (molU).

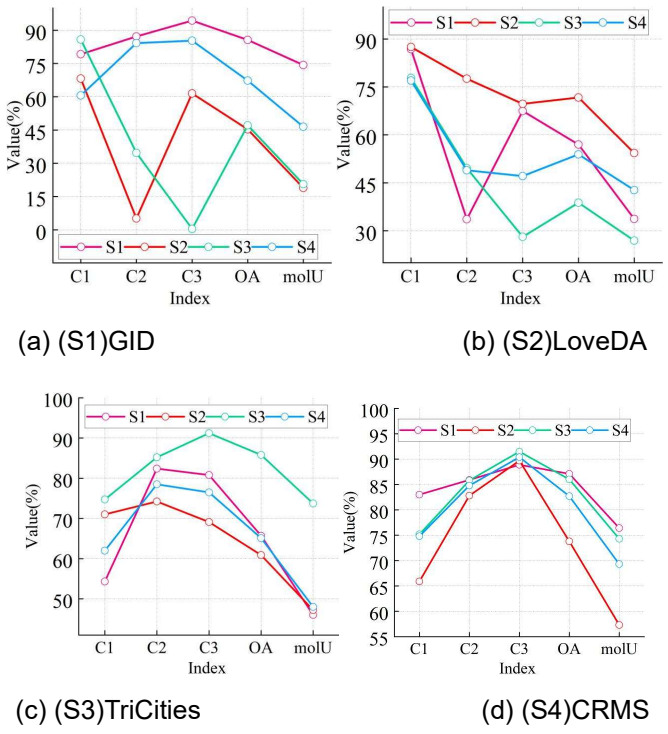
Table 3: The test results of the model trained in this paper on each dataset

| Training set | Test set | All types of ACC(%) | | | OA(%) | molU(%) |
|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | | |
| GID | GID | 79.2 | 87.2 | 94.3 | 85.6 | 74.3 |
| | LoveDA | 68.2 | 5.12 | 61.5 | 45.3 | 18.9 |
| | TriCities | 85.8 | 34.7 | 0.4 | 47.1 | 20.6 |
| | CRMS | 60.6 | 84.2 | 85.3 | 67.3 | 46.5 |
| LoveDA | GID | 86.8 | 33.6 | 67.5 | 57 | 33.7 |
| | LoveDA | 87.5 | 77.6 | 69.7 | 71.7 | 54.3 |
| | TriCities | 77.8 | 49.6 | 28.1 | 38.8 | 26.9 |
| | CRMS | 77 | 48.9 | 47.1 | 53.9 | 42.7 |
| TriCities | GID | 54.3 | 82.4 | 80.8 | 65.7 | 46 |
| | LoveDA | 71 | 74.2 | 69.1 | 60.9 | 47.2 |
| | TriCities | 74.7 | 85.2 | 91.2 | 85.8 | 73.7 |
| | CRMS | 62 | 78.5 | 76.5 | 65.1 | 48 |
| CRMS | GID | 83 | 85.9 | 88.9 | 87.1 | 76.4 |
| | LoveDA | 65.9 | 82.8 | 89.7 | 73.8 | 57.3 |
| | TriCities | 75.2 | 85.8 | 91.5 | 86 | 74.3 |
| | CRMS | 74.8 | 84.8 | 90.4 | 82.7 | 69.3 |

(S3)TriCities due to the property of multi-source and multi-resolution, hence the stronger migration ability and robustness of the models trained on this dataset. The experimental results support this view, showing that the models trained using only the self-annotated dataset maintain high accuracy on both (S1)GID, (S2)LoveDA datasets, concentrating on 70%-90%. In particular, the accuracy of the (C3) cropland category on the (S1)GID dataset decreased by 13.5%, while the loss of the cropland category on the (S2)LoveDA dataset was only 0.6%. This result effectively demonstrates the important role of multi-source image datasets in improving the model's relocatability and robustness. By constructing and utilizing datasets with multi-source multi-resolution remote sensing imagery, the generalization ability of the model algorithm in this paper can be significantly improved to better adapt to different classes of structured data objects.

The OA and mIoU of this paper's model on (S1)GID, (S2)LoveDA datasets are significantly improved after training on (S4)CRMS dataset. On the (S1)GID dataset, the OA of the model is improved by 1.5% and the mIoU is improved by 2.1%. While on the (S2) LoveDA dataset, the OA value is improved by 2.1% and the mIoU is improved by 3%.

In order to demonstrate more intuitively the superiority of this paper's model after dataset training, the experimental results of this paper's model trained on the four test sets are shown in Fig. 11.




(a) (S1)GID (b) (S2)LoveDA

(c) (S3)TriCities (d) (S4)CRMS

Figure 11: The visualization results of the model trained in this paper on each dataset

It can be observed from Fig. 11 that the model of this paper always achieves the optimal OA and mIoU accuracies after training on the (S4) CRMS dataset, regardless of which test set it is trained on. By uniting different dataset characteristics, the accuracy of this paper's model on the other source three datasets is improved.

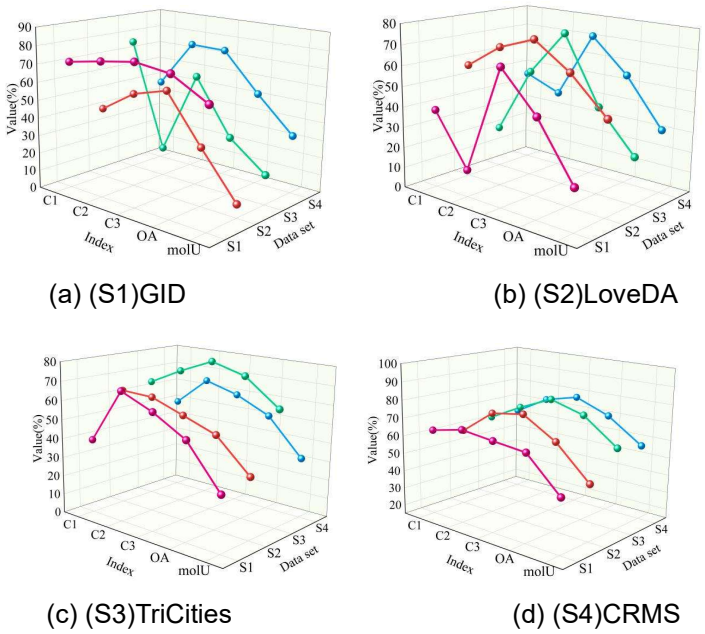
**III. B. 2) Optimized training results of the BANet method**

The experimental results of training BANet with different datasets and testing are shown in Table 4.Comparing with Table 3, it can be seen that the accuracy indexes of BANet trained on a single dataset and tested on the corresponding test set also have better performance. However, the overall accuracy is lower than the model in this paper, and its accuracy is more concentrated in the range of 40%-79%, and the accuracy of (C3) cultivated land image in TriCities dataset after training on CRMS training set is only obtained to reach 80.00%.

Table 4: The test results of the BANet model on each dataset after training

| Training set | Test set | All types of ACC(%) | | | OA(%) | molU(%) |
|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | | |
| GID | GID | 71.2 | 74.8 | 78 | 75.6 | 64.6 |
| | LoveDA | 40.4 | 53.1 | 58.9 | 32.6 | 7.6 |
| | TriCities | 75.8 | 16.4 | 62.5 | 32.4 | 16.8 |
| | CRMS | 47.8 | 74.2 | 73.8 | 52.4 | 32.9 |
| LoveDA | GID | 38.5 | 13.3 | 65.8 | 47.2 | 20.6 |
| | LoveDA | 57 | 68.8 | 75.2 | 63 | 45.8 |
| | TriCities | 20.4 | 53.5 | 75.3 | 43 | 23.3 |
| | CRMS | 45.8 | 38.8 | 71.3 | 54.7 | 31.4 |
| TriCities | GID | 39.3 | 67.9 | 60.7 | 50.8 | 29.2 |
| | LoveDA | 62.2 | 61.6 | 55.4 | 49.1 | 32.4 |
| | TriCities | 64 | 72.7 | 80.1 | 75.2 | 61.6 |
| | CRMS | 49.5 | 64.5 | 59.5 | 51.2 | 32.1 |
| CRMS | GID | 63.1 | 67.2 | 65.2 | 63.4 | 45.3 |
| | LoveDA | 58.7 | 72.4 | 75.4 | 64.3 | 46.4 |
| | TriCities | 62.7 | 72.1 | 80 | 74.6 | 60.5 |
| | CRMS | 62.8 | 73.2 | 77.8 | 70.4 | 57.2 |

Fig. 12 gives the accuracy visualization results of the test on each dataset after training the BANet method on different datasets. It can be seen that because the BANet network lacks a targeted method design for different structural data information, its application on multi-source image datasets instead results in accuracy degradation on some of the images, which is not conducive to adapting to the task of information extraction from multi-source image data.




(a) (S1)GID (b) (S2)LoveDA

(c) (S3)TriCities (d) (S4)CRMS

Figure 12: The visualization results of the BANet model on each dataset after training

## IV. Conclusion

In this paper, under the support of the long and short-term memory network structure, the framework of multimodal information extraction is borrowed from the projection tracing method to complete the establishment of the multilevel information extraction model. The model shows superior performance far beyond similar algorithms in the extraction of multimodal data information, and it performs best in all indicators on the multimodal information dataset-1, with an F1 value of 86.34%, a precision rate of 88.84% in the Marco-F1 value, a recall rate of 88.22%, and an F1 value of 88.52%. On the multimodal information SarcNet dataset, there are three indicators of accuracy and precision, F1 value are the best among the 10 models, which are 80.39%, 85.07% and 82.38%, respectively. And in the optimization training expansion based on multi-source remote sensing dataset, it maintains high accuracy on several indicators in multiple datasets, concentrating on 70%-90%.

## Funding

This work was supported by Hubei Provincial Department of Education Class B Project (Project No: B2013003); Key Teaching and Research Project of Ezhou Vocational University (Project No: J2022ZD08).

## References

[1] Chen, S., Liu, S., & Liu, J. (2024). Type-specific modality alignment for multi-modal information extraction. IEEE Signal Processing Letters.

[2] Yang, L., Liu, B., Lin, H., & Lin, Y. (2016). Combining local and global information for product feature extraction in opinion documents. Information Processing Letters, 116(10), 623-627.

[3] Li, J., Xu, K., Li, F., Fei, H., Ren, Y., & Ji, D. (2021, August). MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 1359-1370).

[4] Ping Tian, D. (2013). A review on image feature extraction and representation techniques. International journal of multimedia and ubiquitous engineering, 8(4), 385-396.

[5] Madduri, A. (2021). Content based image retrieval system using local feature extraction techniques. International Journal of Computer Applications, 183(20), 16-20.

[6] Jing, J., Gao, T., Zhang, W., Gao, Y., & Sun, C. (2022). Image feature information extraction for interest point detection: A comprehensive review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4694-4712.

[7] Jian, M., Yin, Y., Dong, J., & Lam, K. M. (2018). Content-based image retrieval via a hierarchical-local-feature extraction scheme. Multimedia Tools and Applications, 77, 29099-29117.
[8] Barbieri, T. T. D. S., & Goularte, R. (2014, December). KS-SIFT: a keyframe extraction method based on local features. In 2014 IEEE International Symposium on Multimedia (pp. 13-17). IEEE.
[9] Zhang, C., Wang, X., Zhang, H., Zhang, H., & Han, P. (2021). Log sequence anomaly detection based on local information extraction and globally sparse transformer model. IEEE Transactions on Network and Service Management, 18(4), 4119-4133.
[10] Chang, J., & Han, X. (2023). Multi-level context features extraction for named entity recognition. Computer Speech & Language, 77, 101412.
[11] Liu, J., Yang, Y., & He, H. (2020). Multi-level semantic representation enhancement network for relationship extraction. Neurocomputing, 403, 282-293.
[12] Yang, L., Zhang, J., Guo, Y., & Wang, Q. (2017). Bayesian-based information extraction and aggregation approach for multilevel systems with multi-source data. Journal of Systems engineering and Electronics, 28(2), 385-400.
[13] Xuan, Z., Zhao, H., Li, X., & Chen, Z. (2025). Distantly Supervised Relation Extraction Method Based on Multi-Level Hierarchical Attention. Information, 16(5), 364.
[14] Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., & Wang, Z. (2022). A survey of information extraction based on deep learning. Applied Sciences, 12
[15] Chen, X. (2025). Multi-level information extraction based on convolutional neural network and its optimal training method. J. COMBIN. MATH. COMBIN. COMPUT, 127, 699-714.
[16] Chen, G., Zhao, D., & Zeinali, E. (2022, June). Analysis of Rapid Extraction of Data and Information Based on Deep Learning Technology. In International Conference on Applications and Techniques in Cyber Intelligence (pp. 951-959). Cham: Springer International Publishing.