Research on Multi-modal Data Fusion Based on Swarm Learning

Tao Cheng, Haiyan Kang

Abstract— **At present, the data of a single entity is not only increasing in data volume, but also in corresponding types of data modalities. In turn, applications involving multi-modal data fusion are increasing, and the corresponding privacy and security issues cannot be ignored. To this end, a Multi-modal Data Feature Fusion scheme based on Swarm Learning (MDFF-SL) is proposed to protect the privacy of multi-modal data. The SL-MDFF scheme puts all participants into a swarm learning architecture. First, the data features of two different modalities, namely text and image, are locally extracted, and then the respective data feature information is shared through the API interface in the swarm learning architecture for feature fusion. The entire solution ensures data privacy and security by virtue of the decentralization, tamper-proof, traceability and other characteristics of blockchain technology.**

Index Terms—**swarm learning, multi-modal data, data fusion, privacy protection.**

I. Introduction

Modality is a biological concept that refers to the way something happens or exists under sensory conditions. Multimodality refers to the various forms of combinations of two or more modalities [1]. Multimodal data refers to different data sets that cannot be processed directly and can be related to each other. With the rapid development of the Internet, the rapid growth of multimedia data such as images, text, sounds, videos, etc. has given rise to research on multi-modal issues such as image-text pairs and image-sound pairs. How to effectively utilize multi-modal data has become a key and challenging issue.

At the same time, with the arrival of the new industrial revolution, new generation artificial intelligence, big data analysis and processing and other technologies have brought opportunities for the intelligent transformation and development of traditional industries. For example, the manufacturing industry has upgraded from traditional manufacturing models to data-driven intelligent manufacturing [2]. Today, as data has become a factor of production and plays an increasingly important role, new challenges have been posed to data sharing and privacy security. Due to inter-industry competition, monopoly, and the closed and obstructive nature of business between different entities, data and information exchanges between entities are difficult. The phenomenon of "data islands" based on privacy security is becoming increasingly serious, in which multi-modal data accounts for an increasing proportion.

To solve the above problems, Google first defined Federal Learning (FL) in 2017 [3]. FL has received widespread attention from the academic community since it was proposed. The essence of federated learning is a distributed machine learning (ML) framework. Each participant collaboratively trains the machine learning model and uploads the training parameters to the central server without sharing its private data with other participants, thereby meeting the essential privacy protection requirements of "model looking for data". However, there are issues with sending all model data collected from various parties to a central server for computation, such as latency, energy requirements, compliance issues, and ever-increasing transmission and opportunity costs.

To this end, Hewlett Packard Enterprise (HPE) proposed the emerging concept of Swarm Learning (SL) [4] on its official website in 2019. SL is a decentralized, privacy-preserving machine learning framework. The framework leverages computing power at or near distributed data sources to run machine learning algorithms to train models. Each participant uses blockchain technology to share learning results with other participants in a safe and reliable manner, eliminating the need for a central server. On the one hand, it ensures the security and privacy of data, and on the other hand, it greatly improves the data transmission efficiency and system fault tolerance.

Currently, most research on multi-modal data fusion adopts centralized machine learning methods. This paper, for the first time, puts the multi-modal data fusion task under the decentralized swarm learning framework. It proposes a Multi-modal Data Feature Fusion based on Swarm Learning (MDFF-SL), which is aimed at text and image data modalities. First, appropriate deep learning models are used locally to extract the data features of the two different modalities of text and image. Then share their respective data feature information through the swarm learning network for feature fusion. While completing the multi-modal data fusion task, the auditability and reliability of the blockchain are used to improve privacy security in the multi-modal data fusion process.

II. Related work

A. Privacy Protection

In recent years, privacy protection bills have been introduced in various countries one after another. At the same time, people's awareness of data privacy protection has been continuously strengthened, attracting a large number of

Manuscript received February 27, 2024

Tao Cheng, School of Computer Science, Beijing Information Science and Technology University, Beijing, China

Haiyan Kang, School of Information Management, Beijing Information Science and Technology University, Beijing, China

scholars to pay attention to privacy protection issues. In addition to traditional privacy protection methods such as differential privacy [5], homomorphic encryption [6], and secure multi-party computation [7], a number of new privacy computing methods have emerged. Among them, blockchain combined with federated learning is a popular method in the current field of data privacy protection and has broad research value. Preuveneers et al. [8] proposed a blockchain-based federated learning model audit scheme. Model updates uploaded by the client need to be anomaly detected and linked to the distributed ledger. Clients whose detection results are greater than the predefined threshold will be held accountable. At the same time, in order to avoid the single point of failure problem of the central server, the blockchain is used to replace the direct interaction between the central server and the client. Therefore, each node in federated learning has a complete copy of the ledger and can compute aggregated weight updates. Kairouz et al. [9] proposed interactive model updates based on smart contracts to automatically verify model updates to defend against malicious and unreliable participants. At the same time, local differential privacy technology was introduced to prevent member inference attacks and realize privacy-safe federated learning in 5G networks. Maddali et al. [10] introduced VeriBlock, a blockchain-based verification scheme, proposed the Endorser-Verify model, and used verifiable calculations to provide mathematically verifiable execution proofs, reduce redundant execution of smart contracts without affecting the security of the blockchain system, and achieve privacy protection for data on the chain.

B. *Multi-modal Data Fusion*

In many machine learning problems, a sample is often represented by a feature set of multiple modalities, and each feature set can be regarded as a perspective of the original data set. These feature sets can be divided into two parts, namely different types of features and features from different data sources [11]. For example, in content-based web image retrieval, an object can be described by both the image and the text surrounding the image, where each feature describes different independent information about the same sample.

In recent years, many researchers have studied multi-modal fusion technology in deep learning, which has been widely used in tasks such as speech recognition, target detection, and emotion recognition. Bica et al. [12] explored social media images posted during two major earthquakes in Nepal from April to May 2015 by combining text and visual content to provide highly relevant information. Vempala et al. [13] studied the association between tweets and images and their use in classifying visually relevant and irrelevant tweets. They designed a classifier by combining features from text, images, and socially relevant contextual features (e.g., post time, follower ratio, number of comments, retweets), and reported an F1-score value of 70.5% in the binary classification task, which is 5.7% higher than using pure text classification. Similarly, Gautam et al. [14] in 2019 compared single-modal and multi-modal methods on crisis-related social media data, using a decision fusion-based method to classify them into informational and non-informational categories based on text and image features.

III. Related Technologies

A. *Swarm Learning*

Swarm learning [15], inspired by biology, is a decentralized machine learning solution based on blockchain technology that enables participating entities to utilize distributed data while protecting data privacy and security.

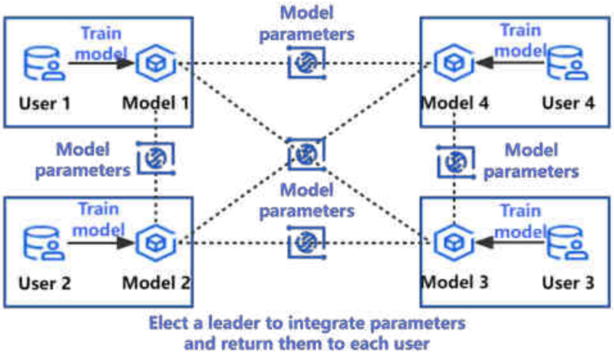

Fig.1 Swarm learning

The SL architecture is shown in Figure 1. As a latest technology, swarm learning does not require a central server for parameter aggregation. Each node independently builds a model locally using its own private data and shares parameters through the swarm network. This provides greater fault tolerance compared to traditional centralized parameter server-based frameworks. Swarm learning provides security measures to support data sovereignty, security, and confidentiality achieved through private blockchain technology. Each participant is clearly defined and only pre-authorized participants can execute transactions. When new nodes join, they need to pass verification measures before they can join the blockchain. New nodes register through blockchain smart contracts, obtain models, and perform local model training. Next, the model parameters are exchanged through the swarm application programming interface (API), and the parameters uploaded by each node are merged before starting a new round of training, and the model is updated with the new parameters.

Collaboration between nodes to train models and data sharing are two important features of SL. A further advantage is that data sharing can be transformed into parameter sharing and applying blockchain technology to ensure the security of each node and the identity authentication of each participating training node, thereby achieving node collaboration with completely confidential data.

Therefore, SL has the following advantages compared to the traditional centralized ML method and the FL method proposed by Google:

(1) Reduce the transmission cost of raw data and eliminate investment in central facilities for centralized storage and processing of aggregated data;

(2) Guarantee data privacy and security compliance;

(3) Effectively avoid single points of failure that threaten business continuity and improve the fault tolerance of the system;

(4) Reduce the delay in data creation and obtaining actionable model parameters from training data.

B. *Blockchain*

Blockchain [16]-[17] is a decentralized distributed database technology, which uses blocks as the basic unit and uses chain connections to connect data in chronological order. Each block contains a certain number of transaction records or other data, and is connected to the previous block through a cryptographic hash function, forming an tamper-proof chain.

The core characteristics of blockchain include decentralization, transparency, security, and traceability. Decentralization means that there is no single central authority that controls the entire network, but is maintained and managed jointly by the nodes in the network. Transparency means that all transaction records are publicly recorded on the blockchain and can be viewed by anyone, thus ensuring openness and transparency of data. Security is guaranteed through cryptography technology. Each block contains the hash value of the previous block. Anyone who wants to tamper with the data must modify all subsequent blocks at the same time, so it is extremely difficult. Traceability means that every transaction can be traced back to its historical record, thus enhancing trust.

Blockchain technology was initially used in the field of cryptocurrencies, such as Bitcoin, to record transaction information. But with the development of technology, it has been applied to many fields such as supply chain management, digital identity verification, smart contracts and other fields. The emergence of blockchain has changed the traditional centralized data management model, providing new ideas for building a more secure, transparent and efficient data exchange and management system.

C. *Multi-modal Fusion*

Multi-modal fusion is mainly divided into early fusion, late fusion, and hybrid fusion [18].

Preliminary fusion, also known as feature-level fusion, is to directly splice the features of each modality, and finally input the fused features into the classifier. The early fusion combines the information of each modality and can provide comprehensive feature information for the classifier. Early fusion has two disadvantages: (1) It ignores the alignment relationship between modalities, which can easily cause information redundancy; (2) Because information from different modalities may come from different representation spaces, fusion requires mapping the features of each modality into a unified space, which may cause information loss.

Late fusion, also known as decision-level fusion, inputs the features of each modality into independent classifiers, and integrates the decision results of each classifier to obtain the final classification result. The two advantages of late fusion are: (1) It is not necessary to convert the features of each modality into the same format, and a classifier suitable for the modality itself can be designed according to the characteristics of the same modality; (2) It is helpful to alleviate problems such as over-fitting of learning of a single classifier. Later fusion also has certain shortcomings. For example, decision-making fusion will accumulate the internal errors of independent classifiers, and fusion will produce larger errors.

Hybrid fusion is a method that combines early fusion and late fusion. Its disadvantage is that it increases the structural complexity and training difficulty of the prediction model. Hybrid fusion is widely used. For example, in image question and answer tasks, literature [19] uses recurrent neural networks and convolutional neural networks to learn question statements and image information respectively, and uses an attention mechanism to achieve the fusion of question statement information and image information.

IV. A MULTI-MODAL DATA FUSION SCHEME BASED ON SWARM LEARNING

A. *Design of the Scheme*

In order to effectively ensure data privacy and security in the multi-modal data fusion process, swarm learning technology is introduced, and a Multi-modal Data Feature Fusion scheme based on Swarm Learning (MDFF-SL) is proposed to achieve on-chain and off-chain collaborative fusion of multi-modal data. MDFF-SL consists of a swarm learning network and a data feature fusion process. The framework is shown in Figure 2.

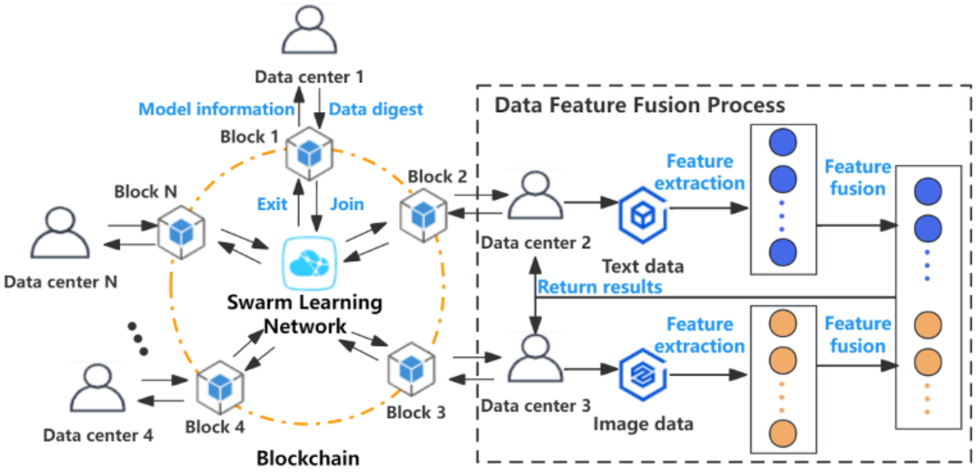

Fig.2 MDFF-SL framework

There are usually two situations where multi-modal data comes from. The first situation is that one data center provides data in different modalities for the same entity at the same time. The second situation is that each data center provides data of the same entity in different modalities. This paper mainly considers the second scenario and focuses on text data and image data.

For a certain data collaboration task, each data center with different modal data is registered through the blockchain smart contract to build a swarm learning network. The local data is not uploaded to the chain, and only the data summary information is uploaded to the blockchain. Each data center performs data fusion through data collaboration, and uses the fused feature vectors as input to build a multi-modal deep learning model for learning the correlation between text and images. Finally, a multi-modal model is obtained that can be used for practical tasks such as prediction and classification, and the model parameters are returned to each data center.

Taking data center 2 and data center 3 as examples to describe the multi-modal data fusion process:

First, data center 2 provides text type data, and data center 3 provides image type data;

Secondly, data center 2 and data center 3 respectively extract features from local data and convert different modal data into high-dimensional feature expressions;

Finally, the middle layer of the model is selected for feature fusion, and the obtained fusion results are sent back to each data center.

B. *Construction of Swarm Learning Network*

The swarm learning network refers to a temporary model training environment composed of various data centers that join the blockchain to participate in a certain data collaboration task. The data center that join the swarm learning network is also called participating node. The construction process of the swarm learning network is divided into the following three main stages:

(1) Initialization and joining

Initialization and joining is an offline process, and each data center is registered in the smart contract. Participating nodes joining the swarm learning network are able to work together to develop the operational and legal requirements of the decentralized system, including data sharing agreements, arrangements to ensure node visibility across entity organizational boundaries, and consensus on the expected outcomes of the model training process. They also agree on the general model to be trained and the reward system.

(2) Installation and configuration

After the initialization and joining process is completed, all participating nodes download and install the SL platform on their respective machines, and at the same time perform the initial configuration of the swarm learning network. Afterwards, start the SL platform and the connections between participating nodes. Booting is an orderly process in which a set of participant nodes designated as peer discovery nodes are started first, followed by the remaining nodes in the network.

(3) Integration and training

After each installation and configuration is completed, each participating node can perform data collaboration based on actual tasks. SL provides a simple set of APIs to enable rapid integration with multiple frameworks. These APIs are merged into existing code bases, enabling rapid conversion of independent machine learning nodes into swarm learning nodes, enabling data fusion, etc. The task proceeds smoothly in the swarm learning network.

C. *Feature Extraction*

Feature extraction involves several steps, including data preparation, data preprocessing, model construction, model training, model evaluation, and feature extraction.

1) Text Data Feature Extraction

Take the AWD-LSTM (ASGD Weight-Dropped LSTM) model as an example to extract features from text data. AWD-LSTM is a neural network model for processing text data, which was proposed by Stephen Merity et al. in 2017 [20]. By combining LSTM (Long Short-Term Memory Network) and a series of regularization techniques, this model can effectively capture long-term dependencies in text data and prevent overfitting to a certain extent. The specific process of text information feature extraction is as follows.

Step1: Data preparation

Collect and prepare text datasets, including training sets and test sets.

Step2: Data preprocessing

Clean text data and remove useless punctuation marks, special characters, etc. Word segmentation, dividing the text into independent words or phrases. Build a vocabulary that maps words in the text to unique integer indices. Convert text data into a sequence of integers, with each word corresponding to its index in the vocabulary. Pad or truncate text sequences so that each sequence is the same length.

Step3: Model Construction

Build a neural network model based on AWD-LSTM, including multi-layer LSTM units, dropout layers, etc. The embedding layer can be initialized using a pre-trained word embedding model, such as Word2Vec, etc.

Step4: Model training

Define a loss function, such as the cross-entropy loss function, to measure the difference between the model prediction results and the real label. Select an optimizer, such as Adam or SGD, etc., to update the model parameters to minimize the loss function. The training set is used to train the model, and the model parameters are continuously updated through the back propagation algorithm, so that the model can better fit the training data.

Step5: Model evaluation

Use the validation set to evaluate the model performance and monitor the model's generalization ability on unseen data. Use various evaluation metrics such as accuracy, precision, recall, F1 score, etc. to evaluate the performance of the model.

Step6: Feature extraction

In the trained model, the feature representation of the text is obtained by extracting the hidden state of the LSTM layer. These feature representations can be used for subsequent tasks such as text classification, sentiment analysis, etc.

2) Image Data Feature Extraction

Take the Resnet50 model [21] as an example to perform feature extraction on image data. ResNet (Residual Network) is a deep residual network structure proposed by Microsoft Research, which is widely used in computer vision tasks such as image classification, target detection, and semantic segmentation. ResNet50 is a deeper and more widely used network model in the ResNet series, which has a depth of 50 layers. The specific process of image information feature extraction is as follows.

Step1: Data Preparation

Collect and prepare image datasets, including training and test sets. Make sure that the label information for the image dataset is correct and that each image has a corresponding label. Divide the data set into a training set and a test set, usually using a ratio of 70-30 or 80-20.

Step2: Data preprocessing

Load the image datasets and perform necessary preprocessing on the images, such as scaling, cropping, etc., to ensure that the images input to the model have the same size. Normalize the image data, scaling the pixel values to the range [0, 1], or use other normalization methods.

Step3: Model construction

Use the deep learning framework to load the pre-trained ResNet50 model, and load it directly through the ready-made model library. After the model is loaded, the top layer of the model (fully connected layer) is removed so that only the convolutional layer and the pooling layer are retained, thereby achieving feature extraction without classification.

Step4: Model training (optional)

For feature extraction tasks, additional model training is usually not required because ResNet50 has been pre-trained on large-scale image data sets.

Step5: Model evaluation (optional)

If fine-tuning or other custom operations are performed, you can use the test set to evaluate the model to ensure the performance of the model on new tasks.

Step6: Feature extraction

Input the preprocessed image data into the ResNet50 model with the top layer removed. The feature representation of each image in the model is obtained through forward propagation, which is usually the feature vector obtained through global average pooling. These feature vectors are the feature extraction results of image information, which can be used for subsequent machine learning tasks or image retrieval applications.

D. Feature Fusion

Feature fusion adopts an intermediate fusion method, taking the output of any layer of the pre-trained model before the softmax layer and concatenating them into a feature vector and returning it to each data center. There is a trade-off between the ease of classification and the completeness of information throughout the process. The features extracted at each stage will lose some information contained in the representation of the previous stage, but the processed data at each stage becomes more separable, thus making classification easier. Therefore, the best option is to concatenate the processed output feature vectors of the penultimate fully connected layers of the two models in an integrated feature representation.

V. Conclusion

In the era of artificial intelligence and big data, data is being presented in different modal forms while increasing in volume. Multi-modal data is playing an increasingly important role in people's lives. In order to effectively utilize multi-modal data to better carry out practical applications such as disease prediction, target detection, emotion recognition, etc., researchers have proposed multi-modal data fusion methods, and related data privacy and security issues have also arisen. This paper proposes a Multi-modal Data Feature Fusion scheme based on Swarm Learning (MDFF-SL) to provide privacy protection for the process of multi-modal data fusion, thereby reducing the risk of multi-modal data sharing leakage and further promoting the development of related industries and applications.

References

[1] Z. Ren, Z. Wang *et al.*, "A review of multimodal data fusion," *Journal of Computer Engineering & Applications*, vol. 57, no. 18, 2021.

[2] F. Jonas *et al.*, "A framework for data-driven digital twins of smart manufacturing systems," *Computers in Industry*, vol.136, 2022.

[3] B. McMahan *et al.*, " Communication-efficient learning of deep networks from decentralized data, " in *Artificial intelligence and statistics*, 2017, pp. 1273-1282.

[4] Swarm Learning, August 3, 2023. [Online]. Available: https://www.hpe.com/us/en/what-is/swarm-learning.html.

[5] X. Wang and H. Kang, "Research on noise addition and precision analysis in differential privacy," *Journal of Lanzhou University of Technology*, vol. 49, no. 3, pp. 94-103, 2023.

[6] M. Albrecht, M. Chase, H. Chen, *et al.*, "Homomorphic encryption standard," *Protecting privacy through homomorphic encryption*, pp. 31-62, 2021.

[7] J. Zhou, Y. Feng, Z. Wang, *et al.*, "Using secure multi-party computation to protect privacy on a permissioned blockchain," *Sensors*, vol. 21, no. 4, pp. 1540-1557, 2021.

[8] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, *et al.*, "Chained anomaly detection models for federated learning: An intrusion detection case study," *Applied Sciences*, vol. 8, no. 12, pp. 2663-2684, 2018.

[9] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1-2, pp. 1-210, 2021.

[10] L. P. Maddali, M. S. D. Thakur, R. Vigneswaran, M. A. Rajan, S. Kanchanapalli and B. Das, "VeriBlock: A Novel Blockchain Framework based on Verifiable Computing and Trusted Execution Environment," in *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, Bengaluru, India, 2020, pp. 1-6,.

[11] R. Wang, W. Ji, M. Liu, *et al.*, "Review on mining data from multiple data sources," *Pattern Recognition Letters*, vol. 109, pp. 120-128, 2018.

[12] M. Bica, L. Palen and C. Bopp, "Visual representations of disaster," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1262-1276.

[13] A. Vempala and D. Preoţiuc-Pietro, "Categorizing and inferring the relationship between the text and image of twitter posts," in *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2019, pp. 2830-2840.

[14] A. K. Gautam, L. Misra, A. Kumar, *et al.*, "Multimodal analysis of disaster tweets," in *2019 IEEE Fifth international conference on multimedia big data (BigMM)*, IEEE, 2019, pp. 94-103.

[15] J. Han, Y. Ma and Y. Han, "Demystifying swarm learning: A new paradigm of blockchain-based decentralized federated learning," *arXiv: 2201.05286*, 2022.

[16] Q. Shao, C. Jin, Z. Zhang, *et al.*, "Blockchain technology: architecture and progress," *Chinese Journal of Computers*, vol. 41, no. 5, pp. 969-988, 2018.

[17] H. Kang and J. Deng, "A review of blockchain data privacy protection research," *Journal of Shandong University (Science Edition)*, vol. 56, no. 5, pp. 92-110, 2021.

[18] Y. R. Pandeya, and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, pp. 2887-2905, 2021.

[19] D. Sharma, S. Purushotham and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, vol. 11, no. 1, pp. 19826, 2021.

[20] S. Merity, N. S. Keskar and R. Socher "Regularizing and optimizing LSTM language models," *arXiv:1708.02182*, 2017.

[21] J. Y. I. Alzamily, S. B. Ariffin and S. S. Abu-Naser, "Classification of Encrypted Images Using Deep Learning-Resnet50," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 21, pp. 6610-6620, 2022.