Computers and Electronics in Agriculture 226 (2024) 109398

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

VEG-MMKG: Multimodal knowledge graph construction for vegetables based on pre-trained model extraction

Bowen Lv [a,b,c,d], Huarui Wu [a,c,d], Wenbai Chen [b], Cheng Chen [a], Yisheng Miao [a,c,d,*], Chunjiang Zhao [a,*]

[a] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[b] College of Automation, Beijing Information Science and Technology University, Beijing 100192, China
[c] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[d] Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100097, China

ARTICLE INFO

Keywords:
Knowledge graph
Multimodal fusion
Image-text pairs
Pre-trained model

ABSTRACT

Knowledge graph technology is of great significance to modern agricultural information management and data-driven decision support. However, agricultural knowledge is rich in types, and agricultural knowledge graph databases built only based on text are not conducive to users' intuitive perception and comprehensive understanding of knowledge. In view of this, this paper proposes a solution to extract knowledge and construct an agricultural multimodal knowledge graph using a pre-trained language model. This paper takes two plants, cabbage and corn, as research objects. First, a text-image collaborative representation learning method with a two-stream structure is adopted to combine the image modal information of vegetables with the text modal information, and the correlation and complementarity between the two types of information are used to achieve entity alignment. In addition, in order to solve the problem of high similarity of vegetable entities in small categories, a cross-modal fine-grained contrastive learning method is introduced, and the problem of insufficient semantic association between modalities is solved by contrastive learning of vocabulary and small areas of images. Finally, a visual multimodal knowledge graph user interface is constructed using the results of image and text matching. Experimental results show that the image and text matching efficiency of the fine-tuned pre-trained model on the vegetable dataset is 76.7%, and appropriate images can be matched for text entities. The constructed visual multimodal knowledge graph database allows users to query and filter knowledge according to their needs, providing assistance for subsequent research on various applications in specific fields such as multimodal agricultural intelligent question and answer, crop pest and disease identification, and agricultural product recommendations.

1. Introduction

With the advancement of agricultural digitization, the concept of precision agriculture has been put forward, and how to efficiently utilize big data in the field of agricultural expertise has become an important issue. Combining knowledge graph technology with agriculture can visualize the complex data in the field of agriculture, which helps to carry out in-depth data mining of agricultural big data, and solve the problems of data dispersion, diversity, and poor utilization of data value in the field of agriculture. On May 17, 2012, (Singhal et al., 2012) proposed the concept of a knowledge graph and announced the construction of the next generation intelligent search engine based on knowledge graph. Traditional knowledge graphs store a large amount of text data in relational triplets, encoding it in the form of < head, relationship, tail > and presenting the knowledge system to users in the form of a network. Text knowledge graph has been widely applied in the field of agriculture, achieving functions such as agricultural fragmented knowledge integration and sharing. (Chen et al., 2019) constructed a knowledge graph of the entire agricultural field using natural language processing technology and automatically extracted entity relationships and linked them to the knowledge base. (Zhang et al., 2022) proposed a BERT based method for extracting crop disease information from China and constructed a domain knowledge graph. In addition to serving as a tool for storing data, knowledge graphs can also serve as knowledge

* Corresponding authors at: National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China.
E-mail addresses: miaoys007@nercita.org.cn (Y. Miao), zhaocj@nercita.org.cn (C. Zhao).

https://doi.org/10.1016/j.compag.2024.109398
Received 14 December 2023; Received in revised form 7 August 2024; Accepted 26 August 2024
Available online 6 September 2024

support for intelligent search, recommendation, Q&A and other application systems, playing an important role in applications and effectively improving the efficiency and quality of public knowledge acquisition (Ji et al., 2021). In the field of intelligent retrieval, (Wang et al., 2023) developed an agricultural knowledge graph containing 13,983 entities by extracting knowledge between the target data domain and the source data domain. They employed Bi-LSTM and CRF models as entity recognition models to achieve knowledge graph entity retrieval. (Chhetri et al., 2023)focused on cassava as the object of research and utilized knowledge graph combined with deep learning to obtain a low latency disease recognition method that mitigates the limitations of semantic technologies. In the realm of agricultural product recommendation, (Xie et al., 2022) put forth an agricultural product recommendation algorithm that merged attention factor decomposer with a knowledge graph. This innovative algorithm furnishes a fresh perspective and approach to pinpointing customized agricultural product recommendations for users of e-commerce platforms. Meanwhile, (Buche et al., 2019)introduced a primary core ontology to structure knowledge as causal relationships, construct expert knowledge bases, and offer technical measures for maintaining food quality recommendations. In terms of question and answer, (Li et al., 2023)created a knowledge graph utilizing a substantial volume of literature and professional knowledge regarding crop diseases and pests. Using this graph, they devised a rapid and precise automatic question answering algorithm and implemented it in a user-friendly prototype system utilizing the Pyqt5 structure. (Nizar et al., 2021)further developed an automated crop selection tool using crop value chain data, enabling global identification of suitable crops based on geographical location.

Previous studies on knowledge graphs did not involve information other than text. As an important way of information management, knowledge graphs have emerged in large quantities with the emergence of multimodal data such as images, videos, and audio. The use of a single data form often leads to the lack of information description and semantic association, and cannot meet the application needs in the context of smart agriculture(Yang et al., 2021). The multimodalization of knowledge graphs holds significant value in enhancing the interactive experience of these graphs and promoting information fusion (Peng et al., 2023) (Wang et al., 2022). Unfortunately, large-scale multimodal knowledge graphs have yet to appear in the agricultural domain due to the absence of a complete multimodal corpus. Compared with traditional knowledge graphs, the data involved in the agricultural vertical field includes spectroscopy, remote sensing, etc., which are more diverse than traditional fields. Data similarity is difficult to measure directly (Wang et al., 2016). At the same time, how to match the most representative image for the knowledge graph node from a large amount of information is a major difficulty in building a multimodal knowledge graph (Chen et al., 2020). In order to solve the problem of constructing an agricultural multimodal knowledge graph, the main contributions of this study are summarized as follows:

(1) A multimodal knowledge graph construction method is proposed. The single-modal model is used to extract knowledge to complete the text knowledge graph construction. The multimodal pre-training method proposed in the article is used for image-text matching. The fused data is completed to complete the vegetable multimodal knowledge graph.

(2) By fine-tuning the pre-training model, image and text knowledge features are extracted from the self-built data set to achieve text and visual retrieval; the complementary method of visual knowledge effectively overcomes the ambiguity problem existing in single modality and provides a new idea for cross-modal knowledge integration in agriculture.

(3) To address the issue of insufficiently fine-grained semantic associations between different modalities, we employed the RoBERTa-wwm network to extract entities from sentences and construct a textual knowledge graph. Additionally, to mitigate the problem of high similarity among small classes of entities in the agricultural vertical, we introduced fine-grained comparative learning to identify phrase and picture regions with the greatest similarity.

(4) This paper takes vegetables and field crops represented by cabbage and corn as research objects, and constructs a text-image pairing corpus in the agricultural vertical field. The corpus contains 10 categories, more than 16,000 entity triples, and more than 8,000 pairs of matching different agricultural entity images and texts. Compared with previous datasets, this dataset has richer entity categories and larger multimodal samples.

The structure of this article is as follows: Section 2 provides detailed information regarding the multimodal knowledge graphs, dataset construction methods, cross-modal model architecture, and metrics used in experiments. Section 3 compares the experimental results of different models and demonstrates the visualization of multimodal knowledge graphs achieved through fine-tuning the pre-trained model. Section 4 discusses the advantages and limitations of this research method compared to other multimodal knowledge graph construction methods. Finally, Section 5 summarizes the entire article and elaborates on future research directions.

2. Materials and methods

The construction steps of the multimodal vegetable knowledge graph in this article include four steps: acquiring multimodal knowledge, extracting text knowledge, learning image-text representation (Chen et al., 2023), and storing multimodal knowledge.

The construction of multimodal knowledge maps requires a large amount of corpus text and image sets. This article uses methods such as web crawlers and on-site collection of agricultural bases to obtain vegetable data. After data cleaning and preprocessing, the data is divided into training sets, testing sets, and validation sets for subsequent map construction. Text knowledge distillation involves using the pre-trained model RoBERTa for vegetable named entity recognition. The extracted entities and relationships are used to construct a text knowledge graph. The image-text representation learning process utilizes a multimodal pre-training model for image-text matching, which includes entity disambiguation, image-text entity matching evaluation, and optimal result selection (Mousselly et al., 2018). The former is used to handle the problem where a word or image in a multimodal knowledge graph contains multiple meanings, and multiple images or texts represent the same meaning. The latter two are used to semantically match each image with the corresponding text entity, helping to extract entities that match the optimal image set. The overall structure and technical roadmap for constructing a multimodal knowledge graph in the field of agricultural vegetables are shown in Fig. 1.

2.1. Data used

Flickr8k-CN dataset: the dataset contains 8000 images, each paired with five different text captions that provide content descriptions of objects and events in the image. The Chinese descriptions are obtained by manual annotation and cover a variety of topics and scenes, such as sports, food, landscapes and people. To validate the performance and accuracy of the model, this dataset will be used in this paper for graphic retrieval testing.

Self-built vegetable dataset: We collected image data of different growth stages of cabbage and corn from Xiao-tangshan National Precision Agriculture Demonstration Base and Beijing Breeding Base on-site, and employed network crawling to augment the dataset with related disease and pesticide data. During preprocessing, the original images were cropped to a standard format and resized to 224 x 224 pixels. Text data corresponding to the images was obtained from the China Agricultural Technology Promotion Information Service Platform using a lightweight crawler framework. The corpus was then classified through data cleaning, denoising, and redundant steps, in combination with domain expert knowledge. The text was subsequently vectorized to a

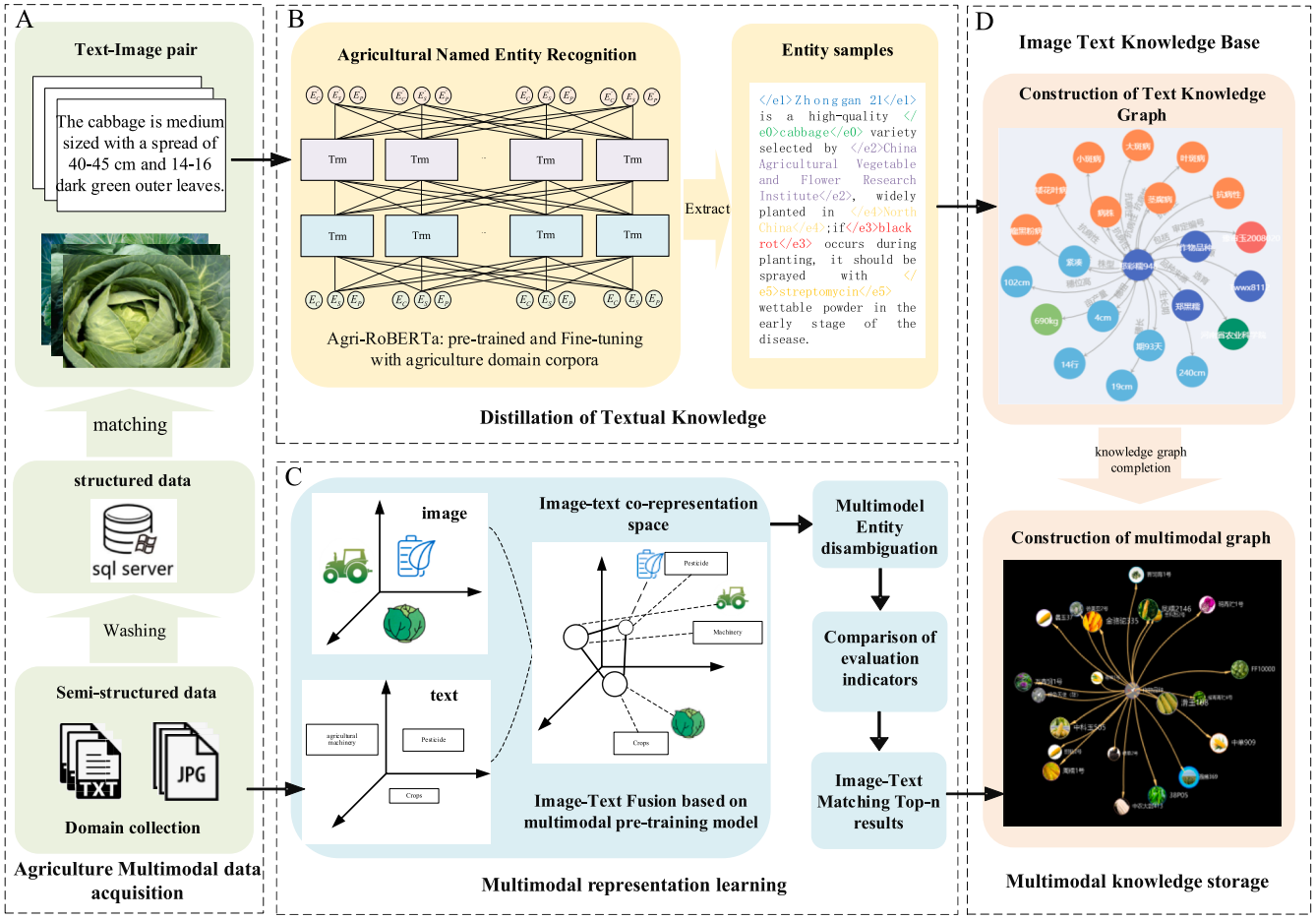


Fig. 1. The overall structure of constructing a multimodal knowledge graph in the field of agricultural vegetables. A dataset was created by cleaning semi-structured data and matching relevant information (A). This dataset was then used to fine-tune the Agri-RoBERTa model for text entity recognition(B). A multimodal representation learning model was employed to fuse image-text pairs(C). The constructed multimodal knowledge graph was integrated with the textual knowledge graph to enhance comprehension(D).

length of 128 characters. If the original text exceeded this length, it was truncated; if the length was insufficient, it was supplemented with [O].

The vegetable corpus encompassed 5295 matched image-text pairs, and subsequent extraction yielded 8644 entities related to cabbage and corn, encompassing 10 types of entities such as crop diseases and pests, pesticide names, agricultural machinery names, and variety names. The proportion of various entities in text and image is presented in Fig. 2. The obtained data was stored as structured triplets and image-text pair styles, and the image-text comparison is provided in Table 1.

2.2. Cross-modal architecture

Compared to the construction process of traditional knowledge graphs, the main difference of multimodal knowledge graphs lies in the representation learning steps of different types of knowledge, aiming to achieve cross modal named entity recognition and alignment matching between data entities. The Transformer structure, due to its excellent global dependency modeling ability, has become the core of multimodal pre-training models.

It has become very mature in the downstream application field of cross modal entity matching. Based on this, this article proposes the use

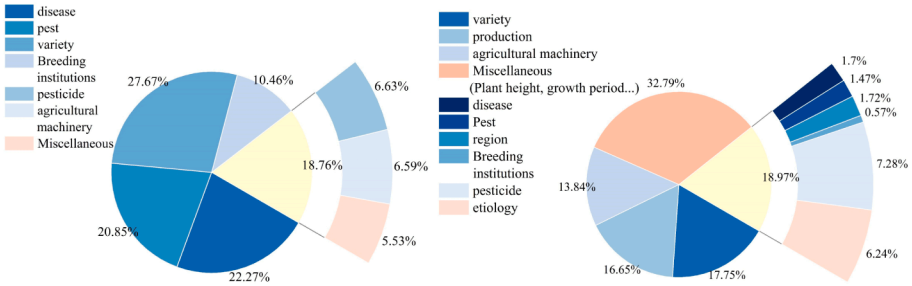
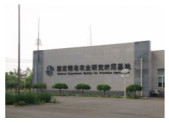


Fig. 2. Fig(a) shows the percentage of image-text pairs for each category and Fig (b) shows the percentage of extracted entities for each category.

Table 1
Examples of image-text pairs and entities.

Entity type	Text description	Image	Entity type	Text description	Image
Disease	Cabbage black rot Yellow lesions on leaf margins		Institutions	Xiaotangshan National Precision Agriculture Research Demonstration Base	
pest	Aphids are yellow green and densely distributed on cabbage		pesticide	Emamectin Benzoate is a new type of highly efficient semi-synthetic antibiotic insecticide	
variety	Cabbage variety Huifeng No.3		Agricultural machinery	Cabbage unmanned seedling throwing machine	

of image-text pre-training model fine-tuning ideas. This approach eliminates the need for huge computational power required for training multimodal models and large annotated data required for traditional semi-supervised training methods. The image and text are encoded separately using a dual stream Chinese pre-training model. This is followed by comparative learning of the encoding to achieve cross-modal fine-grained interaction. The corresponding relationship between the image and text is extracted in the vertical domain. The structure of the image-text pre-training model is shown in Fig. 3.

2.3. Image-text encoding

The vision module employs pre-trained ResNet50, ViT, and Swin-L models (Dosovitskiy et al., 2020) as visual coders to extract image features respectively. The ViT model, specifically, excels in processing images and has found widespread application in agricultural visual coding tasks (Picek et al., 2022) (Li et al., 2023). We opted for the Vision Transformer model because it has been pre-trained on large-scale datasets. This allows us to directly utilize its weight parameters, thus reducing the workload associated with training the visual coder in this paper and effectively preventing the occurrence of overfitting phenomenon. Given o images $l_{e,o}$ of the entity e, we rescale each image to unified $H \times W$ pixels, and the i-th input image $I_i \in R^{C\times H\times W}(0 \leqslant i \leqslant 1)$ is first reshaped into $u = HW/P^2$ flattened 2D patches, then pooled and projected as $X_{pc,i} \in R^{u\times d_v}$, where the resolution of the input image is $H \times W$,C is the number of channels and d_v denotes the dimension of hidden states of ViT. We concatenate the patched embeddings of o images to get the visual sequence patch embeddings $X_{pc} \in R^{m\times d_v}$,where $m = (u \times o)$. The ViT coding structure is shown in Fig. 4.

$$X_{v,o} = X_{pc} + X_{pos} \tag{1}$$

$$\overline{X}_{v,l} = M_A\left(L_N\left(X_{v,o}\right)\right) + X_{v,l-1}, l = 1\cdots L_v \tag{2}$$

$$X_{v,l} = F_N\left(L_N\left(\overline{X}_{v,l}\right)\right) + X_{v,l}, l = 1\cdots L_v \tag{3}$$

where $V_{pos} \in R^{m\times d_v}$ represents the corresponding position embedding layer, embedding, $X_{V,l}$ is the hidden states of the l layer of visual encoder. M_A denotes multi-head attention, and F_N is a fully connected feed-forward network.

The text module is encoded using different parametric quantities BERT and RoBERTa. RoBERTa (Liu et al., 2019) uses dynamic masking training as compared to BERT. While the text batch is larger and the sequences are longer, which achieves better results on datasets such as CLUE, RACE, etc. BERT has been applied in entity relationship extraction in the agricultural field (Chen et al., 2023).The structure vector representation consists of three parts: token embedding, segment

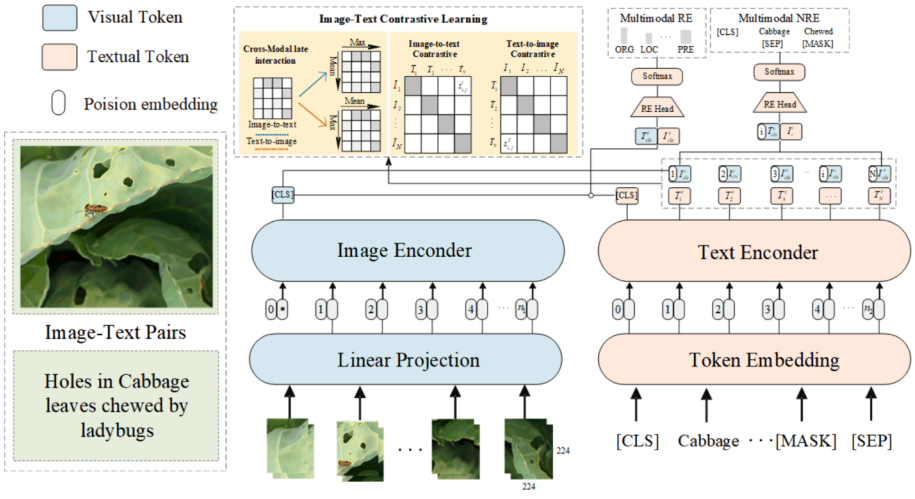


Fig. 3. General structure of the image-text pre-training model.

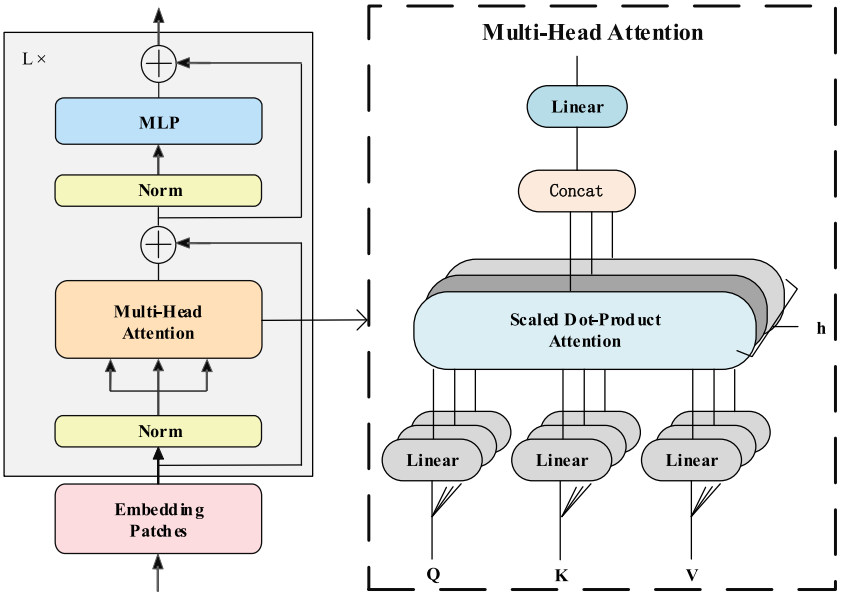


Fig. 4. ViT pre-training model structure.

embedding and position embedding. Each input sentence is pre-processed by inserting two special tokens: [CLS] and [SEP]. [CLS] denotes the start token in the pre-trained model, which indicates that processing of the input is initiated; and [SEP] denotes the delimiter for the input of the model. The structure is shown in Fig. 5.We leverage the first L_T layers of RoBERTa. as the text encoder, which also consists of L_T layers of M_A and F_N blocks similar to the visual encoder except that L_N comes after M_A and F_N. To be specific, a token sequence $\{w_1, \cdots, w_n\}$ is embedded to $X_{wd} \in R^{n \times d_T}$ with a word embedding matrix, and the textual representation is calculated as follows:

$$X_{T,0} = X_{wd} + T_{pos} \tag{4}$$

$$\overline{X}_{T,l} = L_N\left(M_A\left(X_{T,l-1}\right)\right) + X_{T,l-1}, l = 1 \cdots L_T \tag{5}$$

$$X_{T,l} = L_N\left(F_N\left(\overline{X}_{T,l}\right)\right) + X_{T,l}, l = 1 \cdots L_T \tag{6}$$

where $T_{pos} \in R^{(n) \times d_T}$ denotes position embedding, $X_{T,l}$ is the hidden states of the l layer for the output textual sequence.

2.4. Cross-modal fine-grained contrastive learning

If simple global feature encoding is done for each image and text, ignoring the fine-grained interactions of words and patches between the two modalities, the entity differentiation will be greatly reduced not to meet the needs of vegetable entity classification. To alleviate this problem while maintaining the training and inference efficiency of the dual-stream model, token-based cross-modal interactions are modeled. For image data I encoder denoted as f_θ, and for text data T encoder denoted as g_ϕ. Given an image $x^I \in I$ and a text $x^T \in T$,the encoded representations $f_\theta(x^I)$ and $g_\phi(x^T)$ are close if they are related and far apart if not, under a distance metric.

For the k-th visual token, we compute its similarities with all textual tokens of x_j^T, and use the largest one

$$\max_{0 \leqslant r \leqslant n_2} \left[f_\theta\left(x_i^I\right)\right]_k^{\mathrm{T}} \left[g_\phi\left(x_j^T\right)\right]_r \tag{7}$$

As its maximum token similarity with x_j^T. Then, we use the average maximum similarity of all unfilled tags in the similarity between the image and the text (text to image). Therefore, the similarity between the i-th image and the j-th text can be formulated as:

$$s_{i,j}^I\left(x_i^I, x_j^T\right) = \frac{1}{n_1} \sum_{k=1}^{n_1} \left[f_\theta\left(x_i^I\right)\right]_k^T \left[g_\phi\left(x_j^T\right)\right]_{m_k^I} \tag{8}$$

where $m_k^I = \operatorname{argmax}_{0 \leqslant r \leqslant n_2} \left[f_\theta\left(x_i^I\right)\right]_k^{\mathrm{T}} \left[g_\phi\left(x_j^T\right)\right]_r$. Similarly, the j-th text is similar to the i-th image as:

$$s_{i,j}^T\left(x_i^I, x_j^T\right) = \frac{1}{n_2} \sum_{k=1}^{n_2} \left[f_\theta\left(x_i^I\right)\right]_{m_k^T}^T \left[g_\phi\left(x_j^T\right)\right]_k \tag{9}$$

where $m_k^T = \operatorname{argmax}_{0 \leqslant r \leqslant n_1} \left[f_\theta\left(x_i^I\right)\right]_r^{\mathrm{T}} \left[g_\phi\left(x_j^T\right)\right]_k$, Equation $s_{i,j}^I\left(x_i^I, x_j^T\right)$ is unnecessarily equal to $s_{i,j}^T\left(x_i^I, x_j^T\right)$

The tokenized maximum similarity in the equation (13) means that for each image block, find its most similar text token. In each training batch, we sample α image text pairs $\{x_k^I, x_k^T\}_{k=1}^b$. For x_k^I in the image-text

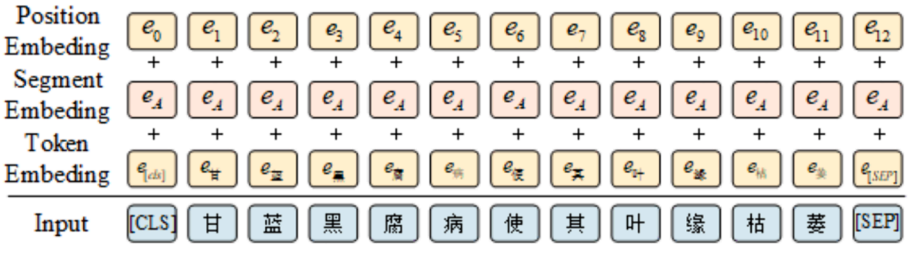


Fig. 5. Text pre-training model input representation.

pair $\{x_k^I, x_k^T\}_{k=1}^b$, x_k^T is its positive sample, while other texts are taken as negative samples within the same batch. The image text contrast loss of m can be formulated as:

$$L_k^I\left(x_k^I, \left\{x_j^T\right\}_{j=1}^{\alpha}\right) = -\frac{1}{\alpha}\log\frac{\exp\left(s_{k,k}^I\right)}{\sum_j \exp\left(s_{k,j}^I\right)} \tag{10}$$

$s_{k,j}^I$ represents the similarity between the k-th image and the j-th text. Similarly, the contrastive loss from text to image x_k^T is:

$$L_k^T\left(x_k^T, \left\{x_j^I\right\}_{j=1}^{\alpha}\right) = -\frac{1}{\alpha}\log\frac{\exp\left(s_{k,k}^T\right)}{\sum_j \exp\left(s_{j,k}^T\right)} \tag{11}$$

The total loss of this mini-batch can be represented by

$$L = \lambda L_k^I + (1-\lambda)L_k^T \tag{12}$$

λ is the balance coefficient of comparative loss. By applying it to (13) and (14) for comparative loss (10), the dual stream model learns fine-grained alignment between each vegetable entity image patch and text marker.

2.5. Evaluation indicators

For the image-text retrieval task, Recall@K and MR are commonly used as evaluation indices. Recall@K measures the proportion of queries where the correct results are ranked in the top K retrieval results. The values of K are typically set to 1, 5, and 10, and the range of values for Recall@K is between 0 and 1. A higher value indicates better performance. The formula is as follows:

$$\mathrm{Re}call@k = \frac{TP@k}{TP@k + FN@k} \tag{13}$$

$$MR = \sum_{i=k}^{n} \mathrm{Re}call@k/n \tag{14}$$

For multimodal named entity recognition (MNER) tasks and multimodal relationship extraction (MRE) tasks, in addition to recall rate, Precision and $F1$ value are generally used as evaluation indicators.

$$\mathrm{Pre}cision = \frac{TP}{TP + FP} \tag{15}$$

$$F1 = \frac{2{*}Precision{*}\mathrm{Re}call@}{Precision + \mathrm{Re}call@} \tag{16}$$

In the formula, True Positive(TP) denotes the number of entities correctly identified, False Positive (FP)denotes the number of misreported non-entities, and False Negative(FN) denotes the number of correct entities missed.

3. Results

3.1. Hardware and software settings

Algorithm debugging was performed using JupyterLab in AutoDL, with an environment of Python 3.9.13 execution, Pytorch Deep Learning Framework version 1.13.1, accelerated via GPU, CUDA, and cuDNN version 11.7. Multimodal mapping construction and fine-tuning experiments were carried out on a Windows 11 Professional Workstation Edition operating system, utilizing an Intel i9-13900KF processor, 72.00 GB RAM, and four NVIDIA RTX4090 graphics cards.

3.2. Text knowledge graph construction

Neo4j supports fast one-time import of a large number of nodes. At the same time, the database has excellent capabilities in processing entity nodes and edges in knowledge graphs. Therefore, this paper uses the text pre-training model to extract vegetable planting related knowledge and stores it in Neo4j text format, and constructs the corresponding knowledge graph. The "LOAD CSV" statement can help batch import CSV format files. This solution can modify the data in the database in real time. In view of the large amount of vegetable planting data used in the text and the complex relationships, the LOAD CSV data import method is adopted for the convenience of data management. First, put the CSV file in the import folder of the local graph, and use the Cypher language LOAD statement to import vegetable planting related entities, relationships and attributes. The steps for storing vegetable planting text knowledge graph data are divided into the following three steps: (1) First, store the 6 types of entity and relationship data mentioned in the text as CSV files and set them to UTF-8 format. Find the import folder in Neo4j and put the CSV file into the folder. (2) Store the entity description data in the form of attribute values in the CSV file and set it to UTF-8 format. The data includes information such as the temperature and humidity conditions of the disease onset. (3) Then create entities and attributes. For example, the statements for creating entity categories and attributes for vegetable varieties, diseases, and planting locations are shown in Table 2.

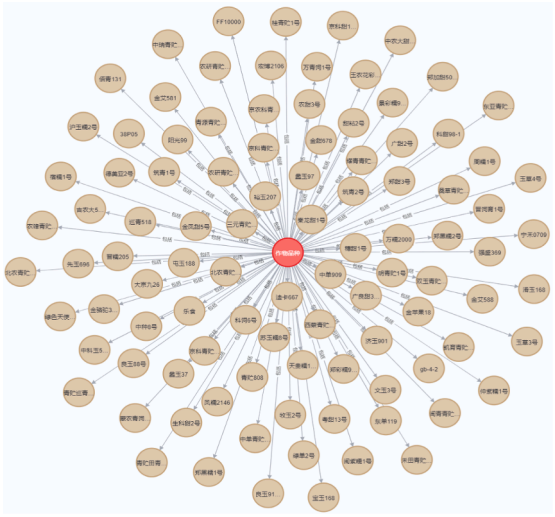
All entities, attributes and relationships are imported to generate a vegetable planting knowledge graph. The visualization results of the four types of entity relationships created in the above table are shown in Fig. 6. The red nodes in the figure represent the major categories of crop varieties, and the corresponding relationships are shown in (a) in the vegetables. The green nodes represent the vegetable disease nodes, and the relationships represent the disease resistance of different varieties of crops to diseases, as shown in (b). Figure (c) shows the R&D institutions corresponding to each variety of vegetables, and Figure (d) shows the provinces most suitable for the growth of different crops.

3.3. Zero-shot experiment

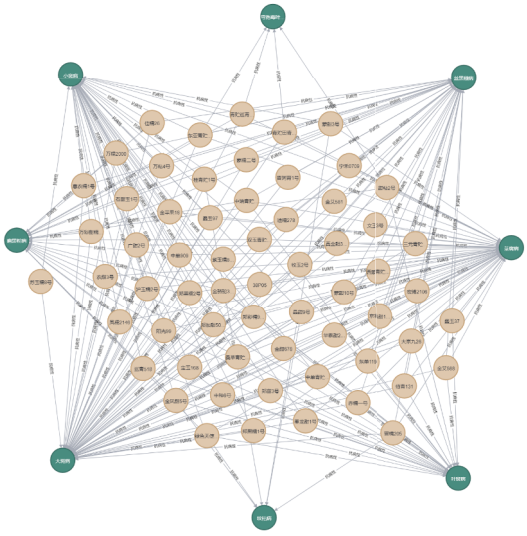
We tested the retrieval performance of the public dataset Flick8k-CN and the self-built dataset on both sides of image-text pairs in a zero-shot scenario. The pre-trained text dataset was entirely composed of Chinese text, and the pre-trained model employed the publicly available Chinese Clip and baseline models. The experiment results in Table 3 revealed that the baseline model achieved a mean rank (MR) of 55.83 % on the Flick8k-CN dataset, while the zero-shot retrieval average accuracy trained using large-scale ViT-L and RoBERTa-wwm joint encoding could reach 73.63 %, an overall improvement of 19.8 %. This indicated that using larger-scale training could effectively enhance the model's generalization ability. Moreover, the image-text retrieval ability R@1 of the pre-trained model employing BERT encoding was 32.5 %, which was approximately 25 % higher than the mean accuracy of the method utilizing the RoBERTa model. This demonstrated that RoBERTa long sequence Chinese encoding training possessed significant advantages in

Table 2
Example of batch import statement for vegetable planting entity.

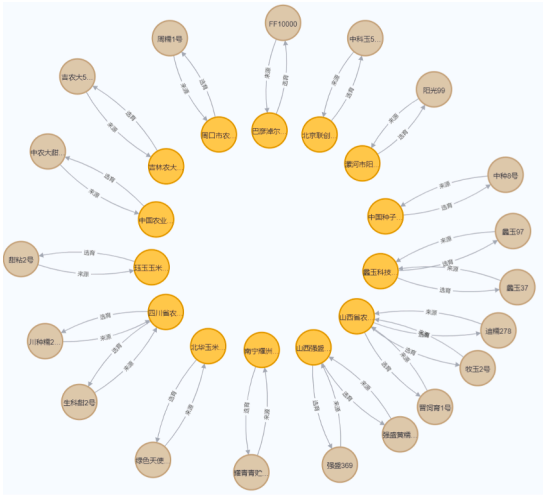
Function	Neo4j create nodes, relationship statements
Creation of vegetable varieties	LOAD CSV WITH HEADERS FROM "file://vegtable.csv" AS line
Vegetable entities and properties	CREATE(vegetable(title: line. title, authorized Number: line. Authorized Number, characteristics: line. characteristics, cultivation Technique: line. Cultivation Technique, opinions: line. opinions;)
Creation of Vegetable Disease Entities	LOAD CSV WITH HEADERS FROM "file: Il/disease.csv" AS line CREATE(dis /title: line. title)
Disease Entity Establishment	CREATE(disease (title: line. title, authorized Number: line. Authorized Number, characteristics: line. characteristics, cultivation Disease: line. Cultivation Pest, opinions: line. opinions;)
Creating location entities	LOAD CSV WITH HEADERS FROM "file: Il/locate. csv" AS line CREATE(locate /title: line. title)



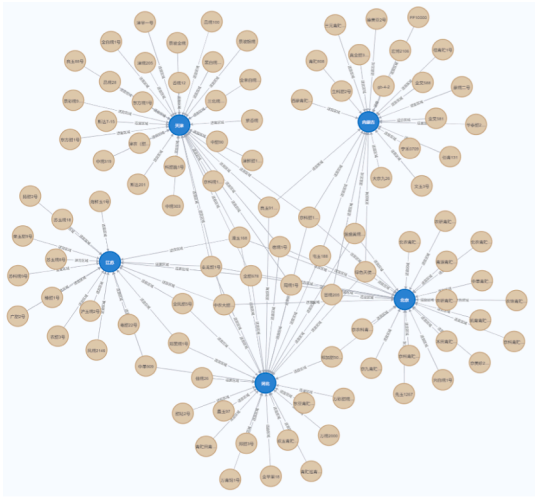
(a)Vegetable species subgraph



(b) Disease resistance of different vegetable species



(C) Breeding organizations corresponding to vegetable species



(d) Vegetable species and suitable growing areas

Fig. 6. Fig. (a) shows the result of text retrieval image recall for self-built dataset and Fig. (b) the result of text recall for image retrieval.

Table 3
Zero-shot retrieval experiment results on Flick8k-CN.

Dataset		Flick8k-CN						
Backbone		Text retrieval		R@10	Image retrieval		R@10	MR
Image encoding	Text encoding	R@1	R@5		R@1	R@5		
ResNet50	RBT3	32.5	58.4	72.7	35.7	58.6	77.1	55.83
ResNet101	BERT-wwm	34.4	49.2	73.1	36.1	55.9	76.4	54.18
Swim-S	RoBERTa-wwm	46.3	57.3	76.8	49.5	64.2	78.0	62.02
Swim-L	RoBERTa-wwm	55.5	67.4	78.1	55.5	68.8	78.4	67.28
ViT	BERT-wwm	43.2	58.5	74.2	39.3	63.6	75.3	59.02
ViT-B	RoBERTa-wwm-B	55.7	68.1	79.2	56.3	69.2	80.1	68.10
ViT-L	RoBERTa-wwm-L	57.9	76.2	84.0	59.9	78.6	85.2	73.63

retrieval.

The average optimal MR value of the zero-shot text retrieval experiments on the self-constructed dataset reached 44.56 %, and the optimal value of the zero-shot image retrieval was 46.87 %. This indicates that the pre-trained model has a certain degree of generalization ability in the agricultural vertical. However, the bilateral retrieval ability was lower than that in the Flick8k-CN dataset. This difference is partially attributed to the fact that the self-constructed dataset mainly consists of Chinese vegetable-related knowledge, which differs from the open-domain-oriented knowledge in the Flick8k-CN dataset. In contrast, pre-training on a large-scale Chinese dataset generates a model with higher similarity to the zero-shot content of the latter. Moreover, we compared the bilateral retrieval model's ability between larger pre-training scale models and found a difference of 15.24 % in retrieval ability, indicating that a larger data volume for pre-training generates a model with better generalization performance for image-text matching downstream tasks. The experimental results on the self-built dataset are shown in Fig. 7.

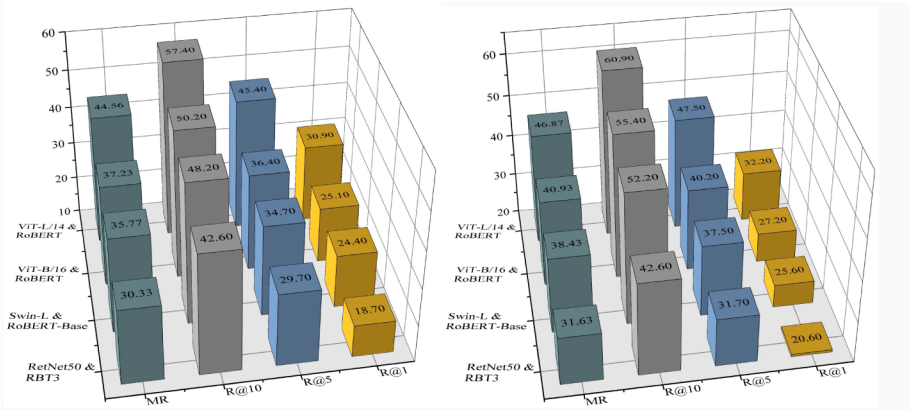


Fig. 7. Fig. (a) shows the result of text retrieval image recall for self-built dataset and Fig. (b) the result of text recall for image retrieval.

3.4. Fine-tuning experiment

In this paper, LoRA (Low-Rank Adaptation of Large Language Models) is used as an example for efficient model fine-tuning, and the parameters are divided into two groups by the get_optimizer_grouped_parameters function, one group using weight attenuation and other group as a control to prevent overfitting from occurring; the choice of the Fused Adam optimizer for optimization. The image resolution is 224*224, the sentence max_seq_length is set to 128, lora_r is set to 32, lora_dropout is set to 0.1, lora_alpha is set to 32, save_steps = 500, learning_rate is set to,and the rest of the parameters such as max_grad_norm, warmup_ratio, etc. are kept the same as the pre-trained model. The self-built dataset is the same as Flickr8k-CN, in which 70 % of the data is randomly used for fine-tuning and 30 % is used for image-text matching experiments.

Using the Flick8k-CN dataset, we conducted fine-tuning experiments on four different sizes of pre-trained models in Table 4.

The average optimal bilateral retrieval increased by 84.87 % compared to the pre-fine-tuning phase, representing an enhancement of 11.24 %. Notably, the ResNet50 and RT3 models, which had fewer pre-training parameters, exhibited the largest performance enhancement of 19.83 % compared to the pre-fine-tuning phase. Additionally, the performance gap between models with different parametric quantities on the Flick8k-CN dataset narrowed by 8.62 % during fine-tuning. These experimental results demonstrate that fine-tuning using a small sample data size can enhance the generalization ability of pre-trained models. Furthermore, fine-tuning can partially mitigate the performance gap between models with varying numbers of parameters, allowing smaller parameter models to perform better in the downstream retrieval task.

To demonstrate the excellent performance of the fine-tuned model on Flick8k-CN and self-built datasets, we compared and considered the previous SOTA models: In addition, we further considered another group of SOTA multimodal methods previously used for MNER and MRE: 1) MEGA (Zheng et al. 2021) proposed multimodal relationship extraction based on efficient graph alignment; 2) EEGA (Yuan et al., 2023) proposed multimodal entity relationship joint extraction of edge-enhanced graphs; 3) HVPNeT (Chen et al., 2022) proposed a new hierarchical visual prefix fusion network; 4) KnowPrompt (Chen et al., 2023) proposed a knowledge-aware prompt adjustment method with collaborative optimization. 5) TMR (Zheng et al. 2023) converts the text and its paired image problems into translation models for each other. The experimental results are shown in Table 5, which show that our model achieved the best results on the vegetable dataset after fine-tuning.

Fig. 8 presents the experimental results of models with varying parameter quantities on the self-built agricultural vegetable dataset. The large-scale ViT-L and RoBERTa-wwm models are applied to this dataset, and the image and text retrieval Mean Reciprocal Rank (MRR) reach 59.67 % and 76.7 % on average, respectively. These results suggest that the model demonstrates a certain degree of generalization ability after fine-tuning on a self-built small sample dataset. It effectively aligns fine-grained cross-modal entities in the vertical domain graph, eliminating the limitations of traditional supervised and semi-supervised methods that rely on a large amount of annotated data. Instead, it pairs unlabeled image-text entities with the input model to achieve semantic alignment. This approach provides a new solution for entity alignment in multimodal knowledge graphs.

3.5. Example of Text-Image retrieval experiment

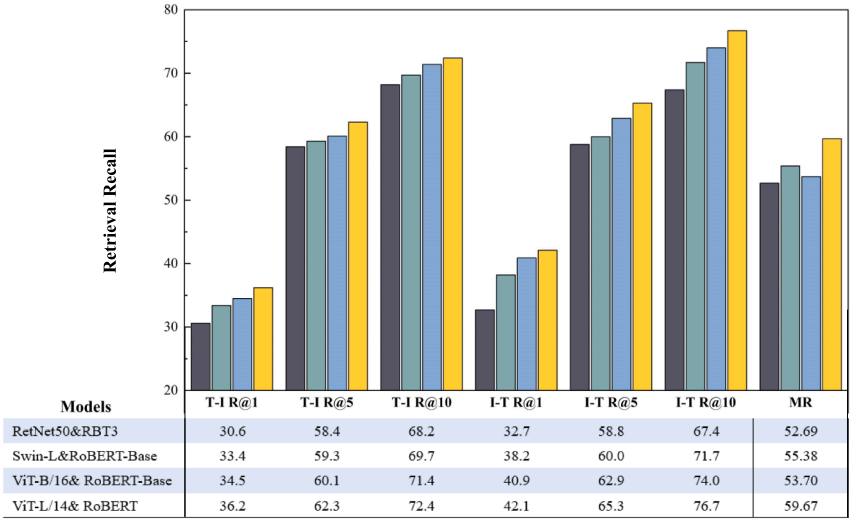
Fig. 9 shows the results of the Top5 image-text retrieval after fine-tuning in four ways on a self-constructed vegetable dataset with cabbage soft rot images, text as an example, and the number of pre-trained parameters for different models. The results of retrieval errors are shown in red dotted wireframes, the results of retrieval matches are shown in blue wireframes, and images without box markers indicate results that

Table 4
Retrieval effect of the fine-tuned model on Flick8k-CN.

Dataset		Flick8k-CN						
Backbone		Text retrieval		R@10	Image retrieval		R@10	MR
Image encoding	Text encoding	R@1	R@5		R@1	R@5		
ResNet50	RBT3	68.6	84.7	85.5	58.4	76.7	80.1	75.66
ResNet101	BERT-wwm	69.5	85.0	85.2	59.9	76.9	82.6	76.52
Swim-S	RoBERTa-wwm	72.4	84.8	86.0	63.8	75.4	80.5	77.15
Swim-L	RoBERTa-wwm-B	73.0	85.2	86.2	64.4	79.3	83.0	78.52
ViT	BERT-wwm	74.6	84.0	85.1	68.2	79.0	82.5	78.90
ViT-B	RoBERTa-wwm-B	75.8	85.3	87.2	70.7	80.0	83.1	80.35
ViT-L	RoBERTa-wwm-L	81.3	86.4	88.2	81.2	84.4	87.2	84.87

Table 5
Performance comparison of different competitive baseline approaches for NER and RE.

Models	Task	Flick8k-CN			self-built dataset		
		Precision	Recall	F1	Precision	Recall	F1
MEGA	MNER	72.1	78.3	78.5	71.7	63.1	68.2
	MRE	78.3	75.4	76.3	73.9	60.8	66.6
EEGA	MNER	74.3	76.2	75.2	75.5	65.6	77.9
	MRE	82.1	85.4	83.3	70.7	69.1	70.3
HVPNeT	MNER	87.4	84.5	86.3	71.2	70.4	74.3
	MRE	88.7	87.7	81.9	73.0	74.3	79.7
KnowPrompt	MNER	90.1	87.1	88.2	74.5	77.9	72.5
	MRE	**90.0**	**88.6**	**88.3**	78.4	71.1	74.4
TMR	MNER	**90.5**	**87.6**	**89.0**	83.4	71.7	81.8
	MRE	88.6	88.3	88.8	84.7	73.0	80.2
Ours	MNER	86.3	88.0	85.2	**88.1**	**88.3**	**88.3**
	MRE	88.2	87.4	88.0	**89.1**	**76.7**	**81.3**



Models	T-I R@1	T-I R@5	T-I R@10	I-T R@1	I-T R@5	I-T R@10	MR
RetNet50&RBT3	30.6	58.4	68.2	32.7	58.8	67.4	52.69
Swin-L&RoBERT-Base	33.4	59.3	69.7	38.2	60.0	71.7	55.38
ViT-B/16& RoBERT-Base	34.5	60.1	71.4	40.9	62.9	74.0	53.70
ViT-L/14& RoBERT	36.2	62.3	72.4	42.1	65.3	76.7	59.67

Fig. 8. Retrieval recall results after fine-tuning using self-built dataset.

are relevant to the query but lack ambiguity in vegetable disease targeting, i.e., the probability of being correct is lower than the blue boxed images. From the retrieval results can be found in the use of ResNet50-RBT3 skeleton retrieval results in the bilateral retrieval process are first returned to the wrong retrieval conclusions, return the correct conclusions of the number of samples is only one, compared to the pre-training model of different scales to ViT as a skeleton correct return value is too small, in the text generation results of the main body of the object of the cabbage will be wrongly considered as a kale or a flowering edible rape, the analysis of the reasons for this may be the use of vegetable The reason for analyzing this may be due to the high frequency of the word "kale" in the fine-tuning of the dataset, the strong similarity between the two images of soft rot disease, and the fact that the foci area in the early stage of soft rot disease accounts for a small area of the overall cabbage bulb, which makes there is an imbalance of information between the images and the text descriptions. The ViT-L/14, RoBERTa-wwm model trained using a larger data volume contains three correct objects in the Top five of the text retrieval process retrieval, and there is no wrong retrieval results, and the return of the two paragraphs of text in the text retrieval explicitly indicates that the appearance of soft-rot disease in cabbage is more accurately identified, and it is easy to observe that the method has the best experimental performance.

3.6. Multimodal knowledge graph construction

Currently, in the agricultural vertical, knowledge graph association research has not yet involved multiple modal types. In order to better serve agricultural users, the image-text entities after matching and retrieval are used to construct a multimodal knowledge graph for Chinese vegetables. The graph contains a total of 8644 corn and cabbage related entities and 16,535 triplets. Vegetable multimodal knowledge map contains ternary specific data statistics and entity relationships as shown in Table 6, and its corresponding knowledge graph is shown in Fig. 10.From it can be seen that the entities are divided into six categories, and each category of entities and the corresponding relationships are represented by different color distinctions. The entity representations include two forms of image and text, so that the user can intuitively access the knowledge. At the same time, in order to enhance the readability of the user, for the miscellaneous category of some entities to give labeling tips, such as detailed description of the North China region, including Beijing, Tianjin, etc.; for some of the nodes that do not match the appropriate picture data to supplement the description, such as supplementing the cabbage blight generally occurs in temperatures higher than 25 degrees Celsius, humidity is higher than 80 % of the environment to help the user for the prevention and control of the disease. This type of knowledge supplementation can enable users to understand the characteristics of vegetables more comprehensively.

To establish a large-scale agricultural database and visualize the multimodal knowledge graph of vegetables, we utilize the GraphXR framework to build a Chinese agricultural vegetable database. The multimodal knowledge graph of vegetables is visualized, and Cypherz serves as a query language to retrieve specific entity nodes based on user needs. Additionally, Corn and Cabbage text entities and relationships are integrated into the Neo4j database to construct a traditional

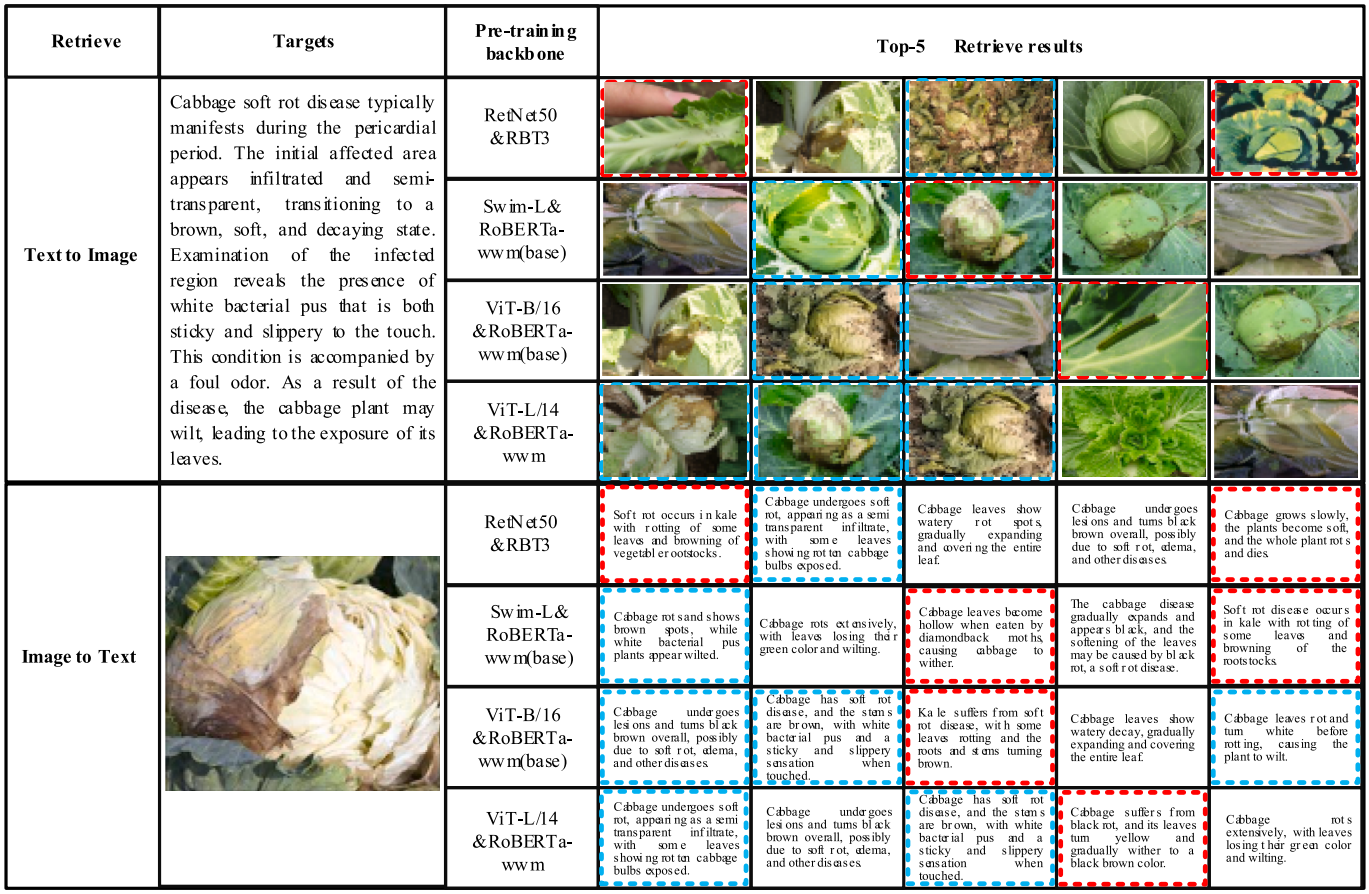


Fig. 9. Bi-directional retrieval instance after fine-tuning self built dataset.(The pre-trained model generates text in Chinese and then translates it into English.).

knowledge graph, which is then extended to achieve the multimodal knowledge graph. The overall database includes vegetable pests and diseases, institution and growth information. The multimodal vegetable knowledge graph's outline is shown in Fig. 11, with nodes containing vegetable pests and diseases, breeding units, and suitable growth information. Different types of relational information are labeled using different colors.

Entity retrieval and filtering are based on the above multimodal knowledge graph of vegetable information. When end users query the knowledge graph, relevant nodes and edges, i.e., entities and relationships, will be recognized and presented, and subgraphs will be generated instantly. Taking the "crop variety" node as an example for searching, the result is shown in Fig. 11(b). The subgraph displays 20 varieties of Corn and Cabbage. Clicking on a specific variety provides its related information, which is succinctly presented in the interface. This allows users to gain a comprehensive understanding of the selected crop variety. Users can reason and make decisions based on multiple pieces of information.

4. Discussion

Currently, pre-trained models are widely used in downstream tasks. Compared to traditional image-text retrieval methods, pre-trained image-text models are experimented on larger scale data with better generalization performance and zero-sample retrieval capability. The results of the graphic matching experiments on the vegetable dataset in this paper validate the effectiveness of pre-trained model migration on agriculture similar to the conclusions of (Dong et al., 2023). The paper presents a solution to address the challenges in related research on multimodal knowledge graphs, such as diverse definitions and poor visualization (Zhu et al., 2022) (Ektefaie et al., 2023). It provides a clear hierarchical structure and relationships for the construction of a vegetable knowledge base. Previously, there had been no medium-sized or larger multimodal knowledge graph database in the agricultural field. (Zhou et al.,2021)used the tomato and cucumber multimodal knowledge graph as a guide for image characterization and disease recognition. However, the construction of knowledge graph relationships mainly relied on expert guidance and manual annotation, which presented problems such as large workload and small graph size. The unsupervised pre-training model used in this paper effectively overcame the aforementioned problems. Some representative multimodal knowledge graphs have been proposed and applied to downstream tasks in the open domain due to the low similarity between entities and relations (Sun et al., 2020). For example, (Liu et al., 2021) proposed MMKG for the first time in 2018 and constructed two knowledge graphs, DBpedia15K and YAGO15K datasets. Richpedia (Wang et al., 2020) aimed to add a large number of pictures to Wikipedia by using textual entities, introduce visual descriptors, calculate the similarity between pictures by synthesizing the distance between different descriptors, and construct a multimodal knowledge graph. However, this approach ignores the influence of textual knowledge and results in missing descriptions.

Although the multimodal knowledge graph constructed in this paper can effectively connect vegetable knowledge together and provide users with a multimodal domain visualization knowledge base for easy retrieval, there are still several problems to be solved:

(1) The multimodal knowledge graph for vegetables constructed in this paper includes only two types of information: image and text. (Wang et al., 2023) proposed a knowledge graph containing four types of information. Unlike the open domain, obtaining highly relevant multi-type information in the agricultural domain is

Table 6
Triple data statistics and entity relationships.

statement	Number of triplets	Example of a triplet		
		Head entity	relation	Tail entity
[Qiugan No.7]is a variety of [cabbage]	3866	Cabbage	variety	Qiugan No.7
[Qiugan No.7] is highly resistant to [virus disease], [blight disease], and [black rot]	2569	Qiugan No.7	disease	virus disease
[virus disease] can be treated with [morpholine] guanidine	2646	virus disease	pesticide	morpholine
[Qiugan No.7] is bred by [National Engineering Research Center For Vegetables]	440	Qiugan No.7	institution	National Engineering Research Center For Vegetables
[Qiugan No. 7] is widely planted in [North China]	3713	Qiugan No.7	region	North China
[Cabbage seedlings] are planted with [unmanned seedling transplanters]	640	Cabbage seedling	machinery	unmanned seedling throwing machine
[Qiugan No.7] takes about [70 days] from planting to harvesting	2861	Qiugan No.7	miscellaneous	70 days

costly, and there is a lack of fusion algorithms for more than three types of information.

(2) The arithmetic cost of experimenting with multimodal pretrained models is prohibitively high, and there is a huge difference between the computational power that a single computing device can provide and the total computation required for a model with a slightly larger pretrained sample size(Zhao et al.,2023). The NVIDIA H100 SXM has a single-card FP16 computation power of only 2,000 TFLOPs, whereas the Visual-Glm requires about 100 ZFLOPs, a difference of eight The difference is 8 orders of magnitude. Even if only model fine-tuning is performed, it can only be experimented on small-scale multimodal models, resulting in low efficiency of image-text pair matching in the vegetable domain.

(3) In actual working conditions, knowledge related to vegetables evolves continuously over time, and vegetable growth knowledge is closely associated with time. However, the current model does not account for newly introduced entities and relationships during the construction of the multimodal knowledge graph, which poses some limitations. Therefore, in the future, we can learn from (Chen et al., 2023) to establish a time-consistent multimodal knowledge graph in the agricultural domain.

5. Conclusions

Text data alone cannot satisfy all the effective information needed for agricultural wisdom services, this paper proposes a novel method for building a vegetable database, fusing image and text information through a pre-training model to represent the fusion of image and text information, constructing a multimodal knowledge graph for vegetables through the aligned and matched knowledge, and realizing the query and filtering of information in the knowledge base according to the users' needs. Using the pre-trained model to learn and fine-tune fine-grained comparisons between image regions and entity vocabularies, a 76.7 % recall is achieved on a self-built dataset of corn and cabbage, and

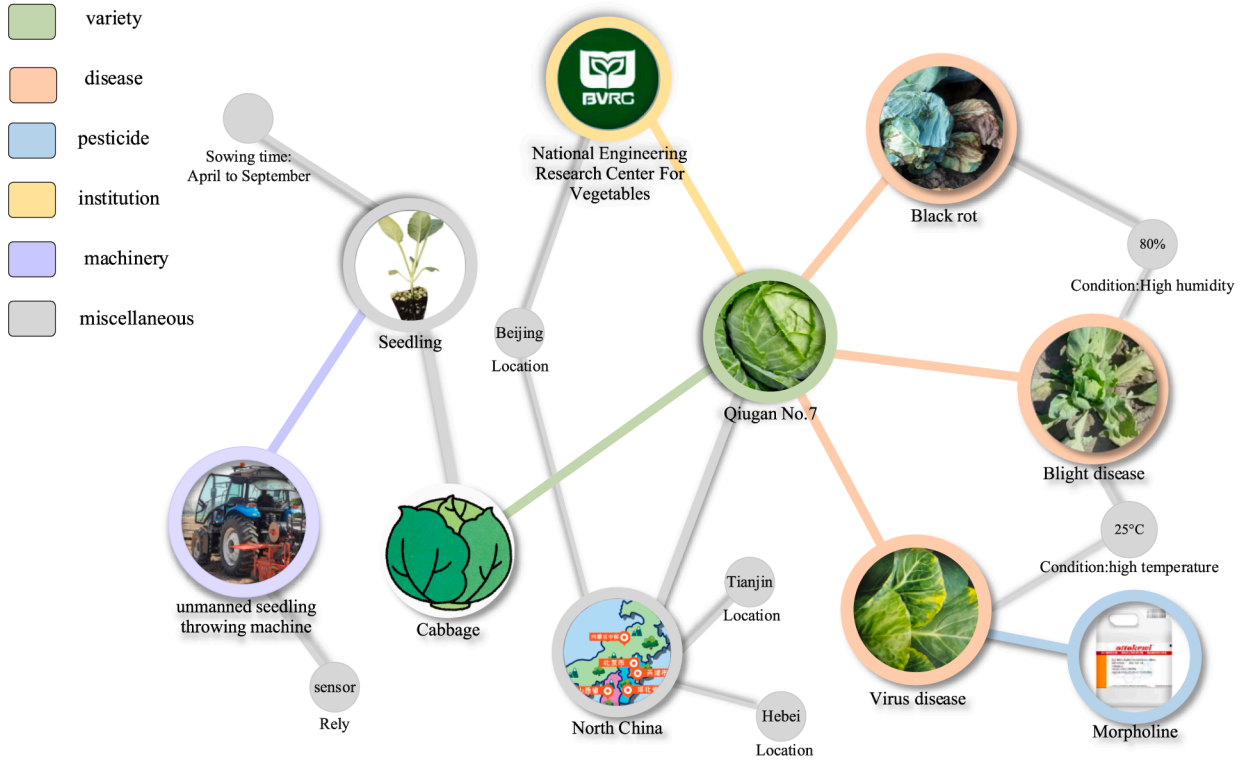
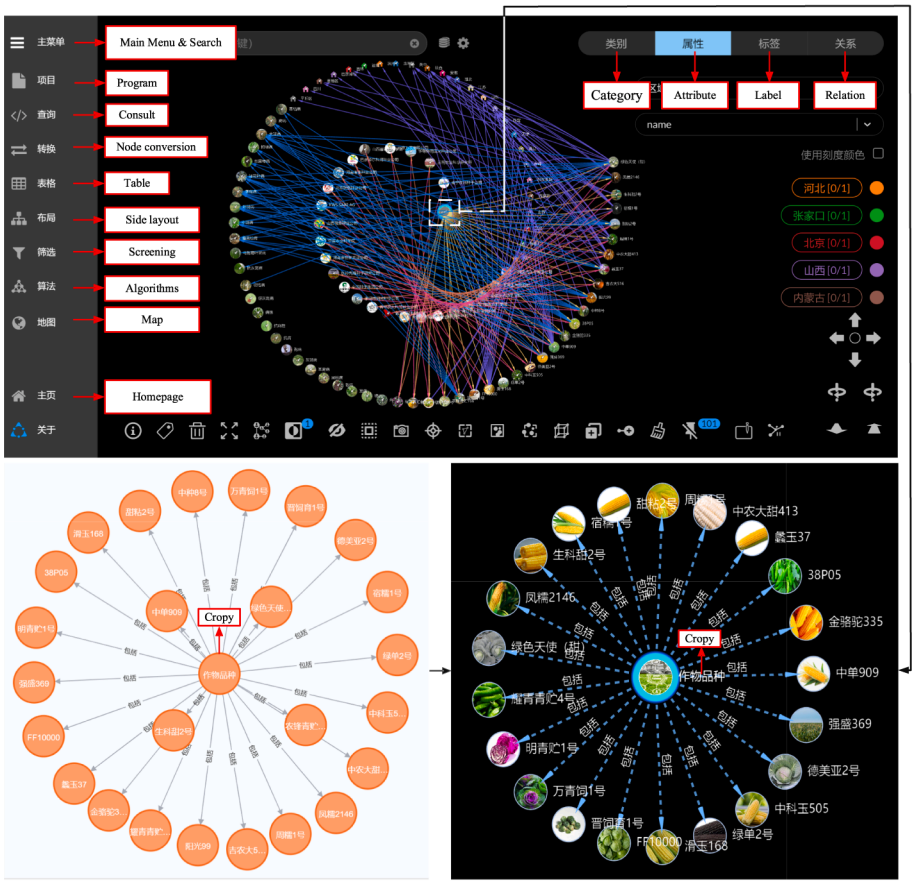


Fig. 10. Information on the subgraphs of the knowledge graph corresponding to the relationships in Table 4.



(a) (b)

Fig. 11. The upper part of the image is an extract of the multimodal knowledge graph and its stored multimodal Chinese agricultural database; The following section is the text knowledge graph (a). The results of querying the "crop varieties" section in the database are used to complete the multimodal knowledge graph (b).

accurate image matching is realized for vegetable-related text entities. In this work, the Chinese large-scale pre-trained model achieves impactful performance for image and text matching, discarding the shortcomings of traditional semi-supervised methods that rely on a large amount of annotated data, and providing an effective and easy method for multimodal knowledge graph construction in the Chinese domain. In future research, the multimodal knowledge graph will continue to expand the number of entities and relationships and provide effective help for various applications such as vegetable-related multimodal intelligent Q&A, crop pest and disease identification, and product recommendation.

CRediT authorship contribution statement

Bowen Lv: Writing – original draft, Validation, Data curation. **Huarui Wu:** Writing – review & editing, Resources, Funding acquisition. **Wenbai Chen:** Writing – review & editing, Supervision. **Cheng Chen:** Funding acquisition. **Yisheng Miao:** Writing – review & editing, Project administration, Investigation. **Chunjiang Zhao:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by National Key Research and Development Program of China under Grant 2022YFD1600602 , in part by Supported by China Agriculture Research System of MOF and MARA Grant CARS-23-D07, in part by Central Guiding Local Science and Technology Development Fund Projects under Grant 2023ZY1-CGZY-01

References

Buche, P., Cuq, B., Fortin, J., Sipieter, C., 2019. Expertise-based decision support for managing food quality in agri-food companies. Comput. Electron. Agric 163, 104843. https://doi.org/10.1016/j.compag.2019.05.052.

Chen, Y., Kuang, J., Cheng, D., Zheng, J., Gao, M., Zhou, A. ,2019. AgriKG: an agricultural knowledge graph and its applications. In Database Systems for Advanced Applications: DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22–25, 2019, Proceedings 24, pp. 533-537. Springer International Publishing. doi: 10.1007/978-3-030-18590-9_81.

Chen, L., Li, Z., Wang, Y., Xu, T., Wang, Z., Chen, E.,2020. MMEA: entity alignment for multi-modal knowledge graph. In Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020, Hangzhou, China, August 28–30, 2020, Proceedings, Part I 13. pp.134-147. Springer International Publishing. doi: 10.1007/978-3-030-55130-8_12.

Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., Chen, H. ,2022. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. arXiv preprint. doi: 10.18653/v1/2022.findings-naacl.121.

Chen, T., Qian, Y., Wang, Y., Chen, X., Ouyang, D., Dong, S., Huang, L.,2023. RoBERT-Agri: An Entity Relationship Extraction Model of Massive Agricultural Text Based on

the RoBERTa and CRF Algorithm. In 2023 IEEE 8th International Conference on Big Data Analytics (ICBDA).pp. 113-120. doi: 10.1109/icbda57405.2023.10105090.
Chen, Y., Ge, X., Yang, S., Hu, L., Li, J., Zhang, J. ,2023. A Survey on Multimodal Knowledge Graphs: Construction, Completion and Applications.Mathematics,11,8, 1815.doi: 10.3390/math11081815.
Chen, X., Zhang, J., Wang, X., Wu, T., Deng, S., Wang, Y., Zhang, N.,2023. Continual Multimodal Knowledge Graph Construction. doi: 10.24963/ijcai.2024/688.
Chhetri, T.R., Hohenegger, A., Fensel, A., Kasali, M.A., Adekunle, A.A., 2023. Towards improving prediction accuracy and user-level explain ability using deep learning and knowledge graphs: A study on cassava disease. Expert Syst. Appl. 233, 120955 https://doi.org/10.1016/j.eswa.2023.120955.
Dong, X., Wang, Q., Huang, Q., Ge, Q., Zhao, K., Wu, X., Hao, G., 2023. PDDD-pre-train: A series of commonly used pre-trained models support image-based plant disease diagnosis. Plant Phenomics 5, 0054. https://doi.org/10.34133/plantphenomics.0054.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. t https://arxiv.org/abs/2010.11929.
Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., Zitnik, M., 2023. Multimodal learning with graphs. Nat. Mach. Intell. 5 (4), 340–350. https://doi.org/10.1038/s42256-023-00624-6.
Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y., 2021. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Trans. Neural Netw. Learn. Syst. 33 (2), 494–514. https://doi.org/10.1109/tnnls.2021.3070843.
Li, X., Yu, M., Xu, D., Zhao, S., Tan, H., Liu, X., 2023. Non-contact measurement of pregnant Sows' backfat thickness based on a hybrid CNN-ViT Model. agriculture 137, 1395. https://doi.org/10.3390/agriculture13071395.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.Stoyanov, V.,2019. RoBERTa: a robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692.
Mousselly-Sergieh, H., Botschen, T., Gurevych, I., Roth, S. 2018, June. A multimodal translation-based approach for knowledge graph representation learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 225-234. doi: 10.18653/v1/s18-2027.
Nizar, N.M.M., Jahanshiri, E., Tharmandram, A.S., Salama, A., Sinin, S.S.M., Abdullah, N.J., Azam-Ali, S.N., 2021. Underutilised crops database for supporting agricultural diversification. Comput. Electron. Agric 180, 105920. https://doi.org/10.1016/j.compag.2020.105920.
Peng, C., Xia, F., Naseriparsa, M., Osborne, F., 2023. Knowledge graphs: Opportunities and challenges. Artif. Intell. Rev. 1–32 https://doi.org/10.1007/s10462-023-10465-9.
Picek, L., Šulc, M., Patel, Y., Matas, J. ,2022. Plant recognition by AI: Deep neural nets, transformers, and kNN in deep embeddings. Frontiers in Plant Science, 2788. doi: 10.3389/fpls.2022.787527.
Singhal, A. ,2012. Introducing the knowledge graph: things, not strings. Official google blog,5,16, 3. https://blog.google/products/search/introducing-knowledge-graph-things-not/.
Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., Zheng, K. ,2020, October. Multi-modal knowledge graphs for recommender systems. In Proceedings of the 29th ACM international conference on information & knowledge management. pp. 1405-1414. doi: 10.1145/3340531.3411947.
Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L. ,2016. A comprehensive survey on cross-modal retrieval. arXiv preprint https://arxiv.org/abs/1607.06215.
Wang, X., Meng, B., Chen, H., Meng, Y., Lv, K., Zhu, W. ,2023, October. TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and Audio. In Proceedings of the 31st ACM International Conference on Multimedia. pp. 2391-2399. doi: 10.1145/3581783.3612266.
Wang, L., Jiang, J., Song, J., Liu, J., 2023. A weakly-supervised method for named entity recognition of agricultural knowledge graph. Intell. Automation Soft Comput. 37, 1. https://doi.org/10.32604/iasc.2023.036402.
Wang, M., Wang, H., Qi, G., Zheng, Q., 2020. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. Big Data Res. 22, 100159 https://doi.org/10.1016/j.bdr.2020.100159.
Wang, E., Yu, Q., Chen, Y., Slamu, W., Luo, X., 2022. Multi-modal knowledge graphs representation learning via multi-headed self-attention. Information Fusion 88, 78–85. https://doi.org/10.1016/j.inffus.2022.07.008.
Xie, H., Yang, J., Huang, C., Wang, Z., Liu, Y. ,2022. Recommendation algorithm for agricultural products based on attention factor decomposer and knowledge graph. In 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML) pp. 626-631. IEEE .doi: 10.1109/cacml55074.2022.00110.
Yang, X., Shu, L., Chen, J., Ferrag, M.A., Wu, J., Nurellari, E., Huang, K., 2021. A survey on smart agriculture: development modes, technologies, and security and privacy challenges. IEEE/CAAJ. Autom. Sin. 8 2, 273-302. https://doi. org/10.1109/JAS.2020.1003536.
Yuan, L., Cai, Y., Wang, J., Li, Q. ,2023, June. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In Proceedings of the AAAI conference on artificial intelligence Vol. 37, No. 9, pp. 11051-11059. doi: 10.1609/aaai.v37i9.26309.
Zhang, W., Wang, C., Wu, H., Zhao, C., Teng, G., Huang, S., Liu, Z., 2022. Research on the Chinese named-entity–relation-extraction method for crop diseases based on BERT. Agronomy 12 (9), 2130. https://doi.org/10.3390/agronomy12092130.
Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y, Wen, J. R. ,2023. A survey of large language models. arXiv preprint https://arxiv.org/abs/2303.18223.
Zheng, C., Feng, J., Fu, Z., Cai, Y., Li, Q. Wang, T. 2021. Multimodal relation extraction with efficient graph alignment. In Proceedings of the 29th ACM International Conference on Multimedia. pp. 5298-5306. doi: 10.1145/3474085.3476968.
Zheng, C., Feng, J., Cai, Y., Wei, X., Li, Q.,2023. Rethinking Multimodal Entity and Relation Extraction from a Translation Point of View. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume1: Long Papers. pp. 6810-6824. doi: 10.18653/v1/2023.acl-long.376.
Zhou, J., Li, J., Wang, C., Wu, H., Zhao, C., Teng, G., 2021. Crop disease identification and interpretation method based on multimodal deep learning. Comput. Electron. Agric 189, 106408. https://doi.org/10.1016/j.compag.2021.106408.
Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., Yuan, N.J., 2022. Multi-modal knowledge graph construction and application: A survey. IEEE Trans. Knowl. Data Eng. https://doi.org/10.1109/tkde.2022.3224228.