Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.035257
Article

Tech Science Press

DCRL-KG: Distributed Multi-Modal Knowledge Graph Retrieval Platform Based on Collaborative Representation Learning

Leilei Li[1], Yansheng Fu[2], Dongjie Zhu[2,*], Xiaofang Li[3], Yundong Sun[2], Jianrui Ding[2], Mingrui Wu[2], Ning Cao[4,*] and Russell Higgs[5]

[1]Artificial Intelligence Academy, Wuxi Vocational College of Science and Technology, Wuxi, 214068, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, 204209, China
[3]Department of Mathematics, Harbin Institute of Technology, Weihai, 264209, China
[4]College of Information Engineering, Shandong Vocational and Technical University of International Studies, Rizhao, 276826, China
[5]School of Mathematics and Statistics, University College Dublin, Dublin, D04 V1W8, Ireland
*Corresponding Authors: Dongjie Zhu. Email: zhudongjie@hit.edu.cn; Ning Cao. Email: ning.cao2008@hotmail.com
Received: 14 August 2022; Accepted: 04 November 2022

Abstract: The knowledge graph with relational abundant information has been widely used as the basic data support for the retrieval platforms. Image and text descriptions added to the knowledge graph enrich the node information, which accounts for the advantage of the multi-modal knowledge graph. In the field of cross-modal retrieval platforms, multi-modal knowledge graphs can help to improve retrieval accuracy and efficiency because of the abundant relational information provided by knowledge graphs. The representation learning method is significant to the application of multi-modal knowledge graphs. This paper proposes a distributed collaborative vector retrieval platform (DCRL-KG) using the multi-modal knowledge graph VisualSem as the foundation to achieve efficient and high-precision multimodal data retrieval. Firstly, use distributed technology to classify and store the data in the knowledge graph to improve retrieval efficiency. Secondly, this paper uses BabelNet to expand the knowledge graph through multiple filtering processes and increase the diversification of information. Finally, this paper builds a variety of retrieval models to achieve the fusion of retrieval results through linear combination methods to achieve high-precision language retrieval and image retrieval. The paper uses sentence retrieval and image retrieval experiments to prove that the platform can optimize the storage structure of the multi-modal knowledge graph and have good performance in multi-modal space.

Keywords: Multi-modal retrieval; distributed storage; knowledge graph

1 Introduction

One of the main directions of artificial intelligence research and development is how efficiently express and expand human knowledge. Knowledge bases are often used as the basis for the realization of artificial intelligence (AI) tasks such as natural language understanding and natural language generation. In recent years, the knowledge graph technology of structured representation of knowledge has received extensive

attention in the field. The knowledge graph uses the structure of triples to describe knowledge facts, and each triple is composed of entities and relationships. The structured nature of the knowledge graph makes the knowledge graph perform well in the directions of knowledge representation, knowledge embedding, and knowledge expansion. Knowledge graph technology has been widely used in artificial intelligence applications such as language representation learning and intelligent language question and answer. To make the knowledge graph have more complete knowledge content and make the knowledge graph more accurate and more efficient in expressing learning embedding, more and more researchers are not limited to structured text knowledge content, but are committed to constructing a multi-modal knowledge graph and bringing high-quality external information into the knowledge graph. This is a hot and cutting-edge research direction for multi-modal knowledge graphs.

In 2017, Xie proposed the Image-embodied knowledge representation learning (IKRL) method [1], which embeds the image information of entities based on the attention mechanism [2], and constructed a multi-modal knowledge graph representation learning method. IKRL collects a collection of image information corresponding to entities in the text knowledge graph, uses convolutional neural networks to process the image information and obtains a collection of image vector representations, and calculates the similarity between the image vector and the corresponding entity embedding vector in the entity space. The attention value of different pictures in the entity picture set is merged with the picture vector corresponding to the entity based on the attention mechanism to obtain the embedding vector of the entity corresponding picture set in the entity space. The image vector and the text vector are used together to form the loss function of the training IKRL model based on the Translating embeddings for modeling multi-relational data (TransE) [3] method in the knowledge graph expression learning method, to complete the task of multi-modal representation learning. IKRL initially constructed a multi-modal knowledge graph, but it still has limitations with very limited coverage. Most of the other multi-modal knowledge graphs that contain pictures built after this also have problems such as narrow knowledge domains or a lot of noise in pictures.

VisualSem [4], proposed in 2021, constructs a high-quality multi-modal knowledge graph composed of image information and language information. VisualSem is also a multi-language-oriented knowledge graph, and the nodes in the graph have different explanations corresponding to different languages. VisualSem contains about 90k nodes, about 1.3M explanations, and about 938k pictures. The prominent feature of VisualSem is that the number of nodes is large and scalable. A survey on the construction and application of multimodal knowledge graphs [5] compared VisualSem with other multimodal knowledge graphs and highlighted the feature of VisualSem. A multimodal knowledge graph representation method in 2022 [6] uses VisualSem as the experimental dataset and conducts extensive comparative experiments. ImageNet [7] is the source of image information in the knowledge graph. The knowledge graph is connected with other knowledge bases such as Wikipedia to make the nodes diversified, and the pictures in the knowledge graph are filtered through multiple filters to solve the challenge caused by image noise. All in all, VisualSem is a high-quality multi-modal intelligent map with great application scenarios and development significance. Compared with other multi-modal knowledge maps such as FB15-IMG [8] and MMKG [9], VisulSem has the advantages of high-quality and diversified data. Therefore, this paper builds a distributed vector retrieval platform for multi-modal multi-channel collaborative representation learning based on VisualSem, aiming to realize reasonable knowledge graph storage, high-quality knowledge graph expansion, and high-accuracy knowledge graph vector entity retrieval. The multimodal retrieval platform based on the multimodal knowledge graph proposed in this paper has the following application scenarios: when the user tries to identify the content in the image or obtain the description information of an entity, the platform can efficiently and accurately identify the images and sentences and returns the entity's rich association description information. In Section 2, this paper mainly describes key technology used in DCRL-KG. In Section 3, this paper describes the design of the platform, including

distributed storage module, knowledge graph extension module and collaborative retrieval module. In Section 4, this paper uses VisualSem as the dataset to check the accuracy of sentence retrieval and image retrieval function. In Section 5, this paper concludes the whole work and points out the future development direction. In summary, this platform mainly includes the following:

(a) Use distributed storage technology to rationally organize the storage structure, improve the retrieval speed of the knowledge map, and increase the scalability of the knowledge graph at the storage space level.
(b) When performing multi-modal knowledge graph expansion, use BabelNet [10] to multiple filters the expanded images, restrict and constrain the added nodes, and then ensure the high quality of the knowledge graph in the process of expanding the knowledge graph.
(c) The core function of this platform is to efficiently and accurately realize sentence retrieval and image retrieval of multi-modal knowledge graphs. Sentence Bert (SBERT) model [11] and CLIP model [12] are used to extract the feature vector of sentence data and image data while model training. In the process of vector retrieval, based on different retrieval principles and methods corresponding to different retrieval models, the k-nearest neighbor index strategy is used to obtain retrieval results using each model, and finally, the linear addition method is used to merge the retrieval results of different models. In this way, the collaborative retrieval results of multiple models can be obtained.

2 Key Technology

This paper will mainly introduce some of the key technologies used in this platform and the reasons why this platform uses this technology, including distributed storage technology, Sentence Bert model and CLIP model for feature vector generation, and retrieval fusion method based on linear addition.

2.1 Distributed Storage Technology

A distributed system refers to a large number of ordinary servers that are connected using the network. When storing data, the data is stored on these servers according to specific rules. These servers as a whole provide data storage functions, as shown in Fig. 1. Distributed storage technology has the characteristics of scalability, availability, reliability, high-tech, easy maintenance, and low cost. This platform uses distributed storage because the amount of data in the multi-modal knowledge graph is large. The multi-modal knowledge graph contains multi-modal data of text and images, and the explanations in the knowledge graph contain multiple languages. In other words, the data types in the knowledge graph are different and easy to categorize. After the distributed storage of the data is classified, the platform can reduce the retrieved data's scope when performing sentence or image retrieval, thereby increasing the speed and efficiency of retrieval. In addition, the high scalability advantage of distributed data storage technology provides broad storage space for the expansion of multi-modal knowledge graph nodes and image information.

2.2 SBert and CLIP for Feature Vector Generation

To obtain more accurate data retrieval results, the platform uses the Sentence Bert model to learn the text data in the knowledge graph and the sentences input by the user into the platform to efficiently obtain the feature vector expression of the sentence. Bert (Bidirectional Encoder Representation from Transformers) [13] is a pre-trained language representation model that uses ML (masked language prediction task) and NSP (next sentence prediction task) for pre-training and uses deep two-way Transformer components to build the entire model, and finally generates a deep two-way language that can integrate left and right contextual information feature. Bert obtained state-of-the-art results in NLP tasks and has now been widely used in many NLP scenarios. Due to its pre-training and fine-tuning SOT effect, Bert has become

a research hotspot in various fields of AI. A large number of Bert-based models have been applied to handle various classic AI tasks including natural language processing, image processing, and cross-modal representation learning. For example Vit [14], VideoBert [15], VisualBert [16], Vl-Bert [17], VilBert [18], Unicoder-VL [19], UNITER [20], Lxmert [21], BLIP [22], B2T2 [23] and Flava [24]. There are also studies using Bert to interact with knowledge graphs, such as ERNIE [25] and KnowBert [26]. However, in tasks such as sentence pair regression, the Bert model requires that two sentences must be input at the same time, resulting in too many traversals and unacceptable time consumption. The proposal of the Sentence Bert model is mainly oriented to solve the problem of semantic similarity search that does not apply to the Bert model, to achieve the effect of keeping the accuracy unchanged while greatly reducing the time consumption. The platform uses the CLIP model to jointly train a large amount of text data and image data in the knowledge graph and uses the trained CLIP model to generate the feature vector expression of the input image on the user side for subsequent vector similarity matching. The model supports data in two modalities of text and image. Using contrastive learning, the problem of image recognition and classification is transformed into an intelligent matching problem between image and text. It can efficiently learn visual concepts from natural language supervision. The CLIP model is widely used in many cross-modal AI tasks.

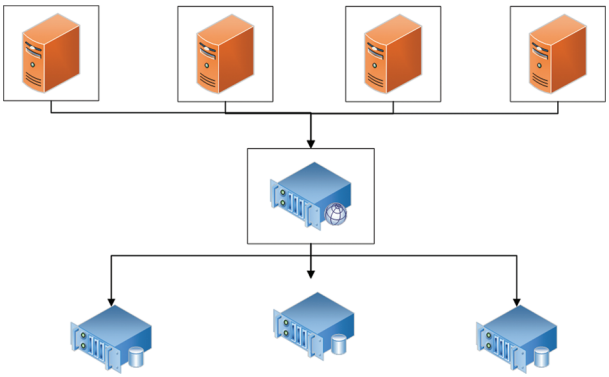

Figure 1: The architecture of the distributed storage system

2.3 Retrieval Fusion Method Based on Linear Addition

To improve the accuracy of sentence retrieval and image retrieval, the distributed collaborative vector retrieval platform (DCRL-KG) builds multiple retrieval models based on different feature vector similarity calculation methods and obtains multiple retrieval results for a single retrieval task. The reason is that the retrieval basis of different retrieval models is different, and the emphasis on the data is not the same. Therefore, different models may have their advantages and disadvantages in different parts of the data set. According to prior knowledge, it can be known that by using a series of different retrieval algorithms and combining the retrieval results of these algorithms, the retrieval performance can usually be improved compared to using only one certain retrieval algorithm. Through this method, the multi-channel retrieval method that combines multiple models can excellently handle a wider range of data, obtain more objective and comprehensive retrieval results, and improve the accuracy of retrieval.

Combining the retrieval results of different retrieval models requires the use of reasonable methods. Using inappropriate result fusion methods will not only make it difficult to improve retrieval performance but will reduce the accuracy of retrieval. DCRL-KG uses the linear addition of the retrieval results of the retrieval model set to form the final sentence retrieval result or image retrieval result. The process of fusion of retrieval results of conventional multiple search models is shown in Fig. 2.

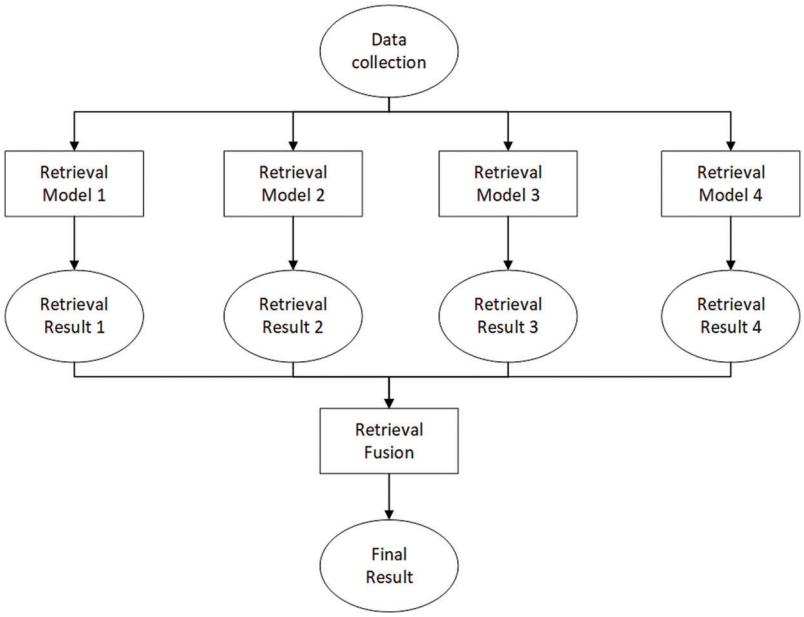

Figure 2: Fusion of search results from multiple search models

3 Design

3.1 Overall Design

In terms of data storage, DCRL-KG uses a distributed storage structure to store the multi-modal knowledge graph of massive data in a distributed manner, optimizes the storage structure of the knowledge graph, and optimizes the storage and retrieval efficiency of the entire platform. In terms of the expansion of the knowledge graph, DCRL-KG adopts the nearest neighbor method to constrain the addition of nodes to the existing knowledge graph and uses multiple filtering methods to add high-quality and highly-related images to the existing knowledge graph to ensure that the quality of knowledge graph will not be damaged in the process of expanding the knowledge graph. In terms of sentence retrieval and image retrieval, DCRL-KG uses the Sentence Bert model and CLIP model for feature vector generation, comprehensively uses a variety of retrieval models based on different vector similarity measurement methods, and finally uses retrieval fusion technology to obtain credible results of multi-model collaborative retrieval. DCRL-KG's multi-channel retrieval function has high data adaptability and accuracy. Fig. 3 shows the overall architecture of DCRL-KG.

3.2 Distributed Storage Module

The main data of the multi-modal knowledge graph is composed of explanations and images corresponding to the nodes, and the explanations corresponding to the nodes are composed of multiple languages. DCRL-KG performs entity retrieval on the multi-modal knowledge graph based on the calculated similarity between the vector representation of the sentence or image used in the retrieval and the vector representation of all nodes in the entity space, using k-nearest neighbors as the retrieval strategy. When using sentences for retrieval, since the explanation in the knowledge graph contains multiple languages, the platform will first convert the sentences used in the retrieval into a set of sentences covering all languages in the knowledge graph. If the distributed storage technology is not used at this time, the sentence corresponding to each language in the sentence set will be retrieved, and the nodes in the knowledge graph will be traversed to determine whether the node has the current language interpretation and if there is, the corresponding similarity is calculated. If distributed storage technology

is used, taking the condition of whether the node contains a certain language interpretation as the basis for the distributed storage technology to divide the data, the size of the traversal entity space can be reduced during the sentence retrieval process. Thereby achieving the effect of increasing the retrieval efficiency. In addition to improving retrieval efficiency, the multi-modal knowledge graph has the characteristic of huge data volume, distributed storage technology also guarantees the good scalability of platform data.

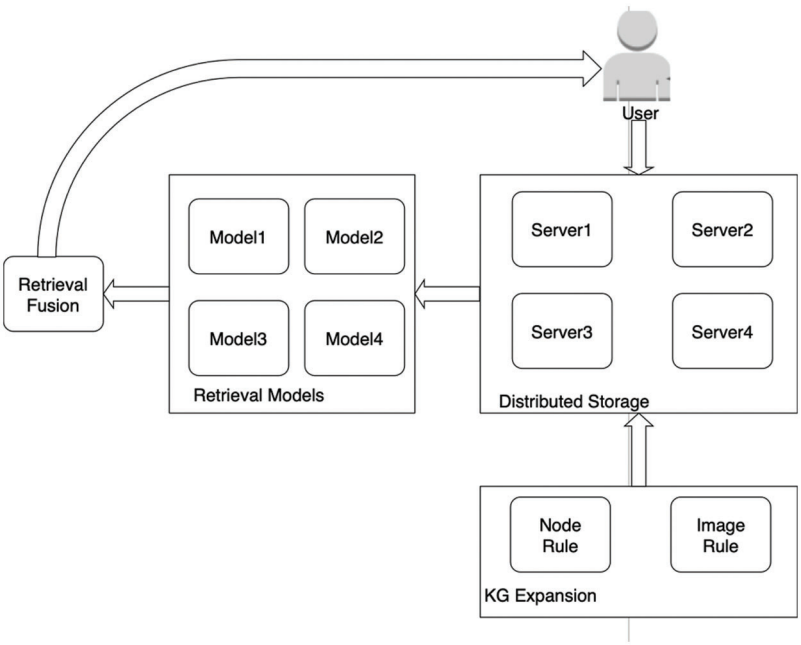

Figure 3: Overall architecture diagram of DCRL-KG

3.3 Knowledge Graph Extension Module

The main function of the DCRL-KG knowledge graph expansion module is to expand the information of nodes in the knowledge graph and obtain high-quality image information under the premise of ensuring the high quality of the multi-modal knowledge graph. The knowledge graph extension module uses the BabelNet v4.0 API to retrieve the expanded nodes based on the existing nodes in the knowledge graph. Knowledge graph expansion of DCRL-KG is carried out iteratively, and the steps of each iteration process are as follows:

Step 1: The nodes in the existing graph will call BabelNet v4.0 API to obtain the neighbor nodes of the existing node. The neighbor node is the node that is directly connected to the existing node through the relationship defined in the BabelNet v4.0 API. The retrieved nodes are multi-source, which can make the multi-modal graph diversified;

Step 2: Check whether there are duplicate nodes in the candidate nodes generated by the BabelNet v4.0 API search. If there are duplicates, remove the duplicate nodes from the candidate nodes;

Step 3: The quality of the image of the candidate node is checked by the method of multiple filters. Check whether the images of the candidate node are available, judge whether the image is a high-quality picture based on an image binary classification model to reduce noise images, and finally use the CLIP model to measure the interpretation of the candidate node and the correlation between the images. The more association performances are stronger, the more the candidate node meets our expectations. Remove nodes that do not meet the above filtering conditions from candidate nodes;

Step 4: The explanation information and image information of the selected nodes are added to the knowledge graph using distributed technology, and an iterative process of the knowledge graph has been completed.

After the above iterative steps, the knowledge expansion module adds multi-source, high-quality nodes to the existing knowledge graph. After multiple constraints and filtering, it can be ensured that the new nodes added are connected to the existing knowledge graph, and the image information of the new node is of high quality and can effectively describe the new node visually. The method of the entire knowledge graph extension module is shown in Fig. 4.

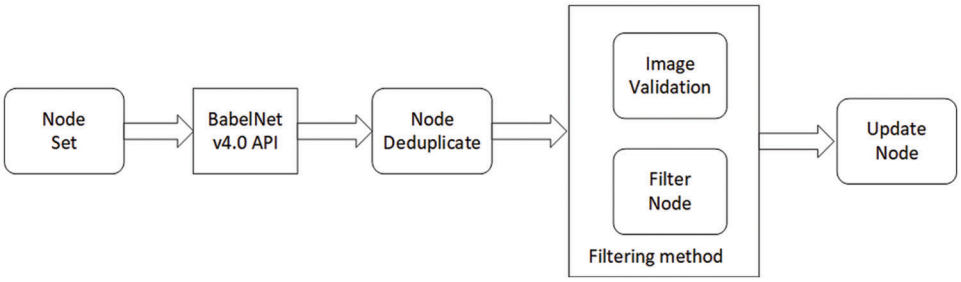

Figure 4: Knowledge graph extension module process

3.4 Collaborative Retrieval Module

In the multi-model collaborative retrieval module, when retrieving sentences or pictures, first use the Sbert model or the CLIP model to generate related sentence feature vectors or image feature vectors, and embed the retrieved sentences or images into the vector representation space of the knowledge graph, so that the problem is transformed into a vector retrieval problem in the vector space. Using multiple vector retrieval methods and fusion to obtain the final retrieval result, the platform determines the node in the knowledge graph corresponding to the retrieval sentence or image according to the result and returns related information and related images of the node in the knowledge graph. The process of multi-modal retrieval example is shown in Fig. 5. In the example of Fig. 5, a user of the platform retrieves image descriptions and sentence descriptions about British short cats. The platform uses the pre-trained CLIP model and SBERT model to extract image features and sentence features respectively. Based on the similarity of the feature vectors, the platform calculates the entity node corresponding to the British short cat in the multimodal knowledge graph through comparison, and at the same time gives the probability that the retrieved content belongs to the entity node. The platform returns the text description and image description of the corresponding nodes in the knowledge graph to the user as the result, thus completing the multimodal retrieval process.

4 Experiment

In this section, the paper uses the VisualSem dataset, which has 89,896 nodes, 13 relations, 1,342,764 sentences and 938,100 images in the multi-modal knowledge graph. In the experiment process, this paper sets below parameters: the batch size of 128 and reliable retrieval result number of 10. And referring to the description in VisualSem, this paper splits 85896 nodes to train dataset and each 2000 nodes to both test and valid dataset with sentences and images related to the nodes.

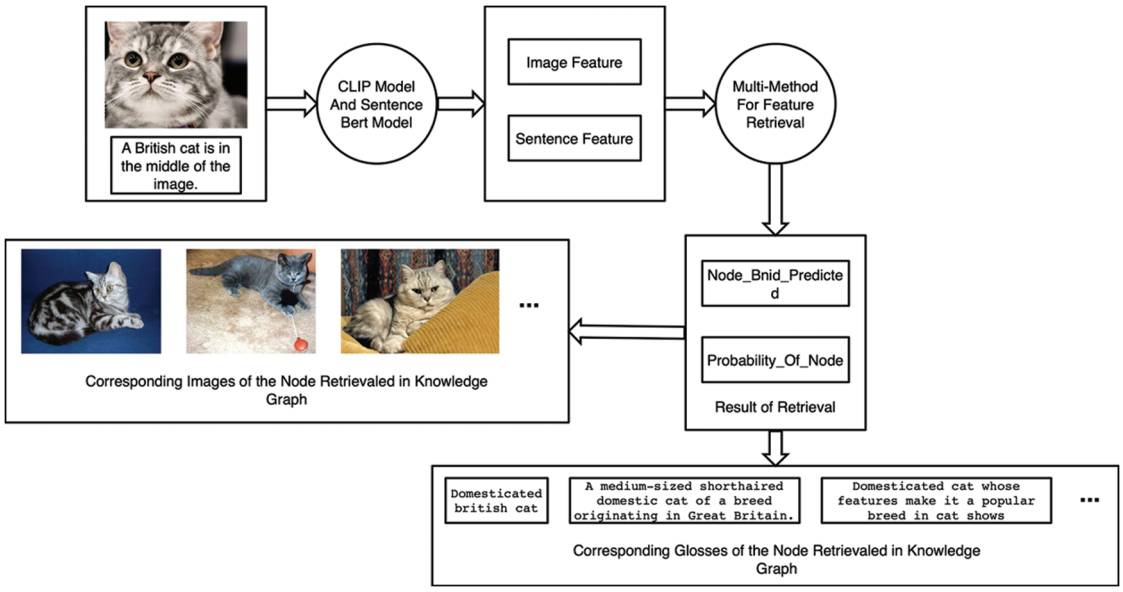

Figure 5: Process of multi-modal retrieval instance

4.1 Experiment Environment

To ensure the storage, expansion, and retrieval capabilities of the graph data in the form of the platform, while considering the scalability of the platform in a balanced manner, this paper constructed an experimental environment as shown in Table 1.

Table 1: Experiment environment

Server	CPU	Memory	Disk
Server-1	3.60 GHz 2 core 4 threads	8 GB	250 GB
Server-2	3.60 GHz 2 core 4 threads	8 GB	500 GB
Server-3	3.60 GHz 2 core 4 threads	8 GB	500 GB
Server-4	3.60 GHz 2 core 4 threads	8 GB	250 GB
Server-5	3.60 GHz 2 core 4 threads	8 GB	500 GB
Server-6	3.60 GHz 2 core 4 threads	8 GB	500 GB
Server-7	3.60 GHz 2 core 4 threads	8 GB	500 GB
Server-8	3.60 GHz 2 core 4 threads	8 GB	250 GB

4.2 Experiment Design

This part describes the experimental method, uses the experimental environment described in the previous part to experiment, and displays and analyzes the experimental results in the next part. The experiment uses the multi-modal knowledge graph VisualSem as the database, uses the sentence data and image data not in the knowledge graph to retrieve the entity nodes existing in the graph, and judges DCRL-KG according to whether the input sentence or image content and the retrieval result are consistent and accurate. In this way, we can figure out whether the sentence retrieval function and image retrieval function of the platform is as expected. The number of nodes in VisualSem is huge, and each

node in the graph contains accompanying multi-source text descriptions and image descriptions, which constitute the data diversity and richness of VisualSem. The following lists the statistical values related to the nature of the VisualSem knowledge graph, and compares it with other multimodal knowledge graphs, including WN9-IMG [1], FB15-IMG [8], DB15k [9], and Yago15k [9], as shown in Table 2. The text data and image data in VisualSem provide rich training data for the cross-modal representation learning model and provide strong data support for the intelligent application of the DCRL-KG platform.

Table 2: Multi-modal knowledge graph statistics

Knowledge graph	#nodes	#relations	#glosses	#images
WN9-IMG	6,555	9	N/A	65,550
FB15-IMG	11,757	1231	N/A	107,570
DB15k	14,777	279	N/A	12,841
Yago15k	15,283	32	N/A	11,194
VisualSem	89,896	13	1,342,764	938,100

In the experiment, the retrieval results of sentence retrieval or image retrieval on the DCRL-KG platform will be recorded and compared with the input sentences or pictures. The recorded indicators include whether the retrieval node is related to the content described by the input sentence or image and whether it is accurate. At the same time, the similarity between the input content calculated by the model and the matching node is recorded as the basis for platform retrieval. We expect that nodes with higher similarity rankings in the retrieval results should meet the requirements of relevance and accuracy, which means that DCRL-KG achieves representation learning for semantic and image modalities as well as matching learning in entity vector space. During the experiment, we found that the performance of DCRL-KG in the image retrieval task is not as good as the sentence retrieval task, because the image contains more abundant information, which increases the difficulty of representing and matching the information, which we will explain in the next part.

4.3 Experiment Result

Sentence Retrieval: This part conducts the verification and analysis of the sentence retrieval function of the DCRL-KG platform. The platform obtains the feature vector from the text information input by the user through the Sbert model and uses the multi-channel retrieval method to output the relevant content of the retrieval target node and the query text and the associated text information in the knowledge graph. In the example, the query sentence entered by the user to the platform is “a commercial airliner flies on a clear bright blue sky day”. The platform returns the ten nodes (top10) that are closely related to the query sentence (measured by vector similarity) in the knowledge graph. The specific search results are shown in Table 3.

The top10 sentence retrieval results in Table 3 include matching nodes in the knowledge graph, the similarity between the feature vectors of the query sentence and the node text description, the main content of the matching nodes in the knowledge graph, and the judgment retrieval based on the comparison of the node content and the meaning of the query sentence whether the results are relevant and accurate. Analyzing the data in the table, the top10 nodes retrieved by the sentence are all related to the query sentence, which is a specific category of airliners. Among them, except for the query node ranked tenth according to the similarity, the remaining nodes are accurately matched with the query target. The retrieval example shows that the sentence retrieval function of this platform has high accuracy. The similarity between different nodes and query sentences is due to the different text descriptions of

different nodes in the knowledge graph. Supplementing and perfecting the text description of the knowledge graph nodes can help the platform to better complete the vector matching task, which further improves the retrieval accuracy on this basis.

Table 3: Sentence retrieval example result

Node_id	Similarity	Node_content	Is_related	Is_accurate
bn:01871529n	0.82669199	Airbus_A300	True	True
bn:03783144n	0.82669199	Boeing_777	True	True
bn:03264239n	0.82514828	NAMC_YS-11	True	True
bn:00048175n	0.79192758	jetliner	True	True
bn:03226093n	0.79104048	Boeing_747	True	True
bn:02420429n	0.79067987	Boeing_787_Dreamliner	True	True
bn:01067844n	0.78690660	De_Havilland_Comet	True	True
bn:02555458n	0.76830590	Avro_Canada_C102_Jetliner	True	True
bn:00149229n	0.75657904	Lockheed_L-1011_TriStar	True	True
bn:03788047n	0.75452554	Airspeed_Envoy	True	False

Image Retrieval: This part conducts the verification and analysis of the image retrieval function of the DCRL-KG platform. The platform obtains the feature vector from the image data input by the user through the CLIP model, calculates the similarity between the image feature vector and the knowledge graph text description through the model, and matches the input image. The node to which the text description belongs is used as the search result, and the relevant content of the search result and the possibility of matching the image with the specific text description are output. In the example, the query picture input by the user terminal to the platform is the example input picture (British short cat) in Fig. 5. The platform returns the ten nodes (top10) in the knowledge graph whose text description is most relevant to the query image content (top10). The specific search results are shown in Table 4.

Table 4: Image retrieval exemple result

Node_id	Probability	Node_content	Is_related	Is_accurate
bn:03213312n	0.000047385693	British_Shorthair	True	True
bn:03482546n	0.000045776367	American_Shorthair	True	False
bn:03573606n	0.000042557716	Exotic_shorthair	True	False
bn:01768680n	0.000022649765	Maru(a cat on YouTube)	True	False
bn:00014737n	0.000020444393	calico_cat	True	False
bn:00427846n	0.000020027161	Burmilla	True	False
bn:00665389n	0.000020027161	Scottish_Fold	True	False
bn:01528722n	0.000019967556	Kurilian_Bobtail	True	False
bn:15776411n	0.000019729137	Asian_cat	True	False
bn:00289092n	0.000019669533	European_Shorthair	True	False

The top10 image retrieval results in Table 4 include the possibility of matching the query image with the specific text description in the knowledge graph, the node corresponding to the text description, and whether the main content of the node and the retrieval result is relevant and accurate. Analyzing the data in the table, the top10 nodes retrieved from the image are all related to the query image, and the node content is of different cat types. Among them, the first search result accurately matches the query target, while the remaining search results only achieve relevant matching without guaranteeing accuracy. It can be seen that image retrieval is inferior to sentence retrieval in accuracy. We think there are two reasons for this phenomenon. Firstly, the information contained in the picture is more concrete and richer than the sentence, which enables image retrieval to obtain more specific retrieval results. For related categories that belong to the same abstract level as the query target, the matching degree with the query image is low. However, the nodes in the knowledge graph describe specific factual content and lack abstract concept nodes, which makes it difficult to obtain retrieval results of different abstract levels when performing image retrieval. Except for the nodes that exactly match the retrieved content, the platform can only output related nodes. Aiming at this point, expanding the multi-modal knowledge map in the direction of the diversity of conceptual levels can further improve the accuracy of image retrieval. Secondly, because images contain more information than sentences, an image often contains more than one entity, and the interference of non-main entities in the image increases the difficulty of image retrieval. In response to this, attention-based strategies enable the model to select key entities for representation and recognition, thereby improving the accuracy of image detection.

The above content gives the example retrieval results of sentence retrieval and image retrieval. The above experiments can intuitively demonstrate the retrieval function of the DCRL-KG platform proposed in this paper. The retrieval accuracy of DCRL-KG has not been proved through experiments with large data volumes. The VisualSem dataset starting with Section 4 is used for both the sentence retrieval and image retrieval experiments respectively, and the mean rank (MR) and Hit@1, Hit@3, and Hit@10 are used as the experimental indicators. The experimental results are shown in Table 5. The experimental results more clearly show that the image retrieval task has a large gap in effect compared with the sentence retrieval task. The above experimental results prove that the method proposed in this paper can realize the functions of sentence retrieval and image retrieval, but the retrieval accuracy still needs further work to continue to improve.

Table 5: Sentence and image retrieval result of VisualSem

	Mean Rank (MR)	Hits@1	Hits@3	Hits@10
Sentence retrieval	3820	51	60	68
Image retrieval	4117	10	16	25

5 Conclusion

Knowledge graphs can efficiently structure knowledge, and make natural language understanding, natural language generation, and other AI fields have further development. Therefore, knowledge graphs have become a hot research direction. The multi-modal knowledge graph introduces external information as a supplement to knowledge based on structured knowledge, which enriches the content of the knowledge graph and obtains a better knowledge expression effect. In the above context, this paper proposes a multi-modal multi-channel collaborative representation learning distributed vector retrieval platform (DCRL-KG) that uses distributed storage technology to improve the retrieval speed of multi-modal knowledge graphs and optimize the storage structure. Use BabelNet v4.0 API to achieve multi-filtered knowledge graph expansion to ensure that the knowledge graph is continuously enriched under

the premise of high quality. Using the multi-channel collaborative retrieval method, high-precision sentence and picture retrieval functions are realized in the multi-modal knowledge graph. In the experiment section, this paper uses VisualSem as the dataset, and proves that the DCRL-KG platform has the function of multimodal content retrieval through sentence retrieval experiments and image retrieval experiments. After analysis, it is concluded that enriching the image information content in the multimodal knowledge graph and optimizing the vector representation learning method are the directions for continued improvement of this work in the future.

Acknowledgement: The authors would like to thank the associate editor and the reviewers for their time and effort provided to review the manuscript.

Funding Statement: This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF.201714), Weihai Science and Technology Development Program (2016DX GJMS15), Weihai Scientific Research and Innovation Fund (2020) and Key Research and Development Program in Shandong Provincial (2017GGX90103).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Xie, Z. Liu, H. Luan and M. Sun, "Image-embodied knowledge representation learning," arXiv preprint arXiv:1609.07028, 2017.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, USA, 2017.

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems*, vol. 26, Lake Tahoe, Nevada, USA, 2013.

[4] H. Alberts, T. Huang, Y. Deshpande, Y. Liu, K. Cho *et al.,* "VisualSem: A high-quality knowledge graph for vision and language," arXiv preprint arXiv: 2008.09150, 2020.

[5] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun *et al.,* "Multi-modal knowledge graph construction and application: A survey," arXiv preprint arXiv:2202.05786, 2022.

[6] N. Huang, Y. R. Deshpande, Y. Liu, H. Alberts, K. Cho *et al.,* "Endowing language models with multimodal knowledge graph representations," arXiv preprint arXiv:2206.13163, 2022.

[7] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li *et al.,* "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, Florida, USA, pp. 248–255, 2009.

[8] H. Mousselly-Sergieh, T. Botschen, I. Gurevych and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proc. of the Seventh Joint Conf. on Lexical and Computational Semantics*, New Orleans, Louisiana, USA, pp. 225–234, 2018.

[9] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio *et al.,* "MMKG: Multi-modal knowledge graphs," in *European Semantic Web Conf.*, Portorož, Slovenia, pp. 459–474, 2019.

[10] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[11] N. Reimers and I. Gurevych, "Sentencebert: Sentence embeddings using siamese bertnetworks," arXiv preprint arXiv:1908.10084, 2019.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.,* "Learning transferable visual models from natural language supervision," in *Int. Conf. on Machine Learning*, Virtual Event, pp. 8748–8763, 2021.

[13] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth 16 × 16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2010.

[15] C. Sun, A. Myers, C. Vondrick, K. Murphy and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, South Korea, pp. 7464–7473, 2019.

[16] L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh and K. W. Chang, "Visualbert: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019.

[17] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu *et al.,* "Vl-bert: Pre-training of generic visual-linguistic representations," arXiv preprint arXiv:1908.08530, 2019.

[18] J. Lu, D. Batra, D. Parikh and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, vol. 32, Vancouver, BC, Canada, 2019.

[19] G. Li, N. Duan, Y. Fang, M. Gong and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11336–11344, 2020.

[20] Y. C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed *et al.,* "Uniter: Universal image-text representation learning," in *European Conf. on Computer Vision*, Glasgow, UK, pp. 104–120, 2020.

[21] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," arXiv preprint arXiv:1908.07490, 2019.

[22] J. Li, D. Li, C. Xiong and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," arXiv preprint arXiv:2201.12086, 2022.

[23] C. Alberti, J. Ling, M. Collins and D. Reitter, "Fusion of detected objects in text for visual question answering," arXiv preprint arXiv:1908.05054, 2019.

[24] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba *et al.,* "Flava: A foundational language and vision alignment model," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 15638–15650, 2022.

[25] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun *et al.,* "Enhanced language representation with informative entities," arXiv preprint arXiv:1905.07129, 2019.

[26] M. E. Peters, M. Neumann, R. L. Logan IV, R. Schwartz, V. Joshi *et al.,* "Knowledge enhanced contextual word representations," arXiv preprint arXiv: 1909. 04164, 2019.