CHAPTER-29

COLLABORATIVE REPRESENTATION LEARNING-BASED DISTRIBUTED MULTI-MODAL KNOWLEDGE GRAPH RETRIEVAL PLATFORM

[1]NAZEER SHAIK, [2]Dr. B. HARICHANDANA, [3]Dr. P. CHITRALINGAPPA
[1,3]Department of CSE (Data Science), Srinivasa Ramanujan Institute of Technology, Anantapur
[2]Department of Computer Science & Engineering, Srinivasa Ramanujan Institute of Technology, Anantapur

Chapter ID: NSP/ICAAR-2023/A-29

ABSTRACT

The knowledge graph with relational abundant information has been widely used as the basic data support for retrieval platforms. Multi-modal knowledge graphs benefit from the addition of images and text descriptions to node information, which contributes to their advantage. In cross-modal retrieval platforms, multi-modal knowledge graphs can help improve retrieval accuracy and efficiency because they provide abundant relational information. For the application of multimodal knowledge graphs, the representation learning method is crucial. Asa foundation for efficient and high-precision multimodal data retrieval, this paper proposes a distributed collaborative vector retrieval platform (DCRL-KG) based on the multimodal knowledge graph VisualSem. To improve retrieval efficiency, use distributed technology to classify and store the data in a knowledge graph. Secondly, this paper uses BabelNet to expand the knowledge graph through multiple filtering processes and increase the diversity of information. As a final step, this paper develops a variety of retrieval models to achieve high- precision language retrieval and image retrieval by fusing retrieval results with linear combination methods. According to the paper, the platform can optimize the multimodal knowledge graph's storage structure and performing well in a multimodal environment.

***Indexed Terms**: Multi-modal retrieval; distributed storage; knowledge graph.*

INTRODUCTION

The goal of artificial intelligence research and development is to efficiently express and expand human knowledge. Artificial intelligence (AI) tasks such as natural language understanding, and natural language generation are often realized using knowledge bases. A large amount of attention has been drawn to the technology of knowledge graphs as a structured representationof knowledge in recent years. To describe knowledge facts, the knowledge graph uses the structure of triples, in which entities and relationships are arranged into triples. Knowledge graphs perform well in terms of representation, embedding, and expansion because of their structured nature. Artificial intelligence applications such as language representation learning, and intelligent language question-and-answer have widely used knowledge graph technology.

Knowledge graphs need to have more complete knowledge content, be more accurate and efficient in expressing learning embedding, and more and more researchers are not only focusing on structured text content, but also incorporating high-quality external information into the knowledge graph and constructing a multi-modal knowledge graph. In the field of multimodal knowledge graphs, this is a hot, cutting-edge research area.

In the year of 2017 Xie proposed the Image-embodied knowledge representation learning (IKRL) method, which embeds image information of entities based on the attention mechanism 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, processes the image information using convolutional neural networks to produce image vector representations, computes the similarity between the image vector and the entity embedding vector in entity space. Based on the attention mechanism, the attention values of different pictures in the entity picture set are merged with the picture vector corresponding to the entity to produce an embedding vector of the entity corresponding picture set. As part of the knowledge graph expression learning method, the image vector and the text vector are combined to form the loss function for the training IKRL model using the Translating embeddings for modelling multi-relational data (TransE) [3] method. A multimodal knowledge graph was initially constructed by IKRL, but it has limitations and is very limited in coverage. In addition to narrow knowledge domains and noise in pictures, most of the other multi-modal knowledge graphs that contain pictures are also subject to similar problems.

As proposed by VisualSem [4], VisualSem [4] constructs a multimodal knowledge graph based on image and language data. Additionally, VisualSem is a multi-language-oriented knowledge graph whose nodes have corresponding explanations in different languages. There are approximately 90k nodes, 1.3M explanations, and about 938k pictures in VisualSem. It is notable that VisualSem has many nodes, which makes it scalable and easy to use. VisualSem was compared with other multimodal knowledge graphs in a survey on the construction and application of multimodal knowledge graphs [5], and its features were highlighted. A multimodal knowledge graph representation method in 2022 [6] uses VisualSem as an experimental dataset and conducts extensive comparative experiments. ImageNet [7] is the source of image information in the knowledge graph. To reduce the challenge posed by image noise, the pictures in the knowledge graph are filtered through multiple filters and connected with other knowledge bases, such as Wikipedia, to make the nodes diversified. VisualSem is an intelligent multi-modal map of high quality with a wide range of applications and development significance. With VisulSem, you get high-quality and diversified data when compared to other multi-modal knowledge maps, such as FB15-IMG [8] and MMKG [9]. Based on VisualSem, this paper builds a distributed vector retrieval platform for multimodal collaborative representation learning, with the aim of achieving reasonable knowledge graph storage, high-quality knowledge graph expansion, and high-accuracy knowledge graph vector entity retrieval. Based on the multimodal knowledge graph proposed in this paper, the multimodal retrieval platform can be used in the following applications: when the user tries to identify the content in the image or obtain information about the entity, the platform can accurately and efficiently identify the images and sentences and return rich association descriptions. This paper mainly describes DCRL-KG's key technologies in Section 2. Section 3 describes the platform's design, including the distributed storage module, knowledge graph extension module, and collaborative retrieval module. As a test dataset, this paper uses VisualSem to evaluate the accuracy of sentence retrieval and image retrieval. This paper concludes the work in Section 5 and points out directions for future development.

As a summary, this platform consists mainly of the following:

1. Distributed storage technology can be used to rationally organize the storage structure, speed up retrieval of knowledge maps, and increase the scalability of knowledge graphs.
2. In the process of expanding the knowledge graph, use BabelNet [10] to multiple filter the expanded images, restrict and constrain the added nodes, and then ensure the high quality of the

knowledge graph.

3. As a core function of this platform, sentence retrieval and image retrieval of multimodal knowledge graphs are efficiently and accurately realized. While training the model, we extract feature vectors from sentence data and image data using Sentence Bert (SBERT) and CLIP models [11, 12]. In the process of vector retrieval, based on different retrieval principles and methods corresponding to different retrieval models, the k-nearest neighbour 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 the different models. This allows us to retrieve multiple models collaboratively.

KEY TECHNOLOGY

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

1. The Technology of Distributed Storage

A distributed system refers to many ordinary servers that are connected using the network. When storing data, the data is stored on these servers according to specific rules when storing. As shown in Figure 1, these servers perform data storage functions. Distributed storage technology has the characteristics of scalability, availability, reliability, high-tech, easy maintenance, and low cost. There is a large amount of data in the multi-modal knowledge graph, so distributed storage is used. The multi-modal knowledge graph contains multi-modal data of text and images, and explanations in the knowledge graph contain multiple languages. As a result, the knowledge graph has a variety of data types that can be categorised easily. Upon categorizing the distributed data, the platform can reduce the scope of the retrieved data during sentence or image retrieval, which makes the retrieval process faster and more efficient. In addition, the high scalability advantage of distributed data storage technology provides broad storage space for the expansion of multimodal knowledge graph nodes and image information.

2. Feature vector generation with SBert and CLIP

For more accurate data retrieval, the platform employs the Sentence Bert model to learn the text data in the knowledge graph and the sentences the users input into the platform. Bert (Bidirectional Encoder Representation from Transformers) [13] is a pre-trained language representation model which uses ML (masked language prediction task) and NSP (next sentence prediction task) for pre-training, then builds the entire model using deep two-way Transformer components, and finally generates a deep two-way language that can integrate left and right contextual information. As a result of its pre-training and fine-tuning SOT effect, Bert has become a research hotspot in various AI fields. Bert achieved state-of-the-art results in NLP tasks and has become widely used in many NLP scenarios. Natural language processing, image processing, and cross-modal representation learning have all been handled with Bert-based models. Vit [14], Video Bert [15], Visual Bert [16], Vl-Bert [17], VilBert [18], Unicoder-VL [19], UNITER [20], Lxmert [21], BLIP [22], B2T2 [23] and Flava [24] are a few examples. Also, Bert has been used in studies to interact with knowledge graphs, such as ERNIE [25] and KnowBert [26]. In tasks such as sentence pair regression, however, the Bert model requires two sentences to be input at the same time, which results in too many traversals and unacceptable time consumption. Sentence Bert is primarily designed to solve the problem of semantic similarity search that does not apply to the Bert model, to keep accuracy while reducing time. For subsequent vector similarity matching, the platform uses the trained CLIP model to generate the feature vector expression of an input image using the combined training of text data and image data in the knowledge graph. Using contrastive learning, the problem of image recognition and classification is transformed into an intelligent matching problem between image and text. This model supports data in two modalities: text and image. In many cross-modal AI tasks, the CLIP model can efficiently learn visual concepts from natural language supervision.

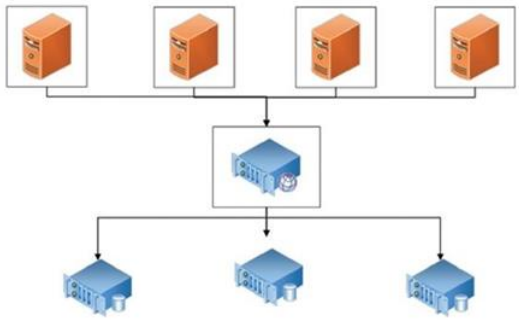

Fig. 1: The architecture of the distributed storage system

3. Fusion Method Based on Linear Addition for Retrieval

Using different feature vector similarity calculation methods, the distributed collaborative vector retrieval platform (DCRL-KG) builds multiple retrieval models to improve sentence retrieval and image retrieval accuracy. For each retrieval task, multiple retrieval results are obtained. Different retrieval models have different retrieval bases, and their emphasis is different on the data. In different parts of the data set, different models may have different advantages and disadvantages. As a result of prior knowledge, it is known that combining the retrieval results of several retrieval algorithms can usually result in better retrieval performance than using only one retrieval algorithm. With this method, a multichannel retrieval method that combines multiple models can effectively handle a wider range of data, obtain more comprehensive retrieval results, and improve retrieval accuracy.

In order to combine retrieval results from different retrieval models, reasonable methods must be used. It will be difficult to improve retrieval performance and less accurate to retrieve data if inappropriate result fusion methods are used. In DCRL-KG, retrieval results are linearly added to form the final sentence retrieval or image retrieval results. Fig. 2 illustrates the process of combining the results of conventional multiple search models.

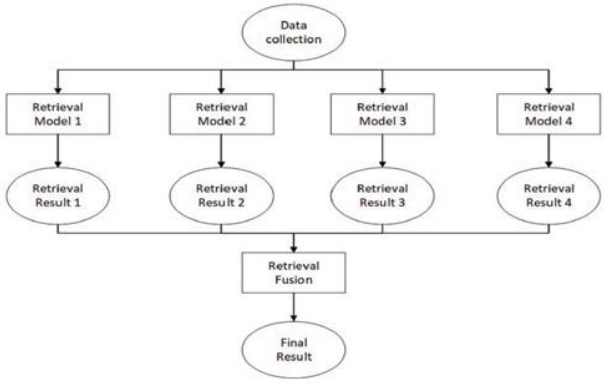

Fig.2: Fusion of Search Results from Multiple Search Modes

DESIGN

1. Overall Design

To store the massive multimodal knowledge graph of massive data in a distributed manner, DCRL-KG uses a distributed storage structure, optimizes the storage structure of the knowledge graph, and optimizes the platform's storage and retrieval efficiency. In terms of the expansion of the knowledge

graph, DCRL-KG adopts the nearest Neighbour method to restrict the addition of nodes to the existing knowledge graph and adds high-quality and highly related images to the existing knowledge graph using multiple filtering methods. By expanding the knowledge graph, the quality will not be compromised. For sentence retrieval and image retrieval, DCRL-KG uses the Sentence Bert model and CLIP model, as well as a variety of retrieval models based on different vector similarity measurement methods to get credible retrieval results from multi-model collaborative retrieval using retrieval fusion. The DCRL- KG retrieval function is highly adaptable and accurate. Figure 3 demonstrates the DCRL-KG architecture.

2. Distributed Storage Module

Multi-modal knowledge graphs consist of explanations and images that correspond to nodes, with explanations in multiple languages. In DCRL-KG, entity retrieval is performed 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 neighbours as a retrieval strategy. Since the explanation in the knowledge graph contains multiple languages, the platform will first convert the sentences used for retrieval into a set of sentences covering all languages. In the absence of distributed storage, 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 if the node has the current language interpretation, and if so, its similarity will be calculated. The size of the traversal entity space can be reduced during sentence retrieval if distributed storage technology is used, using the condition of whether the node contains a certain language interpretation as a basis for dividing the data. This results in an increase in retrieval efficiency. Besides improving retrieval efficiency, the multi-modal knowledge graph has the characteristic of large data volumes. Distributed storage technology also guarantees good platform scalability.

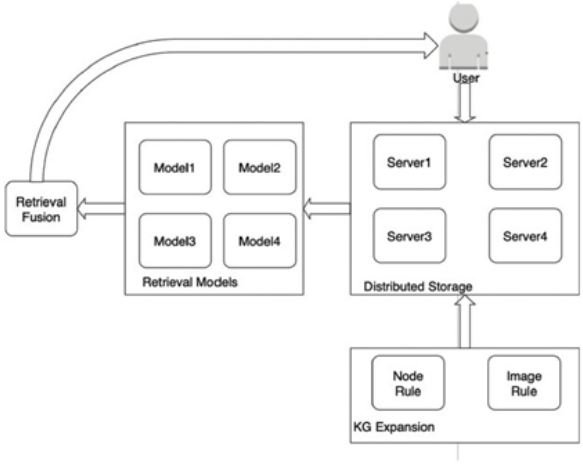

Fig.3: Overall Architecture Diagram of DCRL-KG

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 based on the premise of ensuring the quality of the multi-modal knowledge graph. Based on the existing nodes in the knowledge graph, the extension module retrieves expanded nodes using the BabelNet v4.0 API. In DCRL-

KG, knowledge graph expansion occurs iteratively, and eachiteration includes the following steps:

Step.1: The nodes in the existing graph will call the BabelNet v4.0 API to find their neighbour nodes. In BabelNet v4.0, a neighbour node is a node that is directly connected to an existing node by a relationship defined by the API. The retrieved nodes are multi- source, so the graph can be diversified.

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

Step.3: We use multiple filters to check the image quality of the candidate node at step 3. Check if there are images of the candidate node, judge if it 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 as well as the correlation between images. The stronger the association performances, the more likely the candidate node will meet our expectations. Remove nodes that fail to meet the above criteria.

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

As a result of the above iterative steps, the knowledge expansion module adds high-quality, multi-source nodes to the existing knowledge graph. In addition to ensuring that the new nodes are connected to the existing knowledge graph after multiple constraints and filtering, the image information can effectively describe the new node visually as well as being of high quality. Figure 4 illustrates how the entire extension module works.

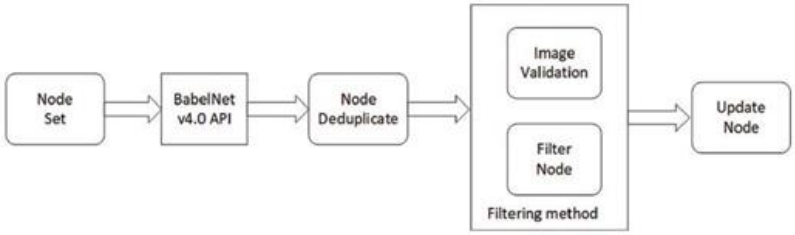

Fig.4: Knowledge Graph Extension Module Process

4. Collaborative Retrieval Module

Using the multi-model collaborative retrieval module, we generate related sentence feature vectors or image feature vectors by using the **Sbert Model** or the **CLIP model** first. The retrieved sentences or images are then embedded into the vector representation space of the knowledge graph, transforming the problem into a vector retrieval problem. 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 based on the result and returns related information and related images of the node. Figure 5 illustrates the process of multi-modal retrieval. A user can retrieve image descriptions and sentence descriptions about British short cats from the platform in Fig. 5. CLIP and SBERT models are pre-trained to extract image features and sentence features. By comparing the feature vectors, the platform determines the entity node that corresponds to the British short cat in the multimodal knowledge graph, and at the same time provides the probability that the retrieved content belongs to that entity node. By returning the text description and image description of the

corresponding nodes in the knowledge graph, the platform completes the multimodal retrieval process.

EXPERIMENT

A multi-modal knowledge graph consisting of 89,896 nodes, 13 relations, 1,342,764 sentences and 938,100 images is used in this section of the paper. The experiment is conducted with the following parameters: batch size of 128 and reliable retrieval result number of 10. Based on the description in VisualSem, 85896 nodes are split into train datasets, and 2000 nodes are split into test and valid datasets, accompanied by sentences and images.

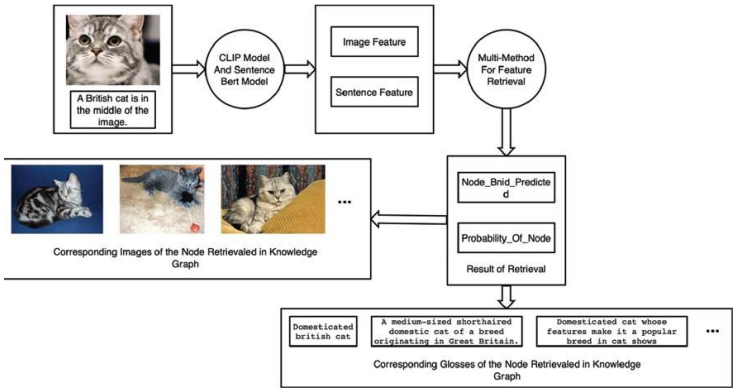

Fig.5: Process of Multi-Modal Retrieval Instance

1. An Environment for Experiments

This paper constructed an experimental environment as shown in Table 1 to ensure the storage, expansion, and retrieval capability of graph data within the platform, while considering the platform's scalability in a balanced manner.

Table 1: Exproment 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

2. Design of Experiments

Next, the experimental results will be displayed and analysed, using the experimental environment described in the previous part. Using VisualSem's multi-modal knowledge graph as a database, DCRL-KG retrieves entity nodes existing in the graph from sentence and image data that are not in the knowledge graph, and judges DCRL-KG based on the consistency and accuracy of the input sentence or image content. By doing so, we will be able to determine if the platform is able to retrieve sentences and images correctly. The number of nodes in VisualSem is huge, and each node in the graph

contains accompanying multi-source text descriptions and image descriptions, providing visual diversity and richness to the data. In Table 2, we compare the VisualSem knowledge graph to other multimodal knowledge graphs, including WN9-IMG [1], FB15-IMG [8], DB15k [9], and Yago15k [9], based on its statistical values. Data from VisualSem's text and image databases provide rich training data for the cross-modal representation learning model and support DCRL-KG's intelligent application.

Table 2: Multi-modal knowledge graph statistice

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

A comparison of the DCRL-KG platform retrieval results with input sentences or pictures will be conducted as part of the experiment. 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. In addition, platform retrieval is based on the similarity between the input content and the matching node calculated by the model. Nodes with higher similarity rankings are expected to meet the requirements of relevance and accuracy, which means that DCRL-KG achieves both representation and matching learning in entity vector space. We found that DCRL-KG performs less well in the image retrieval task than it does in the sentence retrieval task because images contain more information, making it more difficult to represent and matchthe information, which will be explained in the next section.

3. The Results of the ExperimentRetrieval of Sentences:

This part conducts the verification and analysis of the sentence retrieval function of the DCRL-KG platform. By using the Sbert model, the platform obtains the feature vector from the text information entered by the user and then uses the multichannel retrieval method to return the relevant content of the retrieval target node, the query text, and the text information associated with it to the knowledge graph. In the example, the user enters the query sentence "a commercial airliner flies on a clear bright blue-sky day". In the knowledge graph, the platform returns the ten nodes (top10) closely related to the query sentence (measured by vector similarity). Table 3 contains specific search results.

According to Table 3, the top 10 sentence retrieval results include nodes that match the query sentence, 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 judgment retrieval based on whether the results are relevant and accurate based on the comparison of the node content and the query sentence. Using the data in the table, all 10 of the nodes retrieved by the sentence are related to the query sentence, which is a specific class of aircraft. Apart from the query node ranked tenth, all of the remaining nodes match the query target accurately. This platform's sentence retrieval function is highly accurate, as shown by the retrieval example. The similarity between different nodes and query sentences is due to the different text descriptions of different nodes in the knowledge graph. By supplementing and perfecting the nodes' text descriptions, the platform can complete the vector matching task more efficiently, thus improving retrieval accuracy.

Table 3: Sentence retrievel 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

Verification and analysis of DCRL-KG's image retrieval function are conducted in this part. Based on the image data input by the user, the platform calculates the similarity between the image feature vector and the knowledge graph text description and matches the input image based on that similarity. 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 shown. Figure 5 shows the example input picture (British short cat) that is input to the platform by the user terminal. In the knowledge graph, the platform returns the ten nodes (top10) that are most relevant to the query image content. Table 4 showsspecific search results.

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 contents of the node and the retrieval result are relevant and accurate. Based on the data in the table, all the top 10 nodes retrieved from the image are related to the query image and contain different types of cat content. Among them, the first search result accurately matches the query target, while the remaining search results achieve relevant matching without ensuring accuracy. Image retrieval is inferior to sentence retrieval in accuracy. There are two reasons for this. As a result, image retrieval can obtain more specific retrieval results because the picture contains more concrete and rich information than the sentence does. When related categories belong to the same abstract level as the query image,

there is a low degree of matching. An image retrieval knowledge graph lacks abstract concept nodes, which makes it difficult to retrieve results of different abstract levels. Platforms can only output nodes that are directly related to the retrieved content. The accuracy of image retrieval can be further improved by expanding the multimodal knowledge map in a direction that considers the diversity of conceptual levels. Secondly, because images contain more information than sentences, an image often contains more than one entity, and the interference of non-main entities increases the difficulty of retrieving the image. As a result, attention-based strategies improve image detection accuracy by selecting key entities for representation and recognition. This paper presents example retrieval results for sentence retrieval and image retrieval. The above experiments show the retrieval capabilities of the DCRL-KG platform proposed in this paper intuitively. Experiments with large data volumes have not yet been conducted to prove DCRL-KG's retrieval accuracy. In both experiments, the VisualSem dataset starting in Section 4 is used, and the experimental indicators are the mean rank (MR) and **Hit@1, Hit@3, and Hit@10**. Experimental results are presented in Table 5. As shown in the experimental results, the image retrieval task has a significant gap in effect as compared to sentence retrieval. It can be concluded from the above experimental results that the method proposed in this paper can achieve both sentence retrieval and image retrieval functions, but retrieval accuracy still needsto be improved.

CONCLUSION

The use of knowledge graphs has become a hot research area, as they can efficiently structure knowledge, and make natural language understanding, natural language generation, and other AI fields more advanced. Multi-modal knowledge graphs enrich knowledge based on structured knowledge by adding external information, resulting in an increased level of knowledge expression. This paper proposes a distributed vector retrieval platform for collaborative representation learning using distributed storage technology to optimize the storage structure of multimodal knowledge graphs and improve retrieval speed. To ensure that the knowledge graph is continuously enriched under the premise of high quality, use the BabelNet v4.0 API for multi-filtered knowledge graph expansion. Using the multi-channel collaborative retrieval method, high-precision sentence and picture retrieval functions are realized in a multi-modal knowledge graph. In the experiment section, this paper uses VisualSem as the dataset and proves that DCRL-KG has the function of multimodal content retrieval through sentence retrieval experiments and image retrieval experiments. According to the analysis, enriching the image information in the multimodal knowledge graph and optimizing the vector representation learning method are the directions for future improvement.

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: 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 Bert networks," 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, 2018.3306 IASC, 2023, vol.36, no.3*
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 andperformant 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 forvision 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.*