Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29655

Multimodal Knowledge Graph Construction and Enhancement Techniques

Peng Xu

School of Geosciences, China University of Petroleum (East China), Qingdao, China
2673902359@qq.com

Abstract. As society transitions to the digital age, knowledge graphs facilitate the discovery of deep relationships and empower artificial intelligence with capabilities and insights, promoting data analysis and decision support. However, traditional knowledge graphs are mostly text-based, with limited perception and integration of information from other sensory channels, and limited cognition in understanding the material world. Multimodal knowledge graphs, on the other hand, use heterogeneous data from multiple sources as a link to interconnect and enrich knowledge representation, filling the semantic gap between different modalities and better serving the need for a deep understanding of the real world. Based on literature research and discussion, this paper summarizes the key technologies in the field of multimodal knowledge graphs from their technical characteristics, conducts detailed technical analysis and research, and also summarizes the important methods for multimodal knowledge enhancement, including representation learning, cross-modal alignment, and reasoning. Finally, this study summarizes the main challenges faced by current technologies, and provides an outlook for future research directions.

Keywords: multimodel knowledge graph, knowledge graph construction, cross-model alignment, representation learning

1. Introduction

Before delving into the Multimodal Knowledge Graph (MMKG), it is important to understand the core framework of traditional knowledge graphs. Knowledge graphs represent concepts and their relationships in the physical world in a structured form, and aggregate a large amount of knowledge by reducing data granularity. The basic framework of traditional knowledge graph (KG) is based on Resource Description Framework (RDF) as a knowledge representation model [1], with the mathematical expression shown below:

$$KG = \left\{ (e_i, \gamma_k, e_j) | e_i, e_j \in E, r_R \in R \right\} \tag{1}$$

In the formula, (e_i, γ_k, e_j) represents the entities and their relationships in a triple. E represents the set of entities, and R represents the set of relationships. Although this simple triple seems

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29655

uncomplicated, a large number of such triples interconnected form a vast semantic network that can more intuitively connect knowledge directly. This is also one of the basic logical points of the development of MMKG. Since MMKG is a development form based on traditional KG, the expansion content mainly involves adding multimodal information to entities and relationships. Therefore, MMKG can be considered as an expansion of the original triple, and its mathematical expression is as follows.

$$MMKG = (\mathrm{E}, \mathrm{R}, \mathrm{A}, \mathrm{V}, \mathrm{M}) \tag{2}$$

In the formula, E represents the entity set, R represents the relationship set, A represents the attribute set, V represents the multimodal value set, including non-text data objects such as images, videos, and audios, and M represents the mapping function, which connects entities, relationships, or attribute values with multimodal data.

The technological breakthrough of multimodal knowledge graph has greatly benefited from the development of multimodal large models in recent years. This research focuses on the construction and enhancement technology of multimodal knowledge graph, mainly elaborating on the following four aspects:

1) Provide a comprehensive overview of the entire technical stack for constructing and optimizing multimodal knowledge graphs.

2) Detailed technical elaboration on the application of mainstream large models in multimodal knowledge graph enhancement.

3) Introduce the multimodal knowledge storage structure.

This research can provide a reference basis for the vertical application of MMKG.

2. Overview of building intelligent multimodal knowledge graph

The construction process of MMKG integrates both top-down and bottom-up strategies and incorporates multimodal data [2]. The following is a detailed analysis of the construction process:

The bottom-up construction process mainly involves aggregating and analyzing data. Structured data can be analyzed by obtaining data information from data sources such as databases and technical documents. Unstructured data, due to its complex data format, requires more collection sources and a more complex processing process. The current mainstream data processing method is to use the LLMS algorithm to obtain entities, relationships, and attributes from unstructured data. Based on this, the method of entity alignment is used to integrate the paths corresponding to entities from different sources and ensure complementarity between multimodal data, making the knowledge graph very complete. After that, all the refined knowledge will be integrated into a unified storage space, and a comprehensive knowledge system will be established.

3. Cross-modal knowledge acquisition and fusion

The construction of MMKGs often faces a state of knowledge scarcity, requiring different knowledge extraction methods to extract knowledge units such as entities, relationships, and attributes from massive amounts of raw data. Although knowledge from different sources has certain cross-relationships, how to convert these cross-relationships into descriptive information specific to an entity is an urgent problem to be solved. Multimodal knowledge fusion serves as a connector to link entities with the same semantics, breaking information silos and building a unified framework.

DOI: 10.54254/2755-2721/2025.LD29655

3.1. Multimodal knowledge extraction

Multimodal knowledge extraction achieves named entity recognition and relationship extraction by fusing multimodal data such as images and speech, supporting the construction of multimodal knowledge graphs [3]. Image knowledge extraction utilizes feature extraction, statistical modeling, and deep learning techniques to supplement missing text semantics and resolve semantic conflicts; Speech knowledge extraction can be divided into pipeline based and end-to-end based. The former relies on the transcription process, while the latter directly associates the original speech with triplets; Large scale model knowledge extraction achieves cross modal semantic alignment through pre training and fine-tuning, and adapts to multiple scenarios with strong inference generation capabilities, such as EMRE2llm, OpenVek, RiVEG and other frameworks, reducing manual intervention and mining deep semantic connections. It has become the current mainstream technology direction.

3.2. Multimodal entity alignment

The core purpose of multimodal entity alignment is to correlate different knowledge graphs and the various modal information they contain, such as providing two misaligned multimodal knowledge graphs $\mathrm{MMKG}_1 = (\mathrm{E}_1, \mathrm{R}_1, \mathrm{A}_1, \mathrm{V}_1, \mathrm{M}_1)$ and $\mathrm{MMKG}_2 = (\mathrm{E}_2, \mathrm{R}_2, \mathrm{A}_2, \mathrm{V}_2, \mathrm{M}_2)$, as well as a pre-aligned seed entity set $\mathrm{S} = \{(\mathrm{e}_1, \mathrm{e}_2)|\mathrm{e}_1 \in \mathrm{E}_1, \mathrm{e}_2 \in \mathrm{E}_2, \mathrm{e}_1 \equiv \mathrm{e}_z\}$. The task of multimodal entity alignment is to find the entity set A that maps the physical world based on the seed set, where $\equiv$ represents the equivalence relationship and $\mathrm{S} \subseteq \mathrm{A}$. $\mathrm{S} \subseteq \mathrm{A}$[4].

Traditional entity alignment methods usually lack flexibility and require a large amount of human intervention in data processing. The emergence of embedded learning methods makes up for this deficiency. This section will focus on introducing embedded methods of various modalities.

3.2.1. Text feature embedding

Text embedding is the process of learning to convert text into digital vectors, aiming to capture semantics in low-dimensional Spaces and bring the locations of similar semantics closer together. This process includes preprocessing (such as word segmentation, stop word processing, etc.) and noise elimination. Furthermore, although traditional feature representation methods (bag-of-words, TF-IDF) can count word frequencies, they cannot capture semantics and word sequences. Modern pre-trained models (such as Word2Vec) can assign fixed vectors to words, but they have the problem of polyseme. To address this limitation, context models (such as BERT) can dynamically adjust vectors through the Transformer self-attention mechanism and use [CLS] markers or average pools to generate long text embeddings [5]. Finally, the final vector is applied to downstream tasks, such as calculating the semantic correlation of entities through cosine similarity in the knowledge graph, assisting in entity alignment and relationship prediction, etc.

3.2.2. Structure feature embedding

In the construction of MMKGs, structural feature embedding can reflect the similarity of MMKGs and capture the topological relationships between entities. The current mainstream structural feature extraction method mainly adopts the structural feature embedding technique based on the global structure representation of knowledge graph to map entities and relationships to low-dimensional continuous space vectors. This section will introduce representative methods in detail.

DOI: 10.54254/2755-2721/2025.LD29655

TransE is a pioneering KGE model that takes relations as the core, representing each entity and relation as a vector and achieving relational reasoning in KGs through vector operations [4]. Assuming an effective triple (h, r, t) ,, it should satisfy $h + r \approx t$ in the vector space, where h,r and t represent the head entity, relation, and tail entity vectors, respectively (see Figure 1).

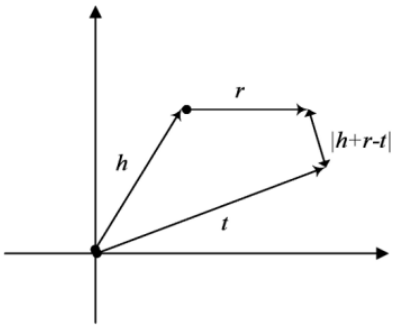

Figure 1. TransE model

Inspired by the Euler formula $e^{i\theta} = \cos\theta + i\sin\theta$, the RotatE model is proposed, as shown in Figure 2. Each relation is represented by a complex vector, and the rotation operation is implemented through complex multiplication. For the triple (h, r, t) , the RotatE model requires that $t = h \circ r$, where $\circ$ denotes the Hadamard product, and each element of the relation vector r is a complex number with a modulus length of 1 (i.e., $|r_i| = 1$)), ensuring the rotation property. Its scoring function is simplified to:

$$f(h, r, t) = -||h \circ r - t|| \tag{3}$$

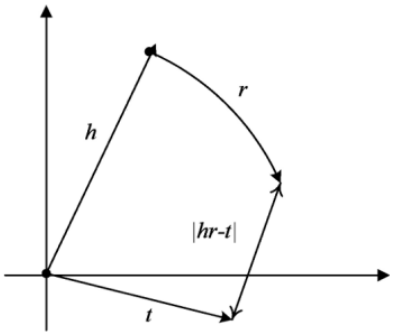

Figure 2. RotatE model

In model training, a loss function for self-adversarial sampling training can be adopted:

$$\mathcal{L} = -\log\sigma(\gamma - f(h, r, t)) - \sum_{i=1}^{n} \omega_i \log\sigma\left(f\left(h_i^{'}, \gamma, t_i^{'}\right) - \gamma\right) \tag{4}$$

Where $w_i = P\left(\left(h_i^{'}, r, t_i^{'}\right) \middle| (h, r, t)\right)$ is the sampling probability of the negative sample, as follows:

DOI: 10.54254/2755-2721/2025.LD29655

$$P\left(\left(h_i^{'}, r, t_i^{'}\right) \middle| (h, r, t)\right) = \frac{\exp\left(\alpha f\left(h_i^{'}, r, t_i^{'}\right)\right)}{\Sigma_i \exp\left(\alpha f\left(h_i^{'}, r, t_i^{'}\right)\right)} \tag{5}$$

Where α is the temperature parameter, $f(\cdot)$ is the model scoring function, the higher the current score of a negative sample, the greater the probability of it being selected.

3.2.3. Vision feature embedding

Visual feature embedding is the process of transforming unstructured image data into low-dimensional vectors, providing visual descriptions for entity vectors in knowledge graphs. The current mainstream visual feature extraction methods are mainly based on deep convolutional neural networks (CNN) and Transformer models. CNN extracts local features and hierarchical representations from input data through a multi-layer structure, and integrates these features into higher-level semantic information. Currently, more efficient neural networks such as DenseNet, VGG, AlexNet, and ResNet are used for visual feature extraction [5]. ResNet-50/ResNet-101 can extract richer image features through residual connections, while VGG-16/VGG-19 models can extract higher-grained image features. Feature extraction methods based on the Transformer model have become mainstream in the field of image processing due to their powerful capabilities, especially with improved models such as ViT and SwinTransformer, which often perform well in capturing global context and long-range dependencies in images due to their global attention mechanism.

3.2.4. Cross-modal semantic alignment

Cross-modal semantic alignment aims to map data of different modalities (which have already undergone conflict detection and resolution) to the semantic space of the same dimension. Currently, the method based on contrastive learning is the mainstream in this field. Its principle is to input the matching samples as positive samples and the unmatched samples as negative samples into the model, pull in the positive samples and push away the negative samples, thereby completing cross-modal semantic alignment [6]. OpenAI's CLIP model is a milestone in this paradigm, successfully aligning visual and text features into the same space [7]. The structure of this model mainly consists of two parts.

Text encoder: Based on Transformer, it is used to process text features.

Both are mapped to a shared embedding space through linear projection, and the Image encoder: Based on ResNet or Vision Transformer (ViT), used for extracting image features.

Both are mapped to a shared embedding space through linear projection, and the contrast loss function (InfoNCE) is utilized to maximize the similarity of matching image-text pairs and minimize the similarity of non-matching pairs. The definition of its contrastive loss function is:

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij} \log \frac{\exp(\mathrm{sim}(i,j))}{\sum_{k=1}^{N}\exp(\mathrm{sim}(i,k))} \tag{6}$$

Here, y_{ij} indicates whether the image matches the text, and sim is the cosine similarity.

In addition, there are currently two new research routes [8]. The first approach is to introduce large-scale parameters and the Transformer architecture, and utilize the attention mechanism to

DOI: 10.54254/2755-2721/2025.LD29655

achieve more efficient semantic alignment, such as models like Flamingo and Perceiver. Another approach is based on the CLIP model and further introduces language decoders and instruction fine-tuning strategies, such as models like BLIP-2 and LLaVA. These improvements enable the model to possess both powerful multimodal data analysis and image-text generation capabilities.

3.3. Cross-modal semantic fusion

On the basis of alignment, the goal of fusion is to integrate information from different modalities into a unified joint representation [6]. The specific methods are divided into the following four types.

·Feature level fusion represents connecting the extracted features into high-dimensional feature vectors. Specifically, the extracted multimodal features need to be fused and input into the MLP and Softmax layers to obtain classification results.

·Decision level fusion is the process of integrating decisions based on each modality by applying algebraic combination rules of multiple predicted class labels. The specific process is to input multimodal features into their respective MLP+Softmax layers, output single modal prediction results, fuse single modal prediction probabilities through linear weighted summation, and finally obtain the fusion results through the Softmax layer.

·Mixed level fusion is a collaborative scheme of feature level fusion and decision level fusion. Its process can be represented as inputting multimodal features into their respective linear regression layers for prediction, concatenating single modal scores with original feature vectors and inputting them into the FC layer, and generating the final result through a nonlinear transformation.

·Model level fusion achieves deeper cross modal information interaction by constructing multimodal joint feature representations. Compared to feature level fusion, its advantage lies in capturing semantic associations and contextual information between modalities, enhancing the processing capability of complex tasks.

4. Multimodal knowledge storage

The storage of multimodal knowledge is not simply about storing data such as images, videos, and texts in a database. Due to differences in data formats, the semantic representations of different modal data are also not the same. The storage system needs to build and maintain this cross modal semantic relationship and ensure the consistency of all associated data. To address these challenges, modern multimodal knowledge graphs typically adopt a hybrid storage architecture, where data from different modalities can be stored in specific databases. The overall framework is as follows:

·Structured knowledge storage layer

For pure textual triplet knowledge, graph databases are preferred, including Neo4j, JanusGraph, Nebula Graph, etc.. This layer can serve as the hub of the overall framework, storing pointers or indexes.

·Multimodal data storage layer

The storage methods for unstructured data are more diverse. For example, Amazon S3 and MinIO are very suitable for storing binary data; MongoDB can store structured documents.

·Vector data storage layer

This layer mainly stores multimodal entity embedding vectors. Because the data is represented by vectors, vector databases such as Milvus, Pinecone, and Weaviate can be used.

DOI: 10.54254/2755-2721/2025.LD29655

5. Conclusion

The research on MMKGs has evolved from early modal concatenation to a new stage of deep fusion, with the core driving force being how to handle semantic gaps between different modalities. At the construction level, pre-trained models represented by CLIP, ViLBERT, etc., provide a strong foundation for achieving fine-grained semantic alignment through comparative learning and deep attention mechanisms. At the enhancement level, representation learning methods that integrate graph structures with multimodal content have been developed, as well as cross modal inference techniques based on this foundation. However, current research still faces many challenges, including insufficient semantic alignment accuracy, scarcity of multimodal data, and low model reliability. Looking ahead to the future, the development of multimodal knowledge graphs will closely interact with the wave of cutting-edge artificial intelligence technologies, and MMKGs will inevitably become an indispensable core component of future artificial intelligence infrastructure.

References

[1] Chen X, Jia S, Xiang Y. A review: Knowledge reasoning over knowledge graph [J]. Expert systems with applications, 2020, 141: 112948.
[2] Tian X, Xu J, Wang S, et al. Construction and Application of a Multi-Modal Knowledge Graph Integrated with Large Language Models in the Field of Manufacturing Processes [C]//2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2025: 0288-0292.
[3] Chen L, Li Z, Wang Y, et al. MMEA: entity alignment for multi-modal knowledge graph [C]//International conference on knowledge science, engineering and management. Cham: Springer International Publishing, 2020: 134-147.
[4] Cong Shuo A review of knowledge graph entity alignment research: from traditional methods to cutting-edge technologies Computer Engineering and Applications. doi: 10.3778/j.issn.1002-8331.2412-0384.
[5] Shi Zhenpu, Lv Xiao, Dong Yanru, etc Research on the Development Status of Multimodal Knowledge Graph Fusion Technology in the Medical Field [J]. Computer Science and Exploration, 2025, 19 (07): 1729-1746
[6] Shan Rongzhen Research and Application of Multimodal Knowledge Graph Fusion Algorithm [D]. University of Electronic Science and Technology of China, DOI: 10.27005/d.cnki.gdzku.2025.002264.
[7] Wu Shiwei Research on Semantic Understanding Methods for Multimodal Video Scenes [D]. University of Science and Technology of China, 2025. DOI: 10.27517/d.cnki.gzkju.2025.000091
[8] Huang Jinbo. Research on Image Description Method Based on CLIP Pre trained Model [D]. Guizhou Normal University, DOI: 10.27048/d.cnki.ggzsu.2025.001345.