



**Michigan  
Technological  
University**

Michigan Technological University  
**Digital Commons @ Michigan Tech**

---

Michigan Tech Publications

---

4-9-2023

## Multi-modal knowledge graph inference via media convergence and logic rule

Feng Lin  
*Beijing Forestry University*

Dongmei Li  
*Beijing Forestry University*

Wenbin Zhang  
*Michigan Technological University, wenbinzh@mtu.edu*

Dongsheng Shi  
*Beijing Forestry University*

Yuanzhou Jiao  
*Beijing Forestry University*

*See next page for additional authors*

Follow this and additional works at: <https://digitalcommons.mtu.edu/michigantech-p>



Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Lin, F., Li, D., Zhang, W., Shi, D., Jiao, Y., Chen, Q., Lin, Y., & Zhu, W. (2023). Multi-modal knowledge graph inference via media convergence and logic rule. *CAAI Transactions on Intelligence Technology*.

<http://doi.org/10.1049/cit2.12217>

Retrieved from: <https://digitalcommons.mtu.edu/michigantech-p/17088>

Follow this and additional works at: <https://digitalcommons.mtu.edu/michigantech-p>



Part of the [Computer Sciences Commons](#)

---

**Authors**

Feng Lin, Dongmei Li, Wenbin Zhang, Dongsheng Shi, Yuanzhou Jiao, Qianzhong Chen, Yiyang Lin, and Wentao Zhu

# *CAAI Transactions on Intelligence Technology*

## Special issue Call for Papers

---

**Be Seen. Be Cited.  
Submit your work to a new  
IET special issue**

Connect with researchers and experts in your field and share knowledge.

Be part of the latest research trends, faster.


[Read more](#)



The Institution of  
Engineering and Technology

## ORIGINAL RESEARCH

# Multi-modal knowledge graph inference via media convergence and logic rule

Feng Lin<sup>1,2</sup>  | Dongmei Li<sup>1,2</sup> | Wenbin Zhang<sup>3</sup> | Dongsheng Shi<sup>1,2</sup> | Yuanzhou Jiao<sup>1,2</sup> | Qianzhong Chen<sup>1,2</sup> | Yiyi Lin<sup>1,2</sup> | Wentao Zhu<sup>1,2</sup>

<sup>1</sup>School of Information Science and Technology, Beijing Forestry University, Beijing, China

<sup>2</sup>Engineering Research Center for Forestry-oriented Intelligent Information Processing, National Forestry and Grassland Administration, Beijing, China

<sup>3</sup>Michigan Technological University, Houghton, Michigan, USA

**Correspondence**

Dongmei Li.  
Email: lidongmei@bjfu.edu.cn

**Funding information**

National College Students' Training Programs of Innovation and Entrepreneurship, Grant/Award Number: S202210022060; the CACMS Innovation Fund, Grant/Award Number: CI2021A00512; the National Nature Science Foundation of China under Grant, Grant/Award Number: 62206021

**Abstract**

Media convergence works by processing information from different modalities and applying them to different domains. It is difficult for the conventional knowledge graph to utilise multi-media features because the introduction of a large amount of information from other modalities reduces the effectiveness of representation learning and makes knowledge graph inference less effective. To address the issue, an inference method based on Media Convergence and Rule-guided Joint Inference model (MCRJI) has been proposed. The authors not only converge multi-media features of entities but also introduce logic rules to improve the accuracy and interpretability of link prediction. First, a multi-headed self-attention approach is used to obtain the attention of different media features of entities during semantic synthesis. Second, logic rules of different lengths are mined from knowledge graph to learn new entity representations. Finally, knowledge graph inference is performed based on representing entities that converge multi-media features. Numerous experimental results show that MCRJI outperforms other advanced baselines in using multi-media features and knowledge graph inference, demonstrating that MCRJI provides an excellent approach for knowledge graph inference with converged multi-media features.

**KEYWORDS**

logic rule, media convergence, multi-modal knowledge graph inference, representation learning

## 1 | INTRODUCTION

Media convergence can take advantage of the multi-media nature of things to provide people with richer information. With the development of multi-media technologies, research on media convergence is actively being carried out. Knowledge is formed in different areas of convergence [1]. Multi-media information essentially refers to multi-modal information provided through text, images, and video in various media. There is also a growing propensity to predict the course of social events or the emotional tendencies of a particular person using the multi-modal features of various media. For this task, multi-modal knowledge graphs (MKGs) with multi-modal information have drawn a lot of attention.

Conventional KG can clearly show the relationships among entities in the real world in the form of triples, but their modalities are single and cannot fully cover real-world knowledge. The concept of MKG has been proposed as shown in Figure 1, where the MKG links multi-modal information of various media on the corresponding entities and addresses this issue to a certain extent. However, practically, each entity gives information of different media very different attention, so simply including media features in the knowledge graph does not guarantee that they will be effectively used. Most KGs combining multi-media features do not currently take this into account. This just adds the multi-media features to the knowledge graph without fully utilising the multi-media information. Besides, as one of the main knowledge graph inference methods

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

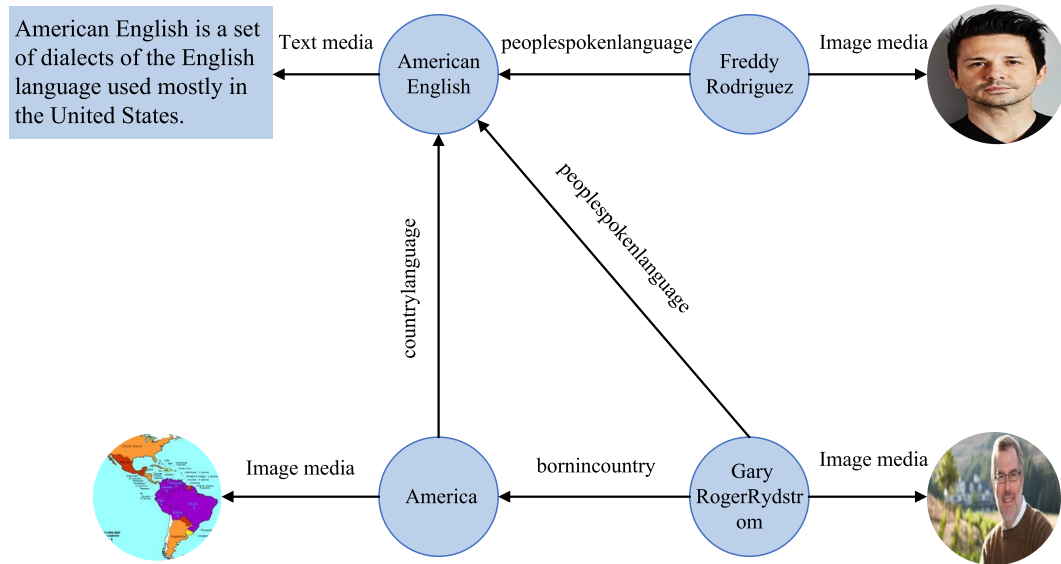


FIGURE 1 An example of multi-modal knowledge graph (MKG).

currently available, inference based on representation learning can map the entities and relations in KGs to a low-dimensional space to obtain the corresponding vectors. Some representation learning methods, such as TransH [2] and TransR [3], can solve one-to-many and many-to-many problems to a large extent. However, these problems are significantly amplified when applying these methods to the inference of MKG. Multi-modal knowledge graphs introduce a large number of other modal features (e.g. pictures and text descriptions), leading to a significant increase in knowledge graph complexity. As a result, the efficiency of the current representation model and the accuracy of inference results are reduced.

Multi-modal knowledge graphs increase the complexity of KGs, but they are unable to make up for the information limitations of conventional KGs. In conventional knowledge graph representation learning, vectors can only represent the semantic relationships between different entities. But in fact, the discovery of new facts often rely on features of things themselves, such as images, and text descriptions. Therefore, multi-media information in the KG can improve the accuracy and effectiveness of reasoning [4]. The introduction of these multi-media features can improve the link prediction task efficiency by enriching the information of the entities, but it does not make it interpretable.

Multi-headed self-attention is used to address the issue of not being able to fully utilise multi-media features and the impact of multi-media feature introduction on the representation model. Additionally, some conventional KG representation learning methods purely consider a single triple. Paths in KGs always play an important role in providing additional relationships between entities [5]. Considering that the accuracy and interpretability of the additional semantic information of the logic rules will greatly improve the effectiveness of our model if we can take advantage of it. Figure 2 illustrates an example of logic rules applied to representation learning; rule R2 ( $x, \text{BornInCountry}, z \wedge (y, \text{CountryLanguage}, z) \rightarrow (x,$

$\text{PersonSpokenLanguage}, y)$  and rule R1 ( $(x, \text{PersonMotherTongue}, y) \rightarrow (x, \text{PersonSpokenLanguage}, y)$ ) can be used to iteratively compose the path into a triple (Freddy Rodriguez, PersonSpokenLanguage, American English). Therefore, we introduce additional semantic information of logic rules and use logic rules to combine relations into paths, which substantially improves both the accuracy of knowledge graph path representation learning and the explainability of representation learning.

This paper proposes the Media Convergence and Rule-guided Joint Inference (MCRJI) model. Different media information is combined by learning how much attention entities pay to different media features through multi-headed self-attention. Besides, logic rules are used to combine paths and association relations at the semantic level for representation learning. Finally, entity embeddings converging multi-media feature information are used for link prediction, making full use of the different media features of entities. In this work, our main contributions can be summarised as follows:

- To the best of our knowledge, this is the first attempt to combine media convergence with logic rules for MKG inference, increasing the available information while improving the explainability of the inference.
- Our proposed MCRJI model fully considers multi-media features. It uses multi-headed self-attention to converge different media features of entities and add them to their vector representation. Finally, the inference is performed based on the new entity representation guided by logic rules. In other words, we make full use of the features and logic rules of different media, thereby the improving link prediction efficiency and interpretability.
- We conduct a large number of experiments on link prediction for MKG, and the MCRJI model achieves good performance. The impact of various rule confidence levels shows how the effective use of rules and multi-media

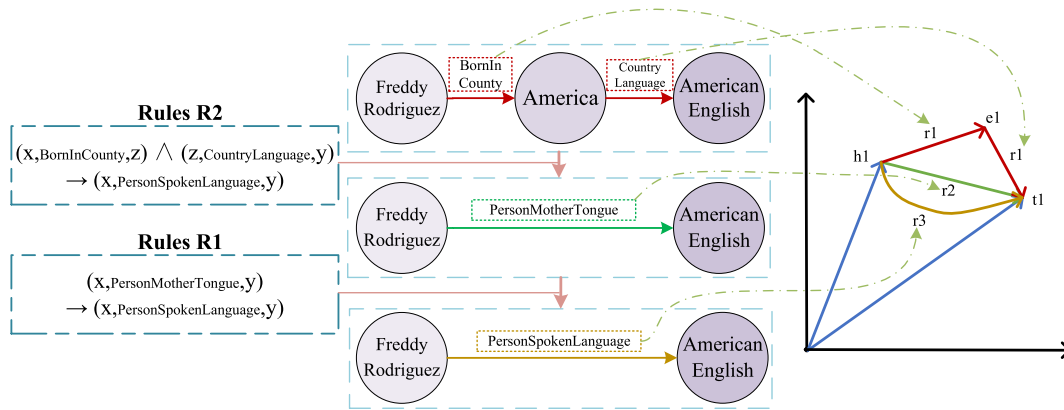


FIGURE 2 An example of logic rules applied to representation learning.

features is ensured by the confidence level of the rules considered in the model. Additionally, our model has good robustness to various confidence levels.

## 2 | RELATED WORK

### 2.1 | Multi-media information convergence

The variety of information that people are exposed to daily is increasing as a result of technological advancement. Media convergence is also gaining traction. Multi-media information convergence allows machines to make full use of multi-modal information, similar to how humans perceive the world using the same multi-sensory information that they use for vision, hearing, smell, and touch. In the field of multi-modal information convergence, numerous models have recently been put forth to predict information, broaden its scope, and improve the accuracy of results and robustness of models. For instance, Moon et al. [6] extract features from images and text using Convolutional Neural Networks [7] and Long Short-Term Memory [8]. Then, simple attention is used to fuse multi-modal information to disambiguate named entities using multi-modal information. Yan et al. [9] propose a video captioning framework based on object-relational graphs and multi-modal feature convergence, which uses a multi-modal feature convergence network to combine features of different modalities.

With the deep application of multi-modal information, Aljunid et al. [10] propose a multi-model deep learning approach for collaborative filtering recommender systems. Sun et al. [11] propose the MKG attention network (MKGAT) to improve the efficiency of recommendations made by recommendation systems. MKG attention network includes MKG embedding and recommendation modules, where the MKG embedding module uses an entity encoder and attention layer to learn a new representation for each entity. In MKG's attention, the add and concatenation aggregation methods are proposed for the convergence of multi-modal information. As a result, it becomes possible for the new entity to fuse the information of nearby entities while retaining its information. This converged modal

entity can be used to express knowledge inference relations. However, this model further increases KG complexity while introducing a large amount of other modal information, which results in reducing the efficiency of representation learning.

To give different attention to the information from different modalities, Wang et al. [4] propose the Multi-modal knowledge graphs representation learning via multi-headed self-attention (MKGRL-MS) model for fusing multi-modal information. The features of image and text modalities are encoded using ResNet and RoBERTa-www-ext. In particular, a multi-headed self-attention is used to obtain the attention of different modal features, and consequently a new entity representation, which is the sum of entity representation and multi-modal feature representation of entities.

However, the above-mentioned models only enrich the information of entities through multi-media features. The fact that only a single triple is still considered in representation learning does not make the multi-media information-based prediction task interpretable.

### 2.2 | Rule employment for knowledge graphs

Logic rules contain rich semantic information and are interpretable. If we want to apply logic rules in KG inference, we must first pre-define a rule set for KGs and use it to infer the facts that are already present in the KGs. However, the set of rules used in this approach is usually incomplete when dealing with KGs with more complex structures, and different rules always infer some utterly contradictory conclusions. Therefore, several methods have been proposed to discover rules from KGs, including AMIE [12], AMIE+ [13], RLvLR [14], and CARL [15]. Richardson and Domingos [16] propose a Markov Logic Network by combining Markov random field networks and first-order logic. It implements uncertainty inference by assigning learnable weights to rules. Not coincidentally, Bayesian Logic Programming [17] uses Bayesian networks to demonstrate logic rules. In addition, they achieve inference by discovering probabilistic relationships between these variables.

The inference results of these methods are usually explainable. However, these techniques are less efficient when the KG is large and complex in structure. More importantly, the sensitivity of the rules results in frequent failure to infer missing valid triples.

Minervini et al. [18] impose equivalence and inverse constraints on relational embedding to improve the efficiency and accuracy of KG inference, but this approach is not universal because it only considers two constraints between relations instead of general rules. Guo et al. in KALE [19] obtain logic rules from t-norm and convert the rules into complex equations formed by triples. However, the interpretability and accuracy of the logic rules are reduced while converting them into complex equations.

## 2.3 | Knowledge graph inference

### 2.3.1 | Conventional knowledge graphs inference

Knowledge inference is the process of inferring unknown facts or relations from known ones in a graph. There are three main forms of inference in KG: representation learning-based, neural network-based, and rule-based inference. Besides, there is also a hybrid inference approach, as the name implies, which combines multiple methods for inference with complementary advantages. Among these methods, representation learning-based inference and hybrid inference have received wide attention because of their effectiveness.

Representation learning-based inference automatically captures the features required for inference without instructing the inference step, so this approach is not interpretable. TransE [20] is widely used in representation learning and is considered the benchmark for KG representation learning. It is assumed that the distance between the tail entity and the head entity embedding is roughly equal to the distance of the relationship embedding. However, TransE cannot accurately represent complex relationships in KG, such as “one-to-many” and “many-to-one”. To address this issue, a series of more advanced models have also been proposed, such as TransH [2], TransR [3], TransD [21], and TransG [22]. TransH [2] is the first to project the entity representation onto the hyperplane of a particular relationship. TransR [3] introduces the space of a particular relationship through a projection matrix. The distance is then computed on the space. TransD [21] makes more improvements. It makes itself more efficient by dynamically generating the projection matrix through two vectors. TransG [22] also considers uncertainty by introducing a Gaussian distribution. These methods are very efficient and scalable but represent unsatisfactory learning results because of their simple loss functions [23].

The logic rules in KG contain a wealth of information that can greatly enhance the effectiveness of representation learning. Therefore, in recent years, several rule-enhanced hybrid approaches have been introduced that can address the drawbacks of both rule-based and representation learning-based approaches. For instance, Guo et al. [24] propose the

rule-guided embedding approach, which iteratively models the observation of triples in a knowledge graph. Similarly, Zhang et al. [25] propose an iterative embedding approach through representation learning, equation induction, and injection. However, these models require the use of methods that approximate embedding results [22] or t-norm fuzzy logic [26] approaches. Therefore, these approaches are not suitable for use in large-scale KGs with complex structures. To address these issues, Niu et al. [5] propose a Rule and Path Joint Embedding model (RPJE), which makes full use of logic rules to enhance the effectiveness and explainability of representation learning. Specifically, the logic rules mined from KG are first encoded as path rules. Paths are then composed using the encoded rules and representation learning is performed to ensure that the logic rules are well interpretable. Tang et al. [27] propose the RULE model, which enables the embedding of pretrained logic rule information into the vector space to improve the reliability of the KG embedding. In addition, RULE improves the inference process by learning the confidence scores of the rules and controlling their weights.

### 2.3.2 | Multi-modal knowledge graphs inference

Currently, the majority of the MKG inference methods use multi-modal knowledge while learning the representation of entities and relations. Multi-modal knowledge graph inference models mainly include translation-based and neural network-based models. The translation-based model introduces multi-modal information based on the conventional translation model for knowledge inference based on representation learning. Xie et al. [28] propose the Image-embodied knowledge representation learning model considering the visual information of entities, which combines images and KG for the first time for knowledge graph representation learning. Hatem et al. [29] propose a translation model that defines the scoring function of a knowledge graph as the sum of three scoring functions: structural knowledge, visual knowledge, and textual knowledge. Wang et al. [30] propose the TransAE, which combines self-encoder and TransE to learn MKG representation for knowledge inference. Lu et al. [31] propose the Multi-modal knowledge graph representation learning model, which introduces a multi-modal knowledge alignment scheme to correlate and merge multi-modal knowledge and uses an adversarial training strategy to enhance its robustness. Ning et al. [32] propose the PDRL model, which combines relational paths in the knowledge graph with entity description information to improve model performance.

A MKG inference model for basal neural networks is based on neural networks that are treated as scoring functions for knowledge graph inference. Zhang et al. [33] propose a multi-modal multi-relational feature aggregation network for medical knowledge graph representation learning. For the multi-modal content of entities, an adversarial feature learning model is used to learn multi-modal entity public representations by mapping text and image information of entities into the same vector space. Tang et al. [34] propose a multisource knowledge graph

representation learning model MKRL that exploits both structural knowledge of KG and multi-modal knowledge of hierarchical types, textual relations, and entity descriptions.

### 3 | METHOD

This section presents in detail our proposed MCRJI model shown in Figure 3. Media Convergence and Rule-guided Joint Inference model consists of four main sub-modules: feature matrix coding of multi-media, media convergence based on multi-headed self-attention, rules employment for representation learning, and KG inference with the feature of multi-media based on representation learning. The upper left part is the feature matrix coding of multi-media, the lower left part is the media convergence based on multi-headed self-attention, the upper right part is the rules employment for representation learning, and the lower right part is the KG inference with the feature of multi-media based on representation learning. These four sub-modules will be presented in detail in the sequel.

#### 3.1 | Feature matrix coding of multi-media

We use different encoders for image and text descriptions to obtain feature vectors for the different media of the entity, as follows:

**Images:** For image encoding, we use the residual network model ResNet50. First, we unify the size of the images and

adjust the last fully connected layer. Finally, the feature vectors of the image media are obtained.

**Text:** The pretrained BERT-large-cased language model is used for text encoding. It is pretrained on a large-scale unlabelled corpus using self-supervised learning methods to capture the rich semantic information in the text.

Through the above operations, we extract the eigenvectors of the image and text description. Meanwhile, to reduce the training overhead and noise, we perform Principal Component Analysis processing on the obtained feature vectors. Finally, we splice the processed feature vectors into a feature matrix of entities.

#### 3.2 | Media convergence based on multi-headed self-attention

In Section 3.1 we completed a simple stitching of various media information features. However, each entity actually pays different attention to different media. While some entities may focus more on visual media features such as image and video media, others may concentrate more on text media, which is a specific description of the entity. Entities in positive triples tend to have different attention to their media features. Therefore, we use multi-headed self-attention [4] to converge multi-media information of entities, which allows the features of different media to interact with each other. As shown in (1), the dot product is used to calculate the attention score as follows:

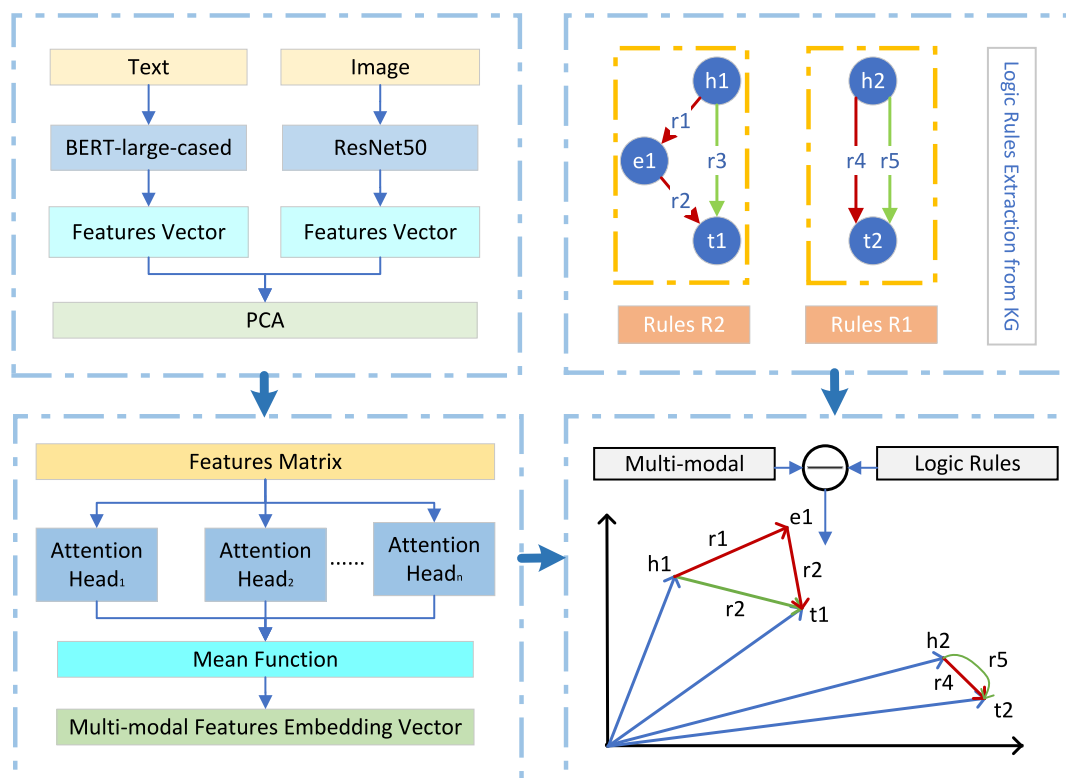


FIGURE 3 Overview of the Media Convergence and Rule-guided Joint Inference model (MCRJI)



$$M^* = \text{softmax} \left( \frac{W_q \cdot m^* \cdot (W_k \cdot m^*)^T}{\sqrt{D_k}} \right) \cdot W_v \cdot m^*, \quad (1)$$

where  $W_q$  is the query matrix,  $W_k$  is the key matrix,  $W_v$  is the value matrix,  $m^*$  is the multi-media feature matrix obtained after feature encoding, and  $M^*$  is the multi-media feature matrix obtained by multi-headed self-attention.

Finally, we further integrate the multi-media information, converging the multi-media feature matrix into a vector that represents a multi-media feature. The advantage of doing so is to give the multi-media features the same dimensionality as the distributed representation of entities, which facilitates the subsequent representation learning. The formal expression is shown in (2):

$$M = \frac{\sum_{n=1}^n M_n^*}{n}, \quad (2)$$

where  $M$  is the multi-media feature vector of the entity,  $n$  is the number of unimodal feature vectors, and  $M_n^*$  is the division of  $M^*$ .

### 3.3 | Rules employment for representation learning

We discover the paths from KG and use AMIE to mine the KG for the implied rules and their confidence levels. To ensure the efficiency of discovering rules, we set the length of the rules to be no greater than 2 and add the inverse version of each relation to the knowledge graph. Thus, we define a new triple  $(t, r^{-1}, h)$  to represent the inverse relationship  $r^{-1}$  in the original triple  $(h, r, t)$ .

#### 3.3.1 | Rule of length 1

The rule of length 1 can directly discover the semantic similarity between relations. There are many relations with semantic similarity implied in KG. Besides, the inverse relations we add may have semantic similarities with the original relations. Therefore, we should make the embedding of these relations closer during the training process. The formal expression is shown in (3):

$$\text{score}_1(r, r_e) = \|r - r_e\|, \quad (3)$$

where  $r$  and  $r_e$  are two relations in the set of relations.  $\text{score}_1(r, r_e)$  denotes the scoring function of the similarity between a relation  $r$  and another relation  $r_e$ . If  $r_e$  and  $r$  are the relations covered by the rule, they should be assigned with a smaller score.

#### 3.3.2 | Rule of length 2

The rule of length 2 can combine two relations at a time by traversing the paths up until the rule is unable to combine the

relations. We implement the PTransE [35] path extraction process on KG. Second, since the rules mined by AMIE cannot be used directly in paths, we convert the original rules into chain rules to be used directly in paths. Taking the original rule  $(c, r1, a) \wedge (c, r2, b) \rightarrow (a, r3, b)$  as an example, we first convert  $(c, r1, a)$  to  $(a, r1^{-1}, c)$ . Then, we exchange the order of two relations to get the rule  $(a, r1^{-1}, c) \wedge (c, r2, b) \rightarrow (a, r3, b)$ , and carry out abbreviation to get the rule  $(r1^{-1}, r2) \rightarrow r3$ . Finally, we can get the path  $r1^{-1} \rightarrow r2$  composed as  $r3$ .

To make full use of the rules, we should traverse the paths and iteratively combine those using rules of length 2 until the rules are unable to combine the relations. This method of obtaining long paths by iterative combination of rules of path 2 is to avoid the unreliable rules caused by too long paths. In addition, when multiple rules can be matched in the path, the rule with the highest confidence level should be selected. The formal expression is shown in (4):

$$\text{score}_{\text{path}}(p, r) = \text{con}(p|h, t) \left( \prod_{u_i \in \text{rule}(p)} u_i \right) \| \text{com}(p) - r \|, \quad (4)$$

where  $\text{score}_{\text{path}}(p, r)$  denotes the scoring function between the path  $p$  and the relation  $r$ ,  $\text{con}(p|h, t)$  denotes the confidence of path  $p$  from entity pair  $(h, t)$ ,  $h$ ,  $r$ , and  $t$  are the respective embeddings of head, relation, and tail entities,  $\text{com}(p)$  is the composition result of path  $p$ , and  $\text{rule}(p)$  denotes the confidence set of the logic rules in the combined path process.

### 3.4 | Knowledge graph inference with feature of multi-media based on representation learning

By converging features of multi-headed self-attention, we propose a distributed representation as a sum of semantic and multi-media feature representations of entities, where the multi-media features of entities are represented in (2) while semantic features of entities are conventional forms of knowledge representation.

Particularly, our model uses logic rules to learn the representation of entities based on the convergence of multi-media features, thereby transforming the scoring function into the following equations:

$$\begin{aligned} \text{score}_2(p, r) &= \text{con}(p|h - F_h, t - F_t) \left( \prod_{u_i \in \text{rule}(p)} u_i \right) \\ &\quad \times \| \text{com}(p) - r \|, \end{aligned} \quad (5)$$

$$\text{score}_3(h, r, t) = \| h - F_h + r - (t - F_t) \|, \quad (6)$$

where  $F_h$  and  $F_t$  are multi-media feature vectors.  $\text{score}_2$  incorporates multi-media features based on  $\text{score}_{\text{path}}(p, r)$ , and  $\text{score}_3$  is a representation learning of these entities and relations using the conventional additive method when there are no

available rules among entities. When no rules are available, the representation learning of KG is similar to transE, but this approach fuses the multi-media feature vectors of entities.

The loss function used to formalise the training objective of the MCRJI model is given by:

$$\text{loss} = \sum_{(h,r,t) \in T} \left[ \text{loss}_1(h,r,t) + \alpha_1 \sum_{p \in P(h,t)} \text{loss}_2(p,r) + \alpha_2 \sum_{r_e \in D(r)} \text{loss}_3(r,r_e) \right], \quad (7)$$

where  $D(r)$  is the set of relations derived from  $r$  on the rule of length 1,  $r_e$  is a relation in  $D(r)$ ,  $P(h,t)$  denotes all paths connecting entity pairs  $(h,t)$ , and  $P$  is one of the paths in  $P(h,t)$ , while  $\alpha_1$  and  $\alpha_2$  are two hyperparameters that weight the influence of paths and embedded constrained pairs of relations, respectively,  $\text{loss}_1(h,r,t)$ ,  $\text{loss}_2(p,r)$ , and  $\text{loss}_3(r,r_e)$  are three loss functions, which are defined according to (3), (5), (6) as follows:

$$\text{loss}_1(h,r,t) = \sum_{(h',r',t') \in S^-} (\gamma_1 + \text{score}_1(h,r,t) - \text{score}_1(h',r',t')), \quad (8)$$

$$\text{loss}_2(p,r) = \sum_{r' \in S^-} (\gamma_2 + \text{score}_2(p,r) - \text{score}_2(p,r')), \quad (9)$$

$$\text{loss}_3(r,r_e) = \sum_{r' \in S^-} (\gamma_3 + \beta \text{score}_3(r,r_e) - \text{score}_3(r,r')), \quad (10)$$

where  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$  are hyperparameters that denote the margins of the loss function in Eqs,  $\beta$  denotes the confidence of the rule that associates  $r$  and  $r_e$  in the rules.  $S$  is a set of positive triples and  $S^-$  is a set of negative triples. The negative triple is created by randomly replacing one entity in a positive triple (i.e.,  $S^- = \{(h',r,t) | h' \in E\} \cup \{(h,r,t') | t' \in E\}$ ).

## 4 | EXPERIMENTS

In this section, four datasets are used to evaluate the MCRJI model. Our datasets and rules are first introduced in Section 4.1, and the experimental setup is then described in detail in Section 4.2. In Section 4.3, we will discuss the experimental results.

### 4.1 | Datasets and rules

We validate MCRJI model on four datasets: FB15K and FB-*img* from large-scale Freebase, DB15K from DBpedia15K after alignment, and WN18 extracted from WordNet. The statistics of the used datasets are shown in Table 1.

TABLE 1 Statistics of datasets.

Datasets	FB15K	FB- <i>img</i>	DB15K	WN18
Entities	14 951	11,757	12,842	40,943
Relationships	1345	1231	279	18
Train	483 142	285,850	79,223	141,442
Valid	50 000	29,580	9903	5000
Test	59 071	34,863	9902	5000
Text	14,951	9683	11,539	38,783
Image	40,237	28,035	30,786	70,651

TABLE 2 Statistics of rules in various confidence.

Datasets	FB15K	FB- <i>img</i>	DB15K	WN18
#0.5	27,879	21,813	1467	105
#0.6	24,807	19,396	1289	97
#0.7	21,685	16,792	1088	97
#0.8	18,046	14,117	867	41
#0.9	14,723	11,850	742	17

Any rule-mining tool can be used to generate the rules applied in our model. We chose AMIE+ [13] because it is convenient and mature enough to mine logic rules with different confidence levels in different KG. Table 2 lists the statistics of rules with different confidence levels mined from FB15K, FB-*img*, DB15K, and WN18 in the range [0.5,0.9] in steps of 0.1, which are encoded for representation learning.

## 4.2 | Experimental setup

### 4.2.1 | Evaluation protocol

Link prediction is considered as an evaluation criterion for model efficiency in the field of KG inference. In this task, we need to replace the entities (h or t) in the test set with entities from the dictionary. These triples that are replaced are called negative triples. Then, the scoring function of the model is used to score these negative triples and sort them in descending order. The higher the rank of the triples consisting of the correct entities, the better the model is at predicting the entities.

There are two main evaluation metrics for the link prediction task: mean rank and Hits@10. After the scoring was completed, we determine the positive triples and their rankings. In the test set, the average rank of positive triples is defined as mean rank. Hits@10 is the probability that a positive triple is ranked in the top 10.

When constructing negative triples, some of them will belong to the KG. This will affect the results of model evaluation. Therefore, we remove these negative triples from the training, valid, and test sets to ensure fairness in evaluation. The evaluation without the above operation is called Raw, and the evaluation with the filtering operation is called Filt.

## 4.2.2 | Baselines for comparison

We establish two baselines to verify the validity of MCRJI.

The first baseline is made up of three other inference models. We select three advanced models for link prediction experiments, including three types of baselines: (1) TransE [20], no consideration of logic rules and multi-modal features, and only a single triple. (2) MKGRL-MS [4], with consideration multi-modal features but not logic rules. (3) RPJE [5], with consideration of logic rules but not multi-modal features. We experiment with their source code on several public datasets.

The second baseline is the confidence level of the logic rules. In the MCRJI model, a higher confidence level of the rules does not necessarily imply a better model because higher confidence rules represent a smaller number of available rules, which can make the MCRJI model less effective when there are too few available rules. Therefore, a trade-off should be made between higher confidence and more rules to achieve the best model effect.

## 4.2.3 | Parameter setting

We set parameters that remain constant throughout the experiment (e.g., dimension = 100, epochs = 1000, norm = 2, weight decay = 0.001, patience = 10, learning rate = 0.01, attention-heads = 2, confidence = 0.7, and stop = 5). Confidence = 0.7 means that the confidence level of the logic rule we use is greater than or equal to 0.7, and patience = 10 indicates that the validity of the model did not improve for 10 consecutive times in the validation set. The learning rate and weight decay will be changed. Stop = 5 indicates that the model stops training and outputs a result when the learning rate and weight decay are changed more than 5 times. In addition, we use the Stochastic Gradient Descent optimiser.

For the hyperparameters, we choose  $\gamma_1 = \gamma_2 = 1$ . Besides, we manually adjust  $\gamma_3$  in the set {1, 1.5, 2, 2.5, 3}, and the weight coefficients  $\alpha_1, \alpha_2$  in the set {0.5, 1, 1.5, 2, 3, 5}. The best model is selected on the validation set. The obtained optimal  $\gamma_3, \alpha_1$ , and  $\alpha_2$  are assigned as:  $\gamma_3 = 1, \alpha_1 = 1$ , and  $\alpha_2 = 3$ .

## 4.3 | Experimental results

We interpret the experimental results in terms of both the validity of MCRJI and the optimal value of the rule confidence level. They correspond to the first baseline and the second baseline, respectively.

### 4.3.1 | Effectiveness of Media Convergence and Rule-guided Joint Inference model

We conduct preliminary experiments on four datasets, FB15K, FB-img, DB15K, and WN18. The experimental results are shown in Table 3 and Table 4.

**TABLE 3** Comparison of TransE, Multi-modal knowledge graphs representation learning via multi-headed self-attention (MKGRL-MS), and Media Convergence and Rule-guided Joint Inference model (MCRJI) on FB15K and FB-img.

Dataset	FB15K				FB-img			
	Mean rank		Hits@10		Mean rank		Hits@10	
	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
TransE	348	271	9.4	16.6	282	227	13.1	20.1
TransH	370	294	10.1	16.7	297	243	13.4	19.9
MKGRL-MS	338	261	9.1	16.2	268	214	12.5	19.7
RPJE	298	219	12.3	20.6	199	135	16.9	26.7
MCRJI	<b>295</b>	<b>216</b>	11.8	20.1	<b>182</b>	<b>117</b>	17.8	28.5

Note: The bold values are the experimental results of our proposed model, which achieve better results than other baselines.

**TABLE 4** Comparison of TransE, Multi-modal knowledge graphs representation learning via multi-headed self-attention (MKGRL-MS), and Media Convergence and Rule-guided Joint Inference model (MCRJI) on DB15K and WN18.

Dataset	DB15K				WN18			
	Mean rank		Hits@10		Mean rank		Hits@10	
	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
TransE	1263	1017	10.9	12.5	463	451	53.6	61.0
TransH	1291	1050	10.5	11.8	715	702	42.8	49.1
MKGRL-MS	1155	922	9.8	11.5	449	437	52.8	59.8
RPJE	1198	949	10.1	12.5	315	303	50.9	61.6
MCRJI	<b>1141</b>	<b>899</b>	8.2	9.9	<b>301</b>	<b>289</b>	51.1	61.4

Note: The bold values are the experimental results of our proposed model, which achieve better results than other baselines.

First, we point out that the two datasets in Table 3 are from the same set, which can be regarded as two different subsets of large-scale Freebase. The datasets in Table 4 are from different sets. Secondly, it can be seen that MCRJI has improved the mean rank in the experiments of each dataset and outperforms other existing models. This also tentatively confirms the validity of MCRJI. In addition, we need to note the following two points:

- (1) On these four datasets, MCRJI has an improvement on the indicator Hits@10 only for FB-img, while the other datasets have varying degrees of decline. We need to conduct more experiments to find out why this is the case.
- (2) For the datasets in Table 4, the results from Filt do not obtain a significant improvement over those from Raw, which indicates that filtering negative triples has little effect on them. For the two datasets from large-scale Freebase in Table 3, the results of Filt and Raw are significantly different.

To address the first point, we think there are two possibilities. First, if some positive triples improve their rankings but do not make to the top 10, there will be no significant impact on Hits@10. However, the indicator mean rank is not too strict for the experimental results. Once the rank of the triple

changes, the mean rank captures the change in the data. Therefore, the mean rank is more representative of changes in the positive triple as a whole than Hits@10. Second, this phenomenon is likely caused by MCRJI's excessive focus on multi-media features, which causes some of the top 10 ranked positive triples to learn too much information about multi-media features and subsequently affect their own semantic information. This situation leads to a drop in the ranking of positive triples. However, this decline is only beyond the top 10 and is not significant. Therefore, the hit rate @10 decreases when the degree of overfitting of MCRJI increases. From this perspective, Hits@10 better reflects the degree of overfitting of the model compared to the mean rank.

As for the second point, it is caused by the characteristics of the dataset itself. Filt actually removes the occasional positive triples from the negative triples. Based on the experimental results, we think that more positive triples are removed in FB15K and FB-img from large-scale Freebase in Table 3, while fewer positive triples are removed from the dataset in Table 4. We find that the number of positive triples deleted by the filtering operation is related to the degree of one-to-many phenomena in the dataset. The more one-to-many phenomena in the dataset, the smaller the value of E/T (number of entities in the dataset/size of the training set) in that dataset, and the more effective the Filt operation is. The FB15K and FB-img

**TABLE 5** E/T value for datasets.

Datasets	FB15K	FB-img	WN18	DB15K
Entities	14,951	11,757	40,943	12,842
Train	483,142	285,850	141,442	79,223
E/T	0.03	0.04	0.29	0.16

datasets belong to the datasets with a lot of one-to-many phenomena, so the effect of Filt enhancement is very obvious, and the E/T values of the four datasets are shown in Table 5.

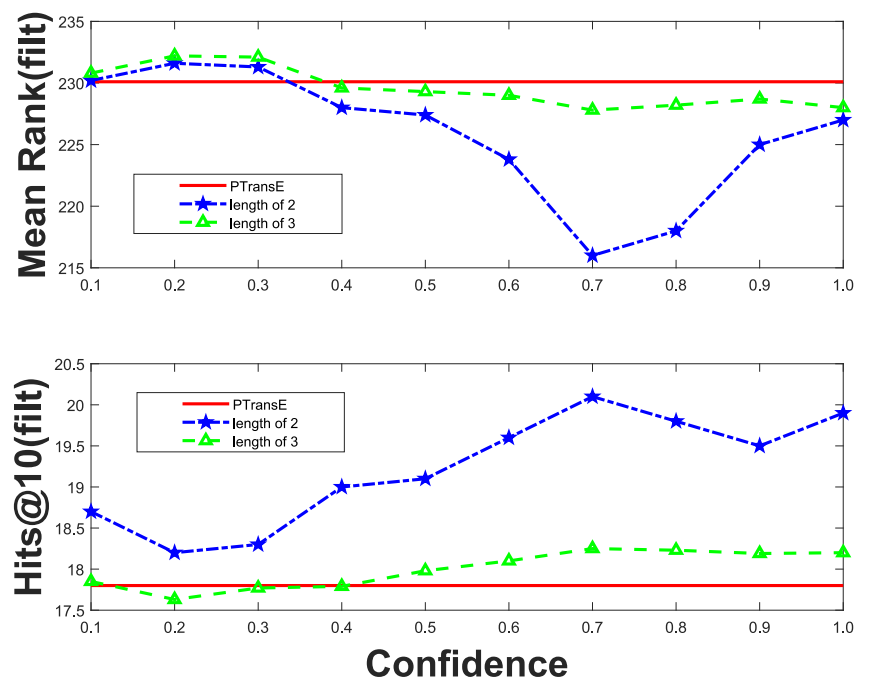
### 4.3.2 | Optimal value of rule confidence

In this section, we will explore the impact of rule confidence levels on our model. This is because we find that rules with various confidence levels have a direct impact on the experimental results in 4.3.1. Second, we will analyse the experimental results obtained at various levels of rule confidence.

Media Convergence and Rule-guided Joint Inference model is a model for representation learning under the joint guidance of logic rules and paths, and for reasoning based on representation learning. The confidence choice of the logic rules will directly affect the effect of representation learning. The rules' confidence level is not as high as it could be; a confidence level that is too high will result in fewer logic rules available, whereas a confidence level that is too low will affect the accuracy of the inference results, all of which will negatively impact the efficiency of our model. Based on this, we carry out additional experiments using various rules to further verify the validity of the MCRJI model. The results of the experiments on FB15K in Figure 4 lead to the following two conclusions:

When the confidence level of the rules is greater than 0.4, the MCRJI model outperforms PTransE for both mean rank and Hits@10, indicating that the rules utilised in MCRJI will be effective as long as the confidence level is in the medium range.

For the path length of the rule, it is found that the path length also has an impact on the effectiveness of the model. The rule with path length 2 consistently outperforms the rule with path length 3, with all other configurations such as confidence level being equal. The reason for this situation is that



**FIGURE 4** Performance comparison for different confidence levels and rule lengths.

longer paths result in less accurate path composition, which indicates poorer learning outcomes.

Therefore, in 4.3.1 we choose a rule confidence level of 0.7 and a path length of the rule no greater than 2 as the optimal settings for the experiments.

The effectiveness of MCRJI is indisputable, which proves that multi-media information can be integrated to offer richer information and a more accurate reasoning basis for knowledge inference. The use of logic rules enables representation learning to be no longer limited to a single triple and improves the interpretability of knowledge reasoning.

## 5 | CONCLUSIONS

In this paper, we propose a new model that uses media convergence techniques to combine different modal information of entities and learn new entity representations by introducing logic rules. Through link prediction experiments, we demonstrate that the introduction of multi-media features and logic rules is important for improving the accuracy and interpretability of knowledge graph inference tasks on multiple datasets.

In the future, MCRJI can be applied to multi-media communication prediction. Meanwhile, we will test whether the rules can be directly applied to other superior media convergence methods, and introduce some other mechanisms that could optimise the rules.

## ACKNOWLEDGEMENTS

This work is supported by National College Students' Training Programs of Innovation and Entrepreneurship (S202210022060), the CACMS Innovation Fund (CJ2021A00512), the National Nature Science Foundation of China under Grant (62206021).

## CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

Feng Lin  <https://orcid.org/0000-0002-5068-9876>

## REFERENCES

- Chung, K., Kim, J.S.: Multi-modal emotion prediction system using convergence media and active contents. *Personal Ubiquitous Comput.*, 1–11 (2021). <https://doi.org/10.1007/S00779-021-01602-8>
- Wang, Z., et al.: Knowledge graph embedding by translating on hyperplanes. *Proc. of the National Conf. on Artificial Intelligence.* 28(1), 1112–1119 (2014). <https://doi.org/10.1609/aaai.v28i1.8870>
- Lin, Y., et al.: Learning entity and relation embeddings for knowledge graph completion. *Proc. of the National Conf. on Artificial Intelligence* 3(1), 345–354 (2015). <https://doi.org/10.1016/j.procs.2017.05.045>
- Wang, E., et al.: Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Information Inf. Fusion* 88, 78–85 (2022). <https://doi.org/10.1016/j.inffus.2022.07.008>
- Niu, G., et al.: Rule-guided compositional representation learning on knowledge graphs. *Proc. AAAI Conf. Artif. Intell.* 34(3), 2950–2958 (2020). <https://doi.org/10.1609/aaai.v34i03.5687>
- Moon, S., Neves, L., Carvalho, V.: 'Multimodal named entity disambiguation for noisy social media posts'. *Annual Meeting of the Association for Computational Linguistics. Proc. of the Conf.* 1, 2000–2008 (2018). *Long Papers.* <https://doi.org/10.18653/v1/P18-1186>
- Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60(6), 84–90 (2017). <https://dl.acm.org/doi/abs/10.1145/3065386>
- Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* 9(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
- Yan, Z., et al.: Multimodal feature fusion based on object relation for video captioning. *CAAI Trans. Intell. Technol.* 8, 1–13 (2022). <https://doi.org/10.1049/cit2.12071>
- Aljunid, M.F., Doddaghatta Huchaiiah, M.: Multi-model deep learning approach for collaborative filtering recommendation system. *CAAI T. Intell. Technol.* 5(4), 268–275 (2020). <https://doi.org/10.1049/trit.2020.0031>
- Sun, R., et al.: Multimodal knowledge graphs for recommender systems. *International Conf. on Information and Knowledge Management, Proc.* 29, 1405–1414 (2020). <https://doi.org/10.1145/3340531.3411947>
- Galárraga, L.A., et al.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. *Proc. of the 22nd international Conf. on World Wide Web*, 413–422 (2013). <https://doi.org/10.1145/2488388.2488425>
- Galárraga, L., et al.: Fast rule mining in ontological knowledge bases with AMIE+. *The VLDB Journal* 24(6), 707–730 (2015). <https://doi.org/10.1007/s00778-015-0394-1>
- Omran, P.G., Wang, K., Wang, Z.: Scalable Rule Learning via Learning Representation, pp. 2149–2155. *IJCAI* (2018)
- Pellissier Tanon, T., et al.: Completeness-aware rule learning from knowledge graphs. In: *International Semantic Web Conf*, pp. 507–525. SpringerCham (2017)
- Richardson, M., Domingos, P.: Markov logic networks. *Mach. Learn.* 62(1), 107–136 (2006). <https://doi.org/10.1007/s10994-006-5833-1>
- Raedt, L.D., Kersting, K.: Probabilistic inductive logic programming. In: *Probabilistic Inductive Logic Programming*, vol. 4911, pp. 1–27. Springer (2008). [https://doi.org/10.1007/978-3-540-78652-8\\_1](https://doi.org/10.1007/978-3-540-78652-8_1)
- Minervini, P., et al.: Regularizing knowledge graph embeddings via equivalence and inversion axioms. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 668–683 (2017)
- Guo, S., et al.: Jointly embedding knowledge graphs and logical rules. In: *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*, pp. 192–202 (2016). <https://doi.org/10.18653/v1/D16-1019>
- Bordes, A., et al.: Translating embeddings for modeling multi-relational data. *Adv. Neural Inf. Process. Syst.* 26, 2787–2795 (2013)
- Ji, G., et al.: Knowledge graph embedding via dynamic mapping matrix. *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conf. on Natural Language Proc.*, 687–696 (2015)
- Xiao, H., Huang, M., Zhu, X.: A generative model for knowledge graph embedding. In: *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 2316–2325 (2016)
- Dettmers, T., et al.: Convolutional 2d knowledge graph embeddings. *Proc. AAAI Conf. Artif. Intell.* 32(1) (2018). <https://doi.org/10.1609/aaai.v32i1.11573>
- Guo, S., et al.: Knowledge graph embedding with iterative guidance from soft rules. *Proc. AAAI Conf. Artif. Intell.* 32(1) (2018). <https://doi.org/10.1609/aaai.v32i1.11918>
- Zhang, W., et al.: Iteratively Learning Embeddings and Rules for Knowledge Graph Reasoning, pp. 2366–2377. *The World Wide Web Conf.* (2019). <https://doi.org/10.1145/3308558.3313612>
- Hájek, P.: *Metamathematics of Fuzzy Logic.* Springer Science & Business Media (2013)

27. Tang, X., et al.: RuleE: Neural-Symbolic Knowledge Graph Reasoning with Rule Embedding (2022). <https://doi.org/10.48550/arXiv.2210.14905>
28. Xie, R., et al.: Image-embodied knowledge representation learning. In: Sierra C Proc. of the Twenty-Sixth International Joint Conf. on Artificial Intelligence, pp. 3140–3146 (2017)
29. Mouselly-Sergieh, H., et al.: A multimodal translation-based approach for knowledge graph representation learning. In: Proc. of the Seventh Joint Conf. on Lexical and Computational Semantics, pp. 225–234 (2018). <https://doi.org/10.18653/v1/S18-2027>
30. Wang, Z., et al.: Multimodal data enhanced representation learning for knowledge graphs. In: International Joint Conf. on Neural Networks, pp. 1–8. IEEE (2019)
31. Lu, X., et al.: MMKRL: a robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.* 52(7), 7480–7497 (2022). <https://doi.org/10.1007/s10489-021-02693-9>
32. Ning, Y., et al.: A knowledge graph representation learning method integrating relation path and entity description information. *J. Comput. Res. Dev.*, 1–14 (2022). <http://kns.cnki.net/kcms/detail/11.1777.TP.20211105.1048.003.html>
33. Zhang, Y., et al.: Multi-modal multi-relational feature aggregation network for medical knowledge representation learning. *Proc. of the 28th ACM International Conf. on Multimedia*, 3956–3965 (2020). <https://doi.org/10.1145/3394171.3413736>
34. Tang, X., et al.: Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. *Inf. Process. Manag.* 56(3), 809–822 (2019). <https://doi.org/10.1016/j.ipm.2019.01.005>
35. Lin, Y., et al.: Modeling relation paths for representation learning of knowledge bases. *Conf. on Empirical Methods in Natural Language Proc.*, 705–714 (2015)

**How to cite this article:** Lin, F., et al.: Multi-modal knowledge graph inference via media convergence and logic rule. *CAAI Trans. Intell. Technol.* 1–11 (2023). <https://doi.org/10.1049/cit2.12217>