Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(8): 14096–14116.
DOI: 10.3934/mbe.2023630
Received: 22 April 2023
Revised: 30 May 2023
Accepted: 04 June 2023
Published: 25 June 2023

Research article

MLSFF: Multi-level structural features fusion for multi-modal knowledge graph completion

Hanming Zhai[1], Xiaojun Lv[2], Zhiwen Hou[1], Xin Tong[1] and Fanliang Bu[1,*]

[1] School of Information Network Security, People's Public Security University of China, Beijing 100038, China

[2] Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China

* **Correspondence:** Email:bufanliang@sina.com.

Abstract: With the rise of multi-modal methods, multi-modal knowledge graphs have become a better choice for storing human knowledge. However, knowledge graphs often suffer from the problem of incompleteness due to the infinite and constantly updating nature of knowledge, and thus the task of knowledge graph completion has been proposed. Existing multi-modal knowledge graph completion methods mostly rely on either embedding-based representations or graph neural networks, and there is still room for improvement in terms of interpretability and the ability to handle multi-hop tasks. Therefore, we propose a new method for multi-modal knowledge graph completion. Our method aims to learn multi-level graph structural features to fully explore hidden relationships within the knowledge graph and to improve reasoning accuracy. Specifically, we first use a Transformer architecture to separately learn about data representations for both the image and text modalities. Then, with the help of multimodal gating units, we filter out irrelevant information and perform feature fusion to obtain a unified encoding of knowledge representations. Furthermore, we extract multi-level path features using a width-adjustable sliding window and learn about structural feature information in the knowledge graph using graph convolutional operations. Finally, we use a scoring function to evaluate the probability of the truthfulness of encoded triplets and to complete the prediction task. To demonstrate the effectiveness of the model, we conduct experiments on two publicly available datasets, FB15K-237-IMG and WN18-IMG, and achieve improvements of 1.8 and 0.7%, respectively, in the Hits@1 metric.

Keywords: knowledge graph completion; multi-modal knowledge graph; link prediction; multi-modal feature fusion; graph neural network; transformer

1. Introduction

The continuous development of deep learning technology has had a significant impact on research in various fields. For instance, in the field of biomedicine, automatic diagnostic techniques based on deep learning have emerged, enabling image recognition and assisting healthcare professionals in the diagnosis and subsequent procedures [1–5]. Furthermore, deep learning has demonstrated a superior performance in scenarios with larger datasets, such as multi-view clustering [6–10]. In order to store and learn from a vast amount of information, knowledge graphs (KG) have emerged.

A knowledge graph can be conceptualized as a large-scale semantic integration network, which represents entities as nodes and relationships as directed edges; thus, it stores a vast amount of human knowledge in the form of a directed graph. The resource description framework (RDF) provides a standard framework for KG representation, wherein fact triples (head, relationship, tail) are employed to describe knowledge [11]. The KG is capable of storing a rich amount of information regarding real-world entities and their relationships and can enable a range of reasoning processes across the graph. The graph-based approach to data processing has demonstrated a superior performance in tasks such as assisting information retrieval, question-answering systems, and recommendation systems, when compared to traditional structured data [12, 13]. However, due to the infinite and constantly evolving nature of real-world knowledge, the incompleteness of the KG has led to the task of knowledge graph completion (KGC).

In the field of natural language processing (NLP), KGC techniques can be broadly categorized into three types: rule-based models, path-based models, and embedding-based models. Rule-based models tend to retain the original semantic information more completely, and therefore offer better interpretability. Path-based models make a better use of and represent the graph structure, enabling guided reasoning through various path-searching mechanisms. Both of these approaches are more interpretable, though their expressiveness is limited by model constraints, and their spatiotemporal complexity is higher. Compared to the first two types of models, embedding-based models typically offer greater expressiveness. With the development of graph neural networks (GNNs), GNN-based models have shown great potential in various graph-based tasks, providing additional ideas for KGC. In recent years, KG has also been studied in computer vision, such as in the context of scene graphs and language and image integration.

In recent years, multi-modal knowledge graphs (MKG) have gained significant attention as an extension to traditional knowledge graphs based on a single modality. MKGs typically augment semantic KGs with additional modality data, such as visual and audio attributes, to provide more physically rich representations of the world [14–16], as illustrated in Figure 1. For a given entity in the knowledge graph, we can use both image and text descriptions to supplement more detailed information that cannot be captured solely by the graph structure. Unfortunately, due to the lack of accumulated multi-modal corpora, existing MKGs often suffer from more severe incompleteness compared to traditional KGs, which greatly reduces their utility and effectiveness. In the task of multi-modal knowledge graph completion (MKGC), we must consider both the issues of multi-modal information fusion and the accuracy and interpretability of knowledge graph completion. In terms of multi-modal information fusion, we need to address issues such as semantic alignment, noise reduction or attenuation, and the realization of unified embeddings. In the process of link prediction, we must not only leverage the semantic richness of multi-modal information to improve accuracy but also enhance the logicality of the algorithm and

improve its interpretability [17, 18].

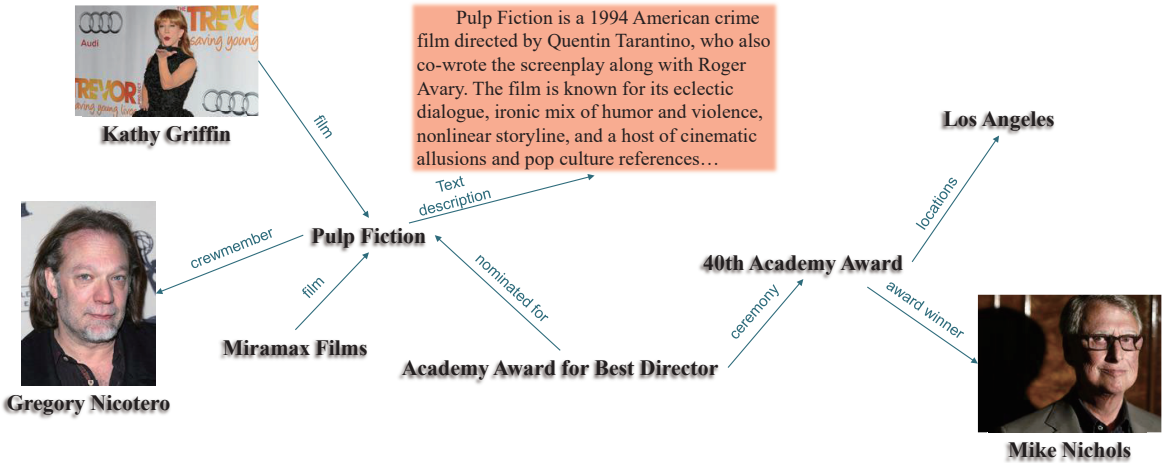


Figure 1. A simple multi-modal knowledge graph example.

Despite the abundance of existing image-text embedding pre-training models, these models often focus on a single pair of corresponding images and text and fail to consider the distinctive structural features of KGs. Therefore, our research builds upon MKGs that contain image-text feature information. In addition to integrating embeddings from different modalities, we also retain local graph features and introduce path features to enhance the interpretability of the reasoning model. Specifically, we propose a method that first utilizes separate modality encoders to learn image and text embeddings, followed by an irrelevant filtering layer to further select semantically relevant key features. Next, we fuse and encode information from different modalities to obtain a multi-modal representation. We then use graph convolution algorithms and path features to extract structural features, and use a scoring function to predict missing triples. Our innovation can be summarized as follows:

1) Designed a structure for extracting image-text information through single-modality encoding, followed by interaction fusion, and improved the semantic similarity through an irrelevant filtering module, thereby enhancing the fusion understanding of different modalities;

2) Proposed a structure feature learning scheme that combines graph convolution and path embedding, thereby enhancing interpretability during the reasoning process;

3) Achieved better results on two public datasets, FB15K-237-IMG and WN18-IMG.

2. Related work

2.1. Knowledge graph completion

The task of knowledge graph completion has been widely studied, with typical sub-tasks including link prediction, entity prediction, and relation prediction, aimed at predicting missing triples (head, relation, tail) in the knowledge graph. Rule-based models such as AMIE and RLvLR utilize symbolic features to perform reasoning through either rule mining or rule searching algorithms [19, 20]. NeuralLP introduced dynamic programming and further optimized rule mining through attention mecha-

nisms and auxiliary memory [21]. Path-based models focus more on the paths between queried head and tail entities, and algorithms such as the path ranking algorithm (PRA) and random walks have been applied and further explored in such models. RNNPRA uses recurrent neural networks (RNN) to better learn path features for reasoning tasks [22]. DIVA proposed a unified reasoning framework that divides multi-hop reasoning into a path search and path inference steps [23]. The continuous development of deep reinforcement learning (DRL) techniques has enabled more effective multi-hop reasoning in sparse graphs. A series of models such as DeepPath and MultiHop have achieved more effective path exploration by designing new reward mechanisms [24, 25].

Currently, the more mainstream methods for solving KGC problems are focused on embedding-based models. Translation-based models such as TransE, TransR, and TransH embed entities and their relations by projection, and use a distance function to score the factual triplets [26–28]. Tensor factorization models such as RESCAL, Tucker, and LowFER use vectors to capture latent semantics through tensor decomposition and continuously improve model efficiency while reducing model size [29–31]. With the continuous improvement in neural networks (NN) in learning and expressing knowledge, additional embedding-based models choose to use neural network architectures to implement KGC. NTN uses neural tensor networks for relation reasoning in KG [32]. ConvE learns deeper features using two-dimensional convolutional layers [33]. InteractE processes more complex semantic information and KG interactions through multiple operations such as feature reshaping, feature permutation, and recurrent convolution [34]. Although CNN-based KGR models generally perform better than traditional NN models, the feature information contained in the graph structure itself has not been well utilized. Therefore, GNNs have been introduced into the KGC field to perform more complex reasoning tasks based on graph structure features. RGCN encodes each entity into a vector, uses specific transformations to aggregate neighborhood information for different relationship categories, and then reproduces facts through a decoder [35]. SACN uses weighted graph convolutional networks (WGCN) to implement the encoder, and then inputs the encoded information into a convolutional network for decoding [36]. NBF-Net and RED-GNN improve on traditional algorithms, choosing Bellman-Ford algorithms and dynamic programming to optimize the propagation strategy in previous GNN models, and achieve efficiency improvements [37, 38].

2.2. Multi-modal task

The traditional tasks in the two major fields of computer vision (CV) and natural language processing (NLP) have been extensively discussed, and more recent research has focused on cross-modal problems. The optimization and development of the Transformer model has led to a series of explorations into visual-text pre-training frameworks. VisualBERT is considered to be the first image-text pre-training model, which uses Faster R-CNN to extract visual features and connects them with text embeddings, which are then input into a transformer initialized by BERT [39]. Inspired by the feature extraction and architecture in the VisualBERT model, more pre-training models have been proposed by adjusting the pre-training tasks and datasets. CLIP uses a dataset of 400 million image-text pairs for pre-training, learning representations by directly matching raw text and corresponding images [40]. METER further explores single-modal feature extraction and processes multi-modal fusion using a dual-stream architecture model, achieving excellent performance on many downstream tasks [41].

Numerous excellent multi-modal pretraining models have adopted the masked language modeling (MLM), masked visual modeling (MVM), and visual-linguistic matching (VLM) tasks as pretrain-

ing objectives; their corresponding downstream tasks are mainly focused on works that deal with the meaning and relationships between text and images, such as visual question answering (VQA), visual commonsense reasoning (VCR), and visual captioning (VC). However, for KGs, their distinguishing feature from semantically structured information is their graph structure. Recently, some studies have recognized the importance of structural features for handling KG-related tasks. DRAGON proposes a deep bidirectional, self-supervised pretraining method for language knowledge models from text and KGs [42]. Knowledge-CLIP takes entities and relations in KGs as inputs and extracts the original features of these entities and relations [43]. Entities can be in the form of images/text, while relations are described using language tokens. These pretraining models with structural features provide better options for MKG-related tasks.

2.3. Multi-modal knowledge graph completion

As an emerging research field, related work in MKGC is not yet systematic, and early MKGC tasks often directly added image information to the input of the original KGR model, which usually led to a suboptimal performance. To address this issue, many studies have made more attempts and explorations in the field of image-text feature fusion in MKG.

IKRL first proposed an attention-based neural network to consider visual information in entity images [44]. TransAE introduced a KG representation learning method that integrates multi-channel (visual and language) information in a translation-based framework, and extended the definition of triple energy to consider new multi-channel representations [45]. MKBE and MRCGN integrated different neural encoders and decoders with relation models to embed learning and multi-modal data for inference [14, 46]. MarT constructed a multi-channel analogical reasoning framework based on structural mapping theory to improve model interpretability [47]. MMKGR used a unified gate attention network to perform an attention interaction and to filter noise for generating more effective and reliable multi-modal complementary feature encoding, and designed a new reinforcement learning framework to predict missing elements in multi-hop reasoning processes [16]. MM-RNS proposed a multi-channel relation-enhanced negative sampling framework that provides bidirectional attention between visual and text features by integrating relation embeddings, and combined it with contrastive learning to construct an effective contrastive semantic sampler to improve MKGC performance [48].

We have conducted a brief overview of the related models in traditional and multimodal KGs, as shown in Table 1.

Table 1. Summarization of existing KGC models.

		Knowledge Graph Completion	Multi-Modal Knowledge Graph Completion
Rule-based Models		AMIE, RLvLR, NeuralLP	-
Path-based Models		RNNPRA, DIVA, DeepPath, MultiHop	-
Embedding-based Models	Translational Models	TransE, TransR, TransH	TransAE
	Tensor Decompositional Models	RESCAL, Tucker, LowFER	-
	Neural Network Models	NTN, ConvE, InteractE, RGCN, SACN, NBF-Net, RED-GNN	IKRL, MKBE, MRCGN, MarT, MMKGR, MM-RNS

In order to provide a clearer demonstration of the effectiveness of the aforementioned work, we have provided a more detailed comparative analysis of selected algorithms in Table 2.

Table 2. Model performance comparison.

Models	Dataset	Technique	Performance(Hits@10)
RLvLR	FB75K	Logic rule	43.4
MultiHop	FB15k-237	Relation path	56.4
TransE	FB15k-237	Translational	47.1
LowFER	FB15k-237	Tensor decompositional	54.4
RED-GNN	FB15k-237	GNN	55.8
TransAE	WN9-IMG	Translational	94.84
MMKGR	WN9-IMG	Attention	92.8

3. Problem formulation

The knowledge graph $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$ is a directed graph, where $\mathcal{E}$ is the entity set, $\mathcal{R}$ is the relation set, and $\mathcal{F} = \{(h, r, t) | h \in \mathcal{E}, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the fact set consisting of fact triples (h, r, t). The head entity $h \in \mathcal{E}$ and tail entity $t \in \mathcal{E}$ are connected by a relation $r \in \mathcal{R}$. For a multi-modal knowledge graph $\mathcal{G}$, the entity e includes two modalities, namely textual information e^t and visual information e^v.

The purpose of multi-modal KGC is to infer incomplete triplets $\mathcal{T} = \{(h, r, t) | h \in \mathcal{E}, t \in \mathcal{E}, r \in \mathcal{R}, (h, r, t) \notin \mathcal{F}\}$ based on known fact triplets (h, r, t). In practice, the incomplete triplets that may appear in our prediction task can take three forms, namely $(h, r, ?)$, $(h, ?, t)$, and $(?, r, t)$. In the implementation process, we input the feature information of entities e and relationships r into an encoder to obtain the corresponding embedding vectors $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$. Then, we use a scoring function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ to evaluate the probability of the truthfulness of inferred triplets. That is, when triplet $(h, r, t) \in \mathcal{G}$is true, $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ scores 1, otherwise, when $(h, r, t) \notin \mathcal{G}$ is true, $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ scores 0. Taking a missing triplet in the form of $(h, ?, t)$ as an example, let us assume the existence of a relationship r_{pd} between the head entity h and the tail entity t, thereby obtaining the complete triplet $\left(h, r_{pd}, t\right)$ with an unknown truthfulness. To evaluate the probability of its actual occurrence, we employ a scoring function, resulting in the output $f\left(\mathbf{h}, \mathbf{r}_{pd}, \mathbf{t}\right)$. The basic terminology definitions are shown in Table 3.

Table 3. Notation summary.

Notation	Explanation
$\mathcal{G}$	Multi-modal knowledge graph
$\mathcal{E}$	Entity set
$\mathcal{R}$	Relation set
$\mathcal{F}$	Fact set
$\mathcal{T}$	Incomplete fact set
(h, r, t)	Fact triplet of the head, relation, tail
$\mathbf{h}$	Embedding of head entity
$\mathbf{t}$	Embedding of tail entity
$\mathbf{r}$	Embedding of relation entity

4. Methodology

The model we proposed, MLSFF, has an overall architecture shown in Figure 2, which consists of three components: 1) single-modality encoders for image and text embedding; 2) a multi-modal feature fusion mechanism with irrelevant filtering to discard interfering information and to reduce noise when the image and text features interact with each other; 3) a reasoning framework that combines the graph structure and path features, introduces a new scoring function containing multi-hop path features, and uses multi-modal features to predict incomplete triplets in KGC processes.

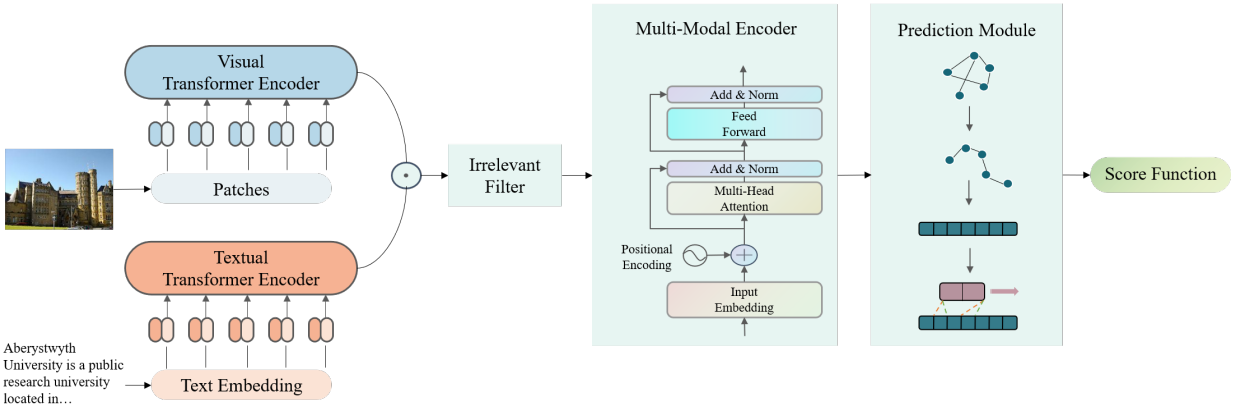


Figure 2. Overview of our model structure.

4.1. Single modal encoder

The emergence of the Transformer model has caused a huge revolution in the NLP field and has been widely used in various tasks. The attempt to introduce the Transformer model into the CV field has not only achieved success, but even achieved astonishing results. Specifically when the pre-training data is large enough, Transformer's performance in CV will be significantly better than CNN, breaking the limitation of the original few inductive biases, and achieving better transfer effects in downstream tasks. We use independent image encoders and text encoders based on the Transformer architecture to extract features from the raw inputs. For a given triple, the entity and relation are sent to the corresponding encoder based on their modality (image or text). The relation represented by language tokens is sent to the text encoder similar to the text entity. The main architecture of our single-modality encoder is illustrated in Figure 3.

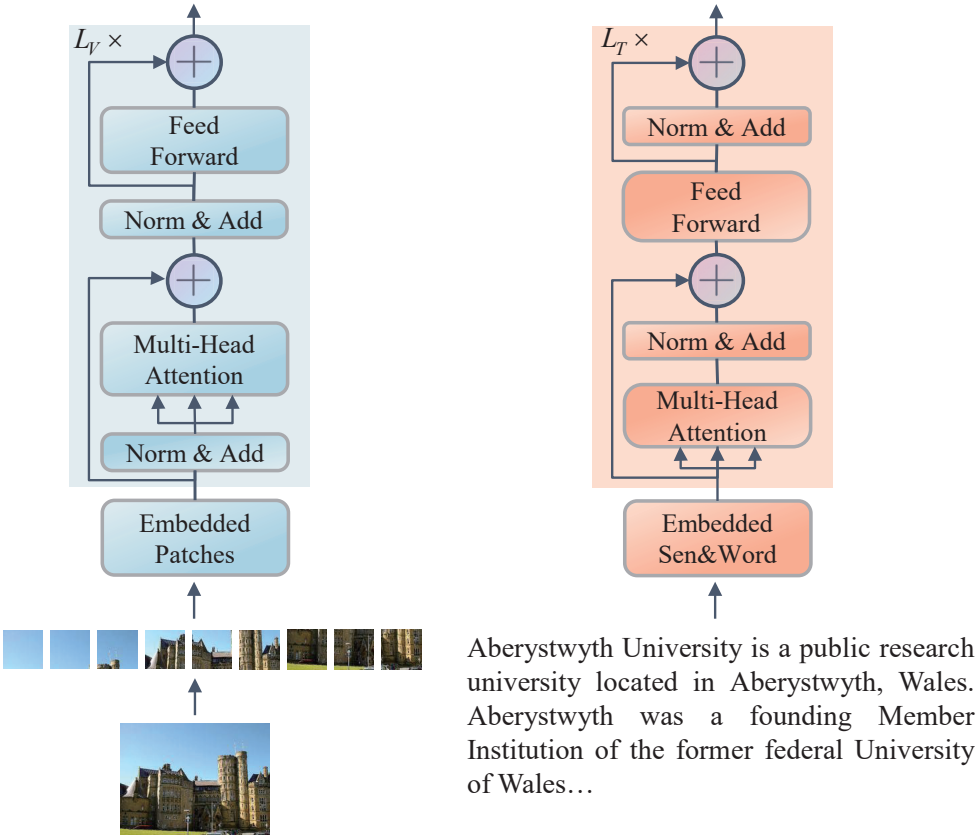


Figure 3. Structure of single modal encoder.

Visual Encoder For image feature extraction, we adopt the embedding layer and Transformer encoder of the pre-trained model ViT as the main architecture [49]. Let C be the number of channels in the image (in RGB images, $C = 3$) and the resolution of each image patch be (P, P). First, we scale the input image I to a unified resolution (A, B), and then divide it into $N = AB/P^2$ patches. We use a linear mapping (i.e., FC layer) to transform each patch into a one-dimensional vector. This completes the embedding of the original image X^v_{pat}. Subsequently, we feed the obtained image embedding and position embedding X^v_{pos} into the Transformer encoder as an input. The overall forward calculation process is as follows:

$$X^v_0 = X^v_{pat} + X^v_{pos} \tag{4.1}$$

$$\hat{X}^v_l = \mathbf{MSA}\left(\mathbf{LN}\left(X^v_{l-1}\right)\right) + X^v_{l-1},\ l = 1, 2, ..., L^v \tag{4.2}$$

$$X^v_l = \mathbf{FFN}\left(\mathbf{LN}\left(\hat{X}^v_l\right)\right) + \hat{X}^v_l,\ l = 1, 2, ..., L^v \tag{4.3}$$

The MSA Block consists of a multi-head attention mechanism, a layer normalization, and a skip connection (Layer Norm & Add), which can be repeated for L^v times, and the output of the $l-$th block is

$\hat{X}_l^v$. The MLP Block consists of feedforward neural network, layer normalization, and skip connection (Layer Norm & Add), which can be repeated for L^v times, and the output of the $l-$th block is X_l^v.

Textual Encoder In NLP tasks, a large number of pre-training models based on the Transformer architecture have emerged, such as BERT, which has recently been widely applied and demonstrated great success in various downstream tasks [50, 51]. In this paper, we use BERT to perform language modeling and feature extraction. Specifically, we divide the complete sentence into a word sequence and perform word embedding to obtain the word embeddings X_{word}^t. In order to preserve sentence-level features, we also embed the entire sentence and align it with the word embeddings to obtain the sentence embeddings X_{sen}^t. Then, we send the word embeddings X_{word}^t, position embeddings X_{pos}^t, and sentence embeddings X_{sen}^t to the encoder.

$$X_0^t = X_{word}^t + X_{sen}^v + X_{pos}^v \tag{4.4}$$

$$\hat{X}_l^t = \mathbf{LN}\left(\mathbf{MSA}\left(X_{l-1}^t\right)\right) + X_{l-1}^t,\ l = 1, 2, ..., L^t \tag{4.5}$$

$$X_l^t = \mathbf{LN}\left(\mathbf{FFN}\left(\hat{X}_l^t\right)\right) + \hat{X}_l^t,\ l = 1, 2, ..., L^t \tag{4.6}$$

The difference between text encoding and visual encoding is that layer normalization (LN) is located after the multi-head self-attention (MSA) and feed-forward network (FFN) layers. Similarly, the output of the $l-$th MSA block is denoted as $\hat{X}_l^t$ and the output of the $l-$th MLP block is denoted as X_l^t. We denote the number of MSA and MLP blocks in the text encoder as L^t.

4.2. Multimodal feature fusion

In the multimodal fusion module, we fuse the separately encoded text and image information. Specifically, since relationships belong to a separate data category with certain label information, although they are usually described using text, their semantic relevance to the text and image descriptions of entities is relatively low. Therefore, we choose to fuse and filter the image and text information separately for relationships, and then introduce the encoded relationship attributes when learning the path features.

To enhance the efficiency of the semantic interaction between the two different modalities of image and text, we adopt an intermediate representation to unify the multimodal information. On one hand, we aim to achieve a more fine-grained interaction between different modal feature information; on the other hand, since images often contain semantically irrelevant information, directly using the complete image embedding in the feature fusion process may introduce noise. Therefore, we feed the learned image and text vectors into a multimodal gated unit for weight learning to achieve the intermediate feature representation.

$$g_f = \sigma\left(X^v W^v \odot X^t W^t\right) \tag{4.7}$$

$$\hat{X}^m = g_f X^v + \left(1 - g_f\right) X^t \tag{4.8}$$

In this equation, σ represents the sigmoid function, X^v and X^t denote the feature vectors outputted by the image and text encoders, respectively, W^v and W^t are parameter matrices, g_f is a scalar within

the range of $[0, 1]$, $\hat{X}^m$ represents the multi-modal embedding vector obtained through the filtering layer, and $\odot$ denotes the element-wise multiplication (i.e., Hadamard product).

Later, we feed the original embeddings $\hat{X}^m$ into the multi-modal encoder to further learn the semantic features.

$$X = Tran\left(\hat{X}^m\right) \tag{4.9}$$

4.3. Prediction block

We have obtained the multi-modal feature embedding of a certain fact description through the previous structure, but this is insufficient for large-scale and complex KGs. Hence, we aim to further learn path features to better accomplish the task of KGC. The overall approach regarding the learning of structural features and completion can be summarized as follows. First, we extract a certain path existing in the MKG, connect the relations in the path, and then divide the path into several shorter components through a sliding window. Then, we select one of the components and use a recurrent attention unit to embed the selected component to obtain a relation vector, which is represented as a weighted combination of existing relations. We recursively merge the divided components of the path, and finally use a scoring function to determine the truthfulness of unknown triplets. The overall process of the prediction block shows in Algorithm 1.

Algorithm 1 Prediction block

Input: the path body $\mathbf{r}_p$
Output: the score of triplet $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$

Initialize the window size w
for all $i = 1, 2, ..., n-1$ **do**
  get path segments $w = \{1, 2, 3\}$ and encoding with LSTM $[\hat{y}_i, \hat{y}_{i+1}] = \mathbf{LSTM}(w_i)$;
  $\mathbf{y}_i = \hat{y}_{i+1}$
end for
$\mu = softmax([\mathbf{FC}(\mathbf{y}_1), \mathbf{FC}(\mathbf{y}_2), ..., \mathbf{FC}(\mathbf{y}_{n+1-w})])$
$Y = \sum_{i=1}^{n+1-w} \mu_i \mathbf{y}_i$
$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \sigma(vec(([X_h; Y] * \omega) W) X_t)$
return $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$

Sliding Window Segmentation To extract fine-grained features from sampled paths, we decompose the sampled paths into combinations of different sizes using sliding windows of varying lengths. In the implementation, we use windows of size $w = \{1, 2, 3\}$. Given the window size, the generated sliding windows traverse the path body $\mathbf{r}_p = \left[r_{p_1}, ..., r_{p_n}\right]$. Then, we use a long short-term memory (LSTM) network as a sequence encoder to conceal the information within the sliding windows. Taking the sliding window of length 2 as an example,

$$[\hat{y}_i, \hat{y}_{i+1}] = \mathbf{LSTM}(w_i) \tag{4.10}$$

Since the final state y_{i+1} usually contains the complete information of the sequence, we select $\mathbf{y}_i = \hat{y}_{i+1}$. $\mathbf{y}_i$ is meaningful to learn the relationship in the window if the relationship segments in

the $i-$th sliding window always appear together in some combination, which is more likely to represent a real "long-distance" relationship. To incorporate this observation into our model, we calculate the probability value of these relationship segments by:

$$\mu = softmax\left(\left[\mathbf{FC}\left(\mathbf{y}_1\right), \mathbf{FC}\left(\mathbf{y}_2\right), ..., \mathbf{FC}\left(\mathbf{y}_{n+1-w}\right)\right]\right) \tag{4.11}$$

where $\mathbf{FC}(\cdot)$ represents a fully connected layer, which is used to learn the probability that the $i-$th window in $\mathbf{y}_i$ represents a meaningful relationship fragment. Finally, we calculate the weighted sum of information from different windows to represent the complete path features:

$$Y = \sum_{i=1}^{n+1-w} \mu_i \mathbf{y}_i \tag{4.12}$$

Scoring Function Considering the excellent performance of graph convolutional models in handling KGC problems, we choose the following scoring function:

$$f\left(\mathbf{h}, \mathbf{r}, \mathbf{t}\right) = \sigma\left(vec\left(\left(\left[X_h; Y\right] * \omega\right) W\right) X_t\right) \tag{4.13}$$

In the proposed scoring function, X_h and X_t represent the multi-modal embeddings of the head and tail entities, respectively, while Y represents the embedding of their relationship, $*$ and ω denote the convolution operation and the convolution kernel, respectively, and $vec(\cdot)$ represents the projection operation from the feature map to the vector space, W is a parameter matrix. With the above method, we can compute whether a fact constructed by a certain relationship between two entities is true or not.

For ease of reference, we summarize the main symbol notations used in this chapter in Table 4.

Table 4. Notation summary.

Notation	Explanation
X	Embedding entity vector
$\hat{X}$	Intermediate state of the embedding entity
$\hat{y}$	Intermediate state of the embedding path
$\mathbf{y}_i$	Embedding path vector
Y	Encoded Complete path vector
W	Parameter Matrix

5. Experiment

5.1. Dataset

We evaluate the effectiveness of the MLSFF model on two publicly available datasets: (i) FB15K-237-IMG: a subset of the large-scale knowledge graph Freebase, where each entity has 10 images, and is a commonly used dataset in KGC tasks; (ii) WN18-IMG: WN18 is a knowledge graph extracted from WordNet. WN18-IMG is an extended dataset of WN18, where each entity has 10 images [52]. These two datasets can be obtained as FB15k-WN18-images. Table 5 shows the statistical information of the datasets.

Table 5. Statistics of datasets.

Datasets	#Entities	#Relations	#Train	#Dev	#Test
FB15k-237-IMG	14,541	237	272,115	17,535	20,466
WN18-IMG	40,943	18	141,442	5000	5000

5.2. Settings

Evaluation Metrics: We adopted classic knowledge graph completion evaluation metrics, including *Hits@k* and mean rank (MR), as shown in Table 6.

Table 6. Summarization of evaluation metrics.

Evaluation metrics	Calculation formula
Hits@k	$Hits@k = \sum_i \frac{1(\text{rank}_i)<k}{Q}$
MR	$MR = \frac{1}{Q}\sum_i \text{rank}_i$

Hits@k: The *Hits@k* metric is defined as the proportion of true entities that appear in the top-k ranked list of entities. It is calculated as follows:

$$Hits@k = \sum_i \frac{1\,(\text{rank}_i) < k}{Q} \tag{5.1}$$

where rank_i represents the rank of the expected entity of the $i-$th incomplete fact triple. Q represents the total number of incomplete fact triples.

Mean Rank (MR): Mean Rank is the arithmetic average of the individual entity ranks, defined as:

$$MR = \frac{1}{Q}\sum_i \text{rank}_i \tag{5.2}$$

Parameter Configuration To consider the model's scale and computational efficiency, we choose the ViT-B/16 pre-trained model for the image encoder. We set the embedding dimensions for both text and image to 768. The number of layers for both the image and text encoders is set to 12, while the number of layers for the modality encoder is set to 3. The graph embedding dimension is set to 200, and the batch size is set to 64. We utilize the Warmup algorithm and the ADAM optimizer to adjust the learning rate of the model parameters. The initial learning rate is set to 0.0005, and the dropout rate is set to 0.1.

Baseline Setup We selected four unimodal methods and four multi-modal methods as baselines to compare with our proposed model. The unimodal methods include the following: 1) TransE [26], a classic translation-based model that encodes entities and relationships into a linear space; 2) DistMult [53], which uses a linear neural network to encode a multi-relation graph for multi-relation learning; 3) ComplEx [54], which solves both symmetric and asymmetric relations by introducing complex methods; and 4) RotatE [55], which defines relations as rotations from the head entity to the tail entity in a complex space to achieve multi-class reasoning. The multi-modal methods include the following: (i) IKRL (UNION) [44], which extends TransE to learn about visual representations of entities and

structural features of KGs; (ii) TransAE [56], which combines multi-modal encoders with TransE to achieve unified representation of visual and textual features; (iii) RSME [57], which uses a forget gate to learn about valuable images for MKG embedding; and (iv) MKGformer [52], which proposes an MKG pre-training model based on a hybrid transformer structure.

5.3. Main results

The experimental results on the two datasets are shown in Table 7, which shows that our model generally outperforms the 8 baseline methods.

Table 7. Results of link prediction on FB15k-237-IMG and WN18-IMG.

Model	FB15k-237-IMG				WN18-IMG			
	Hits@1↑	Hits@3↑	Hits@10↑	MR↓	Hits@1↑	Hits@3↑	Hits@10↑	MR↓
TransE	0.198	0.376	0.441	323	0.40	0.745	0.923	357
DistMult	0.199	0.301	0.466	512	0.335	0.876	0.940	655
ComplEx	0.194	0.297	0.450	546	0.936	0.945	0.947	-
RotatE	0.241	0.375	0.533	**177**	0.942	0.950	0.957	254
IKRL (UNION)	0.194	0.284	0.458	298	0.127	0.796	0.928	596
TransAE	0.199	0.317	0.463	431	0.323	0.835	0.934	352
RSME	0.242	0.344	0.467	417	0.943	0.951	0.957	223
MKGformer	0.256	0.367	0.504	221	0.944	0.961	0.972	28
MLSFF (ours)	**0.274**	**0.411**	**0.552**	193	**0.951**	**0.973**	**0.980**	**22**

Firstly, in all works, the scores on FB15k-237-IMG are generally lower than those on WN18-IMG. The fundamental reason is that the dataset FB15k-237-IMG is more sparse and complex than the dataset WN18-IMG, with a greater variety of relationships between different entities. In addition, our model performs better on Hits@1 than on Hits@3 or Hits@10, indicating a superior discriminative ability in predicting unknown entities. In the MLSFF model, we use two single-modal encoders to extract image and text information, followed by a multi-modal layer for interaction, which enables full learning of semantic information for entity description. We introduce a sliding window in learning the link features, which realizes ”scalable” path sampling and to some extent solves the problem of complex graph structures.

Secondly, some traditional single-modal methods, such as RotatE, even outperform architectures that use multi-modal features in overall performance. This suggests that a well-designed relationship decomposition and learning rule are effective in solving complex graph problems, and fully utilizing structural features can improve prediction accuracy. Therefore, after obtaining multi-modal encoding information, our model not only uses the traditional graph convolutional method to obtain neighbor node information, but also incorporates long-distance path features and borrows from recurrent neural network structures used in processing text information to extract left and right node information from selected paths. By adding certain ”vertical” features during the convolution process, our prediction model can have better interpretability.

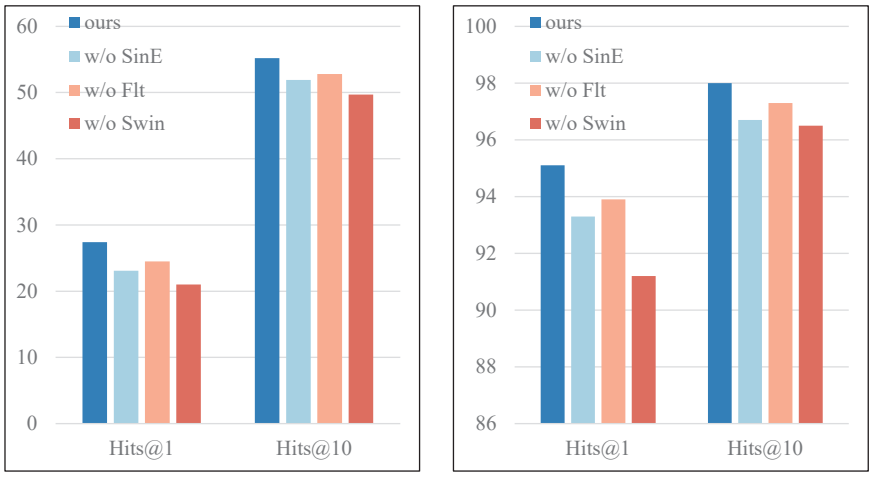
Finally, our model achieved significant improvements of 4.8 and 1.2% on the two datasets, respectively. However, in the FB15k-237-IMG dataset, our model's MR metric results were slightly inferior to those of the RotatE model. This could be attributed to the FB15k-237-IMG dataset containing a

larger number of entities and a more diverse set of relationships, resulting in a sparser and more complex knowledge graph. While our model has improved its ability to learn about multi-hop path relationships to some extent, it lacks similar operations on negative samples, as seen in the RotatE model. As a result, this has impacted the overall accuracy. Overall, the experimental results demonstrate that our model outperforms existing methods on most evaluation metrics, with even more significant improvements observed on more complex knowledge graphs. This is because the MLSFF model learns more comprehensive semantic features by fusing information from both image and text modalities, enabling more comprehensive knowledge extraction from the graph. In addition, we employed convolutional operations that capture neighborhood information and an LSTM structure that learns path-level features to achieve a more comprehensive and three-dimensional feature encoding structure for learning graph structural features, which is highly effective for processing large-scale knowledge graphs.

6. Further analysis

6.1. Ablation study

To investigate the actual effects of each component in the MLSFF model, we conducted ablation studies by removing some of the components.



(a) FB15k-237-IMG (b) WN18-IMG

Figure 4. Ablation on different components of the MLSFF.

w/o SinE: To investigate the effect of the single-modal encoders on understanding image and text semantics, we aligned the one-dimensional vectorized image patches and text embeddings, calculated their Hadamard product, and directly fed them into the multi-modal encoder for learning.

w/o Flt: To further investigate the actual effect of the unrelated filtering layer, we also experimented with the meaning of the multi-modal fusion module by directly fusing the encoded image and text features without the unrelated filtering layer.

w/o Swin: To demonstrate the positive effect of extracting path information on learning graph structure features, we removed the sliding window encoding module and only used graph convolution

operations to obtain structural embeddings.

From Figure 4, it can be seen that using single-modality encoders to extract image and text features can effectively enhance semantic understanding and better learn human knowledge, thereby promoting and improving the performance in KGC tasks. Although image features can assist in text understanding, there is still some noise interference. Filtering out irrelevant information can further enhance the fusion effect between multi-modal features and improve accuracy. In addition, when facing large-scale and complex knowledge graphs, although graph convolutional operations can already fully learn structural information and capture neighbor features, the introduction of path and rule features can further improve model interpretability and prediction ability. Specifically, when dealing with sparse graphs, simple convolutional operations may lead to a certain decrease in accuracy, and learning path features can also help improve model efficiency.

6.2. Hyperparameter analysis

Our connection prediction module is mainly implemented based on the GNN algorithm, which aggregates neighbor information into the target node and then updates the target node based on the integrated information. However, this approach is prone to the problem of over-smoothing, where the representations of different nodes tend to become similar as the number of GNN layers increases during training. To address this issue, we introduce "longer-distance" path embeddings, which combine deep features and breadth features to extract complex graph structure information.

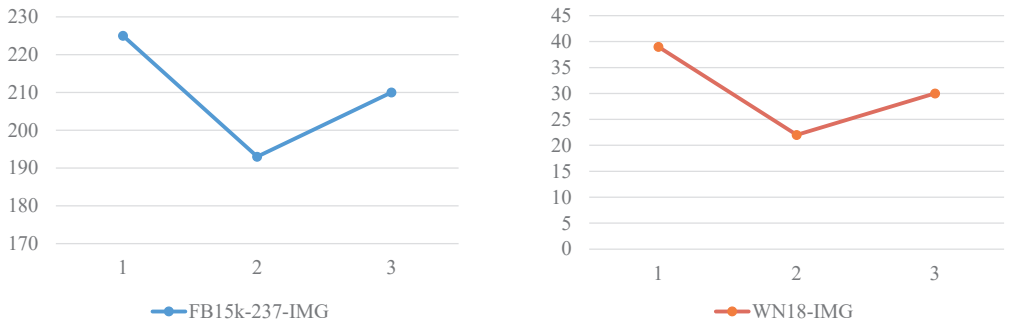


Figure 5. Impacts of the width on FB15k-237 and WN18RR.

We further explore effective graph processing structures by adjusting the number of convolutional layers and the size of the sliding window. In this work, considering memory and computational capacity, we conduct experiments with sliding window widths ranging from 1 to 3. As shown in Figure 5, the model performs better when the sliding window width is set to 2.

When the sliding window width is set to 2, our model can learn more layers of graph structural features and neighbor information. When the sliding window width is too small, that is, when the number of subgraphs learned is too few, the information in the knowledge graph cannot be fully aggregated to learn the structural information of the knowledge graph. In addition, some useful high-order neighbors cannot be captured. When the number of subgraphs is too large, the node representation is overly smoothed due to excessive noise.

6.3. Complexity analysis

MLSFF: Denote the entity embedding dimension as d_e, the structural embedding dimension as d_r, and the number of channels as T. The final output dimension for triplet encoding is denoted as $m \times n$. The main complexity of our model can be represented as $O\left(|\mathcal{E}|\, d_e + |\mathcal{R}|\, d_r + Tmn + Td\,(2d_m - m + 1)\,(d_n - n + 1)\right)$.

TransE: The scoring function of the TransE model is denoted as $\|h + r - t\|$, and as a result, its algorithmic complexity can be represented as $O\left(|\mathcal{E}|\, d + |\mathcal{R}|\, d\right)$.

RED-GNN: As a GNN model in the traditional knowledge graph completion task, the RED-GNN model has an algorithmic complexity denoted as $O\left(d \cdot \min\left(\bar{D}^L, |\mathcal{F}|\, L\right)\right)$. In this context, $\bar{D}$ represents the average degree of the r-directed graph per layer. It can be observed that our model has a slightly higher computational complexity. This is attributed to two main factors: first, the inherent complexity of multimodal knowledge graphs; and second, the decision to incorporate a more extensive graph feature learning scheme to enhance the interpretability of paths.

6.4. Study limitation

Despite the promising results and contributions of our study, there are some limitations that should be acknowledged:

While our model aims to enhance interpretability by incorporating graph features and multi-hop paths, the interpretability of the model's predictions may still be limited. Explaining the reasoning behind specific predictions or understanding the underlying decision-making processes can be challenging, especially in complex multimodal knowledge graphs.

In addition, the proposed model in this paper exhibits high complexity, which results in increased demands for computational resources and significant time consumption. Furthermore, our model does not consider the possibility of negative samples during the sampling process, which has an impact on the overall accuracy of the prediction task.

7. Conclusions

We propose a MLSFF model which first uses two independent single-modality encoders to obtain pre-trained embeddings for both image and text information. Then, after filtering out irrelevant information, the multi-modal features are fused to obtain a unified encoding vector. We utilize graph convolutional algorithms to learn the structural information in the knowledge graph. In addition, we introduce path-based feature information into the graph structural features to obtain richer relationship representations. Our experimental results demonstrate that our model achieves better performance in MKGC tasks. To address the issues of high complexity and the omission of negative samples in our model, we will focus on the following areas for future research: (i) designing simpler and more efficient scoring functions that are more streamlined and computationally efficient; (ii) considering negative sample interference, thereby mitigating their impact on the accuracy of the prediction task; (iii) incorporating additional modalities: to achieve a more comprehensive and diverse multimodal fusion such as numerical features and enhancing the overall performance of the model.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China-China State Railway Group Co., Ltd. Railway Basic Research Joint Fund (Grant No.U2268217) and the Scientific Funding for China Academy of Railway Sciences Corporation Limited (No.2021YJ183).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Shoeibi, N. Ghassemi, M. Khodatars, P. Moridian, A. Khosravi, A. Zare, et al., Automatic diagnosis of schizophrenia and attention deficit hyperactivity disorder in rs-fmri modality using convolutional autoencoder model and interval type-2 fuzzy regression, *Cognit. Neurodyn.*, (2022), 1–23. https://doi.org/10.1007/s11571-022-09897-w
2. A. Shoeibi, M. Khodatars, M. Jafari, N. Ghassemi, P. Moridian, R. Alizadesani, et al., Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review, *Inf. Fusion*, 2022. https://doi.org/10.1016/j.inffus.2022.12.010
3. A. Shoeibi, M. Rezaei, N. Ghassemi, Z. Namadchian, A. Zare, J. M. Gorriz, Automatic diagnosis of schizophrenia in eeg signals using functional connectivity features and cnn-lstm model, in *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications: 9th International Work-Conference on the Interplay Between Natural and Artificial Computation*, (2022), 63–73. https://doi.org/10.1007/978-3-031-06242-1_7
4. P. Moridian, N. Ghassemi, M. Jafari, S. Salloum-Asfar, D. Sadeghi, M. Khodatars, et al., Automatic autism spectrum disorder detection using artificial intelligence methods with mri neuroimaging: A review, *Front. Mol. Neurosci.*, **15** (2022), 999605. https://doi.org/10.3389/fnmol.2022.999605
5. M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, et al., Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review, *Comput. Biol. Med.*, **139** (2021), 104949. https://doi.org/10.1016/j.compbiomed.2021.104949
6. S. Wang, Z. Chen, S. Du, Z. Lin, Learning deep sparse regularizers with applications to multi-view clustering and semi-supervised classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 5042–5055. https://doi.gor/10.1109/TPAMI.2021.3082632
7. S. Du, Z. Liu, Z. Chen, W. Yang, S. Wang, Differentiable bi-sparse multi-view co-clustering, *IEEE Trans. Signal Process.*, **69** (2021), 4623–4636. https://doi.org/10.1109/TSP.2021.3101979

8. Z. Chen, L. Fu, J. Yao, W. Guo, C. Plant, S. Wang, Learnable graph convolutional network and feature fusion for multi-view learning, *Inf. Fusion*, **95** (2023), 109–119. https://doi.org/10.1016/j.inffus.2023.02.013
9. Z. Fang, S. Du, X. Lin, J. Yang, S. Wang, Y. Shi, Dbo-net: Differentiable bi-level optimization network for multi-view clustering, *Inf. Sci.*, **626** (2023), 572–585. https://doi.org/10.1016/j.ins.2023.01.071
10. S. Xiao, S. Du, Z. Chen, Y. Zhang, S. Wang, Dual fusion-propagation graph neural network for multi-view clustering, *IEEE Trans. Multimedia*, 2023. https://doi.org/10.1109/TMM.2023.3248173
11. K. Liang, Y. Liu, S. Zhou, X. Liu, W. Tu, Relational symmetry based knowledge graph contrastive learning, preprint, arXiv:2211.10738. https://doi.org/10.48550/arXiv.2211.10738
12. S. Di, Q. Yao, Y. Zhang, L. Chen, Efficient relation-aware scoring function search for knowledge graph embedding, in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, IEEE, (2021), 1104–1115. https://doi.org/10.1109/ICDE51399.2021.00100
13. Y. Zhang, Q. Yao, W. Dai, L. Chen, Autosf: Searching scoring functions for knowledge graph embedding, in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, IEEE, (2020), 433–444. https://doi.org/10.1109/ICDE48307.2020.00044
14. P. Pezeshkpour, L. Chen, S. Singh, Embedding multimodal relational data for knowledge base completion, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (2018), 3208–3218. https://doi.org/10.18653/v1/D18-1359
15. Y. Zhao, X. Cai, Y. Wu, H. Zhang, Y. Zhang, G. Zhao, et al., Mose: Modality split and ensemble for multimodal knowledge graph completion, preprint, arXiv:2210.08821. https://doi.org/10.48550/arXiv.2210.08821
16. S. Zheng, W. Wang, J. Qu, H. Yin, W. Chen, L. Zhao, Mmkgr: Multi-hop multi-modal knowledge graph reasoning, preprint, arXiv:2209.01416. https://doi.org/10.48550/arXiv.2209.01416
17. Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, Q. Huang, Otkge: Multi-modal knowledge graph embeddings via optimal transport, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 39090–39102.
18. S. Liang, A. Zhu, J. Zhang, J. Shao, Hyper-node relational graph attention network for multi-modal knowledge graph completion, *ACM Trans. Multimedia Comput. Commun. Appl.*, **19** (2023), 1–21. https://doi.org/10.1145/3545573
19. L. A. Galárraga, C. Teflioudi, K. Hose, F. Suchanek, Amie: association rule mining under incomplete evidence in ontological knowledge bases, in *Proceedings of the 22nd international conference on World Wide Web*, (2013), 413–422. https://doi.org/10.1145/2488388.2488425
20. P. G. Omran, K. Wang, Z. Wang, An embedding-based approach to rule learning in knowledge graphs, *IEEE Trans. Knowl. Data Eng.*, **33** (2019), 1348–1359. https://doi.org/10.1109/TKDE.2019.2941685
21. F. Yang, Z. Yang, W. W. Cohen, Differentiable learning of logical rules for knowledge base reasoning, *Adv. Neural Inf. Process. Syst.*, **30** (2017).
22. A. Neelakantan, B. Roth, A. McCallum, Compositional vector space models for knowledge base completion, preprint, arXiv:1504.06662. https://doi.org/10.48550/arXiv.1504.06662

23. W. Chen, W. Xiong, X. Yan, W. Wang, Variational knowledge graph reasoning, preprint, arXiv:1803.06581. https://doi.org/10.48550/arXiv.1803.06581
24. X. V. Lin, C. Xiong, R. Socher, Multi-hop knowledge graph reasoning with reward shaping, preprint, arXiv:1808.10568. https://doi.org/10.48550/arXiv.1808.10568
25. W. Xiong, T. Hoang, W. Y. Wang, DeepPath: A reinforcement learning method for knowledge graph reasoning, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, (2017), 564–573. https://doi.org/10.18653/v1/D17-1060
26. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Adv. Neural Inf. Process. Syst.*, **26** (2013).
27. Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **29** (2015). https://doi.org/10.1609/aaai.v29i1.9491
28. Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **28** (2014). https://doi.org/10.1609/aaai.v28i1.8870
29. S. Amin, S. Varanasi, K. A. Dunfield, G. Neumann, Lowfer: Low-rank bilinear pooling for link prediction, in *International Conference on Machine Learning*, PMLR, (2020), 257–268.
30. I. Balažević, C. Allen, T. Hospedales, Tucker: Tensor factorization for knowledge graph completion, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (2019), 5185–5194. https://doi.org/10.18653/v1/D19-1522
31. M. Nickel, V. Tresp, H. P. Kriegel, A three-way model for collective learning on multi-relational data, in *Icml*, **11** (2011), 3104482–3104584.
32. R. Socher, D. Chen, C. D. Manning, A. Ng, Reasoning with neural tensor networks for knowledge base completion, *Adv. Neural Inf. Process. Syst.*, **26** (2013).
33. T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel, Convolutional 2d knowledge graph embeddings, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **32** (2018). https://doi.org/10.1609/aaai.v32i1.11573
34. S. Vashishth, S. Sanyal, V. Nitin, N. Agrawal, P. Talukdar, Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 3009–3016. https://doi.org/10.1609/aaai.v34i03.5694
35. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, Springer, (2018), 593–607. https://doi.org/10.1007/978-3-319-93417-4_38
36. C. Shang, Y. Tang, J. Huang, J. Bi, X. He, B. Zhou, End-to-end structure-aware convolutional networks for knowledge base completion, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 3060–3067. https://doi.org/10.1609/aaai.v33i01.33013060

37. Z. Zhu, Z. Zhang, L. P. Xhonneux, J. Tang, Neural bellman-ford networks: A general graph neural network framework for link prediction, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 29476–29490.
38. Y. Zhang, Q. Yao, Knowledge graph reasoning with relational digraph, in *Proceedings of the ACM Web Conference 2022*, (2022), 912–924. https://doi.org/10.1145/3485447.3512008
39. L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, K. W. Chang, Visualbert: A simple and performant baseline for vision and language, preprint, arXiv:1908.03557. https://doi.org/10.48550/arXiv.1908.03557
40. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, in *International Conference on Machine Learning*, PMLR, (2021), 8748–8763.
41. Z. Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, et al., An empirical study of training end-to-end vision-and-language transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 18166–18176. https://doi.org/10.48550/arXiv.2111.02387
42. M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, et al., Deep bidirectional language-knowledge graph pretraining, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 37309–37323.
43. X. Pan, T. Ye, D. Han, S. Song, G. Huang, Contrastive language-image pre-training with knowledge graphs, preprint, arXiv:2210.08901. https://doi.org/10.48550/arXiv.2210.08901
44. R. Xie, Z. Liu, H. Luan, M. Sun, Image-embodied knowledge representation learning, in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, preprint, arXiv:1609.07028. https://doi.org/10.48550/arXiv.1609.07028
45. Z. Wang, L. Li, Q. Li, D. Zeng, Multimodal data enhanced representation learning for knowledge graphs, in *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, (2019), 1–8. https://doi.org/10.1109/IJCNN.2019.8852079
46. W. Wilcke, P. Bloem, V. de Boer, R. van t Veer, F. van Harmelen, End-to-end entity classification on multimodal knowledge graphs, preprint, arXiv:2003.12383. https://doi.org/10.48550/arXiv.2003.12383
47. N. Zhang, L. Li, X. Chen, X. Liang, S. Deng, H. Chen, Multimodal analogical reasoning over knowledge graphs, preprint, arXiv:2210.00312. https://doi.org/10.48550/arXiv.2210.00312
48. D. Xu, T. Xu, S. Wu, J. Zhou, E. Chen, Relation-enhanced negative sampling for multimodal knowledge graph completion, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 3857–3866. https://doi.org/10.1145/3503161.3548388
49. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
50. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, preprint, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805
51. G. Jawahar, B. Sagot, D. Seddah, What does bert learn about the structure of language?, in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

52. X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, et al., Hybrid transformer with multi-level fusion for multimodal knowledge graph completion, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2022), 904–915. https://doi.org/10.1145/3477495.3531992

53. B. Yang, W. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, preprint, arXiv:1412.6575. https://doi.org/10.48550/arXiv.1412.6575

54. T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in *International Conference on Machine Learning*, PMLR, (2016), 2071–2080.

55. Z. Sun, Z. H. Deng, J. Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, preprint, arXiv:1902.10197. https://doi.org/10.48550/arXiv.1902.10197

56. H. Mousselly-Sergieh, T. Botschen, I. Gurevych, S. Roth, A multimodal translation-based approach for knowledge graph representation learning, in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, (2018), 225–234. https://doi.org/10.18653/v1/S18-2027

57. M. Wang, S. Wang, H. Yang, Z. Zhang, X. Chen, G. Qi, Is visual context really helpful for knowledge graph? a representation learning perspective, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 2735–2743. https://doi.org/10.1145/3474085.3475470

AIMS Press