Journal of Computers Vol. 37 No. 4, August 2026, pp. 13-27
https://doi.org/10.63367/19911599202608370400 2

Multimodal Knowledge Graph Link Prediction Model Based on Sequential Fusion

Yuhong Zhao* and Yefei Liang

Inner Mongolia University of Science and Technology, Baotou, China
{zhaoyuhong35, lyf_1998aa}@163.com

Received 21 May 2025; Revised 21 October 2025; Accepted 15 November 2025

Abstract. In the research of multimodal knowledge graph link prediction, the existing methods usually use concatenation or attention mechanism to fuse multimodal features. However, these methods fail to fully solve the problem of heterogeneity between different modalities and their complex interactions, which has a certain impact on the accuracy of link prediction. Therefore, in order to solve this problem, this paper proposes a multi-modal knowledge graph link prediction model based on sequential fusion, which uses a sequential fusion method to gradually fuse multi-modal information. In this method, a two-level fusion module is designed to fuse visual information and text information at two levels of coarse granularity and fine granularity, and then the orthogonal fusion module is used to integrate the structural information into the existing visual text features to ensure the synergy between different modalities. Moreover, regularization is added to the loss function to reduce redundant features and improve the performance of link prediction. In order to verify the effectiveness of the proposed method, experiments are carried out on FB15k-237-IMG and WN18-IMG datasets, and compared with the existing multi-modal models, the proposed model achieves advanced results.

Keywords: multimodal knowledge graph, link prediction, sequential fusion, two-level fusion, orthogonal fusion

1 Introduction

Multimodal knowledge graphs store rich information and have many applications in real life such as recommender systems, information retrieval and time series prediction [1]. Multimodal knowledge graphs have more relevant data content, such as textual and visual information. However, due to the insufficient accumulation of multimodal knowledge graph corpus, most multimodal knowledge graphs are incomplete, and there is a problem of missing triples as shown in Fig. 1. Therefore, how to mine and predict these missing information becomes the research focus of the multimodal knowledge graph link prediction task [2].

Multimodal knowledge graphs usually contain three types of information [3], textual information M_t, visual information M_v and graph structure information M_s. As shown in Fig. 1 Multimodal Knowledge Graph can extract visual information and textual information along with structural information, and synthesize different types of data through multimodal learning approach in order to enrich semantic expressions and contextual information, thus solving the problem of missing triples [4]. In multimodal knowledge graph link prediction can be categorized into embedding based approaches as well as pretraining based approaches [5]. Embedding-based methods capture commonalities by projecting modal information into a uniform space with the same relationships. However, this approach is often difficult to fully preserve the unique information of each modality, and also has limitations in modeling the interactions between different modalities. Pretraining based approaches utilize large-scale pretrained models to encode multimodal information, and the link prediction task is accomplished directly or indirectly through the feature representations of these pretrained models. Pre-training methods can better model the interactions between multimodalities and capture rich contextual semantic information, so the research in this paper is based on pretraining methods.

In this paper, a Multi-modal Knowledge Graph Link Prediction model Based on Sequential Fusion (MKGSF) is proposed. Sequential fusion [6] refers to a strategy of gradually and pair-wise fusing multi-modal data together, rather than fusing all modal data simultaneously at once. Specifically, the process of sequential fusion is as follows: first, the data of the two modalities are fused, and then the result is continued to be fused with the data

* Corresponding Author

of the third modality, and so on until all modal data are processed. This method is fundamentally different from the traditional paradigms such as early fusion, late fusion and hybrid fusion. Unlike early fusion, which simply concatenates all modalities at the input layer or shallow layer, it effectively avoids alignment difficulties and information redundancy caused by the heterogeneity of original features by successive pairing. Compared with the late fusion, which only integrates the results of independent processing at the decision level, the sequential fusion performs multiple gradual fusions in the model, which facilitates more adequate and deeper higher-order semantic interactions. Although hybrid fusion also combines fusion strategies at different levels, sequential fusion innovatively proposes an incremental fusion method, which significantly improves the interpretability of the model by systematically building a unified representation [7].

In different scenarios, the fusion order may be different. In this paper, image and text are firstly fused to provide a richer and more comprehensive semantic basis, so as to provide solid semantic support for subsequent structural information fusion and improve the effect of knowledge graph link prediction. MKGSF is composed of three modules, the first is the encoder embedding module, which extracts the feature information of each modality through different encoders. Secondly, the sequential fusion module combines visual and text features through two-level fusion, and then uses the orthogonal method to introduce structural features into the existing multimodal representation, so as to ensure the synergy between different modalities. In addition, a regularization term is added to the loss function to reduce redundant features, so as to enhance the stability and generalization ability of the model when dealing with complex multimodal information, and effectively reduce the risk of overfitting. Finally, the prediction module is used for the link prediction task of multimodal knowledge graph. The contributions of this paper can be summarized as follows:

(1) proposes a sequential fusion method that effectively combines a two-level fusion module with an orthogonal fusion module to effectively fuse textual, visual and structural information.

(2) proposes a visual text fusion module based on two-level fusion, which uses an attention-based knowledge graph alignment method at the coarse-grained level to reduce inter-modal heterogeneity. At the fine-grained level, the module fuses the features of both modalities by calculating the similarity between visual and textual features through correlation-aware fusion.

(3) An orthogonal fusion module is proposed, which maps features into orthogonal ones and incorporates regularization constraints to reduce redundant features, thus enhancing the stability and generalization ability of the model in dealing with complex multimodal information and improving the performance of link prediction.

(4) Comprehensive experiments and extensive analyses were conducted on the multimodal link prediction task. The experimental results show that the model can effectively and robustly model multimodal representations of structure, text, and vision, improving the performance of link prediction.

The contents of this article are divided into the following sections: The second chapter introduces the related work of multimodal knowledge graph, the third chapter analyzes the encoder embedding module and sequential fusion module of multimodal knowledge graph link prediction based on sequential fusion in detail, and the fourth chapter carries out experimental analysis of comparative experiments, ablation experiments and visualization experiments on FB15k-237-IMG and WN18-IMG datasets. The validity of the model is verified. Finally, the conclusion of the paper is summarized.

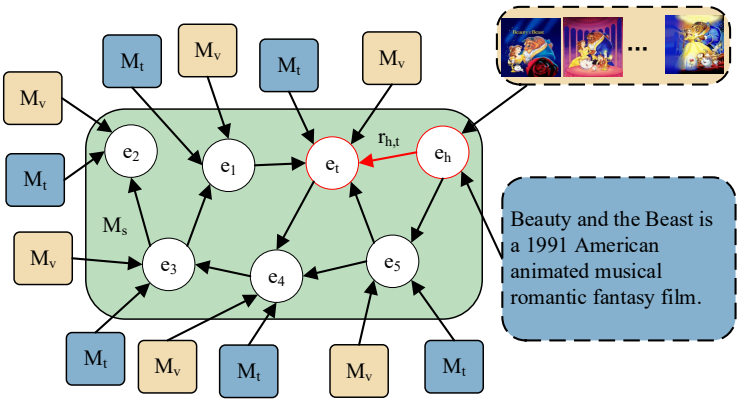

Fig. 1. Example of a multimodal knowledge map

2 Related Work

Models for multimodal knowledge graphs can be categorized into two types, embedding-based methods and pretraining based methods. Embedding based approaches usually project ternary and multimodal information into a low dimensional vector space. IKRL [8] is the first multimodal knowledge representation model based on the TransE model, which improves the embedding accuracy of the knowledge in low dimensional space by introducing visual features into the entity representation through the attentional mechanism. TransAE [9] uses the hidden layer features of an autoencoder as the entity representation. The RSME [10] model designs filter gates, forgetting gates and fusion gates to select important visual information. HRGAT [11] proposes low-rank multimodal fusion to model inter- and intra-modal dynamics. Relational graph attention networks model multimodal information. Pretraining based approaches utilize large-scale pretrained models to encode multimodal information, and subsequent tasks are accomplished directly or indirectly through feature representations of these pretrained models. MKGformer [12] was the first to utilize the Transformer architecture for a variety of multimodal tasks. The model utilizes BERT [13] and Vision Transformer [14] to extract textual and visual features and performs multilevel fusion of these two features. However, the model ignores the importance of structural information. IMF [15] extracts structural features, textual features as well as visual features using bilinear networks and comparative learning for multimodal knowledge graph for link prediction task.

The embedded-based methods pay more attention to how to tap the potential of multi-modal information through the model architecture, while the pre-training-based methods rely on the powerful ability of the pre-trained model and deeply model the multi-modal features through the fusion strategy. In order to more effectively model the interaction between multi-modalities and improve the performance of link prediction, this paper focuses on the pre-training-based method in multimodal knowledge graph. The sequential fusion method is used to solve the problem of heterogeneity and complex interaction between different modalities to further improve the accuracy of link prediction.

3 Multimodal Knowledge Graph Link Prediction Model Based on Sequential Fusion

Encoder Embedding Module In order to efficiently extract information from different modalities, the module employs different pretrained encoders for visual, textual and structural modalities respectively. Specifically, visual features V process the image using Vision Transformer to extract high-level visual features and obtain Visual Hidden States to capture the semantic information inherent in the image. Textual features T encode textual information using the BERT model to generate semantic representations Textual Hidden States with contextual relevance, which fully reflect the deep information of textual semantics. Structural Features S Extracts the topological information of entities and relations in the graph through GAT, and outputs Structure Hidden States containing structural information, which captures the topological properties among entities and relations. Secondly, the sequential fusion module adopts the strategy of combining two-level fusion and orthogonal fusion to realize the effective integration of multimodal features. The two-level fusion mechanism is first utilized to fuse visual information with textual information to generate intermediate fusion features M. Subsequently, the orthogonal fusion module is used to introduce the structural information into the existing modal feature M, which effectively avoids redundancy of information and inter-modal interference, and obtains the final multimodal features Z. Finally, in the prediction module, the multimodal features Z go through the fully connected layer to construct the required score function for the link prediction task. The overall model diagram of MKGSF is shown in Fig. 2.

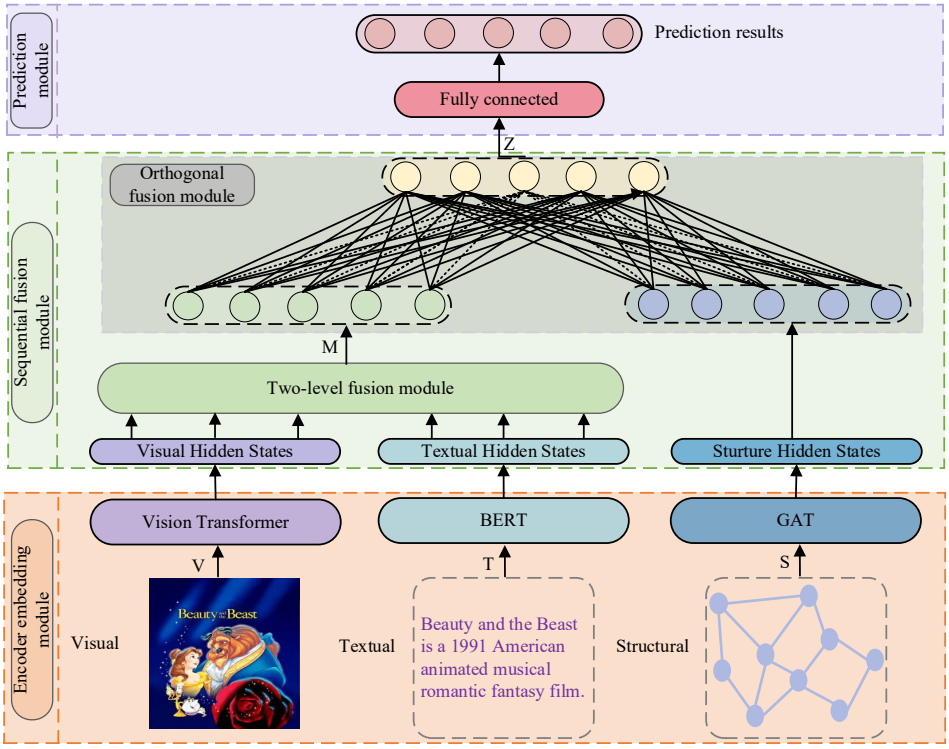

Fig. 2. MKGSF model diagram

3.1 Encoder Embedding Module

(1) The structural features of the knowledge graph are feature extracted using a graph attention network with TransE loss. During the training process, the GAT encoder uses the L_1 distance represented by the neighbor aggregation as the energy function of the triad $\mathrm{E}(h,r,t)=\|h+r-t\|$, in order to minimize the hinge loss and thus optimize the training, and the loss function is shown in equation 1 below:

$$L_{GAT}=\sum_{(h,r,t)\in T}\sum_{(h^{'},r^{'},t^{'})\in T^{'}}\max\{0,\gamma+E(h,r,t)-E(h^{'},r^{'},t^{'})\} \tag{1}$$

where r is the margin hyperparameter, T denotes the set of positive samples of fact triples, and T' denotes the set of negative triples derived from T.

(2) In this paper, a pretrained Vision Transformer model is used as a visual coder to extract image features. Using clip-vit-base-patch32 downloaded from Hugging Face, this pretrained model divides the image into fixed-size blocks and then rearranges the pixel values of each block into vector form as input to the ViT model. These block vectors represent localized regions of the image that are part of the input sequence. The ViT model then extracts the visual features of the image by transforming each block vector into a low-dimensional vector through a fully connected layer. Specifically, the model divides the image into fixed-size blocks, and the original pixel value of a given block is denoted as $I\in R^{C\times \mathrm{H}\times W}$. Where C, H and W are denoted as the number of channels, height and width of the image, respectively. Then $u=HW/P^2$ blocks of pixels are obtained. The image block embeddings of the o images are stitched together to obtain the image block embedding of the visual sequence $X_{pc}\in R^{m\times d_V}$, where $m=u\times o$, d_v denote the hidden state dimensions of the ViT. This linear transformation representation is shown in equation 2.

$$X_{pc} = I \cdot W + b \tag{2}$$

where W is the weight matrix and b is the bias vector. The first L_V layer of ViT is used as a visual coder to extract the basic feature information of the image. In this layer, the image features are processed step by step through operations such as Multi-Head Attention (MHA), Residual Connection, Layer Normalization (LN) and Feed-Forward Network (FFN). The overall process is shown in equations 3-5.

$$X_o^V = X_{pc} + V_{pos} \tag{3}$$

$$\overline{X}_l^V = MHA(LN(X_o^V)) + X_{l-1}^V, l = 1...L_V \tag{4}$$

$$X_l^V = FFN(LN(\overline{X}_l^V)) + \overline{X}_{l-1}^V, l = 1...L_V \tag{5}$$

where $V_{\text{pos}} \in R^{m \times d_V}$ denotes the embedding layer embedding at the corresponding location and X_l^V denotes the hidden state of the visual encoder L layer.

(3) In knowledge graphs, the introduction of entity descriptions can provide models with richer semantic background than the only graph structure information, which is especially important for knowledge representation learning. Entity descriptions usually contain a lot of context information and complex conceptual relationships. bert, as a bidirectional encoding model, can capture complex relationships and semantic details in language by considering context information at the same time. Specifically, the token sequence $\{w_1, ...w_n\}$ is embedded into $X_{wd} \in R^{n \times d_T}$ with a word embedding matrix, and the initial representation of the textual input embedding is shown in equation 6 as follows.

$$X_0^T = X_{wd} + T_{pos} \tag{6}$$

where $T_{pos} \in R^{(n) \times d_T}$ denotes the positional embedding. Utilizing the first L_T layer of BERT as a text encoder, which also consists of several layers similar to the MHA and FFN blocks of the visual coder, the text representation is shown in equations 7 and 8.

$$\overline{\mathrm{X}}_l^T = LN(MHA(X_0^T)) + X_{l-1}^T, l = 1...L_T \tag{7}$$

$$X_l^T = LN(FFN(\overline{X}_l^T)) + \overline{X}_l^T, l = 1...L_T \tag{8}$$

where X_l^T indicates the hidden state of the L layer of the output text sequence.

3.2 Sequential Fusion Module

Sequential fusion refers to a strategy that fuses multimodal data together in a gradual, paired fashion, rather than fusing all modalities at once and simultaneously. The module first uses a two-level fusion module to deeply interact the information from visual and textual modalities in order to fully capture the associated features between them. Subsequently, the orthogonal fusion module is introduced to further integrate the structural information into the existing multimodal representations, thus avoiding information redundancy and interference while maintaining the complementarity of the features of each modality.

Two-level Fusion Module: The module adopts two levels of coarse-grained and fine-grained approaches, aiming to efficiently integrate multi-modal features and improve the quality of information interaction. At the coarse-grained level, the module introduces an attention-based knowledge graph alignment method from the overall semantic level to reduce the heterogeneity between different modalities. At the fine-grained level, the module focuses on local feature interaction and adopts a similarity-aware Aggregator to calculate the Similarity between visual and text features, so as to obtain more accurate fusion features. This strategy can not only strengthen the matching of relevant information, but also effectively weaken the noise interference caused by irrelevant image elements and improve the quality of feature representation. Through the two-level fusion mechanism, the model can fully exploit the complementarity of vision and text, and improve the fusion effect of multi-modal features. The vision-text fusion module constructed based on the proposed method is shown in Fig. 3.

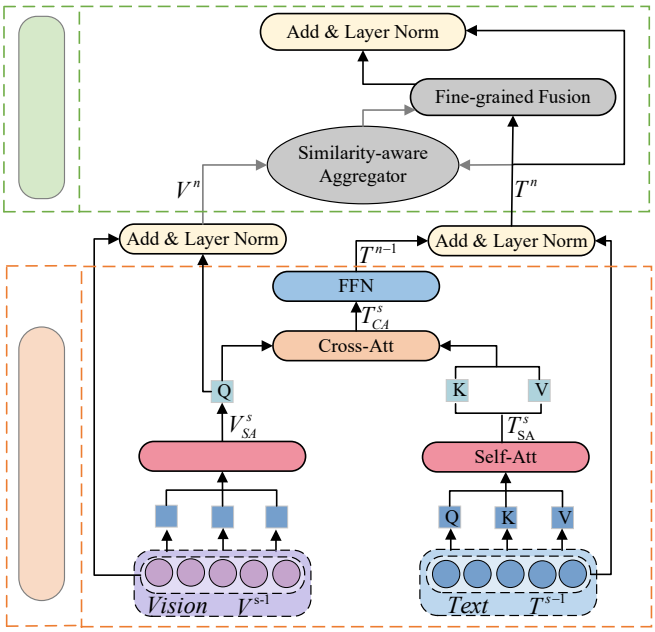

Fig. 3. Visual text fusion module based on two-level fusion

Attention-based knowledge graph alignment module, in the coarse-grained level, visual features and textual features are more focused on inter-modal relationships through different types of attention mechanisms, thus enhancing cross-modal interactions. With this approach, attention can effectively retain more linguistic information and avoid textual information being overwhelmed by visual information, while significantly reducing inter-modal heterogeneity.

The module consists of two attention mechanisms, Self-Attention and Cross-Attention, and FFN. The input visual feature V is passed through Self-Att to get V^s_{SA}, the input text feature T is fed into Self-Att to get T^s_{SA}, then the visual feature V^s_{SA} is fed into Cross-Att as Q, the text feature T^s_{SA} is fed into Cross-Att as K and V to get T^s_{CA}, and finally the obtained T^s_{CA} is fed into the feed-forward neural network layer to get the textual representation joining the visual signals, and the LN is the acronym of normalization as shown in eq. 9-eq. 12.

$$V^s_{SA} = LN(SA(V^{s-1})V^{s-1}) \tag{9}$$

$$T^s_{SA} = LN(SA(T^{s-1}) + T^{s-1}) \tag{10}$$

$$T_{CA}^{s} = LN(CA(T_{SA}^{s}, v^{s-1}) + T_{SA}^{s}) \tag{11}$$

$$T^{s} = LN(FFN(T_{CA}^{s}) + T_{CA}^{s}) \tag{12}$$

where LN stands for Layer Normalization.

At the fine-grained level, in order to reduce the noise interference brought about by irrelevant image elements, the correlation-aware fusion module adopts a weighted fusion strategy to enhance the impact of visual information by directly adjusting the representation of the text, and this approach provides an effective visual enhancement mechanism to ensure that the visual features are adequately reflected in the representation of text features. Specifically, for textual markers, the method calculates the similarity matrix between them and visual markers as shown in equation 13.

$$P = T^{n}(V^{n})^{T} \tag{13}$$

Where, V^{n} denotes the visual vector and T^{n} denotes the text vector, then perform softmax function on the similarity matrix S of the ith text marker and use the average marker aggregator on the visual markers in the image as shown in equation 14 and 15.

$$Agg_{i}(V^{n}) = soft\max(P_{i})V^{n}, (1 \le i < n) \tag{14}$$

$$Agg(V^{n}) = [Agg_{1}(V^{n}); ..., Agg_{n}(V^{n})] \tag{15}$$

Where Agg_{i} represents the similarity-aware aggregated visual representation of the ith text token. The similarity-aware aggregated visual hidden states are incorporated into the text hidden states of the FFN layer and the computation of the FFN process is modified as shown in equation 16 below.

$$M = \mathrm{Re}LU(T^{n}W_{1} + b_{1} + Agg(V^{n})W_{3})W_{2} + b_{2} \tag{16}$$

where $\mathrm{W}_{1} \in R^{d \times d_{n}}$, $\mathrm{W}_{2} \in R^{d_{n} \times d}$, $\mathrm{W}_{3} \in R^{d \times d_{m}}$ denote the new parameters added for the visual hidden state of the aggregation.

Orthogonal Fusion Module: Orthogonality is a core concept in mathematics, linear algebra and signal processing. In simple terms, two vectors are orthogonal when they are perpendicular to each other in vector space, meaning there is no shared direction or overlap between them. This is called orthogonal. This method can not only improve the generalization ability and robustness of the model, but also enhance the interpretability of the features. Like forming an ideal team of experts, orthogonal fusion ensures that each expert contributes his or her unique knowledge, works together and does his or her own job, and ultimately makes the fusion result both comprehensive and efficient. In multimodal learning, orthogonal fusion aims to jointly represent modal information features from different modalities such as visual, textual, structural, while maintaining their independence in the fusion space, thus effectively reducing the interference of redundant information. Specifically, by mapping each modality to an orthogonal space, it can be ensured that each contributes unique and non-duplicated information, thus exploiting the full potential of multi modal data. In this module, the structural information feature S and the visual textual modal information M are mapped into an orthogonalized vector space to obtain two orthogonal vector features f_S and f_M, which are then summed together to obtain the final multimodal information Z. In order to ensure the orthogonality, orthogonalization constraints are introduced into the optimization objective so that f_S and f_M are as independent as possible in the feature space. In this way, orthogonalization fusion can not only improve the representation ability of multimodal knowledge graph, but also enhance the stability and generalization ability of the model when dealing with complex multimodal information. The specific constraints are defined as shown in equation 17.

$$(f_S)^T f_M = 0 \tag{17}$$

Where f_S denotes the feature vector of structural information after orthogonal mapping and f_M denotes the feature vector of visual textual information after orthogonal mapping. This constraint indicates that the features of modality f_S and modality f_M are mathematically orthogonal, the dot product of the two is zero, reducing the overlapping information between modalities.

After obtaining the orthogonalized features f_S and f_M, they are fused together using an additive form as shown in equation 18.

$$Z = f_S + f_M \tag{18}$$

Among them, Z is a feature representation through orthogonal multimodal fusion. This representation ensures that f_S and f_M are as complementary as possible in terms of information, which reduces redundancy, enhances model independence and improves information representation.

Adding two hyperparameters λ_1 and λ_2 to the loss function to implement the L_1 and L_2 regularization constraints can optimize the complex model, prevent model overfitting, and enhance the generalization ability of the model, as shown in equation 19.

$$Loss = (-\frac{1}{N}\sum_{i=1}^{N}(y_i \log(P_i) + (1-y_i)\log(1-P_i)) + \lambda_1 \sum_j \left|\omega_j\right| + \lambda_2 \sum_{p\in w} \omega_j^2 \tag{19}$$

Where y_i is the true label of the sample i, P_i is the predicted probability of the model for the sample i, N is the number of samples, λ_1 and λ_2 are the weight coefficients of the L_1 and L_2 regularizations, respectively, $\sum_j \left|\omega_j\right|$ is the L_1 regularization term that helps the model to generate the coefficient weights, and $\sum_{p\in w} \omega_j^2$ is the L_2 regularization term, which helps to prevent the weights from being too large.

Prediction Module: This module uses the fully connected layer to predict the links of the fused multimodal features Z for knowledge graph, firstly, the fused features Z are linearly transformed $WZ + b$ to extract the correlation information between the features, and then the obtained output is mapped to the probability space using the softmax function, so as to predict the likelihood of each candidate link in the knowledge graph. This method is not only simple and efficient, but also can rank the different links at the output layer and select the most likely links, which is calculated as shown in equation 20.

$$\hat{y} = soft\max(WZ + b) \tag{20}$$

Where, $\hat{y}$ denotes the result obtained from the prediction, W denotes the weight matrix of the fully connected layer, b denotes the bias term, and the softmax function is used to convert the result after linear transformation to a probability distribution.

4 Experimentation

4.1 Multimodal Knowledge Graph Dataset

In order to validate the MKGSF, two publicly available multimodal datasets were used for validation during the experiments, i.e., the WN18-IMG dataset and the FB15k-237-IMG dataset [16]. The specific entity and relationship information of the multimodal datasets are shown in Table 1.

Table 1. Multimodal knowledge mapping dataset

Dataset	Entity	Relation	Train	Valid	Test
WN18-IMG	40943	18	141,442	5,000	5000
FB15k-237-IMG	14541	237	272,115	17,535	20466

4.2 Evaluation Indicators

For the proposed model MKGSF effectiveness is evaluated by four standard evaluation metrics for knowledge graph link prediction: mean rank (MR), Hits@10, Hits@3, and Hits@1.

MR is the average ranking of all correct results in the predicted results, the smaller the value of MR the better, which is calculated as shown in equation 21.

$$MR = \frac{1}{|S|}\sum_{i=1}^{|S|} rank_i = \frac{1}{|S|}(rank_1 + rank_2 + \cdots + rank_i) \tag{21}$$

Hits@K(H@K) denotes the ability of the algorithm to correctly predict the relationship between triples, it evaluates the ability of the knowledge graph link prediction algorithm to predict the correct triples, and it is an indispensable evaluation metric for knowledge graph link prediction algorithms. Generally, take k equal to 1, 3 or 10. the larger the metric, the better. It is the ratio of predictions whose rank is equal to or less than the threshold k as shown in equation 22.

$$H@K = \frac{\left|\{q \in Q : q < k\}\right|}{|Q|} \tag{22}$$

where q denotes the prediction/recommendation item and Q denotes all the predictions/recommendations given by the algorithm.

4.3 Baseline Model

In this paper, the model is compared to several baseline models in order to make a comprehensive comparison to demonstrate the superiority of the model in this paper.

TransAE model [9]: A multimodal feature fusion framework is established to combine the hidden layer representation of the selfencoder with the translation constraints of TransE to realize the unified embedding of graphical features through cross-modal self-supervised learning.

RSME Model [10]: Proposes an adaptive gating mechanism based on modal correlation metrics, using learnable MRP metrics to construct dynamic oblivious gating units to realize a multimodal feature screening strategy.

MKGformer model [12]: A multilevel fusion model is used to fuse visual and textual information, and is the first model to use Transformer.

SGMPT model [17]: A graph structure encoder is designed for extracting rich structural information and adding structural information to textual and visual information for link prediction task.

MMTransformer model [18]: The core of this paper is the proposal of MM-Transformer, a Transformer-based model that effectively fuses structural, visual, and textual features through symmetrical hybrid attention to improve multimodal knowledge graph link prediction.

4.4 Experimental Environment

This experiment uses the Pytorch framework to write on the Pycharm2022.1.3 version of the software, the GPU model is RTX3090, Python3.8.8 version, based on CUDA11.0 version. torchvision==0.8.2, torch==1.7.0,

torchmetrics==0.7.3, and TorchMetrics ==0.8.2. tensorboardX==2.4, pytorch-crf==0.7, pytorch_lighting==1.3.1, flax==0.3.4, etc.

4.5 Comparative Experiments

The MKGSF was compared to several baseline models to demonstrate the validity of the MKGSF model. Since the same dataset was used for each model, most of the baseline model data were derived from existing experiments. Where "-" represents missing experimental data, the best experimental results in each dataset are identified using bold and the next best results are underlined.

Table 2. Knowledge graph link prediction comparison experiment

Model	FB15k-237-IMG				WN18-IMG			
	H@1	H@3	H@10	MR	H@1	H@3	H@10	MR
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	0.243	0.360	0.499	252	0.935	0.958	0.970	**25**
SGMPT	0.252	<u>0.370</u>	0.510	238	0.943	0.966	<u>0.978</u>	<u>29</u>
MMTransformer	**0.259**	0.362	<u>0.511</u>	<u>215</u>	<u>0.948</u>	<u>0.968</u>	0.976	117
MKGSF	<u>0.258</u>	**0.372**	**0.513**	**210**	**0.950**	**0.968**	**0.979**	43

According to Table 2, which reports the results of four evaluation metrics on the FB15k-237-IMG and WN18-IMG datasets, the MKGSF model achieves the best performance on all metrics except H@1 on the FB15k-237-IMG dataset. Specifically, MKGSF outperforms the SGMPT model by 0.6% in H@1, 0.2% in H@3, 0.3% in H@10, and achieves a 28-point lower Mean Rank (MR). Compared to the MM-Transformer model, MKGSF achieves 0.1% lower H@1, but 1.0% higher H@3, 0.2% higher H@10, and 5 points lower MR. Furthermore, MKGSF outperforms the MKGformer model by 1.5% in H@1, 1.2% in H@3, 1.4% in H@10, and 42 points in MR.On the WN18-IMG dataset, MKGSF achieves 0.7% higher H@1, 0.2% higher H@3, and 0.1% higher H@10 than the SGMPT model, although its MR is 14 points higher. Compared with MM-Transformer, MKGSF shows improvements of 0.2% in H@1, 0.3% in H@10, and a substantial 74-point reduction in MR. It is worth noting that a lower MR value indicates better model performance.

Compared with the MKGformer and SGMPT models, MKGSF has a slightly higher MR under the evaluation metrics MR than both of these models, but lower MR than all other models. The reason for the lower MR than the other models may be that the two-level fusion used in the sequential fusion of the model effectively reduces the inter-modal heterogeneity as well as the orthogonal fusion used to achieve effective inter-modal interactions, and the regularization added in the loss function optimizes the model as a whole and reduces the overfitting problem. Therefore, the MR of the MKGSF model is overall better than most of the compared models, and only slightly higher than that of MKGformer and SGMPT.This may be due to the fact that, the MKGformer model designs a model specifically for visual and textual feature fusion, which is more advantageous in fusing the two modalities. While the SGMPT model designs a structural encoder specialized for extracting structural features, which enriches the extraction of structural features, and the accuracy of MR is better than that of MKGSF.

The performance of the model is improved in the following ways: first, in the encoder embedding module, different encoders are used to extract the features of each modality. Second, two-level fusion and orthogonal fusion are used in sequential fusion to gradually fuse each feature. Textual and visual information are effectively integrated in the two-level fusion, and then the structural information is added to the existing visual textual feature information through the orthogonal fusion module, which effectively fuses the three modalities, reduces the inter-modal heterogeneity, and improves the ability of interaction. Finally, the L_1 and L_2 regularization terms are added to the loss function to reduce the redundant features, which improves the stability and generalization ability of the model when dealing with complex multimodal information, and finally significantly improves the performance of the knowledge graph link prediction task.

4.6 Ablation Experiments

In order to validate the effectiveness of the multimodal knowledge graph link prediction model MKGSF, ablation experiments on the validity of modules and ablation experiments on the validity of multimodalities were conducted. The module validity ablation experiment aims to explore the specific contribution of each key module in the model to the overall performance. By gradually removing different modules in the model, we analyze the independent role of each module and its impact on the predicted performance of the final link, so as to verify the rationality and necessity of the module design. The multimodal effectiveness ablation experiments, on the other hand, focus on evaluating the actual contribution of multimodal information. Through the results of single-modal information and modal combination ablation experiments, the model performance under different modes is compared, so as to clarify the key role of multimodal information in improving the model effectiveness.

Experimental Analysis of Ablation of Modules: In order to validate the effectiveness of the different modules of the model, ablation studies are performed on the dataset WN18-IMG. (1) In order to verify the validity of the relevant perception fusion module, the relevant perception fusion module is removed and only the attention-based knowledge graph alignment module is used, denoted as w/CAF. (2) In order to verify the validity of the attention-based knowledge graph alignment module, the module is removed and only the relevant perception module is used, denoted as w/ACA. (3) The removal of the GAT extracts the structural features by verifying the structural information's validity, using the attention-based knowledge graph alignment module and the correlation-aware fusion module, denoted as w/GAT. (4) The orthogonal fusion method is changed to additive fusion only to verify the validity of orthogonalization, denoted as w/OR. as shown in Table 3.

Table 3. Ablation experimental results of different modules in the MKGSF model

Model	WN18-IMG			
	H@1	H@3	H@10	MR
w/CAF	0.944	0.959	0.968	80
w/ACA	0.945	0.956	0.969	85
w/GAT	0.941	0.958	0.972	55
w/OR	0.932	0.954	0.968	69
MKGSF	0.950	0.968	0.979	43

From Table 3. and Fig. 4, it can be seen that each module has some impact on the model performance. The performance of visual text fusion operation in two-level fusion is reduced after removing the correlation perception module because only the global fusion module is used in the fusion process to obtain the global features between the modalities, which reduces the interactive fusion between the modalities and verifies the effectiveness of the correlation perception fusion module. Removing the common attention-based knowledge graph alignment module affects the fusion of information between modalities, making the interaction between modalities become single, resulting in a decrease in accuracy, verifying the effectiveness of the attention-based knowledge graph alignment module. After removing the GAT only the two-level fusion based visual text fusion module remains, which lacks structural information and leads to a decrease in accuracy, verifying the effectiveness of structural information. When orthogonal fusion is changed to fusion by summation only, the accuracy decreases. This is because orthogonality maintains the independence of inter-modal features and avoids information redundancy and mutual interference, thus further confirming the effectiveness of orthogonal fusion. After removing the orthogonal conversion, the multimodal fusion representation utilizing only the summation to predict the results resulted in a significant decrease in the model performance in all the metrics, verifying the effectiveness of orthogonal fusion.

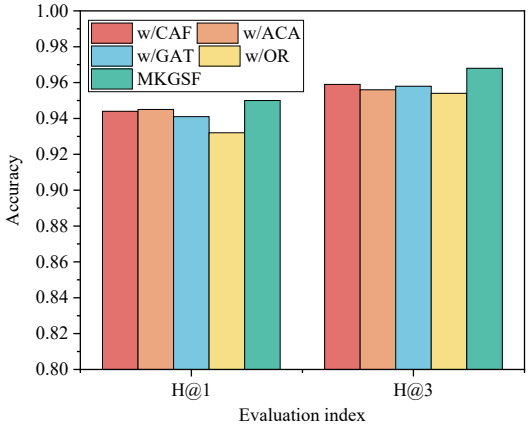

Fig. 4. Comparison plots of ablation experiment accuracy for different modules

Experimental analysis of ablation for different modes: In order to verify the effect of different modalities on this model, ablation experiments with single modality information as well as modality combinations were performed on the FB15K-237-IMG benchmark data. (1) The model is denoted as T when there is only textual information. (2) The model is denoted as S+T when there is only structural and textual information. (3) The model is denoted as V+T when there is only visual and textual information. (4) The model is denoted as S+V+T when structural, textual, and visual information have all been added. as shown in Table 4.

As can be seen from Table 4 and Fig. 5, the accuracy of the model is relatively low when only textual features are used. The model performance is significantly improved when structural information is introduced, indicating that structural information plays a crucial role in representing the attributes of entities and relationships and their intrinsic connections. In addition, when the model utilizes both visual and textual information, the performance also improves significantly, indicating the effectiveness of visual information in processing image features. Ultimately, the model achieves optimal performance when visual features, textual features and structural features work together. This indicates that the model is able to fully integrate information from different modalities and comprehensively capture the interrelationships between different modalities, with each modality embodying an indispensable role, thus achieving excellent performance in the multimodal knowledge graph link prediction task.

Table 4. Results of ablation experiments for different modes in the MKGSF model

Model	FB15k-237-IMG			
	H@1	H@3	H@10	MR
T	0.242	0.346	0.459	242
S+T	0.244	0.355	0.389	230
V+T	0.255	0.368	0.506	219
S+V+T	0.258	0.372	0.513	210

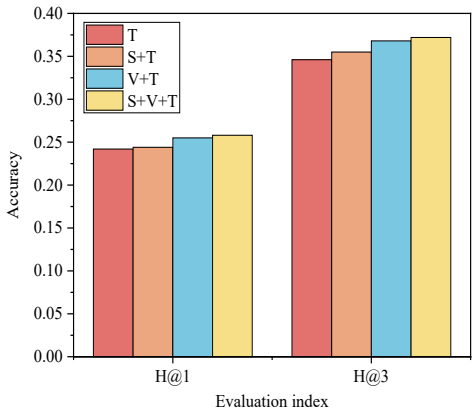

Fig. 5. Comparison plots of the accuracy of ablation experiments for different modalities

4.7 Visualization Experiments

In order to validate the effectiveness of multimodal feature fusion, the example in Fig. 6 was chosen for this experiment to visualize and analyze the predicted scores for each modal feature. The information that LeBron James plays for the Golden State Warriors is obtained from the structural features. Information that LeBron James plays for the Cleveland Cavaliers is available from the textual features, and information that LeBron James plays for the Los Angeles Lakers is available from the visual features. Different modalities provide multifaceted information about LeBron James' career, but the information from a single modality may be one-sided or incomplete. In order to synthesize the information from each modality, this experiment is based on the <LeBron James, plays for, ?> example to visualize the predicted scores for each modality, including visualization of structural features, visualization of visual features, visualization of textual features, visualization of multimodal features, and visualization of the final predicted score.

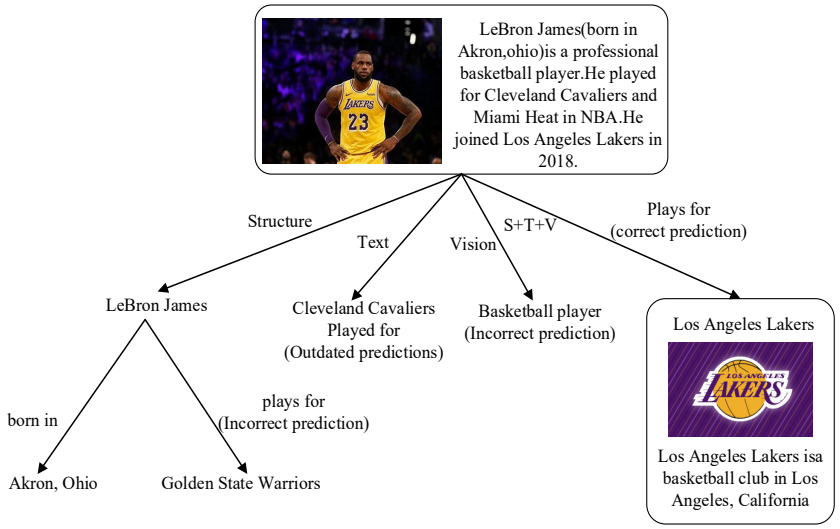

Fig. 6. Example of a multimodal knowledge map

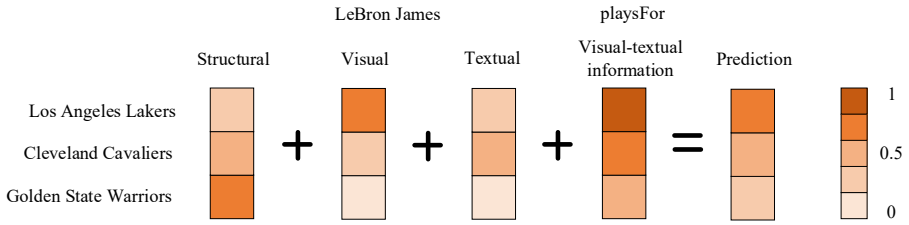

Fig. 7. Visualization of knowledge graph link prediction based on MKGSF

Fig. 7 visualizes the advantages of multimodal feature fusion through the prediction results of different modal features. The unimodal results are error-prone due to the possibility of untimely information updates or other factors in each modality. For example, when structural features are predicted, "LeBron James plays for the Golden State Warriors" has the highest score. When predicting using textual features, the highest score is obtained for "LeBron James plays for the Cleveland Cavaliers". The highest score for the result "LeBron James plays for the Los Angeles Lakers" is obtained when the prediction is made using visual features. Although the visual modality alone predicted the correct result "LeBron James plays for the Los Angeles Lakers" in this example, there is still a high risk of misclassification if the model relies on a single visual modality when the other modalities perform poorly. By integrating the complementary information of structure, text and vision, multimodal fusion can effectively alleviate the inconsistency between modalities and improve the robustness and reliability of the overall prediction.

5 Conclusion

In this paper, we provide an indepth analysis of multimodal knowledge graph link prediction and propose a multimodal knowledge graph link prediction framework, MKGSF, which addresses the problem of heterogeneity and insufficient interaction between different modalities. A sequential fusion method is designed to fuse information from multiple modalities. Sequential fusion is represented as the gradual integration of the features of each modality through a two-level fusion module and an orthogonal fusion module, which gives full play to the advantages of each modality and achieves more efficient feature extraction and fusion during multimodal information processing. MKGSF uses three different pretrained encoders to represent the feature embedding of the textual information, the visual information, and the structural information, respectively. Secondly, the information of each modality is gradually fused by a sequential fusion method, which first fuses visual and textual information using a two-level fusion method, and then adds structural information to the multimodal features using orthogonal fusion, and optimizes the model to prevent overfitting by adding L_1 and L_2 regularization to the loss function. In this paper, comprehensive experiments are conducted on two public datasets, and the results show that the MKGSF model effectively improves the performance of multimodal knowledge graph link prediction. However, MKGSF also has some limitations. On the one hand, the model uses multiple pre-trained encoders, which requires high computing and memory resources. On the other hand, the deviation of data distribution leads to the model tends to learn one-sided rules rather than general logic in data, which weakens its adaptability and robustness to sparse relationships and new fields. Future research will focus on designing more adaptive fusion schemes and optimizing the model efficiency to be validated on more evenly distributed datasets to better adapt to large-scale and diverse application scenarios.

Acknowledgement

This work is supported by the Basic Scientific Research Funds of Universities directly under the Autonomous Region (2024YXXS046), Inner Mongolia Natural Science Foundation (2026MS0457).

References

[1] Y. Xu, Y. Li, M. Xu, Z. Zhu, Y. Zhao, HKA: A Hierarchical Knowledge Alignment Framework for Multimodal Knowledge Graph Completion, ACM Transactions on Multimedia Computing, Communications and Applications 20(8) (2024) 1–19.
https://doi.org/10.1145/3664288

[2] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, O. Hang, Otkge: Multi-modal knowledge graph embeddings via optimal transport, in: Proc. of Advances in neural information processing systems, 2022.

[3] S. Deng, N. Zhang, W. Zhang, J. Chen, J.Z. Pan, H. Chen, Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolutional Network, in: Proc. of Companion of the 2019 World Wide Web Conference, 2019.
https://doi.org/10.1145/3308560.3317701

[4] F. Zhao, C. Zhang, B. Geng, Deep multimodal data fusion, ACM computing surveys 56(9)(2024) 1–36.
https://doi.org/10.1145/3649447

[5] B. Shang, Y. Zhao, J. Liu, D. Wang, LAFA: Multimodal Knowledge Graph Completion with Link Aware Fusion and Aggregation, in: Proc. of the AAAI Conference on Artificial Intelligence, 2024.
https://doi.org/10.1609/aaai.v38i8.28744

[6] S. Labbaki, P. Minary, Orthogonal Sequential Fusion in Multimodal Learning, in: Proc. of 28th International Conference on Information Fusion, 2025.
https://doi.org/10.23919/FUSION65864.2025.11124022

[7] Y. Yuan, Z. Li, B. Zhao, A Survey of Multimodal Learning: Methods, Applications, and Future, ACM Computing Surveys 57(7)(2025) 1–34.
https://doi.org/10.1145/3713070

[8] R. Xie, Z. Liu, H. Luan, M. Sun, Image-embodied knowledge representation learning, in: Proc. of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.

[9] Z. Wang, L. Li, Q. Li, D. Zeng, Multimodal Data Enhanced Representation Learning for Knowledge Graphs, in: Proc. of the International Joint Conference on Neural Networks, 2019.
https://doi.org/10.1109/IJCNN.2019.8852079

[10] 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: Proc. of the 29th ACM International Conference on Multimedia, 2021.
https://doi.org/10.1145/3474085.3475470

[11] S. Liang, A. Zhu, J. Zhang, J. Shao, Hyper-node relational graph attention network for multi-modal knowledge graph completion, ACM Transactions on Multimedia Computing, Communications and Applications 19(2)(2023) 1–21.
https://doi.org/10.1145/3545573

[12] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, H. Chen, Hybrid transformer with multi-level fusion for multimodal knowledge graph completion, in: Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022.
https://doi.org/10.1145/3477495.3531992

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
https://doi.org/10.18653/v1/N19-1423

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. of International Conference on Learning Representations 2021, 2021.
https://openreview.net/forum?id=YicbFdNTTy

[15] X. Li, X. Zhao, J. Xu, Y. Zhang, C. Xing, IMF: Interactive multimodal fusion model for link prediction, in: Proc. of the ACM Web Conference 2023, 2023.
https://doi.org/10.1145/3543507.3583554

[16] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, in: Proc. of Advances in Neural Information Processing Systems 32, 2019.

[17] K. Liang, S. Zhou, Y. Liu, L. Meng, M. Liu, X. Liu, Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. <https://arxiv.org/abs/2307.03591>, 2023.

[18] D. Wang, K. Tang, J. Zeng, Y. Pan, Y. Dai, H. Li, B. Han, MM-transformer: a transformer-based knowledge graph link prediction model that fuses multimodal features, Symmetry 16(8)(2024) 961.
https://doi.org/10.3390/sym16080961