Essay

MM-Transformer: a Transformer-based Knowledge Graph Link Prediction Model by Fusing Multimodal Features

Dongsheng Wang *, Kangjie Tang, Jun Zeng, Yue Pan, Yun Dai, Huige Li, Bin Han

Posted Date: 5 July 2024

doi: 10.20944/preprints202407.0495.v1

Keywords: knowledge graph; multimodal features; link prediction

Article

MM-Transformer: A Transformer-Based Knowledge Graph Link Prediction Model by Fusing Multimodal Features

Dongsheng Wang [1,*], Kangjie Tang [1], Jun Zeng [1], Yue Pan [1], Yun Dai [2], Huige Li [1] and Bin Han [1]

[1] School of computing, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China
[2] Jiangsu Justice police Vocational College, Nanjing, Jiangsu, China
* Correspondence: jsjxy_wds@just.edu.cn

Abstract: Multimodal knowledge graph completion necessitates the integration of information from multiple modalities (such as images and text) into the structural representation of entities to improve link prediction. However, most existing studies have overlooked the interaction between different modalities. To address this issue, this paper proposed a Transformer-based knowledge graph link prediction model (MM-Transformer) that fuses multimodal features. Different modal encoders are employed to extract structural, visual and textual features, and hybrid key-value calculations are performed on features from different modalities based on the Transformer architecture. The similarities of textual tags to structural tags and visual tags are calculated and aggregated respectively, and multimodal entity representations are modeled and optimized to reduce the heterogeneity of the representations. Experimental results demonstrate that, compared to the current multimodal state-of-the-art methods, the proposed method achieves significant performance improvements in knowledge graph link prediction tasks. This proves that the proposed method effectively addresses the problem of multimodal feature fusion in knowledge graph link prediction tasks.

Keywords: knowledge graph; multimodal features; link prediction

1. Introduction

In recent years, knowledge graphs have played a vital role in real-world tasks such as question answering systems [1–3], recommender systems [4,5], and information retrieval [6,7]. However, due to the issue of missing triples, knowledge graphs cannot encompass all knowledge. To address this problem, link prediction techniques have been introduced to predict missing relationship triplets [8]. Traditional link prediction methods, such as translation-based approaches [9] and neural network methods [10,11], often fall short in accurately capturing and processing the relationships between triples due to limitations in model design or the training process. Recently, some studies [12–14] have tackled this issue by enriching datasets and proposing new models to capture multimodal information. However, these methods typically project all modalities into a unified space, failing to effectively capture the complex interactions between modalities, thus limiting their performance. As a result, the multimodal link prediction task has been proposed to overcome these limitations.

Taking Figure 1 as an example, Stephen Curry and LeBron James were both born in Akron, Ohio. This similarity may lead to incorrect graph structure predictions, such as predicting that LeBron James is a player for the Golden State Warriors. Outdated predictions may also arise based on the text description "played for." Visual features alone can only determine that LeBron James is a basketball player. However, by integrating multimodal knowledge such as geographic location and team information, it can accurately predict that LeBron James is a member of the Los Angeles Lakers. Inspired by multimodal representation learning [15], this paper fully leverages the graph structure, visual, and textual features of the knowledge graph and performs multimodal feature fusion based on the Transformer architecture to achieve link prediction on the knowledge graph. The contributions of this paper are summarized as follows:

(1) This paper proposed a method to fuse multimodal features by combining graph structures,

 doi:10.20944/preprints202407.0495.v1

visual elements, and textual information to create a more comprehensive data representation. Compared to methods that utilize single-modal features, this approach provides richer information, thereby significantly enhancing the model's performance and generalization ability.

(2) By comprehensively considering the similarity scores of structural, visual, and textual features, our model demonstrates superiority in the knowledge graph link prediction task. Experimental results show that, compared to existing multimodal methods, our model significantly improves multiple indicators. Additionally, by analyzing the contribution of each feature to the final result, the interpretability and credibility of the model are enhanced.

(3) Through case analysis, multimodal feature fusion effectively reduces the biases that may be caused by single-modal features. By combining information from different modalities, the model can make predictions more accurately and robustly, demonstrating the importance and advantages of multimodal feature fusion in complex tasks.

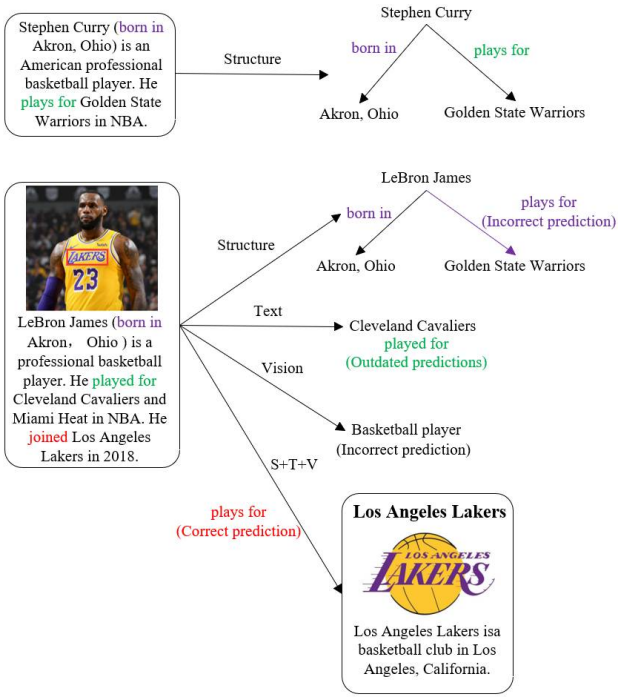

Figure 1. Example of knowledge graph link prediction by fusing multimodal features.

2. 2. Related Work

To achieve multimodal link prediction, researchers have proposed a series of multimodal pre-training methods and pre-trained them on a corpus of image-text pairs, such as ViLBERT [16], VisualBERT [17], and UNITER [18], which effectively improve the processing ability of multimodal information. In addition, Xie et al. [14] proposed to integrate image features into a typical knowledge graph representation learning model for multimodal link prediction. On the other hand, Sergieh et al. [19] and Wang et al. [20] jointly encoded and fused visual and structural knowledge through simple concatenation and autoencoder for multimodal link prediction, respectively. Existing multimodal link prediction methods mainly focus on encoding image features into knowledge graph embeddings. Xie et al. [14] extended TransE to obtain visual representations corresponding to knowledge graph entities and knowledge graph structural information, respectively. Sergieh et al. [19], Wang et al. [20], and Zhao et al. [21] further proposed several fusion strategies to encode visual and structural features into a unified embedding space. Recently, Wang et al. [22] studied the noise introduced by irrelevant images corresponding to entities and designed a forget gate with the MRP metric to select valuable images for multimodal knowledge graph completion.

 doi:10.20944/preprints202407.0495.v1

In order to improve the performance of multimodal feature fusion, researchers have proposed a series of innovative methods. Shankar et al. [23] proposed a progressive fusion method that establishes connections between different layers of the model so that the information of the deep fusion can be used by the shallow layer, avoiding information loss while maintaining the advantages of late fusion. Liang et al. [24] proposed an information theory-based method to define the upper and lower bounds of multimodal interaction and showed how these bounds accurately reflect the true degree of interaction. Finally, they explained how these theoretical results can help evaluate model performance, guide data collection, and select appropriate multimodal models. Jiang et al. [25] proposed a new method called HyperRep, which uses hypergraphs to capture complex high-order relationships between different data types and combines the information bottleneck principle to improve the fusion effect of multimodal data. Golovanevsky et al. [26] proposed a new one-to-many (OvO) attention mechanism, which achieves linear expansion with the increase in the number of modalities and reduces computational complexity. Zhang et al. [27] proposed a method called MLA, which reduces the interference between different modalities by alternately learning a single modality and captures the interaction between different modalities by sharing parts, thereby improving the effect of multimodal representation learning. Experimental results on different tasks demonstrate the effectiveness and versatility of these methods.

However, although previous research on knowledge graph link prediction based on multimodal features has made some progress, it still has obvious limitations such as architectural universality and modal contradictions. It is necessary to propose a unified model to more effectively expand the application of multimodal knowledge graph completion and solve the contradictions in modal fusion.

3. Methodology

Formally, the knowledge graph [28] is defined as $\mathcal{G} = < \mathcal{E}, \mathcal{R}, \mathcal{T} >$, where $\mathcal{E}$ represents the set of entities, $\mathcal{R}$ represents the set of relations, and $\mathcal{T} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ represents the relation triple of the knowledge graph. In the knowledge graph based on multimodal features, each entity is represented not only by text features, but also by structural features and visual features. Therefore, this paper defines the modal feature set of the entity $\mathcal{M} = \{s, v, t\}$, where s, v, t represent the structural modal features, visual modal features, and textual structural features of the entity, respectively.

Multimodal link prediction is one of the knowledge graph completion tasks. It is to predict the tail entity given the entity and relationship, expressed as $(\boldsymbol{e}_h, r, ?)$. For a given entity $\boldsymbol{e}_h$, according to the graph structure S_h and image I_h related to entity $\boldsymbol{e}_h$, this paper models the distribution of the tail entity $\boldsymbol{e}_t$, expressed as $p(\boldsymbol{e}_t | (\boldsymbol{e}_h, r, (S_h, I_h)))$. First, feature extraction is performed on structure, vision and text, then the entity representation of structure-text-image fusion is modeled, and finally the missing entity is predicted based on the multimodal entity representation.

3.1. Overall Architecture

As shown in Figure 2, the model MM-Transformer in this paper uses the basic Transformer model to model the entity representation of structure-text-image fusion. A graph neural network model GraphSAGE is used at the bottom layer to extract the structural features of the knowledge graph, and the pre-trained models ViT and BERT are used to extract visual and text features. In the upper layer, an improved prefix-guided interaction module [29] and a relevance-aware fusion module [29] are used to model the multimodal entity representation, and then link prediction is performed based on the multimodal entity representation.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2024 doi:10.20944/preprints202407.0495.v1

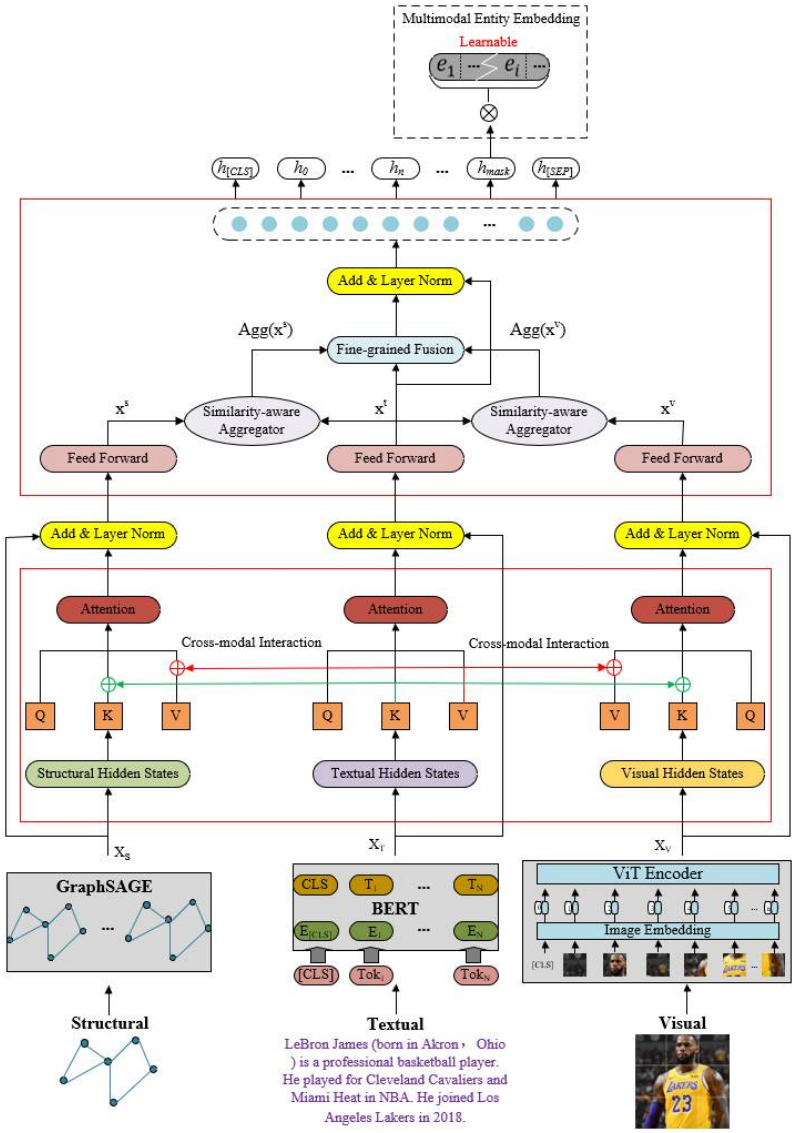

Figure 2. MM-Transformer model overall architecture.

3.2. Structural Feature Extraction

Knowledge graph is a graph-based data structure consisting of nodes and edges. Each node represents an "entity" and each edge is a "relationship" between entities. Knowledge graph is essentially a semantic network. For example, if there is a relationship triple (entity 1, relationship, entity 2), then there will be a directed edge from entity 1 to entity 2 in the graph, indicating that there is a specified relationship between entity 1 and entity 2. Such a graph can capture the relationship and connection between entities well. Through the graph neural network model, we can learn the representation of nodes in the graph and calculate the importance of each node. For example, if a user has many friends, he may be more important in the social network. This can be achieved through the self-attention mechanism, that is, dynamically adjusting the weight according to the relationship between the user and his neighbor nodes, thereby reflecting the importance of the user in the network. In addition to the importance of nodes, we can also calculate the mutual relationship between nodes. For example, whether there is a direct friend relationship between two users, or whether there is an indirect association relationship between them (through common friends, etc.). This can be achieved by calculating the attention weights between nodes, so that the nodes are weighted and aggregated to reflect their mutual relationship in the network.

Assume that the nodes in the knowledge graph are represented as $\{v_1, v_2, \cdots, v_k\}$ and the edges are represented as $\{(v_i, v_j)\}$. GraphSAGE [30] is used as the structural encoder to calculate the structural features of each node v_i. For example, for a node representing a person, its structural features can include its position information in the graph. GraphSAGE (Graph Sample and

 doi:10.20944/preprints202407.0495.v1

Aggregation) is a graph neural network model for learning node representation in graph structures. It adopts a sampling and aggregation strategy to update the node representation by aggregating the node's neighbor information. Each node can be represented as a vector of fixed dimension, which contains the structural information of the node and its position relationship in the graph. Specifically, the feature representation obtained using GraphSAGE refers to the abstract representation of each node in the graph structure. These representation vectors capture the node's neighbor information, which can help this paper understand the relationship between nodes, discover hidden patterns and rules, and capture the relationship and connection between nodes in the graph.

Let x_i represent the hidden state of node v_i in the graph neural network model. The following formula can be used to calculate the hidden state x_i of node v_i:

$$x_i = \sigma(AGG(\{h_j^{(l-1)} | \forall j \in \mathcal{N}(v_i)\}) \cdot W^{(l)}), \quad (1)$$

where σ is the activation function, AGG is the aggregation function, $h_j^{(l-1)}$ is the hidden state of node v_j in the previous layer $l-1$, $\mathcal{N}(v_i)$ represents the set of neighbor nodes of node v_i, and $W^{(l)}$ is the weight matrix. GraphSAGE gradually updates the representation of nodes through multiple layers of aggregation operations to obtain the final structural feature output $X_S \in \mathbb{R}^{k \times d_S}$. In each layer of aggregation operation, the neighbor information of the node will be aggregated and weighted to finally obtain a new table for each node.

3.3. Visual Feature Extraction

This paper adopts ViT [31] as a visual encoder to extract image features. The image representation process involves dividing the image into fixed-size blocks and rearranging the pixel values of each block into a vector form, which is provided as input to the ViT model. These block vectors are regarded as part of the input sequence and represent the local area in the image. The ViT model then transforms each block vector into a low-dimensional vector representation to capture the visual features of the image. This transformation is implemented by a fully connected layer that maps the pixel values of the block vector into a smaller feature space.

Specifically, the image is divided into e blocks of fixed size. The original pixel value of a given block is represented as $Z \in \mathbb{R}^{H \times W \times C}$, where H, W, C represent the height, width, and number of channels of the image, respectively. Then, the original pixel values are rearranged into a one-dimensional vector $z \in \mathbb{R}^{N \times C}$, where $N = H \times W$ represents the total number of pixels in the block. Next, a linear transformation is applied to map the block vector z into a low-dimensional feature space to obtain the embedded block vector $x \in \mathbb{R}^{N \times d_V}$. This linear transformation is expressed as:

$$x = z \cdot W + b, \quad (2)$$

where W is the weight matrix and b is the bias vector. In this paper, the tiles of a images are embedded and concatenated to obtain the visual feature output $X_V \in \mathbb{R}^{m \times d_V}$, where $m = (e \times a)$.

3.4. Text Feature Extraction

This paper uses BERT [32] as a text encoder to extract semantic features of text, including word-level and sentence-level semantic representations, as well as context-aware and multi-level semantic information. The text is fed into the BERT model for processing, and each input text is converted into a sequence of words (tokens), represented as $X = \{x_1, x_2, \cdots, x_n\}$, where x_i represents the i-th word. For each input word x_i, an embedding operation is performed to convert it into a vector representation. A word embedding matrix E is used, in which each row corresponds to the embedding vector of a word. The word embedding operation is expressed as:

$$Embedding(x_i) = E_{x_i}, \quad (3)$$

where E_{x_i} represents the embedding vector corresponding to word x_i. After the word embedding operation, the text feature output $X_T \in \mathbb{R}^{n \times d_T}$ is obtained.

3.5. Multimodal Feature Fusion

Existing multimodal feature fusion methods mainly focus on projecting the representations of different modalities into a unified space and exploiting the commonalities between modalities for

 doi:10.20944/preprints202407.0495.v1

prediction, but this approach may fail to preserve modality-specific knowledge. This paper alleviates this problem by fusing features from different modalities at multiple layers, thereby further leveraging their complementarity.

This paper uses the first L_S layer of GraphSAGE as a structural encoder to obtain the basic graph structure information of the knowledge graph, that is, the structural feature output X_S. X_S is used as the input of Multi-Head Self-Attention (MHA), and the structural representation is calculated as follows:

$$X_0^S = X_S + S_{pos},$$

$$\overline{X_l^S} = MHA\left(LN(X_0^S)\right) + X_{l-1}^S, \quad l = 1 \cdots L_S, \tag{4}$$

$$X_l^S = FFN\left(LN\left(\overline{X_l^S}\right)\right) + \overline{X_l^S}, \quad l = 1 \cdots L_S,$$

where $S_{pos} \in \mathbb{R}^{k \times d_S}$ represents the corresponding position embedding and X_l^S is the hidden state of the l-th layer of the structure encoder.

This paper uses the first L_V layer of ViT as a visual encoder to obtain the basic feature information of the image, that is, the visual feature output X_V. X_V is used as the input of Multi-Head Self-Attention (MHA), and the visual representation is calculated as follows:

$$X_0^V = X_V + V_{pos},$$

$$\overline{X_l^V} = MHA\left(LN(X_0^V)\right) + X_{l-1}^V, \quad l = 1 \cdots L_V, \tag{5}$$

$$X_l^V = FFN\left(LN\left(\overline{X_l^V}\right)\right) + \overline{X_l^V}, \quad l = 1 \cdots L_V,$$

where $V_{pos} \in \mathbb{R}^{m \times d_S}$ represents the corresponding position embedding and X_l^V is the hidden state of the l-th layer of the structure encoder.

This paper uses the first L_T layer of BERT as a visual encoder to obtain the basic feature information of the text, that is, the text feature output X_T. X_T is used as the input of Multi-Head Self-Attention (MHA), and the text representation is calculated as follows:

$$X_0^T = X_T + T_{pos},$$

$$\overline{X_l^T} = MHA\left(LN(X_0^T)\right) + X_{l-1}^T, \quad l = 1 \cdots L_T, \tag{6}$$

$$X_l^T = FFN\left(LN\left(\overline{X_l^T}\right)\right) + \overline{X_l^T}, \quad l = 1 \cdots L_T,$$

where $T_{pos} \in \mathbb{R}^{m \times d_S}$ represents the corresponding position embedding and X_l^T is the hidden state of the l-th layer of the structure encoder.

In the MHA of the Transformer [33] model, the input consists of query (Q), key (K), and value (V) vectors obtained by three linear transformations. Therefore, this paper performs multi-head attention calculations on mixed keys and values in each layer of the attention layer to control the degree of interaction between structural features and visual features and text features to reduce modal heterogeneity in advance. The structural head head_i^S, visual head head_i^V, and text head head_i^T in the original $\text{head}_i = Attention(Q^i, K^i, V^i)$ are redefined, and the calculation process is as follows:

$$Q^i, K^i, V^i = xW_q^i, xW_k^i, xW_v^i,$$

$$head_i^T = Attention(x^t W_q^t, x^t W_k^t, x^t W_v^t),$$

$$head_i^S = Attention(x^s W_q^s, [x^s W_k^s, x^t W_k^t], [x^s W_v^s, x^t W_v^t]), \tag{7}$$

$$head_i^V = Attention(x^v W_q^v, [x^v W_k^v, x^t W_k^t], [x^v W_v^v, x^t W_v^t]),$$

This paper also derives a variant formula of Equation 7:

 doi:10.20944/preprints202407.0495.v1

$$\begin{aligned} head_i^S &= Attention(x^s W_q^s, [x^s W_k^s, x^t W_k^t], [x^s W_v^s, x^t W_v^t]) \\ &= softmax(Q_s[K_s; K_t]^T)\begin{bmatrix} V_s \\ V_t \end{bmatrix} \\ &= \big(1-\lambda(x^s)\big)softmax(Q_s K_s^T)V_s + \lambda(x^s)softmax(Q_s K_t^T)V_t \\ &= \big(1-\lambda(x^s)\big)\underbrace{Attention(Q_s, K_s, V_s)}_{standard\ attention} + \lambda(x^s)\underbrace{Attention(Q_s, K_t, V_t)}_{cross-modal\ interaction} \end{aligned} \tag{8}$$

$$\lambda(x^s) = \frac{\sum_i exp\,(Q_s K_t^T)_i}{\sum_i exp\,(Q_s K_t^T)_i + \sum_j exp\,(Q_s K_s^T)_j}, \tag{9}$$

where $\lambda(x^s)$ represents the mixing parameter of the normalized attention weight, which is a scalar used to weight the structural key and value vectors. This parameter is used to control the degree of interaction between structural features and textual features.

$$\begin{aligned} head_i^V &= Attention(x^v W_q^v, [x^v W_k^v, x^t W_k^t], [x^v W_v^v, x^t W_v^t]) \\ &= softmax(Q_v[K_v; K_t]^T)\begin{bmatrix} V_v \\ V_t \end{bmatrix} \\ &= \big(1-\lambda(x^v)\big)softmax(Q_v K_v^T)V_v + \lambda(x^v)softmax(Q_v K_t^T)V_t \\ &= \big(1-\lambda(x^v)\big)\underbrace{Attention(Q_v, K_v, V_v)}_{standard\ attention} + \lambda(x^v)\underbrace{Attention(Q_s, K_t, V_t)}_{cross-modal\ interaction} \end{aligned} \tag{10}$$

$$\lambda(x^v) = \frac{\sum_i exp\,(Q_v K_t^T)_i}{\sum_i exp\,(Q_v K_t^T)_i + \sum_j exp\,(Q_v K_v^T)_j}, \tag{11}$$

where $\lambda(x^v)$ represents the mixing parameter of the normalized attention weights, a scalar used to weight the visual key and value vectors. This parameter is used to control the degree of interaction between visual features and textual features.

In order to mitigate the negative impact of irrelevant structures and images on text, this paper conducts cross-modal interaction based on tags to achieve alignment between structure, text and image. This paper uses k, m and n to represent the sequence lengths of the structural vector $x^s \in \mathbb{R}^{k\times d}$, the visual vector $x^v \in \mathbb{R}^{m\times d}$ and the text vector $x^t \in \mathbb{R}^{n\times d}$, which are the corresponding output features of multimodal features after multi-head self-attention (MHA) and fully connected feedforward network (FFN). For text tags, this paper calculates the similarity matrix with all structural tags and visual tags respectively, and then performs softmax function processing, and performs average tag aggregation on the structural tags and visual tags. The specific process is as follows:

$$S = x^t (x^s)^T, \quad S = x^t (x^v)^T, \tag{12}$$

$$Agg_i(x^s) = softmax(S_i)x^s, \quad (1 \le i \le n),$$

$$Agg(x^s) = [Agg_1(x^s), \cdots, Agg_n(x^s)], \tag{13}$$

$$Agg_i(x^v) = softmax(S_i)x^v, \quad (1 \le i \le n),$$

$$Agg(x^v) = [Agg_1(x^v), \cdots, Agg_n(x^v)], \tag{14}$$

where $\mathrm{Agg_i}(x^s)$ represents the structural representation of the similarity-aware aggregation of the i-th text token, and $\mathrm{Agg_i}(x^v)$ represents the visual representation of the similarity-aware aggregation of the i-th text token. The structural and visual hidden states of the similarity-aware aggregation are then merged into the text hidden state in the FFN layer, calculated as follows:

$$FFN(x^t) = ReLU(x^t W_1 + b_1 + Agg(x^s)W_3 + Agg(x^v)W_4)W_2 + b_2, \tag{15}$$

 doi:10.20944/preprints202407.0495.v1

where $W_1 \in \mathbb{R}^{d \times d_n}$, $W_2 \in \mathbb{R}^{d_n \times d}$, $W_3 \in \mathbb{R}^{d \times d_k}$ represent the newly added parameters of the structural hidden states for aggregation, and $W_4 \in \mathbb{R}^{d \times d_m}$ represents the newly added parameters of the visual hidden states for aggregation.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

This paper uses two publicly available multimodal link prediction datasets, namely: WN18-IMG: This dataset is an extended version of the knowledge graph WN18 [8] extracted from WordNet [34]. Different from the original WN18, each entity in WN18-IMG is accompanied by 10 images. This multimodal dataset provides richer information to support the link prediction task. FB15K-237-IMG: This dataset is a subset of Freebase [35], which is a large-scale knowledge graph that provides rich multimodal information to help the model understand the relationship between entities. Compared with the regular FB15K-237, each entity in FB15K-237-IMG [8] also has 10 images. Detailed statistics of these datasets can be found in Table 1.

Table 1. Dataset statistics for Multimodal Link Prediction.

Dataset	Ent	Rel	Train	Dev	Test
FB15K-237-IMG	14,541	237	272,115	17,535	20,466
WN18-IMG	40,943	18	141,442	5,000	5,000

During the evaluation process, we consider four effective entity metrics to measure the performance of the model, which follow previous studies: (1) mean rank (MR); (2) hit rate (Hits@1, Hits@3 and Hits@10).

4.1.2. Baselines

This paper compares with several baseline models to fully demonstrate the superiority of this model. The baselines selected in this paper include:

- VisualBERT [17], a pre-trained vision-language model with a single-stream structure.
- ViLBERT [16], a pre-trained vision-language model with a two-stream structure.
- IKRL [36], which extends TransE to learn the visual representation of entities and the structural information of knowledge graphs respectively.
- TransAE [20], combines a multimodal autoencoder with TransE to encode visual and texture knowledge into a unified representation and uses the hidden layers of the autoencoder as the representation of entities in the TransE model.
- RSME [22], which designs a forget gate with MRP metric to select valuable images for multimodal knowledge graph embedding learning.

4.1.2. Experiment Details

This paper adopts a training method based on the Lightning framework. In order to ensure the reproducibility of the experimental results, this paper sets the random seed to 42. The task of this paper is based on the knowledge graph, so the corresponding data class is selected. For regularization, this paper chooses a smaller dropout, set to 0.1. During the training process, this paper adopts a smaller batch size of 96 and sets the gradient accumulation step to 1. In order to improve the generalization ability of the model, this paper adopts the label smoothing technique and sets the label smoothing parameter to 0.3. This paper selects the pre-trained models ViT and BERT and uses the binary cross entropy loss function (BCE). This paper also configures the model to save the checkpoint after each cycle and check the performance of the validation set.

5. Experimental Results

5.1. Overall Performance

The experimental results in Table 2 show that it is necessary to effectively integrate structural features, visual features and text features to improve the link prediction performance of the

 doi:10.20944/preprints202407.0495.v1

knowledge graph. This shows that the model proposed in this paper is superior, especially in capturing multimodal information.

First, judging from the results on the FB15K-237-IMG dataset, MM-Transformer performs significantly better than other methods. Compared with the current multimodal SOTA method RSME, MM-Transformer improves by 7.02%, 5.23%, and 9.42% on the three evaluation indicators of Hits@1, Hits@3, and Hits@10, respectively. In addition, MM-Transformer also performs better on the MR (Mean Rank) indicator, which is reduced to 215. Secondly, the performance on the WN18-IMG dataset is also outstanding. MM-Transformer's scores on Hits@1, Hits@3, and Hits@10 are 0.948, 0.968, and 0.976, respectively, which are higher than RSME and other comparison methods. This further proves the universality and robustness of the model proposed in this paper on different datasets. These results show that MM-Transformer can more effectively integrate multimodal features, thereby improving the accuracy of link prediction.

Table 2. Results of the 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↓
VisualBERT_base [17]	0.217	0.324	0.439	592	0.179	0.437	0.654	122
ViLBERT_base [16]	0.233	0.335	0.457	483	0.223	0.552	0.761	131
IKRL [36]	0.194	0.284	0.458	298	0.127	0.796	0.928	596
TransAE [20]	0.199	0.317	0.463	431	0.323	0.835	0.934	352
RSME [22]	0.242	0.344	0.467	417	0.943	0.951	0.957	223
MM-Transformer	**0.259**	**0.362**	**0.511**	**215**	**0.948**	**0.968**	**0.976**	**117**

Compared with multimodal methods that process different modal features separately, MM-Transformer combines multimodal features for joint learning, which helps to simultaneously consider the common characteristics and complementary relationships between different modal features. The experimental results clearly show that MM-Transformer has significant advantages in capturing and fusing multimodal features. This not only verifies the effectiveness of the model design, but also provides strong support for the use of multimodal fusion methods in knowledge graph link prediction tasks in the future.

5.2. Ablation Study

Table 3 presents the evaluation results of fusing different modal features for the FB15K-237-IMG dataset. Here, this paper uses S to represent structural features, V to represent visual features of images, and T to represent text features of descriptions. It can be observed that when structural, visual or text features are introduced, the performance of the model is significantly improved. This shows the effectiveness of multimodal feature fusion.

As can be seen from Table 3, when text features (T) are used alone, the performance of the model is relatively low. This shows that when only text features are used for link prediction, the accuracy and ranking performance of the model are limited. When structural features (S) are added, the model performance is significantly improved. This shows that structural features play an important role in providing basic information about entities and relationships, which helps to improve the link prediction performance of the model. Similarly, when visual features (V) are added, the model performance is also significantly improved. This shows that the advantage of visual features in capturing image detail information enables the model to utilize more useful information in link prediction, thereby improving performance. Most notably, when structural features, visual features, and text features (S + V + T) are fused simultaneously, the model achieves the best performance. This shows that by fusing multimodal features, the model can make full use of the complementary information between different modal features, thereby significantly improving the accuracy and effect of link prediction.

Table 3. Evaluation results with different modality feature combinations on FB15K-237-IMG.

 doi:10.20944/preprints202407.0495.v1

FB15K-237-IMG				
	Hits@1↑	Hits@3↑	Hits@10↑	MR↓
T	0.241	0.345	0.457	248
S + T	0.242	0.351	0.386	232
V + T	0.256	0.367	0.504	221
S + V + T	**0.259**	**0.362**	**0.511**	**215**

Overall, the experimental results in Table 3 fully demonstrate the effectiveness of multimodal feature fusion. Structural features provide basic information about entities and relationships, visual features capture details in images, and text features provide semantic understanding. By jointly learning these features, the model can more comprehensively capture the interaction between different modal features, reflecting both commonalities and complementarity, thereby significantly improving the link prediction performance of the knowledge graph.

5.3. Visual Analysis

To further verify the effectiveness of multimodal feature fusion, this paper selects a case to visualize the prediction scores of each modal feature as well as the prediction scores after multimodal feature fusion. As shown in Figure 3, by comparing the prediction results of different modal features, we can more intuitively understand the advantages of multimodal fusion. When using structural features for prediction, the system obtains the highest score for "LeBron James plays for the Golden State Warriors." This is because structural features mainly rely on the existing entity and relationship information in the knowledge graph. If the structural information in the knowledge graph is influenced by outdated data or other factors, it may cause deviations in the prediction results. When using text features for prediction, the system obtains the highest score for "LeBron James plays for the Cleveland Cavaliers." Text features mainly rely on descriptive text information and may be affected by the source and context of the text. If the text information is outdated or contains misleading descriptions, the prediction results are also prone to errors. When using visual features for prediction, the system obtains the highest score for "LeBron James plays for the Los Angeles Lakers." Visual features rely on image information. In this case, by identifying relevant images of LeBron James wearing a Los Angeles Lakers jersey, the system can make more accurate predictions. However, if the image data is not comprehensive or is noisy, the prediction results may be affected.

After fusing the features of different modalities, the system's prediction score changed, accurately predicting "LeBron James plays for the Los Angeles Lakers." The effects of multimodal feature fusion are 1) Reducing bias. Single-modal features are often limited by the information bias inherent to that modality. Fusion of multimodal features integrates the information from each modality, reducing the bias impact of single-modal features; 2) Enhanced robustness. By combining structural, textual, and visual features, the complementarity of different modal features makes the model more robust when facing various types of data. For instance, when there is a conflict between structural and textual features, visual features can provide additional judgment basis, thereby improving prediction accuracy; 3) Improved accuracy. By comprehensively considering the information from multimodal features, the model can better understand the connections between entities and relationships, leading to more accurate predictions. This has been verified in this case. The prediction result "LeBron James plays for the Los Angeles Lakers" after multimodal fusion is consistent with the actual situation, demonstrating the effectiveness of the model in this paper for multimodal feature fusion.

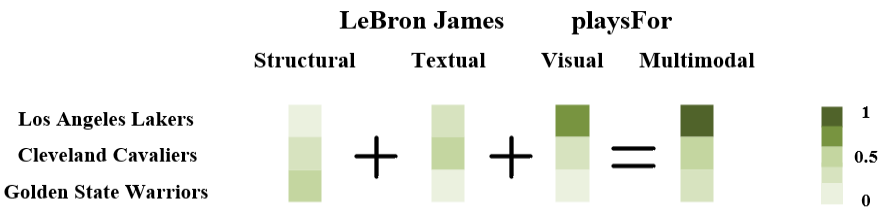

Figure 3. Example visualization of prediction scores in multimodal feature fusion.

 doi:10.20944/preprints202407.0495.v1

Through the analysis of this case, the limitations of single-modal features and the advantages of multimodal feature fusion are clearly demonstrated. Figure 3 intuitively displays the prediction scores of each modal feature and the prediction scores after fusion, further verifying that multimodal fusion can significantly enhance the performance of knowledge graph link prediction.

6. Conclusions and Future Work

This paper studies the problem of link prediction on knowledge graphs. Inspired by multimodal representation learning, we propose an effective model called MM-Transformer for knowledge graph link prediction. The model comprehensively utilizes structural, visual, and textual features. Specific modality encoders are employed to extract these features, and Transformer is utilized for multimodal feature fusion to reduce the heterogeneity of multimodal entity representations. Furthermore, multi-head attention calculations are conducted on mixed keys and values at each layer within the attention mechanism to proactively reduce modality heterogeneity. In the feed-forward network (FFN) layer, token-level fine-grained fusion is implemented to mitigate the adverse impact of irrelevant structures and images on textual information. This research is highly significant for advancing the field of knowledge graph completion and offers valuable insights for addressing other knowledge-intensive tasks.

In the future, we plan to 1) extend the model in this article to pre-training tasks for multimodal knowledge graph construction. By pre-training on large-scale multimodal data, the generalization and robustness of the model can be further improved, providing richer and more accurate representations for knowledge graph completion; 2) further optimize the model architecture and training strategy of MM-Transformer to improve its efficiency and performance on large-scale datasets. Specifically, explore more effective fusion methods and attention calculation methods in terms of multimodal feature fusion and attention mechanisms; 3) cooperate with researchers in other fields to apply the idea of multimodal representation learning to more practical scenarios, such as medical diagnosis, autonomous driving, and intelligent recommendation systems. Through cross-domain cooperation, the scope of application of the model can be broadened and the development of related fields can be promoted.

Author Contributions: Conceptualization, D.W.; methodology, D.W., K.T., J.Z., and Y.P.; software, Y.D., H.L., and B.H.; validation, K.T., J.Z., and Y.P.; formal analysis, H.L., and B.H.; investigation, D.W., Y.D., H.L., and B.H.; resources, D.W., H.L., and B.H.; data curation, D.W., and K.T.; writing—original draft preparation, D.W.; writing—review and editing, K.T., J.Z., and Y.P.; supervision, Y.D., H.L., and B.H.; project administration, D.W., H.L., and B.H.; funding acquisition, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (No.61702234) and Open Fund for Innovative Research on Ship Overall Performance (No.25422217).

Data Availability Statement: Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huang, X.; Zhang, J.; Li, D.; Li, P. Knowledge graph embedding based question answering. In Proceedings of the twelfth ACM international conference on web search and data mining, 30 Jan 2019; pp. 105-113.
2. Yih, SW.; Chang, MW.; He, X.; Gao, J. Semantic parsing via staged query graph generation: Question answering with knowledge base. In Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP, 28 Jul 2015; pp. 1321-1331.
3. Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; Zhu, X. Commonsense knowledge aware conversation generation with graph attention. In IJCAI, 13 Jul 2018; pp. 4623-4629.
4. Huang, J.; Zhao, WX.; Dou, H.; Wen, JR.; Chang, EY. Improving sequential recommendation with knowledge-enhanced memory networks. In The 41st international ACM SIGIR conference on research & development in information retrieval, 27 Jun 2018; pp. 505-514.
5. Zhang, N.; Jia, Q.; Deng, S.; Chen, X.; Ye, H.; Chen, H.; Tou, H.; Huang, G.; Wang, Z.; Hua, N.; Chen, H. Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 14 Aug 2021; pp. 3895-3905.

 doi:10.20944/preprints202407.0495.v1

6. Dietz, L.; Kotov, A.; Meij, E. Utilizing knowledge graphs for text-centric information retrieval. In The 41st international ACM SIGIR conference on research & development in information retrieval, 27 Jun 2018; pp. 1387-1390.
7. Yang, Z. Biomedical information retrieval incorporating knowledge graph for explainable precision medicine. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 25 Jul 2020; pp. 2486-2486.
8. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013, 26, 2787-2795.
9. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI conference on artificial intelligence, 21 Jun 2014; pp. 1112-1119.
10. Nathani, D.; Chauhan, J.; Sharma, C.; Kaul, M. Learning attention-based embeddings for relation prediction in knowledge graphs. arxiv preprint arxiv:1906.01195, 4 Jun 2019.
11. Nguyen, DQ.; Nguyen, TD.; Nguyen, DQ.; Phung, D. A novel embedding model for knowledge base completion based on convolutional neural network. arxiv preprint arxiv:1712.0212, 6 Dec 2017.
12. Pezeshkpour, P.; Chen, L.; Singh, S. Embedding multimodal relational data for knowledge base completion. arxiv preprint arxiv:1809.01341, 5 Sep 2018.
13. Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; Roth, S. A multimodal translation-based approach for knowledge graph representation learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, Jun 2018; pp. 225-234.
14. **e, R.; Liu, Z.; Luan, H.; Sun, M. Image-embodied knowledge representation learning. arxiv preprint arxiv:1609.07028, 22 Sep 2016.
15. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. Ieee Access, 15 May 2019; pp. 63373-63394.
16. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 6 Aug 2019.
17. Li, LH.; Yatskar, M.; Yin, D.; Hsieh, CJ.; Chang, KW. Visualbert: A simple and performant baseline for vision and language. arxiv preprint arxiv:1908.03557, 6 Aug 2019.
18. Chen, YC.; Li, L.; Yu, L.; El, Kholy.A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In European conference on computer vision, 30 Aug 2020; pp. 104-120.
19. Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; Roth, S. A multimodal translation-based approach for knowledge graph representation learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, Jun 2018; pp. 225-234.
20. Wang, Z.; Li, L.; Li, Q.; Zeng, D. Multimodal data enhanced representation learning for knowledge graphs. In 2019 International Joint Conference on Neural Networks (IJCNN), 14 Jul 2019; pp. 1-8.
21. Zhao, Y.; Cai, X.; Wu, Y.; Zhang, H.; Zhang, Y.; Zhao, G.; Jiang, N. MoSE: modality split and ensemble for multimodal knowledge graph completion. CoRR abs/2210.08821, 17 Oct 2022.
22. Wang, M.; Wang, S.; Yang, H.; Zhang, Z.; Chen, X.; Qi, G. Is visual context really helpful for knowledge graph? A representation learning perspective. In Proceedings of the 29th ACM International Conference on Multimedia, 17 Oct 2021; pp. 2735-3743.
23. Shankar, S.; Thompson, L.; Fiterau, M. Progressive fusion for multimodal integration. arxiv preprint arxiv:2209.00302, 1 Sep 2022.
24. Liang, P.P.; Ling, C.K.; Cheng, Y.; Obolenskiy, A.; Liu, Y.; Pandey, R.; Salakhutdinov, R. Quantifying Interactions in Semi-supervised Multimodal Learning: Guarantees and Applications. In The Twelfth International Conference on Learning Representations. 2023.
25. Jiang, Y.; Gao, Y.; Zhu, Z.; Yan, C.; Gao, Y. HyperRep: Hypergraph-Based Self-Supervised Multimodal Representation Learnin, 22 Sept 2023.
26. Golovanevsky, M.; Schiller, E.; Nair, AA.; Singh, R.; Eickhoff, C. One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data. In ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery, 18 Jun 2024.
27. Zhang, X.; Yoon, J.; Bansal, M.; Yao, H. Multimodal representation learning by alternating unimodal adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024; pp. 27456-27466.
28. Li, X.; Zhao, X.; Xu, J.; Zhang, Y.; **ng, C. IMF: interactive multimodal fusion model for link prediction. In Proceedings of the ACM Web Conference 2023, 30 Apr 2023; pp. 2572-2580.
29. Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; Chen, H. 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, 6 Jul 2022; pp. 904-915.
30. Gu, W.; Gao, F.; Lou, X.; Zhang, J. Link prediction via graph attention network. arxiv preprint arxiv:1910.04807, 10 Oct 2019.

 doi:10.20944/preprints202407.0495.v1

31. Alexey, D. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, 3-7 May 2021.
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 24 May 2019; pp. 4171-4186.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, AN.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 30 Jun 2017; pp. 5998-6008.
34. Miller, GA. WordNet: a lexical database for English. Communications of the ACM, 1 Nov 1995; pp. 39-41.
35. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 9 Jun 2008; pp. 1247-1250.
36. **e, R.; Liu, Z.; Luan, H.; Sun, M. Image-embodied knowledge representation learning. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 19-25 August 2017; pp. 3140-3146.