LAFA: Multimodal Knowledge Graph Completion with Link Aware Fusion and Aggregation

Bin Shang[1,2], **Yinliang Zhao**[1,2*], **Jun Liu**[1,2], **Di Wang**[3*]
[1]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering,
School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, China
[3]School of Computer Science and Technology, Xidian University, China
binshang0329@163.com, {zhaoy, liukeen}@xjtu.edu.cn, wangdi@xidian.edu.cn

Abstract

Recently, an enormous amount of research has emerged on multimodal knowledge graph completion (MKGC), which seeks to extract knowledge from multimodal data and predict the most plausible missing facts to complete a given multimodal knowledge graph (MKG). However, existing MKGC approaches largely ignore that visual information may introduce noise and lead to uncertainty when adding them to the traditional KG embeddings due to the contribution of each associated image to entity is different in diverse link scenarios. Moreover, treating each triple independently when learning entity embeddings leads to local structural and the whole graph information missing. To address these challenges, we propose a novel link aware fusion and aggregation based multimodal knowledge graph completion model named LAFA, which is composed of link aware fusion module and link aware aggregation module. The link aware fusion module alleviates noise of irrelevant visual information by calculating the importance between an entity and its associated images in different link scenarios, and fuses the visual and structural embeddings according to the importance through our proposed modality embedding fusion mechanism. The link aware aggregation module assigns neighbor structural information to a given central entity by calculating the importance between the entity and its neighbors, and aggregating the fused embeddings through linear combination according to the importance. Extensive experiments on standard datasets validate that LAFA can obtain state-of-the-art performance.

Introduction

Knowledge graphs (KGs) represent real-world data as fact triples (head entity, relation, tail entity), which have shown great research value and application prospect. KGs are broadly used in many downstream tasks, such as multimedia reasoning (Li, Wang, and Zhu 2020), question answering (Huang et al. 2019), objective detection (Yang et al. 2023), and recommendation system (Guo et al. 2020; Wu et al. 2022). Since existing KGs typically contain structural and visual data, multimodal knowledge graphs (MKGs) have recently attracted great attention in the fields of natural language processing and multimedia (Chen et al. 2022). Generally, multiple images associate an entity to describe the behaviors and appearances of it. Even though the scale of many public MKGs is noticeably large, they are still confronted with incompleteness because the insufficient accumulation of multimodal corpus and the emerging entities with complicated relations. In this case, many researches on multimodal knowledge graph completion (MKGC) have been generalized to find out missing triples automatically (also called link prediction) by extracting knowledge from multimodal data (Wang et al. 2021; Chen et al. 2022; Xu et al. 2022; Shang et al. 2023b). Specifically, images (visual data) can be considered as supplementary information to enhance entity embddings for the MKGC task.

Multimodal knowledge graph completion (MKGC) approaches complete MKGs by projecting entities and relations to latent space as well as learning the dense and low-dimensional vectors (embeddings) of them according to visual and structural information, and predict missing triples by scoring function based on the learned embeddings. Specifically, for MKGC task, one entity generally has multiple associated images, and they can improve the representation quality of the entity embedding. Therefore, it is necessary to fuse the structural properties of KG entities and various images with matched semantics in integrated embeddings. On this account, IKRL (Xie et al. 2017) firstly attempt to fuse visual information to the existing knowledge graph embedding (KGE) models to predict missing triples in MKGs. Mousselly et al. (Mousselly-Sergieh et al. 2018) propose to use Imagined, DeViSE, and simple concatenation to fuse multimodal information. TransAE (Wang et al. 2019) presents an specific auto-encoder module.

Although existing studies for MKGC have shown promising improvements, these approaches are still afflicted by several noticeable limitations as follows: (1) **Modality contradiction**. Many existing MKGC approaches substantially ignore that visual information may lead to uncertainty and introduce noise when adding them to the traditional KG embeddings, which could bring on modality contradiction. Particularly, an entity usually has different attributes in various triples (link information), and the contribution of each associated image to this entity is disparate in diverse link scenarios. For instance in Figure 1, entity *Taylor Swift* has many associated images, but the contribution of them is different in the two links. Although recent works such as RSME (Wang et al. 2021) and MKGformer (Chen et al. 2022) take into ac-

*Corresponding authors.

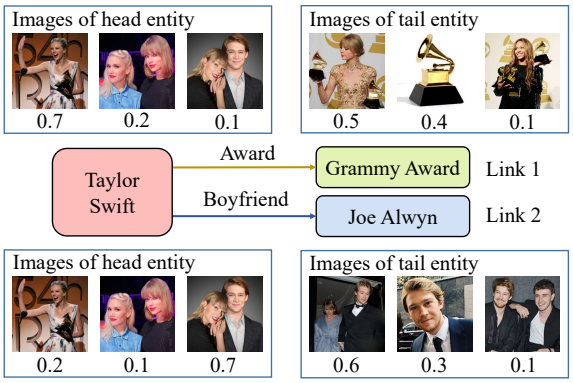


Figure 1: Illustration of the contribution of each associated image to entity *Taylor Swift* in different links. The numbers below the image represent its importance to the entity.

count the noise of images, they target independent entities rather than link information when fusing visual information. (2) **Structural information missing**. Most existing MKGC models attend to treat each triple independently when learning entity and relation embeddings, which result in structural information missing. Since an entity in a KG is often linked with multiple neighbor entities, which can provide rich structural information for the embedding of this entity. Therefore, treating each triple independently when learning entity embeddings leads to missing information about its neighborhood and whole structure of KGs.

Motivated by the above analysis, in this paper, we propose a novel link aware fusion and aggregation based multimodal knowledge graph completion model named LAFA. LAFA is composed of two modules link aware fusion and link aware aggregation to generate entity embeddings, and a decoder for link prediction. In order to alleviate the problem of **modality contradiction**, the link aware fusion module calculates the importance between an entity and its associated images based on link information, and then fuses the visual and structural embeddings by our proposed modality embedding fusion mechanism, which performs linear combination on visual embeddings according to their importance and fuses them with structural embeddings. In order to alleviate the problem of **structural information missing**, the link aware aggregation module calculates the importance between a given central entity and its neighbors, then aggregates the neighbor embeddings with visual information according to the importance through linear combination to assign structural information to the central entity. Our main contributions are summarized as follows:

- We propose a link aware fusion module to alleviate noise of irrelevant visual information by calculating the importance between an entity and its associated images in different link scenarios, and fusing the visual and structural embeddings according to the importance by our proposed modality embedding fusion mechanism. To the best of our knowledge, this work is the first to consider the role of images to entity based on link information.
- We propose a link aware aggregation module to assign neighbor structural information to a given central entity by calculating the importance between the entity and its neighbors, and aggregating the embeddings with visual information by linear combination according to the importance. To the best of our knowledge, this work is the first to aggregate neighbor structural information of entities in MKGC task.
- We conduct comprehensive experiments and extensive analysis on real-world benchmark multimodal datasets. Results and analysis illustrate that our proposed LAFA can effectively model the multimodal representations and substantially outperform the current state-of-the-art (SOTA) models under appropriate circumstances.

Related Work

Our work addresses multimodal knowledge graph completion task, which is relevant to multimodal data and multimodal NLP community. In this section, we briefly introduce the existing unimodal knowledge graph completion (UKGC) methods and multimodal knowledge graph completion (MKGC) approaches.

Unimodal Knowledge Graph Completion

TransE (Bordes et al. 2013) is the first UKGC model, which assumes the triples to satisfy the assumption that $\boldsymbol{h} + \boldsymbol{r} = \boldsymbol{t}$, where $\boldsymbol{h}$, $\boldsymbol{t}$ and $\boldsymbol{r}$ are the embeddings of the head entity, tail entity and relation, respectively. Based on TransE, there are a range of improved models such as TransH (Wang et al. 2014), TransR (Lin et al. 2015), and TransD (Ji et al. 2015). RotatE (Sun et al. 2019) encodes entities and relations into the complex space, allowing them to have more flexible representations. RESCAL (Nickel, Tresp, and Kriegel 2011) encodes entities into vectors and relations into matrices, and then designs a bilinear function to score the triples. Based on RESCAL (Nickel, Tresp, and Kriegel 2011), there are a range of improved models such as NTN (Socher et al. 2013), DistMult (Yang et al. 2015), ComplEx (Trouillon et al. 2016), and TuckER (Balažević, Allen, and Hospedales 2019). ConvE (Dettmers et al. 2018) first uses convolutional neural networks (CNNs) to explore the interaction between entity embeddings and relation embeddings. ConvKB (Dai Quoc Nguyen, Nguyen, and Phung 2018) simplifies ConvE. CompGCN (Vashishth et al. 2020) introduces a graph convolutionl network (GCN) based model. LTE-ConvE (Zhang et al. 2022) introduces a simple linear transformation of entity representation to enhance UKGC models. CompoundE (Ge et al. 2023) extends the distance-based scoring functions to relation-dependent compound operations. Recently, some neural network based models have been proposed such as MRGAT (Dai et al. 2022), HADC (Shang et al. 2023a), ConKGC (Shang et al. 2023c), and GreenKGC (Wang et al. 2023).

Multimodal Knowledge Graph Completion

Existing multimodal knowledge graph completion models focus on encoding image features in KG embeddings. IKRL (Xie et al. 2017) extend TransE (Bordes et al. 2013) to obtain visual embeddings that correspond to the KG entities

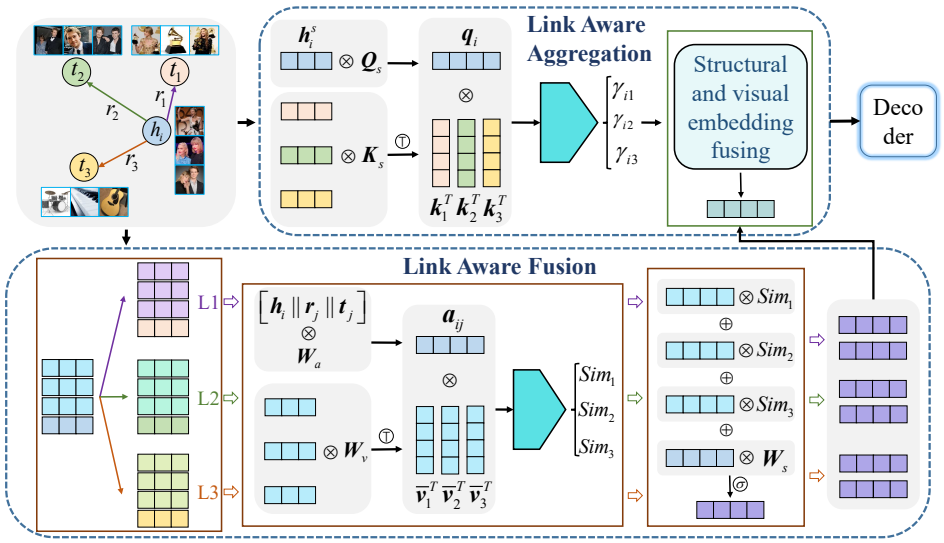


Figure 2: The overall framework of our proposed model LAFA. The lower part represents the link aware fusion module, which first calculates the attention score between an entity and its associated images according to each link by our proposed modality interaction attention mechanism, and then fuses the embeddings of the entity and images based on different links by our proposed modality embedding fusion mechanism. The upper part represents the link aware aggregation module, which calculates the attention score of each neighbor entity to the central entity, and aggregates the embeddings obtained from the link aware fusion module based on structural information.

and structural information of the KG separately. Mousselly et al. (Mousselly-Sergieh et al. 2018) propose to integrate multi-modal information. TransAE (Wang et al. 2019) learns the visual and structural features jointly into unified knowledge embeddings by an auto-encoder. RSME (Wang et al. 2021) automatically encourages or filters the influence of additional visual context during the representation learning. MKGformer (Chen et al. 2022) presents M-Encoder with multi-level fusion at the last several layers of ViT and BERT to conduct image-text incorporated entity modeling. Xu et al. (Xu et al. 2022) propose a multimodal relation-enhanced negative sampling framework to figure out hard negative samples for knowledge graph completion. HRGAT (Liang et al. 2023) incorporates different modal information with graph structure. MoSE (Zhao et al. 2022) designs a modality split representation learning and ensemble inference framework. OTKGE (Cao et al. 2022) models the multi-modal fusion procedure as a transport plan moving different modal embeddings to a unified space.

Although existing MKGC models have shown promising performance, they target independent entities while exploring the contribution of images on entity embeddings without considering the impact of link information on them. Furthermore, they ignore the effect of the structural information of the KG on the entity embeddings, which leads to missing information from their neighbors and the structure of KG.

Methodology

In this section, we will show the formal description and implementation details of our model. First, we introduce the problem formulation of MKGC task. Then we describe the details of each module in LAFA. Finally, we show the decoder and loss function. The overall framework of LAFA is shown in Figure 2. Specifically, LAFA follows Encoder–Decoder framework. The encoder generates entity embeddings containing multimodal and neighborhood structural information, which contains two components. 1) The link aware fusion module can find the noise of irrelevant visual information by calculating importance scores between an entity and its associated images in different link scenarios, then fuses the visual and structural embeddings based on the attention scores and link information. 2)The link aware aggregation module can find the noise of irrelevant neighbors by calculating importance scores between the central entity and its neighbors, based on which the neighbor information with fused multimodal embeddings are aggregated. The learned embeddings from the encoder are fed to the decoder to predict missing triples. The decoder can be implemented by many existing UKGC models, such as DistMult (Yang et al. 2015), ComplEx (Trouillon et al. 2016), and ConvE (Dettmers et al. 2018).

Problem Formulation

A knowledge graph ($\mathcal{G}$) is a directed graph, which can be formulated as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, where $\mathcal{E}$ and $\mathcal{R}$ represents the set of entities (nodes) and relations (edges), respectively. $\mathcal{T} = \{(h, r, t) \mid (h, t \in \mathcal{E}), r \in \mathcal{R}\}$ is triple set in $\mathcal{G}$, and $r \in \mathcal{R}$ is the relation between head entity h and tail entity t. Multimodal knowledge graphs contain visual information based on the above structural information, that is, each entity is associated with multiple corresponding images. Multimodal knowledge graph completion (MKGC) approaches aim to learn the multimodal fused embeddings of entities

and relations by compressing them into a continuous low-dimensional vector space, and conduct link prediction based on the embeddings to predict tail entity for triple $(h, r, ?)$ or head entity for triple $(?, r, t)$.

Link Aware Fusion

An entity h_i usually has multiple images in multimodal knowledge graphs, which can be represented as $\mathcal{V}_{h_i} = \{ v_{i,1}, v_{i,2}, ..., v_{i,N_i^v} \}$, where N_i^v denotes the number of images of entity h_i. To extract image features, the pre-trained ViT (Dosovitskiy et al. 2021) on ImageNet-1k (Touvron et al. 2021) is adopted as the visual encoder.

The images for each entity are from different aspects in various scenarios, which may mislead the entity embedding since parts of them may be irrelevant to this entity due to different link information will give entities different attributes. Therefore, we argue that the contribution of images to entity embedding needs to be based on link information. To be specific, an entity usually has different attributes in different triples, and the contribution of each associated image to this entity is also different in diverse triples. To this end, we propose a link aware fusion module, which designs a modality interaction attention mechanism to dynamically measure the contribution of images to entity embedding based on link information so as to judge which images are noise, and a modality embedding fusion mechanism to fuse visual and structural embeddings.

Modality Interaction Attention Since an entity in a KG has distinct heterogeneous connections with its neighbors, the idea of modality interaction attention comes from an intuition that the contribution of images to entity embedding is different in diverse link scenarios, and whether an image is noise needs to be judged based on triple information. Given a central entity h_i and its neighbor entity set $\mathcal{N}_i^s = \{t_j \in \mathcal{E} \mid (h_i, r_j, t_j) \in \mathcal{T}, r_j \in \mathcal{R}\}$, we firstly randomly initialize the structural embeddings of entities and relations. Then we define a visual matrix to project visual embeddings into hidden embeddings for dimensional unification and similarity matching, as follows:

$$\overline{\boldsymbol{v}}_{i,k} = \boldsymbol{v}_{i,k} \boldsymbol{W}_v, \tag{1}$$

where $\boldsymbol{v}_{i,k} \in \mathbb{R}^{d_v}$ represents the initial visual embedding of image $v_{i,k}$ associated entity h_i from ViT, d_v is the dimension of initial visual embedding, $\boldsymbol{W}_v \in \mathbb{R}^{d_v \times d_h}$ is a trainable visual matrix, and d_h is the dimension of hidden embedding. Triples connected to entity h_i contain different semantic and link information, based on which we calculate the importance between the images with entity h_i for a given triple (h_i, r_j, t_j) to find noisy images, as follows:

$$\boldsymbol{a}_{i,j} = \left[\boldsymbol{h}_i^s \parallel \boldsymbol{r}_j^s \parallel \boldsymbol{t}_j^s\right] \boldsymbol{W}_a, \tag{2}$$

$$b_{i,k} = \frac{\boldsymbol{a}_{i,j} \overline{\boldsymbol{v}}_{i,k}^\top}{\sqrt{d_h}}, \tag{3}$$

$$\alpha_{i,k} = \text{softmax}_k(b_{i,k}) = \frac{\exp(b_{i,k})}{\sum_{m \in \mathcal{V}_{h_i}} \exp(b_{i,m})}, \tag{4}$$

where $\boldsymbol{h}_i^s \in \mathbb{R}^{d_s}$, $\boldsymbol{r}_j^s \in \mathbb{R}^{d_s}$ and $\boldsymbol{t}_j^s \in \mathbb{R}^{d_s}$ represent initial structural embeddings of h_i^s, r_j^s and t_j^s respectively, d_s is the dimension of initial structural embedding, $\boldsymbol{W}_a \in \mathbb{R}^{3d_s \times d_h}$ is a trainable linear transformation matrix, $\alpha_{i,k} \in [0, 1]$ represents the importance of image $v_{i,k}$ to entity h_i, $\parallel$ denotes the concatenation of embeddings. The value of $\alpha_{i,k}$ indicates whether the image $v_{i,k}$ is noise for the entity h_i. In particular, we consider image $v_{i,k}$ to be noise when $\alpha_{i,k} \leq \xi$, where ξ is a predefined threshold.

In addition, in order to find the noisy images of tail entity t_j, the importance of images in set $\mathcal{V}_{t_j} = \{ v_{j,1}, v_{j,2}, ..., v_{j,N_j^v} \}$ associated with the tail entity t_j also need to be measured based on the link information, as follows:

$$c_{j,k} = \frac{\boldsymbol{a}_{i,j} \overline{\boldsymbol{v}}_{j,k}^\top}{\sqrt{d_h}}; \quad \overline{\boldsymbol{v}}_{j,k} = \boldsymbol{v}_{j,k} \boldsymbol{W}_v, \tag{5}$$

$$\beta_{j,k} = \text{softmax}_k(c_{j,k}) = \frac{\exp(c_{j,k})}{\sum_{m \in \mathcal{V}_{t_j}} \exp(c_{j,m})}, \tag{6}$$

where $\boldsymbol{v}_{j,k} \in \mathbb{R}^{d_v}$ represents visual embedding of image $v_{j,k} \in \mathcal{V}_{t_j}$ associated with the tail entity t_j, $\beta_{j,k}$ represents the importance of image $v_{j,k}$ to entity t_j. And the image $v_{j,k}$ is considered to be noise when $\beta_{j,k} \leq \xi$.

Modality Embedding Fusion The modality interaction attention mechanism finds noisy images based on link information. Unlike existing MKGC approaches, we do not directly remove these noisy images, but perform linear combination on visual embeddings of them based on the importance calculated above, then the visual and structural embeddings are fused. The motivation for this is that image information will always be helpful for learning entity embeddings. For entities h_i and t_j, the visual information of them for triple (h_i, r_j, t_j) can be aggregated as follows:

$$\widetilde{\boldsymbol{v}}_{i,j} = \sum_{k \in \mathcal{V}_{h_i}} \alpha_{i,k} \overline{\boldsymbol{v}}_{i,k}, \quad \widetilde{\boldsymbol{v}}_j = \sum_{k \in \mathcal{V}_{t_j}} \beta_{j,k} \overline{\boldsymbol{v}}_{j,k}. \tag{7}$$

To facilitate the fusion of visual and structural embedding, we define a structural matrix to project structural embeddings of entities h_i and t_j into hidden embeddings as follows:

$$\overline{\boldsymbol{h}}_i = \boldsymbol{h}_i^s \boldsymbol{W}_s, \quad \overline{\boldsymbol{t}}_j = \boldsymbol{t}_j^s \boldsymbol{W}_s, \tag{8}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{d_s \times d_h}$ is a trainable structural matrix, and d_s is the dimension of structural embedding. Then the new embedding of entities h_i and t_j containing visual and structural information for triple (h_i, r_j, t_j) can be fused as follows:

$$\widetilde{\boldsymbol{h}}_{i,j} = \sigma\left(\overline{\boldsymbol{h}}_i + \widetilde{\boldsymbol{v}}_{i,j}\right), \quad \widetilde{\boldsymbol{t}}_j = \sigma\left(\overline{\boldsymbol{t}}_j + \widetilde{\boldsymbol{v}}_j\right), \tag{9}$$

where $\sigma(\cdot)$ is sigmoid activation function. In this way, we obtain the updated embeddings of entities by fusing visual and structural embeddings according to diverse link scenarios. Moreover, in order to improve and stabilize the effectiveness of LAFA and the learning procedure, we apply multi-head attention mechanism for capturing subspace information from different parameters. Specifically, P independent attention heads are applied to learn embeddings, and their outputs are combined to generate the unified represen-

tation. The formula is defined as follows:

$$\widehat{\boldsymbol{h}}_{i,j} = \left[\widetilde{\boldsymbol{h}}_{i,j}^{(1)} \parallel \widetilde{\boldsymbol{h}}_{i,j}^{(2)} \parallel ... \parallel \widetilde{\boldsymbol{h}}_{i,j}^{(P)}\right] \boldsymbol{W}_P, \tag{10}$$

$$\widehat{\boldsymbol{t}}_{j} = \left[\widetilde{\boldsymbol{t}}_{j}^{(1)} \parallel \widetilde{\boldsymbol{t}}_{j}^{(2)} \parallel ... \parallel \widetilde{\boldsymbol{t}}_{j}^{(P)}\right] \boldsymbol{W}_P, \tag{11}$$

where $\widetilde{\boldsymbol{h}}_{i,j}^{(p)}$ and $\widetilde{\boldsymbol{t}}_{j}^{(p)}$ mean the embeddings generated by the p-th attention head respectively, $\parallel$ represents the concatenation of embeddings, and $\boldsymbol{W}_P \in \mathbb{R}^{Pd_h \times d_h}$ is a trainable linear transformation matrix. $\widehat{\boldsymbol{h}}_{i,j}$ and $\widehat{\boldsymbol{t}}_j$ are called multimodal fusion embeddings.

Link Aware Aggregation

Knowledge graph is a special graph based dataset, in which a central entity is often connected with multiple tail entities. Existing MKGC models usually focus on the interaction between multimodal data but ignore the structural information of KGs. For an entity h_i, we argue that aggregating the information of its neighbor entities to its embedding is help for improving representation quality. Therefore, we propose a link aware aggregation module to aggregate the neighborhood structural information to the central entity. For a central entity h_i, we define a query matrix to project the initial structural embedding of h_i into a query vector, and a key matrix to project the initial structural embedding of a neighbor entity into a key vector, as follows:

$$\boldsymbol{q}_i = \boldsymbol{h}_i^s \boldsymbol{Q}_s, \quad \boldsymbol{k}_j = \boldsymbol{t}_j^s \boldsymbol{K}_s, \tag{12}$$

where $\boldsymbol{Q}_s \in \mathbb{R}^{d_s \times d_h}$ and $\boldsymbol{K}_s \in \mathbb{R}^{d_s \times d_h}$ are trainable query and key matrices, $t_j \in \mathcal{N}_i^s$ is a neighbor entity, and $\mathcal{N}_i^s$ is the neighbor entity set of h_i. Then the softmax normalization is conducted on the dot product of query vector $\boldsymbol{q}_i$ and key vector $\boldsymbol{k}_j$ to calculate the attention score between them as follows:

$$g_{i,j} = \frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d_h}}, \tag{13}$$

$$\gamma_{i,j} = \text{softmax}_j(g_{i,j}) = \frac{\exp(g_{i,j})}{\sum_{l \in \mathcal{N}_i^s} \exp(g_{i,l})}, \tag{14}$$

where $\gamma_{i,j}$ represents the importance of the neighbor entity t_j to the central entity h_i. And the entity t_j is considered to be noise when $\gamma_{i,j} \leq \xi$. Since the link aware fusion module will calculate multiple multimodal fusion embedding $\widehat{\boldsymbol{h}}_{i,j}$ based on each triple connected to the central entity h_i, we aggregate them with structural information as follows:

$$\boldsymbol{b}_j = \left[\widehat{\boldsymbol{h}}_{i,j} \parallel \widehat{\boldsymbol{t}}_j\right] \boldsymbol{W}_b, \quad \boldsymbol{e}_i = \sum_{j \in \mathcal{N}_i^s} \gamma_{i,j} \boldsymbol{b}_j, \tag{15}$$

where $\widehat{\boldsymbol{h}}_{i,j}$ is the multimodal fusion embedding of central entity h_i calculated according to j-th triple (h_i, r_j, t_j), $\widehat{\boldsymbol{t}}_j$ is the multimodal fusion embedding of entity t_j, $\boldsymbol{W}_b \in \mathbb{R}^{2d_h \times d_h}$ is a trainable linear transformation matrix.

To enable the model can concentrate on information from various subspaces and extract richer feature information, we apply multi-head attention mechanism. Specifically, we use Q independent attention heads to learn embeddings, and then combine them to generate the final embedding $\boldsymbol{h}_i'$ of the central entity, the formula is as follows:

$$\boldsymbol{h}_i' = \left[\boldsymbol{e}_i^{(1)} \parallel \boldsymbol{e}_i^{(2)} \parallel ... \parallel \boldsymbol{e}_i^{(Q)}\right] \boldsymbol{W}_Q, \tag{16}$$

where $\boldsymbol{e}_i^{(q)}$ denotes the embedding learned by the q-th attention head, $\boldsymbol{W}_Q \in \mathbb{R}^{Qd_h \times d_h}$ is a trainable linear transformation matrix. By stacking link aware aggregation, the neighborhood information of each entity can be explored. Therefore, the structural and visual information of the entire KG is aggregated and the highly multimodal contextually relevant embeddings can be generated.

Decoder

MKGC models usually require a basic embedding model for link prediction, which is called decoder. In this paper, we build the decoder based on ConvE (Dettmers et al. 2018). Specifically, The input of the decoder are entity and relation embeddings $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{E}| \times d_h}$ and $\boldsymbol{R} \in \mathbb{R}^{|\mathcal{R}| \times d_h}$. Entity embeddings $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{E}| \times d_h}$ are generated by aforementioned steps and have fused multimodal and structural information of entity neighbors. $|\mathcal{R}|$ denotes the number of relation set, and $|\mathcal{E}|$ represents the number of entity set, d_h is the dimension of embeddings. Then the decoder outputs the score of each triple calculated by the scoring function, which represents the probability that the triple is valid. For a triple (h, r, t), the scoring function of our model can be defined as follows:

$$\Psi(h, r, t) = \sigma\left(\text{vec}\left(\sigma\left(\left[\widehat{\boldsymbol{h}'} \| \widehat{\boldsymbol{r}}\right] * \omega\right)\right) \boldsymbol{W}'\right) \boldsymbol{t}', \tag{17}$$

where $\boldsymbol{h}'$ and $\boldsymbol{t}'$ are the updated embeddings (Eq. (16)) of head entity h and tail entity t respectively, $\boldsymbol{r}$ is the embedding of relation r, $\sigma(\cdot)$ represents a non-linear function, $\widehat{\boldsymbol{h}'}$ and $\widehat{\boldsymbol{r}}$ denote 2D reshaping of $\boldsymbol{h}'$ and $\boldsymbol{r}$ respectively, ω is the convolution filter, $*$ denotes the convolution operation, and $\boldsymbol{W}'$ is a trainable transformation matrix. Then the score is activated by the sigmoid function:

$$p(h, r, t) = \text{sigmoid}(\Psi(h, r, t)). \tag{18}$$

It should be noticed that LAFA can be easily adapted to various decoders such as DistMult (Yang et al. 2015) and ComplEx (Trouillon et al. 2016).

Training and Optimization

We use the cross-entropy loss function as the loss of the entire model, which is defined as follows:

$$\mathcal{L} = \sum_{(h,r,t) \in \mathcal{T}} -\frac{1}{|\mathcal{E}|} \sum_{s=1}^{|\mathcal{E}|} (y(h, r, t_s) \cdot \log(p(h, r, t_s)) + (1 - y(h, r, t_s)) \cdot \log(1 - p(h, r, t_s))), \tag{19}$$

where $y(h, r, t_s) \in \{0, 1\}$ is the label of the triple (h, r, t_s), $|\mathcal{E}|$ is the total number of all candidate tail entities, $\mathcal{T}$ is the set of true triples. We use Adam (Kingma and Ba 2014) as optimizer, and use label smoothing (Szegedy et al. 2016), Dropout (Srivastava et al. 2014), and Batch normalization (Ioffe and Szegedy 2015) to avoid overfitting.

Model	FB15k-237-IMG				WN18-IMG			
	Hits@1↑	Hits@3↑	Hits@10↑	MR↓	Hits@1↑	Hits@3↑	Hits@10↑	MR↓
Unimodal approaches								
TransE	0.198	0.376	0.441	323	0.040	0.745	0.923	357
DistMult	0.199	0.301	0.446	512	0.335	0.876	0.940	655
ComplEx	0.194	0.297	0.450	546	0.936	0.945	0.947	-
ConvE	0.237	0.356	0.501	256	0.937	0.947	0.951	294
LTE-ConvE	0.245	0.377	0.535	169	0.943	0.953	0.961	189
MRGAT	0.266	0.386	0.542	159	0.932	0.946	0.971	38
GreenKGC	0.265	0.369	0.507	241	0.937	0.946	0.950	266
CompoundE	0.264	0.393	0.545	151	0.942	0.952	0.972	36
Multimodal approaches								
IKRL(UNION)	0.194	0.284	0.458	298	0.127	0.796	0.928	596
TransAE	0.199	0.317	0.463	431	0.323	0.835	0.934	352
RSME(ViT-B/32+Forget)	0.242	0.344	0.467	417	0.943	0.951	0.957	223
MKGformer	0.256	0.367	0.504	221	0.944	0.961	0.972	28
LAFA-DistMult	0.264	0.392	0.546	150	0.943	0.958	0.971	29
LAFA-ComplEx	0.262	0.386	0.540	157	0.945	0.962	0.973	27
LAFA-ConvE	**0.269**	**0.398**	**0.551**	**136**	**0.947**	**0.965**	**0.977**	**25**

Table 1: Link prediction results on FB15k-237-IMG and WN18-IMG datasets. The best score is in bold.

Datasets	Entities	Relations	Train triples	Validation triples	Test triples
FB15k-237-IMG	14,541	237	272,115	17,535	20,466
WN18-IMG	40,943	18	141,442	5,000	5,000

Table 2: Statistics of the datasets.

Experiments

Experimental Setup

Datasets We evaluate our proposed model by two publicly available multimodal datasets: 1) FB15k-237-IMG (Bordes et al. 2013; Chen et al. 2022) and WN18-IMG (Bordes et al. 2013). The details of them are summarized in Table 2. FB15k-237-IMG is a subset of the large-scale knowledge graph Freebase (Bollacker et al. 2008), which has 10 images for each entity. WN18 (Bordes et al. 2013) is a knowledge graph originally extracted from WordNet (Miller 1995). WN18-IMG is an extended dataset of WN18 (Bordes et al. 2013) with 10 images for each entity.

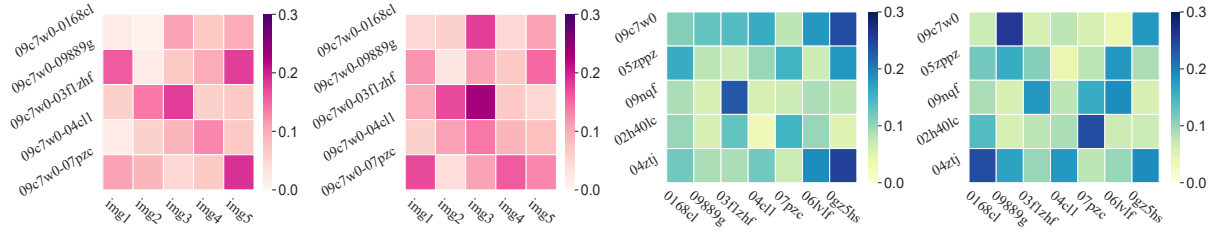
Evaluation Protocol Following previous work (Dettmers et al. 2018), our model is evaluated with link prediction task: ranking all entities to predict the tail entity in query $(h, r, ?)$ or the head entity in query $(?, r, t)$. We adopt four evaluation metrics: the mean rank of correct entities (MR), and the proportion of correct entities ranked in top k Hits@k ($k \in \{1, 3, 10\}$). A small MR or a big Hit@k indicates a good result. And we follow the standard evaluation protocol in the filtered setting (Bordes et al. 2013): all true triples in the KG are filtered out during evaluation, since predicting a low rank for these triples should not be penalized.

Baselines We compare results with the following SOTA models: Unimodal KGC approaches TransE (Bordes et al. 2013), DistMult (Yang et al. 2015), ComplEx (Trouillon et al. 2016), ConvE (Dettmers et al. 2018), LTE-ConvE (Zhang et al. 2022), MRGAT (Dai et al. 2022), GreenKGC (Wang et al. 2023), and CompoundE (Ge et al. 2023). Multimodal KGC approaches IKRL (Xie et al. 2017), TransAE (Wang et al. 2019), RSME (Wang et al. 2021), and MKGformer (Chen et al. 2022).

Implementation Details We define the threshold $\xi = 0.1$. For all MKG datasets, the best performing hyper-parameters are found by grid search on the validation set. And the candidate hyper-parameters are selected in the following ranges: batch size {128, 512, 1024}, number of epochs {500, 1000, 2000}, dropout rate {0.1, 0.2, 0.3}, learning rate {0.001, 0.002, 0.003}, embedding dimensions {100, 200, 300, 400, 500}, attention head number P and Q {1, 2, 3, 4}. The experiments are implemented using the PyTorch (Paszke et al. 2017) framework, and are performed on single NVIDIA GeForce RTX2080Ti GPU.

Main Results

Table 1 presents the link prediction results on FB15k-237-IMG and WN18-IMG datasets. We strictly follow the experimental setting and data splitting of the previous works (Wang et al. 2021; Chen et al. 2022) and report the results in the original papers for some baselines. The results show that LAFA have the best performance compared with existing SOTA unimodal and multimodal approaches, which demonstrate that fusing the visual and structural information to entity embeddings according to link information is generally helpful for MKC tasks. Specifically, the Hits@k ($k \in \{1, 3, 10\}$) are improved by 1%-2% on FB15k-237-IMG dataset. Particularly, Hits@3 and Hits@10 are improved from 0.393 to 0.398 and 0.545 to 0.551 respectively, MR is declined from 151 to 136. Compared with SOTA MKGC method MKGformer, LAFA improves Hits@3 and Hits@10 from 0.367 to 0.398 and 0.504 to 0.551 respec-



(a) Entities-imgs (Head 1) (b) Entities-imgs (Head 3) (c) Entities-entities (Head 1) (d) Entities-entities (Head 3)

Figure 3: Attention matrices of images and entities on FB15k-237-IMG dataset. The darker colored blocks in individual heads represent a higher attention score.

tively. The reason is that LAFA can find noise images and fuse multimodal embeddings by the proposed link aware fusion, and aggregate neighbor structural information by the proposed link aware aggregation.

The lower part of Table 1 shows the performance of LAFA with three different decoders, i.e., DistMult (Yang et al. 2015), ComplEx (Trouillon et al. 2016), and ConvE (Dettmers et al. 2018). Obviously, the decoders do have impact on the performance of the model, but not absolutely because the difference in their link prediction results is not very large. ConvE works best as a decoder for LAFA. DistMult and ComplEx are simple and effective models, but they result in a decline in the model performance when they are used as decoders. One possible reason is that their scoring functions weaken the importance of relations.

Verification of Importance

To explore the potential patterns of our innovations and verify that LAFA can give different importance to images and neighbor entities, we analyze the attention matrices generated according to α (Eq. (4)) and γ (Eq. (14)). Figure 3 shows the attention matrices of images and neighbor entities to the central entity respectively. It can be found that the color distribution is not uniform on all attention matrices, which is in line with our expectation that LAFA can assign different importance score to the specific images or neighbor entities. Furthermore, the consistent distribution of regions with higher importance scores illustrates that attention heads follow the same pattern in capturing subspace semantics. Specifically, we can find from Figure 3 (a) and Figure 3 (b) that the importance of each image associated with entity *09c7w0* is different in different links, which verifies our hypothesis. From Figure 3 (c) and Figure 3 (d), we can observe that the importance of each neighbor entity to the central entity *09c7w0* is also different, the reason is that the semantics of it vary greatly in different links. To this end, the results demonstrate that the proposed LAFA can effectively assign different importance scores to images and neighboring entities of the central entity.

Ablation Study

We conduct the ablation studies by removing the corresponding parts to construct variants of LAFA as follows: (1) LAFA^{-MIA} replaces the modality interaction attention (MIA) module with the traditional vector similarity matching when calculating the importance of the image to the entity. (2) LAFA^{-LAF} removes link aware fusion (LAF) module, in which the visual and structural embeddings are fused only by attention mechanism without link information; (3) LAFA^{-LAA} removes link aware aggregation (LAA) module. The ablation studies results in Table 3 indicate that our proposed MIA, LAF, and LAA are all valid, that is, removing any of them will make the model less effective. MIA assigns different importance score to images and can judge which of them is noise, LAF exploits the influence of images to entity, and LAA assigns neighbor structural information to the central entity. The experimental results prove that the proposed innovations are effective and contribute significantly to the performance of the model.

Model	FB15k-237-IMG			
	Hits@1↑	Hits@3↑	Hits@10↑	MR↓
ConvE	0.237	0.356	0.501	256
LAFA^{-MIA}	0.258	0.384	0.538	164
LAFA^{-LAF}	0.259	0.386	0.540	161
LAFA^{-LAA}	0.257	0.384	0.537	168
LAFA	**0.269**	**0.398**	**0.551**	**136**

Table 3: Ablation study results on FB15k-237-IMG dataset.

Conclusion

In this paper, we present a novel link aware fusion and aggregation multimodal knowledge graph completion model named LAFA. The link aware fusion module calculates the importance between an entity and its associated images in different link scenarios and fuses the visual and structural embeddings according to the importance through our proposed modality embedding fusion mechanism to alleviate noise of irrelevant visual information. The link aware aggregation module calculates the importance between a given central entity and its neighbors, and aggregates the embeddings of them through linear combination according to the importance to assigns neighbor structural information to this entity. Empirical experimental evaluations on well-established multimodal datasets show that LAFA can achieve the state-of-the-art performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62137002, 62192781, and 62072354), the Fundamental Research Funds for the Central Universities (QTZX23084).

References

Balažević, I.; Allen, C.; and Hospedales, T. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5185–5194.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Cao, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, 35: 39090–39102.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. 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*, 904–915.

Dai, G.; Wang, X.; Zou, X.; Liu, C.; and Cen, S. 2022. MR-GAT: Multi-Relational Graph Attention Network for knowledge graph completion. *Neural Networks*, 154: 234–245.

Dai Quoc Nguyen, T. D. N.; Nguyen, D. Q.; and Phung, D. 2018. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In *Proceedings of NAACL-HLT*, 327–333.

Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*.

Ge, X.; Wang, Y. C.; Wang, B.; and Kuo, C.-C. J. 2023. Compounding Geometric Operations for Knowledge Graph Completion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6947–6965.

Guo, Q.; Zhuang, F.; Qin, C.; Zhu, H.; Xie, X.; Xiong, H.; and He, Q. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8): 3549–3568.

Huang, X.; Zhang, J.; Li, D.; and Li, P. 2019. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 105–113.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456.

Ji, G.; He, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 687–696.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, G.; Wang, X.; and Zhu, W. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1227–1235.

Liang, S.; Zhu, A.; Zhang, J.; and Shao, J. 2023. Hyper-node relational graph attention network for multi-modal knowledge graph completion. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–21.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; and Roth, S. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 225–234.

Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 809–816.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *Advances in neural information processing systems*, 1–12.

Shang, B.; Zhao, Y.; Liu, J.; Liu, Y.; and Wang, C. 2023a. A contrastive knowledge graph embedding model with hierarchical attention and dynamic completion. *Neural Computing and Applications*, 35(20): 15005–15018.

Shang, B.; Zhao, Y.; Liu, Y.; and Wang, C. 2023b. Attention-based exploitation and exploration strategy for multi-hop knowledge graph reasoning. *Information Sciences*, 653: 119787.

Shang, B.; Zhao, Y.; Wang, D.; and Liu, J. 2023c. Relation-Aware Multi-Positive Contrastive Knowledge Graph Completion with Embedding Dimension Scaling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 878–888.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems*, 26.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, 2071–2080. PMLR.

Vashishth, S.; Sanyal, S.; Nitin, V.; and Talukdar, P. 2020. Composition-based Multi-Relational Graph Convolutional Networks. In *Proceedings of the 7th International Conference on Learning Representations*, 1–16.

Wang, M.; Wang, S.; Yang, H.; Zhang, Z.; Chen, X.; and Qi, G. 2021. Is visual context really helpful for knowledge graph? A representation learning perspective. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2735–2743.

Wang, Y.-C.; Ge, X.; Wang, B.; and Kuo, C.-C. J. 2023. Greenkgc: A lightweight knowledge graph completion method.

Wang, Z.; Li, L.; Li, Q.; and Zeng, D. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Wu, Y.; Liao, L.; Zhang, G.; Lei, W.; Zhao, G.; Qian, X.; and Chua, T.-S. 2022. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*.

Xie, R.; Liu, Z.; Luan, H.; and Sun, M. 2017. Image-embodied knowledge representation learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3140–3146.

Xu, D.; Xu, T.; Wu, S.; Zhou, J.; and Chen, E. 2022. Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3857–3866.

Yang, A.; Lin, S.; Yeh, C.-H.; Shu, M.; Yang, Y.; and Chang, X. 2023. Context Matters: Distilling Knowledge Graph for Enhanced Object Detection. *IEEE Transactions on Multimedia*.

Yang, B.; Yih, S. W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the 3rd International Conference on Learning Representations*, 1–12.

Zhang, Z.; Wang, J.; Ye, J.; and Wu, F. 2022. Rethinking graph convolutional networks in knowledge graph completion. In *Proceedings of the ACM Web Conference 2022*, 798–807.

Zhao, Y.; Cai, X.; Wu, Y.; Zhang, H.; Zhang, Y.; Zhao, G.; and Jiang, N. 2022. MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10527–10536.