Unleashing the Power of Imbalanced Modality Information for Multi-modal Knowledge Graph Completion

Yichi Zhang♣♡, Zhuo Chen♣♡, Lei Liang◇, Huajun Chen♣♡♠, Wen Zhang♣♡∗
♣Zhejiang University, ◇Ant Group
♠Alibaba-Zhejiang University Joint Institute of Frontier Technology
♡Zhejiang University-Ant Group Joint Laboratory of Knowledge Graph
{zhangyichi2022, zhuo.chen, zhang.wen}@zju.edu.cn

Abstract

Multi-modal knowledge graph completion (MMKGC) aims to predict the missing triples in the multi-modal knowledge graphs by incorporating structural, visual, and textual information of entities into the discriminant models. The information from different modalities will work together to measure the triple plausibility. Existing MMKGC methods overlook the imbalance problem of modality information among entities, resulting in inadequate modal fusion and inefficient utilization of the raw modality information. To address the mentioned problems, we propose **Ada**ptive **M**ulti-modal **F**usion and **M**odality **A**dversarial **T**raining (AdaMF-MAT) to unleash the power of imbalanced modality information for MMKGC. AdaMF-MAT achieves multi-modal fusion with adaptive modality weights and further generates adversarial samples by modality-adversarial training to enhance the imbalanced modality information. Our approach is a co-design of the MMKGC model and training strategy which can outperform 19 recent MMKGC methods and achieve new state-of-the-art results on three public MMKGC benchmarks. Our code and data have been released at https://github.com/zjukg/AdaMF-MAT.

Keywords: Multi-modal Knowledge Graph, Knowledge Graph Completion, Adversarial Training

1. Introduction

Knowledge graphs (KGs) (Wang et al., 2017) model the world knowledge as structured triples in the form of (*head entity, relation, tail entity*). Multi-modal knowledge graphs (MMKGs) Liu et al. (2019) further extend KGs with representative multi-modal information (e.g. textual descriptions, images, etc.) and have become the new infrastructure in many artificial intelligence tasks (Sun et al., 2020; Yasunaga et al., 2021; Cao et al., 2022a; Zhang et al., 2023b; Liang et al., 2023b).

However, KGs often suffer from incompleteness as there are many unobserved triples in the KGs. The incompleteness of KGs limits their usage in the downstream tasks. Therefore, knowledge graph completion (KGC) (Wang et al., 2017) has attracted wide attention as a crucial task to complete the missing triples in the KGs. Mainstream KGC methods employ knowledge graph embedding (KGE) to model the structural information in the KGs, which embeds the entities and relations into a low-dimensional continuous space.

As for the MMKGs, the KGC models (Wang et al., 2021; Zhao et al., 2022) incorporate rich-semantic multi-modal information such as images and text descriptions, referred to as multi-modal knowledge graph completion (MMKGC). Mainstream MMKGC methods (Wang et al., 2021) usually focus on designing elegant methods to achieve multi-modal fusion among structural, visual, and textual infor-
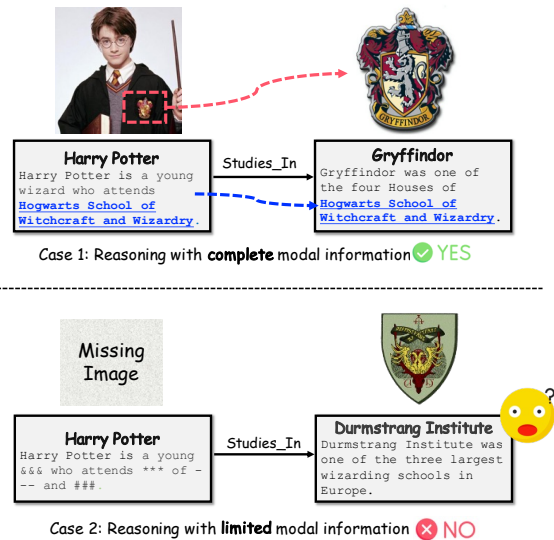

Figure 1: A simple example to show that knowledge graph reasoning with limited modal information may lead to wrong prediction.

mation, which is a common way of thinking for multi-modal tasks. Some other methods (Xu et al., 2022) attempt to enhance the negative sampling (NS) process with the multi-modal information.

However, existing methods neglect the **imbalance of modality information** among entities, which can be observed from two perspectives. Firstly, as for knowledge graph reasoning, different modal information plays distinctive roles and should be adaptively considered. However, the

---
∗ Corresponding author

modal fusion problem is inadequately addressed by existing methods, as the modal information is often inflexibly incorporated into the representation space of structural information uniformly. Secondly, the efficacious features in the images and textual descriptions are usually limited and challenging to extract. In practical scenarios, the KGs constructed from multiple heterogeneous data sources are even modality-missing, further constraining the utilization of modality information in MMKGC. In practice, the KGs constructed with multiple heterogeneous data sources are even modality-missing (Wang et al., 2021), which further limits the usage modality information in MMKGC. Figure 1 reveals that limited or missing multi-modal information significantly hampers modal performance. To achieve better performance for MMKGC, it is crucial to effectively utilize the essential information as well as get higher quality multi-modal information. This need can be summarised as unleashing the power of imbalanced modality information for MMKGC.

To address this problem, we propose a novel MMKGC framework with **Ada**ptive **M**ulti-modal **F**usion and **M**odality **A**dversarial **T**raining (AdaMF-MAT) to augment and effectively utilize the multi-modal information. We introduce AdaMF, an adaptive multi-modal fusion (AdaMF) module that selectively extracts essential multi-modal features from entities to generate representative joint embeddings. We further propose a modality-adversarial training (MAT) strategy to generate synthetic multi-modal embeddings and construct adversarial examples, aiming to enhance the limited multi-modal information during training. While existing NS methods design complex strategies to sample in the given KG, our method directly creates synthetic samples with semantic-rich multi-modal information to enhance the multi-modal embedding learning. The two modules are designed to synergistically interact and mutually complement each other to unleash the power of imbalanced modality information. Our approach is a co-design of the MMKGC model and training strategy, which can bring outperforming empirical results against 19 existing unimodal KGC, multi-moda KGC, and NS methods. Our contribution is three-fold:

- We propose adaptive multi-modal fusion (AdaMF) for MMKGC, to fuse the imbalanced modality information of three modalities (structural, visual, and textual) and produce representative joint embeddings.
- We propose a modality adversarial training strategy (MAT) to utilize imbalanced modality information. MAT aims to generate adversarial examples with synthetic multi-modal embeddings too enhance the MMKGC training.
- We conduct comprehensive experiments with further explorations on three public benchmarks to evaluate the performance of AdaMF-MAT. The empirical results demonstrate that AdaMF-MAT can outperform 19 recent baselines and achieve new SOTA MMKGC results. Moreover, we illustrate the suitability of AdaMF-MAT for the modality-missing scenarios.

2. Related Works

2.1. Multi-modal Knowledge Graph Completion

Existing multi-modal knowledge graph completion (MMKGC) methods (Liang et al., 2022; Chen et al., 2024) are mainly concerned with extending the general knowledge graph embedding (KGE) methods (Bordes et al., 2013; Yang et al., 2015) to the multi-modal scenario. In MMKGC, visual and textual information are delicately considered as well as the structural information of the entities to make better predictions.

A mainstream design approach of MMKGC methods is to enhance the entity representations with their multi-modal information. IKRL (Xie et al., 2017) first introduces the visual information of entities into the TransE (Bordes et al., 2013) score function. TBKGC (Sergieh et al., 2018) later extends IKRL with both visual and textual information. TransAE (Wang et al., 2019) employs the auto-encoder framework to fuse the multi-modal information with the entity embeddings. RSME (Wang et al., 2021) designs several gates to select and keep the useful visual information of entities. VBKGC (Zhang and Zhang, 2022) uses Visual-BERT to extract deep fused multi-modal information to achieve better modality fusion. OTKGE (Cao et al., 2022b) achieve multi-modal fusion via optimal transport. MoSE (Zhao et al., 2022) considered different modal information in a split-and-ensemble manner.

Besides, some methods propose enhanced negative sampling (Bordes et al., 2013) methods to improve the MMKGC performance. MMRNS (Xu et al., 2022) utilizes the modal information to enhance the negative sampling to train better KGE models. MANS (Zhang et al., 2023a) proposes a modality-aware negative sampling method to align the structural and multi-modal information.

2.2. Adversarial Training in Knowledge Graph Completion

Adversarial training (AT) (Kurakin et al., 2017; Goodfellow et al., 2014, 2015; Shafahi et al., 2019) is a powerful technology widely used in many machine learning fields (Mariani et al., 2018; Miyato et al., 2017). AT aims to train a pair of generator

and discriminator with a minimax game adversarially, which can enhance both the generator and the discriminator.

In the knowledge graph community, a few methods (Cai and Wang, 2018; Wang et al., 2018; Tang et al., 2022) employ AT to generate harder negative samples (Bordes et al., 2013) by reinforcement learning to enhance the embedding training process and improve the prediction results. But these methods are designed for unimodal KGC.

3. Methodology

In this section, we will present our MMKGC framework AdaMF-MAT. An overview of AdaMF-MAT is shown in Figure 2. We will introduce the main components in the following paragraphs, which are feature encoders, adaptive multi-modal fusion, and modality adversarial training respectively.

3.1. Preliminaries

A knowledge graph (KG) can be denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}, \mathcal{R}$ are the entity set, relation set. $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the triple set. For multi-modal KGs (MMKGs), each entity $e \in \mathcal{E}$ has an image set $\mathcal{V}_e$ and a corresponding textual description.

An MMKGC model embeds the entities and relations into a continuous vector space. For each entity $e_i \in \mathcal{E}$, we define its structural/visual/textual embedding as e_s, e_v, e_t respectively to represent its different modal features. For each relation $r \in \mathcal{R}$, we denote its structural embedding as r. Besides, an MMKGC model can measure the plausibility of each triple $(h, r, t) \in \mathcal{T}$ with a score function $\mathcal{F}$, which can be calculated by the defined embeddings. In the inference stage of the KGC task, for a given query $(h, r, ?)$ or $(?, r, t)$, the model ranks the corresponding triple scores of each candidate entity and make the prediction.

3.2. Multi-modal Feature Encoding

We first introduce the encoding process to extract the multi-modal features from the original images and textual descriptions of entities. It is a necessary step for all MMKGC methods.

As for the visual modal, we first apply a pre-trained visual encoder (PVE) (Simonyan and Zisserman, 2015; Bao et al., 2022) to extract the visual feature f_v for each entity e with mean-pooling and project it to the embedding space to get the visual embedding, which can be denoted as:

$$f_v = \frac{1}{|\mathcal{V}_e|} \sum_{img_i \in \mathcal{V}_e} \text{PVE}(img_i) \qquad (1)$$

$$e_v = \mathbf{W}_v \cdot f_v + \mathbf{b}_v \qquad (2)$$

where the $\mathbf{W}_v, \mathbf{b}_v$ is the parameters of the visual project layer.

For the textual embedding, we similarly extract the textual feature of each entity with its description and the pre-trained textual encoder (PTE) (Reimers and Gurevych, 2019). The special [CLS] token (Devlin et al., 2019) is applied to capture the sentence-level textual feature f_t, which is a common technology. Then we project the textual feature f_t to the embedding space with another project layer parameterized by $\mathbf{W}_t, \mathbf{b}_t$ to get the textual embedding e_t, which can be denoted as:

$$e_t = \mathbf{W}_t \cdot f_t + \mathbf{b}_t \qquad (3)$$

We maintain consistency with previous methods (Sergieh et al., 2018; Xu et al., 2022) and employ the same PVE and PTE of these methods for a fair comparison. Besides, the project layers are designed to project the multi-modal features from different representation spaces into the same space of the structural embeddings.

3.3. Adaptive Multi-modal Fusion

In MMKGC, the information of three modalities $m \in \mathcal{M} = \{s, v, t\}$ should be carefully considered to measure the triple plausibility. As existing methods (Xie et al., 2017; Sergieh et al., 2018) usually treat them separately or simply put them together, we propose an adaptive multi-modal fusion (AdaMF for short) mechanism to achieve adaptive modality fusion. For each entity e and its embeddings e_m where $m \in \mathcal{M}$, AdaMF works in the following way:

$$\alpha_m = \frac{\exp(w_m \oplus \tanh(e_m))}{\sum_{n \in \mathcal{M}} \exp(w_n \oplus \tanh(e_n))} \qquad (4)$$

$$e_{joint} = \sum_{m \in \mathcal{M}} \alpha_m e_m \qquad (5)$$

where α_m is the adaptive weight, $\oplus$ is the point-wise product operator and w_m is the learnable vector of the modality m. e_{joint} is the joint embedding of entity e. Through such a design of AdaMF, we can learn different modality weights for different entities, thus achieve adaptive modal information fusion to produce representative joint embeddings.

Besides, we apply the RotatE model (Sun et al., 2019) as our score function to measure the triple plausibility. It can be denoted as:

$$\mathcal{F}(h, r, t) = ||h_{joint} \circ r - t_{joint}|| \qquad (6)$$

where $\circ$ is the rotation operation in the complex space. The joint embeddings of the head and tail entity are employed in the score function.

Aiming to assign the positve triples with higher scores, we train the embeddings with a sigmoid
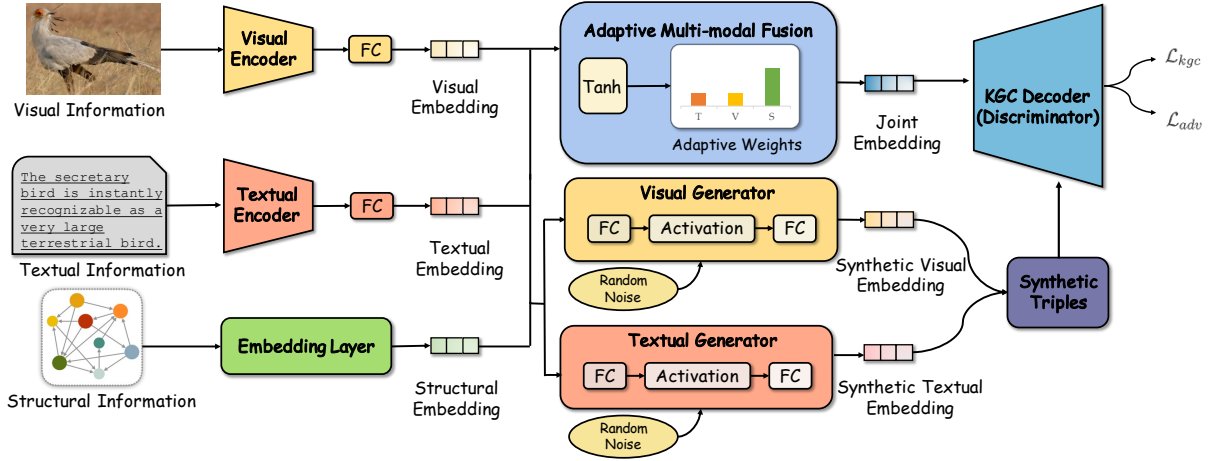
Figure 2: Overview of our method AdaMF-MAT. The feature encoders are designed to encode different modal features (visual/textual/structural) respectively. Each FC represents a fully-connected projection layer. The adaptive multi-modal fusion module is designed to get the fused joint embedding adaptively. The modality adversarial training module employs generators to generate synthetic multi-modal embeddings to construct adversarial examples. The KGC decoder serves as the discriminator and will be enhanced by these adversarial examples during training.

loss function (Sun et al., 2019). The loss function can be denoted as:

$$
\mathcal{L}_{kgc} = \frac{1}{|\mathcal{T}|} \sum_{(h,r,t) \in \mathcal{T}} \Big(- \log \sigma(\gamma - \mathcal{F}(h,r,t)) \\
- \sum_{i=1}^{K} p_i \log \sigma(\mathcal{F}(h'_i, r'_i, t'_i) - \gamma) \Big)
\tag{7}
$$

where σ is the sigmoid function, γ is the margin and p_i is the self-adversarial weight for each negative triple (h'_i, r'_i, t'_i) generated by negative sampling (Bordes et al., 2013). The self-adversarial weight p_i can be denoted as:

$$
p_i = \frac{\exp(\beta \mathcal{F}(h'_i, r'_i, t'_i))}{\sum_{j=1}^{K} \exp(\beta \mathcal{F}(h'_j, r'_j, t'_j))}
\tag{8}
$$

where β is the temperature parameter and K is the number of negative samples. A main training objective of our model is to minimize $\mathcal{L}_{kgc}$.

3.4. Modality Adversarial Training

Previously, we noted that the multi-modal embeddings of entities contain limited essential information and face the imbalance problem. Though AdaMF can adaptively select and fuse the multi-modal embeddings into a joint one, it can not augment the existing multi-modal embeddings to provide more semantic-rich information.

Therefore, inspired by adversarial training (AT) (Mariani et al., 2018), we design a modality adversarial training (MAT) mechanism for MMKGC to enhance the imbalanced multi-modal information.

MAT employs a generator $\mathbf{G}$ that produces adversarial samples, and a discriminator $\mathbf{D}$ that measures their plausibility, which follows the paradigm of general AT. Below we describe the design of the generator $\mathbf{G}$ and the discriminator $\mathbf{D}$.

3.4.1. Generator

Further, we design a modality adversarial generator (MAG) that assumes the role of $\mathbf{G}$ by generating synthetic multi-modal embeddings conditioned on the structural embedding e_s, which aims to construct adversarial samples. The MAG is a two-layer feed-forward network which can be denoted as:

$$
\mathbf{G}_m(e_s, z) = \mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot [e_s; z] + \mathbf{b}_1) + \mathbf{b}_2
\tag{9}
$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ are the parameters of the two feed-forward layers and δ is the LeakyReLU (Maas et al., 2013) activation function. $m \in \{v, t\}$ is the visual/textual modal and $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is the random noise.

For each triple (h, r, t), we generate synthetic multi-modal embeddings for the head and tail entities and construct the synthetic entities. The synthetic head entity h^* has three embeddings h_s, h_v^*, h_t^* where h_v^*, h_t^* are generated by the MAG $\mathbf{G}_v$ and $\mathbf{G}_t$. It is the same for the tail entity.

Further, we can construct three synthetic triples with the two synthetic entities, which are $\mathcal{S}(h, r, t) = \{(h, r, t^*), (h^*, r, t), (h^*, r, t^*)\}$ for a triple (h, r, t). Here, $\mathcal{S}$ contains a group of synthetic triples for (h, r, t), which consists of 3 adversarial samples. In practice, we generate L groups of synthetic triples for each (h, r, t), and $\mathcal{S}$ will contain $3L$ synthetic triples. Since the generator contains

17123

Algorithm 1: Pseudo-code for training AdaMF-MAT

Input: A batch of training triple $\mathcal{B}$ sampled from $\mathcal{T}$, the multi-modal information of the entities', the AdaMF model $\mathbf{D}$, the generator $\mathbf{G}$.

Output: The MMKGC model $\mathbf{D}$ trained with MAT.

1 **for** *each triple* $(h, r, t) \in \mathcal{B}$ **do**
2   // Training $\mathbf{D}$
3   Get the joint embeddings $\boldsymbol{h}_{joint}, \boldsymbol{t}_{joint}$.
4   Calculate the triple score $\mathcal{F}(h, r, t)$ and the nagative triple scores $\mathcal{F}(h'_i, r'_i, t'_i)$.
5   Calculate the kgc loss $\mathcal{L}_{kgc}$.
6   Generate the adversarial example set $\mathcal{S}$ with $\mathbf{G}$.
7   Calculate the adversarial loss $\mathcal{L}_{adv}$.
8   Calculate the overall loss $\mathcal{L}_{kgc} + \lambda \mathcal{L}_{adv}$.
9   Back propagation and optimize $\mathbf{D}$.
10   // Training $\mathbf{G}$.
11   Get the joint embeddings $\boldsymbol{h}_{joint}, \boldsymbol{t}_{joint}$.
12   Generate the adversarial example set $\mathcal{S}$ with $\mathbf{G}$.
13   Calculate the adversarial loss $\mathcal{L}_{adv}$.
14   Back propagation and optimize $\mathbf{G}$.
15 **end**

Table 1: Statistical information of the benchmarks.

Dataset	$|\mathcal{E}|$	$|\mathcal{R}|$	#Train	#Valid	#Test
DB15K	12842	279	79222	9902	9904
MKG-W	15000	169	34196	4276	4274
MKG-Y	15000	28	21310	2665	2663

et al., 2022), and KBGAN (Cai and Wang, 2018) is that MAT generates new hard negative samples by adversarial training instead of sampling from the existing KGs. Such a design of MAT can directly augment the multi-modal embedding learning and enhance the performance of $\mathbf{D}$ without the need for reinforcement learning applied in KBGAN (Cai and Wang, 2018). Besides, MMRNS (Xu et al., 2022) and KBGAN (Cai and Wang, 2018) just employ the unimodal score functions but our approach introduces the multi-modal information with adaptive multi-modal fusion, which is a co-design of the MMKGC model and training strategy.

3.4.3. Training Objective

During training, we iteratively train the $\mathbf{D}$ and $\mathbf{G}$ in an adversarial manner while keeping to optimize the general objective $\mathcal{L}_{kgc}$ for MMKGC. The overall training objective can be represented as:

$$\min_{\mathbf{D}} \mathcal{L}_{kgc} + \min_{\mathbf{D}} \max_{\mathbf{G}} \lambda \mathcal{L}_{adv} \tag{11}$$

where λ is the coefficient of the adversarial loss. The pseudo-code of mini-batch training for AdaMF-MAT is shown in Algorithm 1.

During training, the AdaMF model $\mathbf{D}$ and the generator $\mathbf{G}$ will be optimized iteratively in an adversarial setting.

4. Experiments

In this section, we evaluate our method with the link prediction task, which is a mainstream task in KGC. We first introduce the experiment settings and then present our experimental results. We mainly explore the following four research questions (RQ) about AdaMF-MAT:

- **RQ1**: How does the performance of AdaMF-MAT compare to the existing MMKGC methods in the link prediction task?

- **RQ2**: How does the performance of AdaMF-MAT in the modality-missing scenario?

- **RQ3**: How much do the design of AdaMF and MAT contribute to the performance?

- **RQ4**: can we find some intuitive cases to explain the performance of our method?

random noise, these synthetic triples are different from each other.

3.4.2. Discriminator

With the generated adversarial examples with synthetic multi-modal embeddings, $\mathbf{D}$ aims to distinguish the positive (real) triples from the synthetic ones while $\mathbf{G}$ aims to generate realistic triples and deceive the discriminator $\mathbf{D}$. In MMKGC, we let the score function $\mathcal{F}$ serve as the $\mathbf{D}$ in the AT setting because $\mathcal{F}$ is the triple plausibility discriminator and the target to be adversarially enhanced. Accordingly, the loss of the adversarial training can be designed as:

$$\mathcal{L}_{adv} = \frac{1}{|\mathcal{T}|} \sum_{(h,r,t) \in \mathcal{T}} \Big(-\log \sigma(\gamma - \mathcal{F}(h, r, t))$$
$$- \frac{1}{|\mathcal{S}|} \sum_{\substack{(h^*, r^*, t^*) \\ \in \mathcal{S}(h,r,t)}} \log \sigma(\mathcal{F}(h^*, r^*, t^*) - \gamma) \Big) \tag{10}$$

We keep the format of the loss function similar to $\mathcal{L}_{kgc}$ and contrast the positive triples and the adversarial examples, as such a format is more suitable for MMKGC. From another perspective, MAT can be regarded as an enhanced negative sampling method for MMKGC, as we generate high-quality negative examples in an adversarial manner. The main difference between MAT and existing negative sampling methods like MMRNS(Xu

	Model	DB15K				MKG-W				MKG-Y			
		MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Unimodal KGC	TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
	TransD	21.52	8.34	29.93	44.24	25.56	15.88	32.99	40.18	26.39	17.01	33.60	40.31
	DistMult	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86	25.04	19.33	27.80	35.95
	ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73	28.71	22.26	32.12	40.93
	RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
	PairRE	31.13	21.62	35.91	49.30	34.40	28.24	36.71	46.04	32.01	25.53	35.84	43.89
	GC-OTE	31.85	22.11	36.52	51.18	33.92	26.55	35.96	46.05	32.95	26.77	36.44	44.08
Multi-modal KGC	IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
	TBKGC	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07
	TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
	MMKRL	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69	36.81	31.66	39.79	45.31
	RSME	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09
	VBKGC	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88	37.04	33.76	38.75	42.30
	OTKGE	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
Negative Sampling	KBGAN(TransE)	25.73	9.91	36.95	51.93	29.47	22.21	34.87	40.64	29.71	22.81	34.88	40.21
	KBGAN(TransD)	23.74	9.34	33.51	47.94	29.67	22.38	35.24	40.80	28.73	20.99	34.64	40.76
	MANS	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78	29.03	25.25	31.35	34.49
	MMRNS(RotatE)	29.67	17.89	36.66	51.01	34.13	27.37	37.48	46.82	35.93	30.53	39.07	45.47
	MMRNS(SOTA)	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47
Ours	AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21	38.06	33.49	40.44	45.48
	AdaMF-MAT	**35.14**	**25.30**	**41.11**	**52.92**	**35.85**	**29.04**	**39.01**	**48.42**	**38.57**	**34.34**	**40.59**	**45.76**

Table 2: Link prediction performance on DB15K, MKG-W, and MKG-Y. The best results are marked **bold** and the second-best results are underlined in each column. The results of unimodal KGE and MMRNS are from (Xu et al., 2022) while other results are based on our reproduction. MMRNS(SOTA) is the state-of-the-art results of MMRNS among different score functions.

4.1. Experiment Settings

4.1.1. Datasets

The experiments are conducted on three public benchmarks. They are DB15K, MKG-W, and MKG-Y. DB15K Liu et al. (2019) is constructed from DB-Pedia Lehmann et al. (2015) with images crawled from a search engine. MKG-W and MKG-Y are proposed by (Xu et al., 2022), which are the subsets of Wikidata Vrandecic and Krötzsch (2014) and YAGO Suchanek et al. (2007) knowledge bases respectively. The statistical information about the three datasets is shown in Table 1. We reuse the multi-modal features captured with pre-trained models (Simonyan and Zisserman, 2015; Devlin et al., 2019) released by the original papers.

4.1.2. Task and Evaluation Protocols

To evaluate our method, we conduct link prediction (Bordes et al., 2013) task on the three datasets. Link prediction is a significant task of knowledge graph completion, which aims to predict the missing entity of a given query $(h, r, ?)$ or $(?, r, t)$. The two parts of the link prediction task is called head prediction and tail prediction respectively.

Following the existing works, we use rank-based metrics (Sun et al., 2019) like mean reciprocal rank (MRR) and Hit@K(K=1, 3, 10) to evaluate the results. MRR and Hit@K can be calculated as :

$$\mathbf{MRR} = \frac{1}{|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} \left(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}} \right) \quad (12)$$

$$\mathbf{Hit@K} = \frac{1}{|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} \left(\mathbf{1}(r_{h,i} \leq K) + \mathbf{1}(r_{t,i} \leq K) \right) \quad (13)$$

where $r_{h,i}$ and $r_{t,i}$ are the results of head prediction and tail prediction respectively.

Besides, filter setting (Bordes et al., 2013) is applied to all the results to avoid the influence of the training triples for fair comparisons.

4.1.3. Baselines

To demonstrate the effectiveness of our approach, we select several MMKGC methods as our baselines for comparisons and they can be divided into three categories:

(i) **Uni-modal KGC methods**, including TransE (Bordes et al., 2013), TransD (Ji et al., 2015), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2017), RotatE (Sun et al., 2019), PairRE (Chao et al., 2021), and GC-OTE (Tang et al., 2020). These methods only use the structural information of the KGs to learn structural embeddings.

(ii) **Multi-modal KGC models**, including IKRL (Xie et al., 2017), TBKGC (Sergieh et al., 2018), TransAE (Wang et al., 2019), MMKRL (Lu et al., 2022), RSME (Wang et al., 2021), VBKGC (Zhang and Zhang, 2022), OTKGE (Cao et al., 2022b). These methods utilize both the structural information and multi-modal information in the KGs.

(iii) **Negative sampling methods**, including KBGAN (Cai and Wang, 2018), MANS (Zhang et al., 2023a) and MMRNS (Xu et al., 2022). KBGAN is the first work for adversarial negative sample in KGC based on reinforcement learning. MANS

and MMRNS design new negative sampling strategies enhanced by the multi-modal information to generate high-quality negative samples.

Besides, we re-implement some baselines that have no officially released codes and conduct the link prediction experiments according to the detailed experimental settings in the original paper.

4.1.4. Implemention Details

We implement our method based on OpenKE (Han et al., 2018) and conduct experiments on Ubuntu 22.04.1 operating system with NVIDIA GeForce 3090 GPUs. We employ VGG (Simonyan and Zisserman, 2015) for DB15K and BEiT (Bao et al., 2022) for MKG-W/MKG-Y as PVEs. As for the textual modal, we employ sentence-BERT (Reimers and Gurevych, 2019) following the setting in MMRNS (Xu et al., 2022).

For the hyper-parameters, we tune the embedding dimension in $\{200, 250\}$ while the negative sample number K is tuned in $\{32, 64, 128\}$. The dimension of the random noise in the MAT is 64. The group number of adversarial samples L is set to be 1. The training batch size is fixed to 1024. We search the margin $\gamma \in \{1, 2, 4, 8, 12\}$, the temperature $\beta \in \{0.5, 1.0, 2.0\}$ and the learning rate $\eta_d, \eta_g \in \{10^{-3}, 10^{-4}, 10^{-5}\}$ for $\mathbf{D}$ and $\mathbf{G}$. The coefficient λ of the adversarial loss is searched in $\{0.1, 0.01, 0.001\}$.

4.2. Main Results (RQ1)

To address **RQ1**, we conduct the main link prediction experiments and present the results in Table 2. Our method AdaMF-MAT can outperform the existing 19 MMKGC baselines, with an average improvement of 6.0% in MRR, 8.0% in Hit@1, 5.5% in Hit@3, and 2.1% in Hit@10. These results indicate that AdaMF-MAT shows more significant improvements, particularly on strict metrics like Hit@1, which means that AdaMF-MAT can make more exact predictions.

Compared with MMRNS(RotatE) which also employs RotatE (Sun et al., 2019) as the score function, our AdaMF (w/o MAT) can achieve better results on all metrics. Such a result demonstrates that AdaMF can do better than the existing methods to utilize the multi-modal information in the MMKGs. Additionally, we can observe that MAT can bring full range improvement of metrics by generating high-quality synthetic triples for positive-negative contrast during training, compared with other negative sampling methods like MANS and MMRNS[2]. Some old adversarial methods like KB-GAN may even lead to a decrease in performance, which means MAT is more stable.
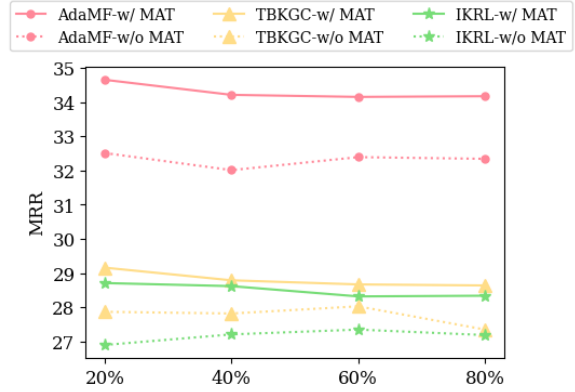

Figure 3: Link prediction results on modality-missing DB15K dataset. The x-axis represents the modality-missing ratios. We report the MRR results of three various MMKGC models (AdaMF, TBKGC, IKRL). The missing modal information is randomly initialized first, as commonly done in existing methods (Sergieh et al., 2018). w/ MAT and w/o MAT indicate the MMKGC model trained with MAT and without MAT respectively.

4.3. Modality-missing Results (RQ2)

To answer **RQ2**, we conduct modality-missing link prediction experiments on DB15K datasets. In the experiments, the multi-modal information is dropped with a given modality-missing ratio. We first complete the missing information with random initialization as commonly done in existing methods (Sergieh et al., 2018) and make comparisons of the link prediction performance among different MMKGC models (AdaMF, TBKGC, IKRL) and training strategies (w/ MAT and w/o MAT). The MRR results are shown in Figure 3. We can draw the following three conclusions from the results:

(i) AdaMF can achieve better link prediction performance in the modality-missing KGs compared with other existing MMKGC baselines like IKRL and TBKGC. This suggests that the adaptive multi-modal fusion performs better than the existing modal fusion methods when the entity modal information is missing.

(ii) MAT can also enhance the MMKGC model to learn better multi-modal embeddings by adversarial training and can be applied to different MMKGC models as a general framework. The results indicate that MAT can improve the performance of all three MMKGC models compared with randomly initializing the missing modal information, which is a vanilla approach in the modality-missing scenario wide used by existing methods (Sergieh et al., 2018; Wang et al., 2021).

(iii) For the modality-missing KGs, the link prediction results of the MMKGC model using random initialization are not stable. Figure 3 shows that

sometimes the experimental results increase even when the missing ratio decreases. We think this is because the various MMKGC models can only make modal fusion but can not augment the limited modal information. When the missing modal information exceeds a certain range, such modal fusion may fail and the results might be unstable. However, MAT resolves such a problem and enables MMKGC models to achieve better performance and make full use of the modal information, as the performance is negatively correlated with the missing ratio. When the modal information is more sufficient, the model's capability will be further enhanced by utilizing more modal information.

In summary, the experimental results prove that our AdaMF-MAT is a better choice for MMKGC in the modality-missing scenario.

4.4. Ablation Study (RQ3)

Model		MRR	Hit@1	Hit@3	Hit@10
AdaMF-MAT		35.14	25.30	41.11	52.92
AdaMF (w/o MAT)	S+V+T(Adaptive)	33.19	23.08	40.34	52.47
	S+V+T(Mean)	32.57	21.45	39.71	51.68
	S+V(w/o T)	32.34	21.84	38.90	50.76
	S+T(w/o V)	31.82	19.63	39.69	52.51
	V+T(w/o S)	31.01	18.45	39.38	52.27
MAT	w/o (h^*, r, t)	34.64	24.52	40.98	52.49
	w/o (h, r, t^*)	34.65	24.49	41.13	52.61
	w/o (h^*, r, t^*)	34.61	24.36	40.98	52.65

Table 3: The ablation study results. S/V/T represent structural/visual/textual embeddings respectively. Adaptive and mean represent the different settings of the weights of each modality in the joint embeddings.

To demonstrate in more detail the effectiveness of each module design and answer **RQ3**, we conduct the ablation study as shown in Table 3. We validate the effectiveness of AdaMF by replacing the adaptive weights with mean weights for different modalities. Besides, we remove the embeddings of each modality to validate the contribution of modal information. As for MAT, we remove each kind of adversarial example to check their effectiveness.

From the ablation study results we can find that the link prediction results decrease when each module or each modal information of our method is removed, which demonstrates their effectiveness.

To better understand MAT, we conduct parameter analysis on the group number L of the adversarial examples. As shown in Figure 4, we can observe that the model performance can be affected to some extent by L, but there is not a very conspicuous change. As L increases, the link prediction performance first increases a few and then decreases.
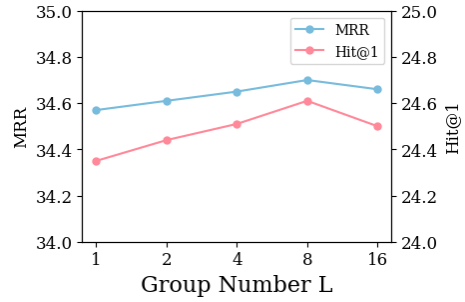

Figure 4: Parameter analysis results about the number of adversarial examples. We report the MRR and Hit@1 results.
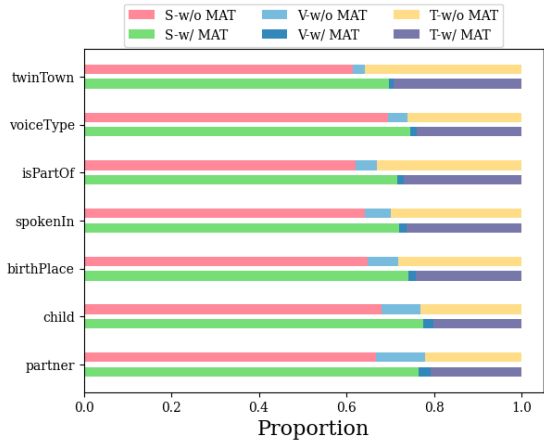

Figure 5: Adaptive weight visualization results of AdaMF for different relations. We consider the AdaMF models trained w/ and w/o MAT. We partitioned the test triples by relation and calculated the average modality weights among the entities.

4.5. Case Study (RQ4)

To demonstrate the effectiveness of AdaMF-MAT in an intuitive view, we visualize the percentage of modality weights in AdaMF under several relations. As shown in Figure 5, we can observe that the modality weights of entities in different relational contexts vary, and the importance of the three modalities is often $s > v > t$. This reflects that AdaMF can extract diverse modal features for different entities and the modality importance is also consistent with the previous research results (Wang et al., 2021; Xu et al., 2022).

Meanwhile, with the enhancement of MAT, the modality weights are changed compared with a vanilla AdaMF model. The overall trend is a further decrease in the weight of visual/textual embeddings especially visual modality, which indicates that MAT can augment the multi-modal embeddings and achieve better adaptive modal fusion. With MAT, the truly essential part of the multi-modal information is augmented and retained to participate in the score function.

5. Conclusion

In this paper, we mainly discuss the problem of utilizing modal information in MMKGC and propose a novel MMKGC framework called AdaMF-MAT to address the limitations of the existing methods. Existing methods for utilizing modal features are relatively crude and treat the modal information in a one-size-fits-all manner. Meanwhile, the modality-missing problem is ignored by existing methods, which we think is also a significant problem in MMKGC. Our method AdaMF-MAT employs adaptive modal fusion to utilize the multi-modal information diversely and augment the multi-modal embeddings through modality-adversarial training. Experiments demonstrate that AdaMF-MAT can outperform all the existing baseline methods and achieve SOTA results in MMKGC tasks.

We think that the utilization problem of the multi-modal information in the MMKGs is still considerable. The combination of large language models and MMKGs is a new and popular approach to MMKGC and the downstream usage of MMKGs.

Limitations and Ethics

Our work mainly proposes an adaptive fusion and adversarial training framework for MMKGC. We think the limitation of this work is that our settings for modality-missing scenarios are not close enough to real production scenarios. Besides, we need to design a more effective approach to utilize the multi-modal information of MMKGs. All experiments were conducted on publicly available datasets, with no violation of scientific ethics or invasion of privacy involved.

Acknowledgements

References

Maksym Andriushchenko and Nicolas Flammarion. 2020. Understanding and improving fast adversarial training. In *Proc. of NeurIPS*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. Beit: BERT pre-training of image transformers. In *Proc. of ICLR*.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proc. of NeurIPS*.

Liwei Cai and William Yang Wang. 2018. KBGAN: adversarial learning for knowledge graph embeddings. In *Proc. of NAACL*.

Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. 2022a. Cross-modal knowledge graph contrastive learning for machine learning method recommendation. In *Proc. of ACM MM*.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022b. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Proc. of NeurIPS*.

Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. Pairre: Knowledge graph embeddings via paired relation vectors. In *Proc. of ACL*.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *CoRR*, abs/2402.05391.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. of NeurIPS*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proc. of ICLR*.

Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proc. of EMNLP*.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *ACL (1)*, pages 687–696. The Association for Computer Linguistics.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial machine learning at scale. In *Proc. of ICLR*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*.

Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. 2023a. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. 2023b. Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1559–1568.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *arXiv preprint arXiv:2212.05767*.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorovz, Slovenia, June 2-6, 2019, Proceedings*.

Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.*

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*.

Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and A. Cristiano I. Malossi. 2018. BAGAN: data augmentation with balancing GAN. *CoRR*.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proc. of ICLR*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. of EMNLP*.

Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proc. of AACL*.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! In *Proc. of NeurIPS*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*.

Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proc. of CIKM*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *Proc. of ICLR*.

Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *Proc. of ACL*.

Zhenwei Tang, Shichao Pei, Zhao Zhang, Yongchun Zhu, Fuzhen Zhuang, Robert Hoehndorf, and Xiangliang Zhang. 2022. Positive-unlabeled learning with adversarial data augmentation for knowledge graph completion. In *Proc. of IJCAI*.

Théo Trouillon, Christopher R. Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. 2017. Knowledge graph completion via complex tensor factorization. *J. Mach. Learn. Res.*

J. D. Tygar. 2011. Adversarial machine learning. *IEEE Internet Comput.*

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is visual context really helpful for knowledge graph? A representation learning perspective. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*.

Peifeng Wang, Shuangyin Li, and Rong Pan. 2018. Incorporating GAN for negative sampling in knowledge representation learning. In *Proc. of AAAI*.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *Proc. of IJCNN.*

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proc. of IJCAI.*

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proc. of ACM MM.*

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proc. of ICLR.*

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In *Proc. of NAACL.*

Yichi Zhang, Mingyang Chen, and Wen Zhang. 2023a. Modality-aware negative sampling for multi-modal knowledge graph embedding. *CoRR.*

Yichi Zhang, Zhuo Chen, Yin Fang, Lei Cheng, Yanxi Lu, Fangming Li, Wen Zhang, and Huajun Chen. 2023b. Knowledgeable preference alignment for llms in domain-specific question answering. *CoRR*, abs/2311.06503.

Yichi Zhang, Zhuo Chen, Wen Zhang, and Huajun Chen. 2023c. Making large language models perform better in knowledge graph completion. *CoRR*, abs/2310.06671.

Yichi Zhang and Wen Zhang. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *CoRR.*

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *Proc. of EMNLP.*

Language Resource References

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web.*

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *ESWC*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*, pages 697–706. ACM.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.