MULTIFORM: Few-Shot Knowledge Graph Completion via Multi-Modal Contexts

Xuan Zhang, Xun Liang ✉, Xiangping Zheng, Bo Wu, and Yuhui Guo

Renmin University of China, Beijing, China
{zhangxuanalex,xliang,xpzheng,wubochn,yhguo}@ruc.edu.cn

Abstract. Knowledge Graphs (KGs) have been applied to many downstream applications such as semantic web, recommender systems, and natural language processing. Previous research on Knowledge Graph Completion (KGC) usually requires a large number of training instances for each relation. However, considering the accelerated growth of online information, there can be some relations that do not have enough training examples. In fact, in most real-world knowledge graph datasets, instance frequency obeys a long-tail distribution. Existing knowledge embedding approaches suffer from the lack of training instances. One approach to alleviating this issue is to incorporate few-shot learning. Despite the progress they bring, they sorely depend on entities' local graph structure and ignore the multi-modal contexts, which could make up for the lack of training information in the few-shot scenario. To this end, we propose a multi-modal few-shot relational learning framework, which utilizes the entities' multi-modal contexts to connect few instances to the knowledge graphs. For the first stage, we encode entities' images, text descriptions, and neighborhoods to acquire well-learned entity representations. In the second stage, our framework learns a matching metric to match the query triples with few-shot reference examples. The experimental results on two newly constructed datasets show the superiority of our framework against various baselines.

Keywords: Few-shot learning · Meta-learning · Knowledge graphs · Attention aggregation function · Multi-modal contexts

1 Introduction

Knowledge Graphs (KGs) encode structured information of entities and their relations in the form of triples (h, r, t), where h represents some head entity and r represents some relation that connects h to some tail entity t. For example, a statement like "*Isaac Newton worked at the University of Cambridge*" can be represented as (*Isaac Newton*, *Work location*, *University of Cambridge*). KGs are the key components of various practical applications such as visual transfer learning [19], recommender systems [33] and so on. Despite their usefulness and popularity, KGs are often highly incomplete. Extensive research, termed as knowledge embedding [2,30,25], has made great progress in automatically completing missing links in KGs.

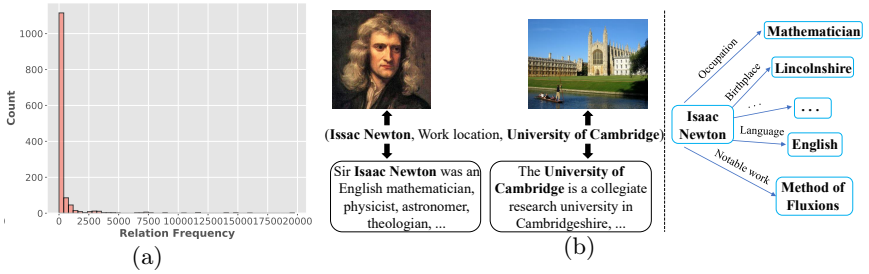

Fig. 1: (a) The distribution of relation frequencies in FB15K. (b) An example of multi-modal contexts of KGs: The left presents the images and textual descriptions of the entities in the triple (*Issac Newton*, *Work location*, *University of Cambridge*); The right presents the one-hop graph structure of the entity *Issac Newton*.

However, research on Knowledge Graph Completion (KGC) for KGs usually assume that sufficient training examples for each relation are available and cannot cope with few-shot relations. In the real world, the KGs evolve quickly with new entities and relations being added by the second and some new relations may not have enough training examples. Even in the classic knowledge graph FB15K, long-tail relations (few-shot relations), which have very few training triples, are actually very common as shown in Figure 1 (a). To be more specific, FB15K contains 1345 relations and about 0.6 million instances, but over 36% of these relations contain no more than 10 instances.

There are also some few-shot learning methods, such as GMatching [38] and FAAN [21], concentrating on alleviating the challenge of the lack of training examples for the long-tail relations. These models aim at predicting new links given only few training triples in a meta-learning scenario. Their main ideas are devising a neighbor encoder to acquire well-represented entities from the neighbors, and then represent few-shot relations with the learned entities. One of the key challenges is to learn the accurate entity representations with very few training information available.

While the few-shot learning models focus on developing various complicated algorithms, they depend on limited training information sorely from the entities' neighborhoods and ignore other crucial multi-modal contexts widely existing in KGs and Freebase [1], such as images and the text descriptions. As Figure 1 (b) shows, these additional multi-modal contexts contain abundant information, which could be helpful during training and make up for the lack of training information in the few-shot scenario.

With the aforementioned statements, we go back to the original KGs, and extract the entities' images, text descriptions and neighborhoods as additional visual, textual and topological information respectively. To predict new links with only few-shot given instances, we propose a MULTI-modal Few-shOt Relational learning fraMework (MULTIFORM). In contrast to previous few-shot learning models sorely depending on entities' neighborhoods, MULTIFORM is able to

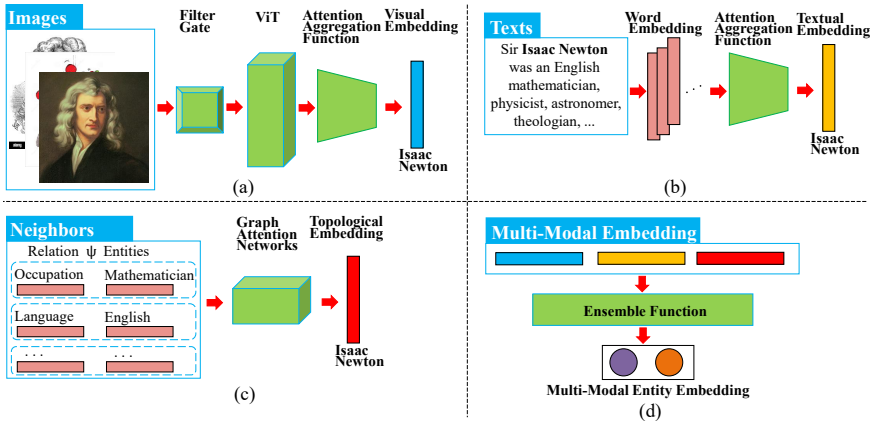

Fig. 2: Multi-modal context encoder for entities: (a) Image encoder; (b) Text encoder; (c) Neighbor encoder; (d) Multi-modal embedding fusion model.

benefit from all multi-modal contexts. MULTIFORM consists of a multi-modal context encoder and a metric learning module. The multi-modal context encoder produces well-learned representations of entities via multi-modal contexts. We separately encode image embedding, text descriptions and one-hop neighbors of entities and leverage an ensemble function to produce new accurate embeddings containing multi-modal information of entities. Our metric learning module aims at learning a matching function that can be used to discover more similar triples given few-shot reference triples. With two newly constructed datasets, i.e., MM-FB15K and MM-DBpedia, we show that our model can achieve consistent improvements over various state-of-the-art baselines on the few-shot KGC task. In summary, the present work makes the following contributions:

- As far as we know, this paper is the first to study few-shot KGC tasks with multi-modal contexts. We design three encoders to extract crucial information from different multi-modal data.
- We explore the impact of different multi-modal contexts, which is empirically important but ignored by the previous studies on multi-modal KGs.
- We construct two new datasets MM-FB15K and MM-DBpedia from FB15K and DBpedia for multi-modal few-shot KGC evaluation. We evaluate our model in the few-shot scenario and the experimental results show the superiority of our model against various state-of-the-art baselines.

2 Related Work

Here we survey three topics relevant to our research: unimodal knowledge embedding models, multi-modal knowledge embedding models, and few-shot learning.

2.1 Unimodal Knowledge Embedding Models

Unimodal knowledge embedding models aim at modeling multi-relational data and automatically inferring missing facts in KGs. Many of them encode both entities and relations into a continuous low dimensional vector space. RESCAL [17] utilizes tensor operations to model relations. TransE [2] is a classic work that encodes both entities and relations into a 1-D vector space. Following this line of research, more effective models such as DistMult [39], ComplEx [30], ConvE [5], Rotate [26], and Rot-Pro [25] have been proposed for further improvements. These embedding-based models heavily rely on extensive collections of training examples, and they are not qualified to deal with sparse triples, as presented in [2] and [38].

2.2 Multi-modal Knowledge Embedding Models

Multi-modal knowledge embedding models mainly focus on encoding visual and structural contexts. IKRL [37] separately trains visual information and structural information on TransE [2]. Mousselly et al. [15] uses three different ensemble function, i.e., simple concatenation, DeViSE [8], and Imagined [4] to fuse multi-modal context embeddings. TransAE [35] utilizes an auto-encoder to integrate them. RSME [34] evaluates different image encoders for multi-modal KGC and verify the effectiveness of Visual Transformer (ViT), so we adopt ViT as image encoder in this paper. There are several models [22,36] taking rich text descriptions into consideration to handle unseen entities.

2.3 Few-Shot Learning

Few-shot learning methods seek to learn novel concepts with only a small number of labeled examples. Recent deep learning based few-shot learning models can be classified into three groups. The first group is *model-based approaches*, which depend on a specially designed part like memory to quickly optimize the model parameters given few-shot training examples. MetaNet [16], a typical model-based approach, learns meta knowledge across tasks and generalizes rapidly via its fast parameterization. The second group is *metric-based approaches*, which try to learn a generalizable metric and the corresponding matching functions among a set of training examples. For example, prototypical networks [24] classify each instance by calculating the similarity to prototype representation of each class, whose idea is similar to some nearest neighbor algorithms. GMatching [38], FSRL [40], and FAAN [21] can also be considered as a metric-based approach. The third group is *optimization-based approaches* [20,7,13], which aim to learn faster by changing the optimization methods on few-shot reference instances. One example is model-agnostic meta-learning (MAML) [7], which first proposed the framework of updating parameters of a task-specific learner and performing meta optimization across tasks by using the above updated parameters. MetaR [3], which transfers relation-specific meta information from support set to query set, can also be regarded as an optimization-based approach for knowledge graph.

Previous few-shot learning research mainly focuses on vision [28], sentiment analysis [12] domains. As for few-shot learning on KGC, Bordes et al. [2] first realized the number of training examples for each relation in KGs have a great impact on the accuracy of the embedding model. However, he did not formulate it as a few-shot learning task. Existing few-shot learning models [38,40,21] on KGC tasks all sorely depend on local graph structures. In contrast to their approaches, we intend to leverage visual, textual and topological context to improve the quality of entity embeddings.

3 Preliminaries

3.1 Task Formulation

Here we give the definition of the few-shot KGC task via multi-modal contexts as follows:

Definition 1. *Given an incomplete KG* $\mathcal{G} = (\mathbf{E}, \mathbf{R}, \mathbf{T})$*, where* $\mathbf{E}$*,* $\mathbf{R}$ *and* $\mathbf{T}$ *are the entity set, relation set, and triple set, respectively, the few-shot KGC task completes* $\mathcal{G}$ *by finding a set of missing triples* $\mathbf{T}' = \{(h, r, t) \mid (h, r, t) \notin \mathbf{T}, h, t \in \mathbf{E}, r \in \mathbf{R}\}$ *when only few-shot entity pairs* (h, t) *and their multi-modal contexts are known for each relation* r*.*

In Definition 1, it is also called the K-shot KGC task when K training examples are given for each relation. In contrast to previous work, which usually assumes the availability of enough triples for training, this work studies the case where only few training triples are available. To be more specific, the goal is to rank the true tail entity higher than other candidate entities, given only K example triples $(h'_i, r, t'_i)_{i=1}^{K}$ for relation r. The candidate set is constructed using the entity type constraint [29].

3.2 Few-Shot Learning Settings

Following the standard meta-learning pipelines [20,7], we describe the settings for training and evaluation of our few-shot learning model. We have different sets for meta-training, meta-validation, and meta testing ($D_{\text{meta–train}}$, $D_{\text{meta–validation}}$, and $D_{\text{meta–test}}$) respectively. Note that none of the above share the same relation label space. On $D_{\text{meta–train}}$, we are interested in training a learning procedure (the meta-learner) that can take few examples as input and produce a matching metric (the learner) that could be used to predict new facts. Using $D_{\text{meta–validation}}$ we can perform hyper-parameter selection of the meta-learner and evaluate its generalization performance on $D_{\text{meta–test}}$.

More specifically, a $D_{\text{meta–train}}$ corresponding to a certain relation $r \in \mathcal{R}$, consists of support and query triples: $D_r = \{D_{\text{s\_r}}, D_{\text{q\_r}}\}$. There are K triples in $D_{\text{s\_r}}$ for K-shot KGC tasks. $D_{\text{q\_r}} = \{h_i, r, t_i, C_{h_i,r}\}$ consists of the query triples of r with ground-truth tail entities t_i for each query (h_i, r), and the corresponding tail entity candidates $C_{h_i,r} = \{t_{ij}\}$ where each t_{ij} is an entity in the KGs. Then

the metric model can be tested on this set by ranking the candidate set $C_{h_i,r}$, given the test query (h_i, r) and the labeled triple in $D_{\text{s\_r}}$. $D_{\text{meta–validation}}$ and $D_{\text{meta–test}}$ are composed of $D_{\text{s\_r}}$, $D_{\text{q\_r}}$. We denote the ranking loss of relation r as $\ell_\theta\left(h_i, r, t_i \mid C_{h_i,r}, D_{\text{s\_r}}\right)$, where θ represents the parameters of our model. Thus, the objective of model training can be defined as:

$$\min_{\theta} \mathbb{E}_{D_r}\left[\sum_{\left(h_i,r,t_i,C_{h_i,r}\right)\in D_{\text{q\_r}}} \frac{\ell_\theta\left(h_i, r, t_i \mid C_{h_i,r}, D_{\text{s\_r}}\right)}{\left|D_{\text{q\_r}}\right|}\right] \tag{1}$$

where D_r is sampled from the meta-training set $D_{\text{meta-train}}$ and $\left|D_{\text{q\_r}}\right|$ denotes the number of tuples in $D_{\text{q\_r}}$.

After sufficient training, we are able to predict facts of each new relation $r' \in \mathcal{R}'$. Due to the assumption of K-shot learning, the relation label space of the above meta-sets is disjoint with each other, i.e., $\mathcal{R} \cap \mathcal{R}' = \phi$. Otherwise, the metric model will actually see more than K-shot labeled data during meta-testing, thus the few-shot assumption is violated. Finally, we construct a subset $\mathcal{G}^*$ from $\mathcal{G}$ by removing all relations in $D_{\text{meta–train}}$, $D_{\text{meta–validation}}$ and $D_{\text{meta–test}}$ to construct entities' neighborhoods.

4 Model

Our model MULTIFORM consists of two modules: a multi-modal context encoder and a metric learning module. The core of our proposed model is a similarity function $f_{\mathcal{S}}\left((h,t),(h',t') \mid \mathcal{V}^*, \mathcal{T}^*, \mathcal{G}^*\right)$, where $\mathcal{V}^*, \mathcal{T}^*, \mathcal{G}^*$ is the set of entities' visual context, textual context, and topological context, respectively. Given K known facts $(h'_i, r, t'_i)_{i=1}^K$ for any query relation r, the model could predict the likelihood of testing triples $\{(h_i, r, t_{ij}) \mid t_{ij} \in C_{h_i,r}\}$, based on the matching score between each (h_i, t_{ij}) and its semantic average of $(h'_i, t'_i)_{i=1}^K$. The implementation of the above matching function involves two sub-tasks: (1) the representations of entity pairs; and (2) the comparison function between two entity-pair representations.

4.1 Multi-Modal Context Encoder

Multi-modal context encoder aims at utilizing the multi-modal contexts to learn well-represented entities. Specifically, it can be split into four parts: an image encoder, a text encoder, a neighbor encoder and a multi-modal embedding fusion model as illustrated in Figure 2. The image encoder aims to extract the visual representations of entities' images and acquire visual embeddings for entities. The text encoder takes textual descriptions as input and output entities' textual embeddings. The neighbor encoder learns from entities' neighborhoods and produces topological embedding. The multi-modal embedding fusion model concatenates on integrating various multi-modal context embeddings and acquiring the accurate entity embeddings.

Image Encoder. Since most entities have more than one image collected in various scenarios, the image set is very possible to contain wrong images, which do not match the corresponding entities. It is essential to find out which images better represent their corresponding entities and filter out the noisy images. [34] shows that incorrect images account for only a small proportion of all images in KGs. Inspired by [34], we utilize a filter gate based on the empirical analysis that the incorrect images have low similarity with the right images. To be more specific, given an entity h, its multiple images can be presented as $V = \{v_1, v_2, \ldots, v_n\}$, where $V \in \mathcal{V}^*$. The filter gate selects the image with the highest similarity to the other images of the given entity to learn the visual embeddings:

$$v_h = \arg\max_{v_i \in V} \left\| \sum_j S\left(v_i, v_j\right) \right\|, \tag{2}$$

where S is the function to measure the visual similarity of two images. We adopt pHash [18] for simplicity. As ViT achieves the best performance over the Convolutional Neural Network (CNN) based models according to [34], we adopt ViT to encode the selected right images to obtain the corresponding embeddings of images in V as $\{z_{v_1}, z_{v_2}, \ldots, z_{v_n}\}$. Finally, we devise an attention aggregation function f_aggre to model representations of different images of the given entity and obtain the visual embedding z_V:

$$f_{aggre}(V) = \sigma\left(\sum_i \alpha_i z_{v_i}\right), \tag{3}$$

$$\alpha_i = \frac{\exp\left\{u_v^T \left(W_v z_{v_i} + b_v\right)\right\}}{\sum_j \exp\left\{u_v^T \left(W_v z_{v_j} + b_v\right)\right\}}, \tag{4}$$

where *sigma* denotes activation unit (we use Tanh); $z_{v_i} \in \mathbb{R}^{d\times 1}$ is the output representations of ViT and d is dimension of representation vectors; $u_v \in \mathbb{R}^{d\times 1}, W_v \in \mathbb{R}^{d\times d}, b_v \in \mathbb{R}^{d\times 1}$ are learnable parameters.

Text Encoder. Given a certain entity and its text description $X = \{x_1, x_2, \ldots, x_n\}$ where x is the word in the sentence, we first use BERT [6] to generate the word embedding $\{z_{x_1}, z_{x_2}, \ldots, z_{x_n}\}$. Similarly, we adopt the attention aggregation function f_aggre to obtain the textual embedding z_X:

$$f_{aggre}(X) = \sigma\left(\sum_i \beta_i z_{x_i}\right), \tag{5}$$

$$\beta_i = \frac{\exp\left\{u_x^T \left(W_x z_{x_i} + b_x\right)\right\}}{\sum_j \exp\left\{u_x^T \left(W_x z_{x_j} + b_x\right)\right\}}, \tag{6}$$

where $z_{x_i} \in \mathbb{R}^{d\times 1}$ is the output representations of BERT and d is dimension of word embedding vectors; $u_x \in \mathbb{R}^{d\times 1}, W_x \in \mathbb{R}^{d\times d}, b_x \in \mathbb{R}^{d\times 1}$ are learnable parameters.

Neighbor Encoder. Recently, Xiong et al. [38] and Zhang et al. [40] have demonstrated the effectiveness of encoding local graph structures as entity representations. Following their researches and inspired by the progress in Graph Convolutional Network (GCN), we consider CompGCN [31] to model the local heterogeneous feature of the neighborhoods. Specifically, for each given head entity h, its neighborhoods forms a set of {*relation*, *tail entity*} tuples. As shown in Figure 2 (c), for the entity *Issac Newton*, one of such tuples is {*Occupation*, *Mathematician*}. Thus, the neighbor set can be denoted as $\mathcal{N}_h = \{r_i, t_i\}_{i=1}^{I}$, where r_i and t_i represent the i-th relation and corresponding tail entity of h, respectively. I is the number of such neighbors and $(h, r_i, t_i) \in \mathcal{G}^*$.

Our CompGCN-based neighbor encoder aims at encoding $\mathcal{N}_h$ and generating a well-learned vector as the feature representation of local connections of h. The details are as follows:

$$y_h^{(k)} = \sigma\left(\sum_{(r_i,t_i)\in\mathcal{N}_h} W_{\lambda(r)}^{(k)} \psi\left(y_{r_i}^{(k-1)}, y_{t_i}^{(k-1)}\right)\right), \tag{7}$$

where $W_{\lambda(r)}^{(k)}$ is a relation-specific shared parameter to learn; ψ a composition function of the relation r_i with its respective tail entity t_i. The composition $\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ can be any entity-relation function akin to TransE [2] or RotatE [26](We choose RotatE according to experimental results); y_h, y_r, y_t is the embeddings of h, r, t respectively and can random initialized or pretrained by existing embedding-based models; $y_h^{(k)}$ is the final topological embedding.

Multi-Modal Entity Embedding Fusion Model. With multi-modal context information encoded, an embedding fusion model is developed to improve the representations of the given entity. Among various ensemble functions, [15] point out that simple concatenation works better than DeViSE [8] and Imagined [4] on multi-modal KGC tasks, and taking limited computational resources and scalability of MULTIFORM, we use simple concatenation to aggregate the visual embedding, textual embedding and topological embedding.

4.2 Metric Learning Module

This module is designed to do effective similarity matching given the output of feature fusion module. For K-shot learning scenario, we get two sets of entity pairs: the query entity pair set (h_i, t_{ij}) and the support pair set $(h_i', t_i')_{i=1}^{K}$. We obtain well represented entity embeddings for each set: $\left[o\left(\mathcal{N}_{h_i}\right); o\left(\mathcal{N}_{t_{ij}}\right)\right]$ and $[o\left(\mathcal{N}_{h'}\right); o\left(\mathcal{N}_{t'}\right)]$ via the multi-modal context encoder. When $K > 1$, we employ a simple sematic averaging function to get $\mathcal{N}_{h'}$ and $\mathcal{N}_{t'}$:

$$\mathcal{N}_{h'} = \frac{\sum_{i=1}^{K} \mathcal{N}_{h_i'}}{K} \tag{8}$$

$$\mathcal{N}_{t'} = \frac{\sum_{i=1}^{K} \mathcal{N}_{t_i'}}{K}. \tag{9}$$

Table 1: Statistics of the Datasets. # Entities denotes the number of unique entities and # Relations denotes the number of all relations. # Tasks denotes the number of relations we use as few-shot tasks.

Dataset	#Entities	# Relations	# Triples	# Tasks
MM-FB15K	14951	1345	592,213	356
MM-DBpedia	12842	279	297,084	69

We can simply concatenate $o(\mathcal{N}_h\prime)$ and $o(\mathcal{N}_t\prime)$ and calculate similarity between pairs in the two sets. For our model's scalability, we use the same multi-step matching processor as [38]. Every process step is defined as follows:

$$h'_{k+1}, c_{k+1} = \text{LSTM}\left(p, [h_k \oplus s, c_k]\right) \tag{10}$$

$$h_{k+1} = h'_{k+1} + p \tag{11}$$

$$score_{k+1} = \frac{h_{k+1} \odot s}{|h_{k+1}|\,|s|} \tag{12}$$

where $s = o\left(\mathcal{N}_{h'}\right) \oplus o\left(\mathcal{N}_{t'}\right)$, $p = o\left(\mathcal{N}_{h_i}\right) \oplus o\left(\mathcal{N}_{t_{ij}}\right)$ are concatenated well-learned embeddings of the support pair and query pair. After n processing steps, we use $score_k$ as the final similarity score between the query and support entity pair.

4.3 Loss Function

For a selected query relation r and its support triples $(h'_i, r, t'_i)_{i=1}^K$, we employ negative sampling methods to construct query triples, i.e., we collect a group of positive query triples $\left\{\left(h_i, r, t_i^+\right) \mid \left(h_i, r, t_i^-\right) \notin \mathcal{G}\right\}$ and corrupt the tail entities to construct another group negative query triples $\{\left(h_i, r, t_i^-\right) \mid \left(h_i, r, t_i^-\right) \notin |\mathcal{G}\}$. Following previous few-shot learning models, we utilize a hinge loss function for our model:

$$l_\theta = \max\left(0, \gamma + score_\theta^- - score_\theta^+\right) \tag{13}$$

where $score_\theta^+$ and $score_\theta^-$ are scalars calculated by comparing the query triple $\left(h_i, r, t_i^+ / t_i^-\right)$ with the support triples $(h'_i, r, t'_i)_{i=1}^K$ using our metric learning model, and the margin γ is a hyperparameter to tune. For each training episode, we first sample D_r from the meta-training set $D_{\text{meta–train}}$. Then we sample K triple as the support triple $D_{\text{s\_r}}$ and a batch of other triples as the positive query/test triples $D_{\text{q\_r}}$ from all known triples in D_r.

5 Experiments

With MULTIFORM, we investigate three issues: (1) Will the incorporation of multi-modal contexts help the few-shot KGC tasks? (2) How much visual context, textual context and topological context contribute to MULTIFORM's performance, respectively? (3) Does the number of multi-modal training triples affect the performance of MULTIFORM? To explore these questions, we conduct

a series of experiments on two few-shot multi-modal knowledge graph datasets and systematically analyze the corresponding results.

5.1 Datasets

Our constructed multi-modal datasets MM-FB15K and MM-DBpedia are based on FB15K [2,1] and DBpedia [14,11,23]. The dataset statistics are shown in Table 1. Figure 1 (b) shows an example of visual and textual contexts. Each entity in MM-FB15K and MM-DBpedia has at least one image and a description of no less than 15 words. Following [38], we construct few-shot multi-modal KGs by selecting those relations that do not have too many training triples. Specifically, to guarantee enough triples for evaluation, we select the relations with less than 500 but more than 50 triples as few-shot tasks, i.e., we obtain 356 and 69 few-shot relations in MM-FB15K and MM-DBpedia, respectively. The rest of the relations are referred to as background relations and their triples provide neighborhoods to learn topological information. In addition, For MM-FB15K, we use 267/18/71 and 51/6/12 task relations for training/validation/testing in MM-FB15K and MM-DBpedia, respectively. The division ratio is about $15:1:4$, similar to the data split in [38,40].

5.2 Baseline Methods

For fair comparison, we select three kinds of baseline methods including unimodal knowledge embedding models, multi-modal knowledge embedding models, and few-shot learning models.

- **Unimodal Knowledge Embedding Models.** This line of research models multi-relational structures in KGs and encodes both entities and relations into a continuous low dimensional vector space. We consider the four widely used baseline methods as follows: TransE [2], DistMult [39], ComplEx [30] and Rot-Pro [25]. For implementation, we use an Open Toolkit [9] released by Xu Han et al. which provides the above knowledge embedding models. We also select RotatE [26], which has been reported very robust under different evaluation protocols in the extensive conducted experiments, comparing with a series of state-of-the-art knowledge embedding methods [27]. For fair comparison, all triples of background relations, training triples, and support triples of validation and test relations, are used during training.
- **Multi-modal Knowledge Embedding Models.** The models mainly focus on encoding visual and structural contexts. We select two state-of-the-art methods, i.e., TransAE [35] and RSME [34] as our baselines.
- **Few-Shot Learning Models.** This type of model concentrates on predicting new facts in KGs with only few-shot reference triples. For fair comparison, we select three typical neighbor encoder based models, i.e., GMatching [38], FSRL [40], FAAN [21].

Table 2: The 5-shot KGC results on the testing dataset. The best baseline results are indicated by underline and the best results of all methods are highlighted in bold.

	MM-FB15K				MM-DBpedia			
Model	MRR	Hits@10	Hits@5	Hits@1	MRR	Hits@10	Hits@5	Hits@1
TransE	0.116	0.164	0.139	0.089	0.103	0.155	0.120	0.077
DistMult	0.083	0.132	0.095	0.037	0.091	0.141	0.118	0.088
ComplEx	0.067	0.147	0.089	0.05	0.121	0.17	0.123	0.109
RotatE	0.131	0.189	0.160	0.101	0.150	0.242	0.179	0.120
Rot-Pro	0.099	0.145	0.112	0.061	0.139	0.200	0.154	0.107
TransAE	0.130	0.243	0.155	0.116	0.156	0.237	0.185	0.131
RSME	0.188	0.308	0.249	0.152	0.177	0.280	0.219	0.145
GMatching	0.261	0.377	0.340	0.189	0.176	0.293	0.231	0.116
FSRL	0.162	0.289	0.197	0.085	0.158	0.304	0.220	0.071
FAAN	0.341	0.458	0.382	0.279	0.195	0.310	0.217	0.136
MULTIFORM	**0.437**	**0.550**	**0.461**	**0.305**	**0.303**	**0.425**	**0.334**	**0.279**

Table 3: Results of model variants on MM-FB15K dataset. The best results are highlighted in bold.

Model Variants	MRR	Hits@10	Hits@5	Hits@1
AS_1	0.401	0.499	0.450	0.293
AS_2	0.383	0.482	0.443	0.288
AS_3	0.351	0.472	0.397	0.282
MULTIFORM	**0.437**	**0.550**	**0.461**	**0.305**

5.3 Implementation Details

The embedding size d is set to 128 and 256 for MM-FB15K and MM-DBpedia datasets, respectively. The number of local neighbors used in the neighbor encoder is set to 45, which works the best for both datasets. As for image encoder and text encoder, we use the open resource from huggingface to implement ViT [1] and BERT [2] and keep their default settings about transformer layers. Besides, the LSTM cell is utilized in the matching function as a matching processor. The dimension of LSTM's hidden state is set to 128 and 256 for MM-FB15K and MM-DBpedia datasets, respectively. The optimal matching step is 2. For parameter updates, we use Adam [10] with the initial learning rate of 0.001 and we have the learning rate decay 0.2 for each 50k training step. The margin γ used in the base loss function is 5.0.

[1] https://huggingface.co/docs/transformers/model_doc/bert
[2] https://huggingface.co/docs/transformers/model_doc/vit

5.4 Results

We first evaluate our model on the few-shot KGC task, which predicts new facts on a query set given only few support triples and their multi-modal contexts. As shown in Table 2, MULTIFORM shows a significant margin over all three types of baselines in the 5-shot scenario. Taking the experimental results (testing MRR and Hits@10) on MM-FB15K as an example, the relative improvement (%) of MULTIFORM against RotatE (the best-performing knowledge embedding models) is up to 233.59% and 191.01% ; MULTIFORM outperforms RSME (the best-performing multi-modal knowledge embedding models) by 132.45% and 78.37%; MULTIFORM shows a significant improvement margin over FAAN (the best-performing few-shot learning models) by 28.15% and 20.09%. These results, to some extent, confirm the effectiveness of the idea that incorporating multi-modal contexts can be helpful to few-shot KGC tasks since multi-modal contexts shape more accurate and well-represented entities' embeddings. Thus, we have so far answered the first question, i.e., MULTIFORM can be well adapted into the few-shot KGC task and produce consistent improvements over all types of baselines by incorporating multi-modal contexts. We also observe that most multi-modal knowledge embedding models have better performance than unimodal knowledge embedding models, which verifies the benefit of utilizing multi-modal contexts. We also noticed that unimodal/multi-modal knowledge embedding models have a big gap in performance compared with few-shot learning models. We guess unimodal/multi-modal knowledge embedding models are designed for transductive learning with sufficient training data and can not be adapted into the few-shot scenario where only few training data are available. By the way, this demonstrates that the few-shot KGC task is a very challenging problem.

5.5 Ablation Study

Here We seek the answer to our second question in this section, i.e., investigating the effectiveness of each context of the proposed model. We consider the following ablation studies:

- **(AS_1)** We evaluate the effectiveness of images. We use randomly initialized vectors as visual embeddings and keep the other two encoders.
- **(AS_2)** We use randomly initialized vectors as the output of the text encoder to verify the effectiveness of text descriptions.
- **(AS_3)** We use randomly initialized vectors as topological embeddings to evaluate the effectiveness of entities' graph structure.

As shown in Table 3, our model has better performance than all model variants. The comparison between MULTIFORM and AS_1, AS_2, and AS_3 indicates that all visual context, textual context and topological context contribute to improvements of our model. By comparison among AS_1, AS_2, and AS_3, we also notice that topological context contributes most to the model's performance,

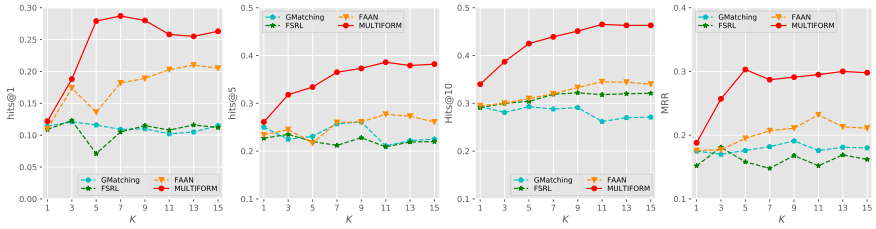

Fig. 3: Impact of few-shot size K.

since MULTIFORM shows the largest decrease when randomly initializing topological embeddings (refer to AS_3); We think it is because the knowledge of the same modality can be absorbed by neural networks more easily. The next largest contribution is made by textual context (refer to AS_2). We guess it is because KGs are originally extracted from the text so there exists semantic similarity. In summary, these results demonstrate that all contexts are important and contribute to MULTIFORM.

5.6 Impact of Few-Shot Size

Since this work studies few-shot learning for KGC tasks, we conduct experiments to analyze the impact of few-shot size K. MULTIFORM consistently outperforms all few-shot baselines under different K, indicating the effectiveness of our model on few-shot link prediction on KGs. We also notice that as K increases, MULTIFORM gets relatively stable improvements compared to GMatching and FSRL, which demonstrates MULTIFORM's stability and robustness.

6 Conclusion and Future Work

In the present work, we introduce a multi-modal few-shot learning framework named MULTIFORM for KGC tasks. MULTIFORM aims at predicting new facts with only several training data and their multi-modal contexts, which is a challenging problem. MULTIFORM leverages visual, textual, and topological information of entities to produce well-learned representations and uses a metric learning method to match entity pairs. The experiment results demonstrate that MULTIFORM can outperform the state-of-the-art baselines. We also analyze the impact of few-shot size and conduct ablation studies on multi-modal contexts, which verify the effectiveness of each context. The goal of our future work is to incorporate external text content of relations and try more feature fusion methods to extend our model in the zero-shot scenario.

Acknowledgements This work was supported by the National Natural Science Foundation of China (62072463, 71531012), and the National Social Science Foundation of China (18ZDA309).

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of data (2008)
2. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States (2013)
3. Chen, M., Zhang, W., Zhang, W., Chen, Q., Chen, H.: Meta relational learning for few-shot link prediction in knowledge graphs. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)
4. Collell, G., Zhang, T., Moens, M.: Imagined visual representations as multimodal embeddings. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA (2017)
5. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018 (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research (2017)
8. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States (2013)
9. Han, X., Cao, S., Lv, X., Lin, Y., Liu, Z., Sun, M., Li, J.: OpenKE: An open toolkit for knowledge embedding. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2018)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
11. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web (2) (2015)
12. Li, Z., Li, X., Wei, Y., Bing, L., Zhang, Y., Yang, Q.: Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)
13. Li, Z., Zhou, F., Chen, F., Li, H.: Meta-sgd: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835 (2017)
14. Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., Rosenblum, D.S.: Mmkg: multi-modal knowledge graphs. In: European Semantic Web Conference. Springer (2019)
15. Mousselly-Sergieh, H., Botschen, T., Gurevych, I., Roth, S.: A multimodal translation-based approach for knowledge graph representation learning. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (2018)
16. Munkhdalai, T., Yu, H.: Meta networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research (2017)
17. Nickel, M., Tresp, V., Kriegel, H.: A three-way model for collective learning on multi-relational data. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011 (2011)
18. Niu, X.m., Jiao, Y.h.: An overview of perceptual hashing. Acta Electronica Sinica (7) (2008)
19. Peng, Z., Li, Z., Zhang, J., Li, Y., Qi, G., Tang, J.: Few-shot image recognition with knowledge transfer. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019 (2019)
20. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)
21. Sheng, J., Guo, S., Chen, Z., Yue, J., Wang, L., Liu, T., Xu, H.: Adaptive Attentional Network for Few-Shot Knowledge Graph Completion. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
22. Shi, B., Weninger, T.: Open-world knowledge graph completion. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018 (2018)
23. Shi, B., Weninger, T.: Open-world knowledge graph completion. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018 (2018)
24. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (2017)
25. Song, T., Luo, J., Huang, L.: Rot-pro: Modeling transitivity by projection in knowledge graph embedding. Advances in Neural Information Processing Systems (2021)
26. Sun, Z., Deng, Z., Nie, J., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019)

27. Sun, Z., Vashishth, S., Sanyal, S., Talukdar, P., Yang, Y.: A re-evaluation of knowledge graph completion methods. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
28. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)
29. Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., Gamon, M.: Representing text for joint embedding of text and knowledge bases. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015)
30. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings (2016)
31. Vashishth, S., Sanyal, S., Nitin, V., Talukdar, P.P.: Composition-based multi-relational graph convolutional networks. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 (2020)
32. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain (2016)
33. Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., Guo, M.: Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018 (2018)
34. Wang, M., Wang, S., Yang, H., Zhang, Z., Chen, X., Qi, G.: Is visual context really helpful for knowledge graph? a representation learning perspective. In: Proceedings of the 29th ACM International Conference on Multimedia (2021)
35. Wang, Z., Li, L., Li, Q., Zeng, D.: Multimodal data enhanced representation learning for knowledge graphs. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE (2019)
36. Xie, R., Liu, Z., Jia, J., Luan, H., Sun, M.: Representation learning of knowledge graphs with entity descriptions. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA (2016)
37. Xie, R., Liu, Z., Luan, H., Sun, M.: Image-embodied knowledge representation learning. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017 (2017)
38. Xiong, W., Yu, M., Chang, S., Guo, X., Wang, W.Y.: One-shot relational learning for knowledge graphs. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)
39. Yang, B., Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
40. Zhang, C., Yao, H., Huang, C., Jiang, M., Li, Z., Chawla, N.V.: Few-shot knowledge graph completion. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020 (2020)