Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective

Meng Wang
School of Computer Science and Engineering, Southeast University
Nanjing, China
meng.wang@seu.edu.cn

Sen Wang
The University of Queensland
Brisbane, Australia
sen.wang@uq.edu.au

Han Yang
Peking University
Beijing, China
captain@pku.edu.cn

Zheng Zhang
Harbin Institute of Technology
Shenzhen, China
darrenzz219@gmail.com

Xi Chen
Tencent
Shenzhen, China
jasonxchen@tencent.com

Guilin Qi
Southeast University
Nanjing, China
gqi@seu.edu.cn

ABSTRACT

Visual modality recently has aroused extensive attention in the fields of knowledge graph and multimedia because a lot of real-world knowledge is multi-modal in nature. However, it is currently unclear to what extent the visual modality can improve the performance of knowledge graph tasks over unimodal models, and equally treating structural and visual features may encode too much irrelevant information from images. In this paper, we probe the utility of the auxiliary visual context from knowledge graph representation learning perspective by designing a Relation Sensitive Multi-modal Embedding model, RSME for short. RSME can automatically encourage or filter the influence of visual context during the representation learning. We also examine the effect of different visual feature encoders. Experimental results validate the superiority of our approach compared to the state-of-the-art methods. On the basis of in-depth analysis, we conclude that under appropriate circumstances models are capable of leveraging the visual input to generate better knowledge graph embeddings and vice versa.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Semantic networks.*

KEYWORDS

Multi-modal, Knowledge graph, Representation learning

ACM Reference Format:
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 *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475470

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475470

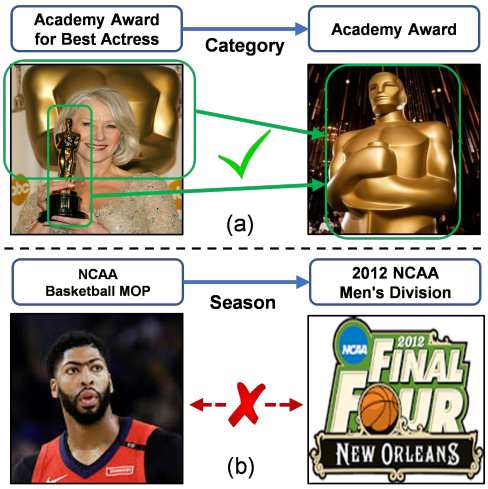

Figure 1: Examples of multi-modal knowledge graph facts. Images of entities around the relation *Category* have common visual features. As a contrast, images have little visual similarity on *Season*.

1 INTRODUCTION

Knowledge graphs (KGs), e.g., Wikidata [30], Freebase [3], and DBpedia [2], contain relational facts in the form of (head) —relation→ (tail), which have been widely used in various kinds of tasks, such as question answering [13], recommendation system [11], and multimedia reasoning [17, 27]. Apart from relations to a fixed set of entities and structural attributes, KGs often include rich visual context, usually images (profile photos, thumbnails, posters, etc.). Figure 1 demonstrates image examples of entities in KGs. Each entity has multiple images which describe the appearances and behaviours of this entity. Therefore, standing on the advance of multi-modal representation learning techniques [1] in recent years, many practitioners believe that visual modality is crucial and beneficial for improving the conventional KG based applications [5, 19], most of which only rely on KG's structural context as inputs. However, it is still not clear to what extent truly multimodal reasoning is required for the current KG tasks and datasets. For instance, it has

been pointed out that many unimodal natural language processing models can perform the same well without any understanding of the visual content than the multi-modal counterpart [9]. Analogously, is visual context really helpful for KG problems? Apparently, how to fully exploit visual information is one of the core issues in multi-modal KG scenarios, which directly impacts the model performance. We will focus on answering this question from the KG representation learning perspective in this work.

KG representation learning [31] aims to encode entities and relations into a low-dimensional, continuous vector space. The learned dense vector representations, a.k.a. embeddings, of entities and relations that mathematically support various machine learning models to perform KG completion and link prediction, in turn, they can be used in multi-modal reasoning tasks. However, most of the existing KG representation learning methods only consider the KG structural context, ignoring the visual information of the entities. Therefore, to achieve better performance, it is a promising pathway to project the heterogeneous features of KG entities into a common space, where the multi-modal information with similar semantics will be fused by unified embeddings. To this end, IKRL [34] started to integrate image features into the translation-based KG representation learning models, such as TransE [4]. IKRL generated two separate representations for each entity, i.e., one is based on KG structures and the other is based on visual context. Differently, Mousselly et al [22] and TransAE model [32] jointly encoded the visual and structural knowledge at the same time. Mousselly et al used three different methods, i.e., simple concatenation, DeViSE [10], and Imagined [6] to integrate multi-modal information, and TransAE utilized an auto-encoder to fuse them.

While current multi-modal KG representation learning methods have brought extent improvement, they assumed that the learned embeddings are expected to be better since the visual modality intuitively contributes to the complementary or supplementary in contents. This assumption may be challenged because images may also introduce noise and lead to uncertainty about whether the visual context really improves embedding quality. For instance, as shown in Figure 1(a), the images of (Academy Award for Best Actress) and (Academy Award) will be beneficial for learning the representation of $\xrightarrow{\text{Category}}$. However, as a contrast, in Figure 2(b), images of (NCAA MOP) and (2012 NCAA Men's Division) do not have any visual similarity or semantic correlation, which will have unwelcome influence on the representation learning of $\xrightarrow{\text{Season}}$. Therefore, directly adding visual information to the traditional KG embeddings may cause negative effects and corruptions. Moreover, most of the above embedding methods use the hidden embeddings of the convolutional neural network (CNN) as the initial representations of the visual information. We argue that the effect of different visual feature encoders is also worth to be discussed by comparing the pre-trained VGGNET to other image encoder models, such as vision transformer [7].

Motivated by the above analysis, in this paper, we focus on exploring the impact of visual context on KG representation learning, and aim to design an embedding model to automatically enhance beneficial visual context and simultaneously filter the noise from images. To be more specific, we propose a novel Relation Sensitive Multi-modal KG Embedding model, RSME for short, as illustrated in Figure 2. RSME is composed of three gates and a basic KG embedding model. The filter gate removes noise at the dataset level and, for each KG entity, only keeps the image with the highest average similarity to the other images. The forget gate utilizes a reductive link prediction mechanism to measure the effect of image information under a specific relation, and then forgets the visual information of the entities that do not match the current relation circumstance. Ultimately, the fusion gate fuses the visual context and structural information to learn KG entities and relations into low-dimensional vectors. On real-world datasets, link prediction experiments are conducted to investigate the quality of embeddings generated by our model. Experiment results demonstrate that RSME outperforms the existing methods and achieves state-of-the-art performance. It is worth noting that: 1) RSME can selectively choose the image information of the entity and ignores the other irrelevant ones; 2) A mean rank proportion (MRP) is proposed to judge the value of image information; 3) Different image encoders are employed and tested in RSME. We conclude that visual context in KG is not always beneficial for representation learning. A good visual context-aware mechanism and appropriate relation circumstance are necessary. We release both code and datasets on Github[1] and hope this work presented here are helpful for multi-modal KG research work in the future.

In summary, the contributions of this work are as follows:

- As far as we know, this paper is the first to study what extent and circumstance that visual context is needed in KG problems. We design a novel model, namely RSME, which takes relation circumstances into account and utilizes an MRP metric in a forget gate to selectively filter visual information during the KG embeddings learning.
- We explore the impact of different visual feature encoders for the multi-modal KG representation learning, which is empirically important but ignored by the previous embedding models. To the best of our knowledge, there is still no similar work.
- We perform comprehensive experiments and sensitivity analysis on real-world benchmark datasets. Results and analysis show that the model can adaptively make use of visual information and substantially outperform the current state-of-the-art models under appropriate circumstances.

2 PROBLEM FORMULATION

In this section, we introduce the notations used in this paper and formulate the multi-modal KG representation learning problem.

Knowledge graph (KG) is defined as a directed graph, which can be represented as a collection of triples expressed as (h, r, t), where h represents the head entity, t represents the tail entity and r represents the relation between h and t.

Knowledge graph embedding aims to compress both entities and relations into a continuous, low-dimensional vector space. Given a triple (h, r, t), A loss function $f(h, r, t)$ is defined on triples to reflect the probability of the relation between h and t. By minimizing the loss function, we can finally get the KG embeddings.

[1] https://github.com/wangmengsd/RSME

A metric space is an ordered pair (G, d) where G is a set and d is a metric on G, i.e., a function $d : G \times G \rightarrow \mathbb{R}$ such that for any $x, y, z \in G$, the following holds:

1. $d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x)$
3. In Euclidean space, $d(x, z) \leq d(x, y) + d(y, z)$

KG embedding group representation: The KG embedding process is divided into three steps. The first step is doing the group operation between the head entity h and the relation r on the group $< G, * >$, thereby generating a characteristic entity $\tilde{t}$ of h and r:

$$\tilde{t} = h * r, \tag{1}$$

where $*$ is the group operation and $h, r \in G$. The second step is to calculate the distance between $\tilde{t}$ and the tail entity t in the metric space $< G, *, d >$:

$$d(\tilde{t}, t), d : G \times G \rightarrow \mathbb{R}. \tag{2}$$

The third step is to calculate the loss function $f(d)$.

Fusion function, i.e., the fusion gate is the mechanism for fusing structural information and the image information:

$$\Phi : G \times G \rightarrow G. \tag{3}$$

Multi-modal KG embedding group representations can be regarded as KG embedding that introduces a fusion gate, Φ . The embedding process is divided into 3 steps. The first step is doing the group operation between $\Phi(h_s, h_i)$ and the relation r on the group $< G, * >$, where h_s is the structural embedding of the head entity and h_i is the image embedding of the head entity. Therefore, a characteristic entity $\tilde{t}$ of $\Phi(h_s, h_i)$ and r can be written as:

$$\tilde{t} = \Phi(h_s, h_i) * r, \tag{4}$$

where $h_s, h_i, r \in G$. The second step is to calculate the distance between $\tilde{t}$ and $\Phi(t_s, t_i)$ in the metric space $< G, *, d >$:

$$d(\tilde{t}, \Phi(t_s, t_i)), d : G \times G \rightarrow \mathbb{R}, \tag{5}$$

where t_s is the structural embedding of the tail entity and the t_i is the image embedding of the tail entity. The third step is to calculate the loss function $f(d)$.

3 METHOD

This section details our proposed relation sensitive multi-modal KG embedding model (i.e., RSME), which can automatically encode the selective visual information. As shown in Figure 2, RSME is composed of four parts, i.e., a basic KG embedding model and three gates (i.e., filter gate, forget gate, and fusion gate). RSME first uses a filter gate to automatically filter irrelevant images instead of feeding all visual context indiscriminately. Once the visual information is selected, the image will pass through a forget gate to enhance beneficial features, and the noise with a small MRP score will be ignored. Following the forget gate, the visual information and KG structural information are fused in a fusion gate and finally the embeddings of entities and relations are obtained by minimizing the loss function.

3.1 Image Encoder

The image encoder aims to extract the visual representations of KG entities. The convolutional neural network (CNN) based models (such as VGGNET [26] and AlexNet [14]) are the most frequently used encoders in previous multi-modal KG embedding work [22, 32, 34]. This causes the different effects of various image encoders in the KG embedding model ignored. Therefore, to investigate the impacts of different image encoders in KG embedding, we employ three kinds of image encoders in this paper, i.e., the CNN encoders (VGG16 [26] or Resnet50 [12]) in the forget gate, the perceptual hash algorithms (pHash) in the filter gate and forget gate, and the vision transformer [7] in the forget gate. The perceptual hash algorithms are commonly used in producing snippets or fingerprints of various forms of multimedia information. Compared with CNN, the vision transformer model needs less induced bias and is more suitable to capture global visual information.

Table 1: The comparison of basic embedding models.

Model	Loss Function	G	$*$	d
TransE	$\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$	$\mathbb{R}^n$	Addition	Euclidean
ComplEx	$\text{Re}\left(\sum_{k=1}^{n} \mathbf{h}_k \mathbf{r}_k \bar{\mathbf{t}}_k\right)$	$\mathbb{C}^n$	Complex product	Inner product
DistMult	$\sum_{k=1}^{n} \mathbf{h}_k \mathbf{r}_k \mathbf{t}_k$	$\mathbb{R}^n$	Hadamard product	Inner product

3.2 Basic Embedding Model

Table 1 compares three popular basic embedding models for KG structural features encoding in multi-modal learning under the group representation. Different from the existing work, we do not use the translation model TransE as the basic model in our framework, but utilize the semantic matching ComplEx model. The reason is that the loss function of ComplEx is $\text{Re}\left(\sum_{k=1}^{n} \mathbf{h}_k \mathbf{r}_k \bar{\mathbf{t}}_k\right)$ which uses the inner product as a metric function and facilitate unifyiung the structural model and the visual information. DistMult is the most concise approach among these models, we hence use it in the forget gate which requires circumventing the interference of structural information.

3.3 Filter Gate

The filter gate is designed to remove noise at the dataset level and, for each KG entity, only keeps the image with the highest average similarity to the other images. In contrast, previous methods directly utilized visual information in the KG embedding task and may involve noise caused by images. Note that the noise of image information mainly comes from two aspects, one is from the inaccuracy of datasets (addressed in this gate), the reason is that some entities have incorrect images, which are automatically downloaded from the Internet. The other noise is at the task level, i.e., the poor visual semantic similarity between related KG entities (addressed in forget gate).

Most entities have more than one image from different aspects in various scenarios. Therefore, it is essential but also challenging to find out which images are better to represent their corresponding entities and filter the irrelevant images. To solve the problem, a filter gate is proposed based on the empirical analysis that incorrect images account for only a big proportion of all the images. These few incorrect images have low similarity with other pictures. To be specific, given an entity e, it contains multiple images in KG which can be represented as $I = \{img_1, img_2, img_3, ..., img_n\}$. The filter

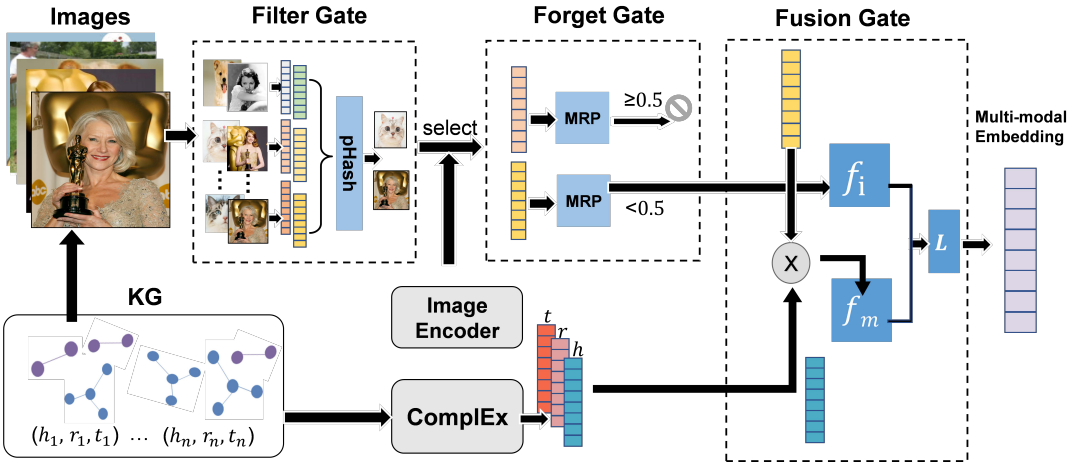

Figure 2: Schematic illustration of the proposed RSME model.

gate selects the image with the highest similarity to the other images of the given entity to perform further representation learning, denoted as img_e:

$$img_e = \arg\max_{img_\theta \in I} \| \sum_{i=1}^{n} S(img_\theta, img_i) \|. \quad (6)$$

where S represents the function to measure the visual similarity of two images. For simplicity and efficiency, pHash [25] is used in the filter gate.

3.4 Forget Gate

In forget gate, the beneficial visual information of entities will be enhanced while the noise with poor visual similarity will be eliminated according to different relation circumstances. As shown in Figure 3, the forget gate mainly includes two key parts, i.e., structural information omitting and mean rank proportion.

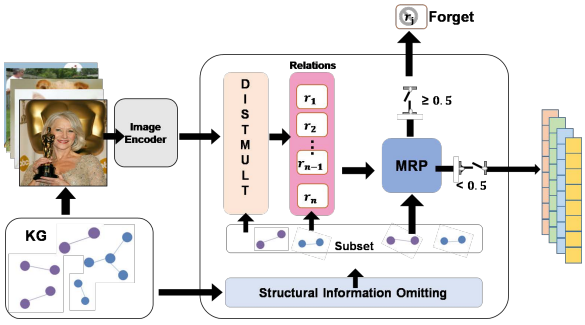

Figure 3: Overall architecture of Forget Gate.

After a preliminary empirical analysis of the dataset, we find the images in the KG can be roughly divided into two categories, i.e., visually checkable images and deep semantic images. The visually checkable images refer to the image pairs of two corresponding entities in a KG triple that has extent visual similarity by color, line composition, or other visual information, such as images in Figure 1(a). Deep semantic images refer to images whose visual relevance can only be recognized with the human experience and knowledge, such as images in Figure 1(b). Therefore, in this work, visual information in visually checkable images are relatively easy to be captured by our forget gate while deep semantic images are relatively difficult to be detected by the model.

By further analysis, we find that the types of images are highly related to the relations. This means that some relations tend to be around by visually checkable images, such as $\xrightarrow{\text{hyponym}}$ or $\xrightarrow{\text{part\_of}}$ while other relations tend to have deep semantic images, such as $\xrightarrow{\text{judge}}$. Therefore, the forget gate determines the relation circumstances in which visual information will be enhanced.

Structural Information Omitting: In order to determine the effectiveness of visual context for KG embedding learning, we need to ensure that the forget gate is not interfered by the structural information, i.e., the independence of visual information should be only maintained. Thus, the first step of the discriminator in the forget gate is to omit the KG structural information.

It is worth noting that most of the existing embedding work only considers the visual information of entities, while the relation is still only encoded by structural information. However, given a head entity and a tail entity, there may be multiple relations between them in the real-world KGs. In this case of multiple relations, since the head entity and the tail entity are not in one-to-one correspondence, the relation between them cannot be directly omitted. Therefore, for the sake of simplification, we follow the conventional settings and extract subsets of the original datasets, i.e., WN18-IMG-S and FB15K-IMG-S, which only contain KG triples with one-to-one corresponding entities.

MRP: For each given relation r, we design a metric mean rank proportion (*MRP*) to determine the value of visual information. If *MRP* of r is greater than a threshold, the image information of the relation r will be forgotten. Otherwise, the image information is retained. Specifically, the *MRP* score of an r can be computed by a reductive link prediction process. For convenience, we use the

relatively simple DistMult model (refer to Table 1) as the KG embedding model. Since the structural information has been omitted, we set fusion function $\Phi(h_s, h_i) = h_i$ and $\Phi(t_s, t_i) = t_i$ and the group operation, $h_i * r = h_i$, where h_i and t_i are visual embeddings encoded by image encoders. Then we directly test the model on the non-structural dataset without training. Finally, we separately count the *MRP* for each given relation r, where the *MRP* can be calculated by the mean rank:

$$MRP = MR/MAX\_RANK. \tag{7}$$

In general, we use 0.5 as the threshold. Since $MRP = 0.5$ is equivalent to randomly selecting a tail entity from all alternative entities. If the MRP of the given relation is greater than and equal to 0.5, we consider that the visual information has a negative effect on the link prediction, then, we will forget it. Otherwise, the visual information of this relation will be passed to the fusion gate.

3.5 Fusion Gate

In this paper, two fusion mechanisms are designed. The first one is the concatenation and the second one is a relation sensitive linear combination. The concatenation is the most common method for feature fusion, which directly links structural embedding and the projection of image embedding as follows:

$$\mathbf{e} = \Phi(e_s, e_i) = [\mathbf{e}_s : \mathbf{W}\mathbf{e}_i], \tag{8}$$

where Φ is the fusion gate, $\mathbf{W}$ is the projection matrix and $\mathbf{e}$ is the final entity embedding. Analogously, the relation sensitive linear combination can be defined as:

$$\mathbf{e} = (1 - \alpha)\mathbf{e}_s + \alpha\mathbf{W}\mathbf{e}_i, \tag{9}$$

where $\mathbf{W}$ is the projection matrix, α is the ratio of visual information which is up to the relation type of the corresponding KG triples as follows:

$$\alpha = \sigma(-MRP) = \frac{1}{1 + e^{MRP}}. \tag{10}$$

As the *MRP* becomes smaller, the importance of the visual information becomes greater. This means the weight we give to the visual information i.e. α increases.

3.6 Loss Function

We use the fusion gate to combine structural information and visual information together. Inspired by the ComplEx embedding model, the overall loss of the proposed embedding RSME model consists of two parts, i.e., the merged energy function and the image energy function. The subscript m represents merged information, the subscript s represents structural information, and the subscript i represents image information, respectively.

The merged energy function of a KG triple (h, r, t) can be expressed as:

$$f_m(h, r, t) = ComplEx(h_m, r_m, t_m) = \sum_{k=1}^{n} \Phi(h_s, h_i)_k \mathbf{r}_k \overline{\Phi(t_s, t_i)_k}, \tag{11}$$

where Φ is the fusion gate, $\mathbf{r}$ is the structural embedding of the relation r. We hope that the head and tail entity visual information can also be consistent. Hence, we define the visual energy function as follows:

$$f_i(h, r, t) = SUM(\mathbf{h}_i \circ \mathbf{r}_i \cdot \mathbf{t}_i), \tag{12}$$

where $\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i$ are the visual embeddings of the entities and the relations, $\circ$ is the group operation, i.e., Hadamard product. However, there is no image for a given relation in KG. To solve the issue, we simply use an identity vector, $\mathbf{1}$, as the embedding of a relation. Base on this, we can define the overall energy function as follows:

$$f_o = f_m + \beta f_i, \tag{13}$$

where β is the hyperparameter. Finally, the overall loss is:

$$L = \sum_{(h,r,t)\in\mathbb{C}} \sum_{(h',r,t')\in\mathbb{C}'} \left[\Omega + f_o(h, r, t) - f_o\left(h', r, t'\right)\right]_+, \tag{14}$$

where $\mathbb{C}$ is the set of training triples. $\mathbb{C}'$ is the negative sampling set of $\mathbb{C}$ and Ω is the slack variable. $\mathbb{C}'$ is constructed by randomly replace the head, tail entity, or the relation.

4 EXPERIMENTS

In this section, we evaluate the performance of the proposed RSME model on real-world benchmark datasets, as well as sensitivity analysis on related parameters and settings to understand the effect of visual context in multi-modal KG embedding.

4.1 Datasets

In this paper, we conduct experiments on two publicly available multimedia retrieval datasets, which are widely used for performance evaluation of multi-modal KG embedding models.

WN18-IMG: WN18 [4] is a well-known KG which is originally extracted from WordNet [21]. WN18-IMG is an extended dataset of WN18 [4] which prepares 10 images for each entity.

FB15K-IMG: FB15K [4] is a widely used dataset in KG embedding link prediction tasks. FB15K-IMG is an extended dataset of FB15K, which prepares 20 images for each entity. We rebuild FB15K-IMG using the script provided by MMKG [20].

WN18-IMG-S: WN18-IMG-S is a subset of WN18-IMG, which only contains the triples with one-to-one correspondence entities.

FB15K-IMG-S: FB15K-IMG-S is a subset of FB15K-IMG, which only contains the triples with one-to-one correspondence entities.

FBX% : FBX is a subset of FB15K-IMG. We create it by randomly selecting 10% of the triples. The same as FB40%, FB60%, FB80%.

4.2 Link Prediction and Training Data Masking

The task of link prediction aims to complete a triple when one of (h, r, t) is missing based on minimizing the loss function. The following measures are used as our evaluation metrics: (1) MR: mean rank of correct entities; (2) Hit@k: proportion of valid entities ranked in top 1, 3, and 10 respectively.

Experimental Setup: We conduct link prediction experiments as other KG embedding models. During the training process, we first generate a corrupted triple for each triple in the test dataset by randomly replacing the head or tail entity of the triple, denoted as $(h^{'}, r, t)$ or $(h, r, t^{'})$. The embedding of entities and relationships is then obtained by minimizing the loss function of L in Equation (14). For the KG triples in the test dataset, we replace the tail entities with all entities separately and then sort each triple in an ascending order by f_o to analyze the mean rank and the Hit@k. A small mean rank or a big Hit@k indicates a good result. RSME(No Image) refers to representations based on structural information only. RSME(VIT) is

Table 2: Results of the link prediction.

Models	FB15K-IMG				WN18-IMG			
	MR	Hits@1	Hits@3	Hits@10	MR	Hits@1	Hits@3	Hits@10
TransE [4]	-	0.247	0.534	0.688	-	0.040	0.745	0.923
DistMult [35]	-	0.218	0.404	0.582	-	0.335	0.876	0.940
ComplEx [29]	-	0.599	0.759	0.840	-	0.936	0.945	0.947
RotatE [28]	43	0.750	0.829	0.884	254	0.942	0.950	**0.957**
TorusE [8]	-	0.674	0.771	0.832	-	0.943	0.950	0.954
TransAE [32]	53	-	-	0.645	-	-	-	-
RSME(No Img)	37.18	0.724	0.824	0.885	555	**0.945**	0.950	0.956
RSME(VIT)	35.76	0.794	0. 867	0.908	514	0.943	0.950	0.954
RSME(VIT+Forget)	**25.48**	**0.802**	**0.881**	**0.924**	**223**	0.943	**0.951**	**0.957**

calculated by structural information and visual information, which is encoded by the vision transformer. RSME(VIT+Forget) means a forget gate is applied to RSME(VIT).

Hyperparameter Setting: RSME is optimized by Adam. We conduct a grid search to find suitable hyperparameters.

Result: Table 2 provides the experimental results on the link prediction. It can be seen that the performance of RSME outperforms all the other models. The difference between the RSME(VIT) and RSME(No Img) groups is significant, which shows the introduction of visual context does help. In addition, comparisons between RSME(VIT) and RSME(VIT+Forget) also imply that forget gate does have a further improvement in most cases. However, we find that some structure-based embedding models based also have a good performance, such as RotatE. We think this may be because the dataset already has very rich structural information and does not need to rely on too much visual information. Therefore, we can not directly conclude that visual information played a key role in the improvement of RSME. Because the model is also affected by hyperparameters, variable initialization, and other factors. In order to further analyze the influence of image information, we mask a part of the train data and experiment with RSME.

We believe that the current dataset FB15K contains enough structural information to make predictions, which interferes with the analysis of visual information. In order to highlight the role of visual information, we occlude a part of the training data and create the data set, FB1%, FB20%, FB40%, FB60%, FB80%. Then, we perform the link prediction experiment again.

RSME(No Image) VS RSME(VIT+Forget): In order to further explore the role of image information, we compare the experimental results of link prediction between RSME and RSME (VIT+Forget) on FBX% with respect to different sizes. Results are shown in Figure 4. Surprisingly, the introduction of visual information is not always beneficial on FBX%. We find that the visual information helps when the structural training dataset is small. With the increase of structural data, the benefits of image information gradually diminish until it disappears.

RSME(VIT+Forget) VS RSME(VIT+Random): In order to verify that the selectivity of visual information brings benefits, we design a random gate. The random gate randomly selects the same number of images as the forget gate does, and then combines them

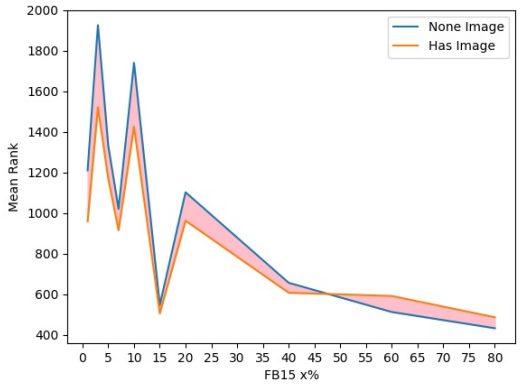

Figure 4: Link Prediction Results on FBX%.

with the structural information. The comparison results on FBX% are shown in Figure 5. The overall MR of RSME(VIT+Forget) is lower than RSME(VIT+Random), which validates that it is necessary to make a targeted selection of image information before fusing with structural information. In addition, on FB80%, RSME(VIT+Forget) gets a worse result than RSME(VIT+Random). We think this is due to the interference of structural information, as mentioned in the previous section.

4.3 Sensitivity of Relation Circumstance

In this subsection, we mainly discuss when it is advantageous to introduce visual information in KG embedding methods and mainly discuss the sensitivity of different relations to the proposed model. We first observe that visual information is not beneficial for all relations. Then, we analyze the results of each relation separately on the basis of link prediction.

Experimental Setup: Following the way described in Section 4.2, we first compare RSME's prediction results for different relationships on WN18-IMG and FB15K-IMG, then calculate the MRP on the datasets WN18-IMG-S and FB15K-IMG-S. The vision transformer is chosen as the image encoder. The weights of the vision

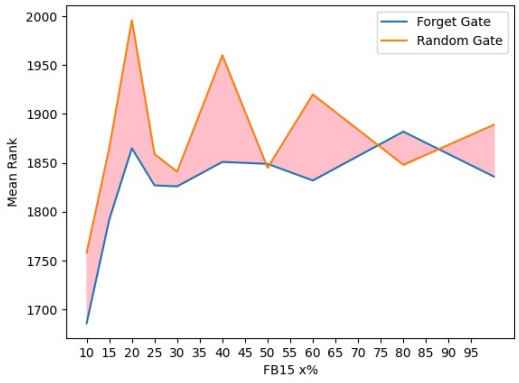

Figure 5: Forget Gate vs Random Gate on FBX%.

Table 3: Statistics of mean rank of different relations.

Relations	RSME(No Image)	RSME(VIT)
/medicine/drug/active_moieties	430.04	1
/fantasy_football/player/nfl_team	21.01	1
/dated_money_value/currency	1.27	4749.67
/tv/tv_program/country_of_origin	44.14	12740.75

transformer are pre-trained. And each image is uniformly resized to 384x384. Each entity only uses one image selected by the filter gate to participate in the calculation.

Result: Table 3 shows that image information is not always conducive to the embedding of all the relationships. Some relations with image information, such as $\xrightarrow{\text{/dated\_money\_value/currency}}$, have worse results than those without image information. This validates our assumption that the noise in the visual information may have an adverse effect on the model.

Table 4 presents the results of all the 10 relationships in WN18-IMG-S and 10 relationships sampled from the FB15K-IMG-S. The average MRP is 0.104 for WN18-IMG-S and the average MRP is 0.386 for FB15K-IMG-S. This result is significant at the MRP <0.5 level, which means the visual information can indeed help establish the correct connection between the head and tail entities. The data in Table 4 also shows that the MRP in WN18-IMG-S reported significantly better than that in the FB15K-IMG-S. We think this is because many of the images in WN18-IMG-S are from ImageNet, where the vision transformer is pre-trained. The same training dataset obviously makes sense to improve the performance of the model. Interestingly, the average MRP of FB15K-IMG-S is observed at 0.386 although the images of the entities in FB15K-IMG-S are not in ImageNet. This shows that vision transformer does have unexpected effectiveness as an image metric. In addition, we notice that FB15K-IMG-S still has a lot of relations with MRP greater than and equal to 0.5. It indicates that the visual information of these entities does not bring negative effects. The model should forget this type of visual information.

Table 4: The MRP on different relations.

WN18-IMG-S		FB15K-IMG-S	
relationship	MRP	relationship	MRP
_hyponym	0.094	active_moieties	0.000
_hypernym	0.132	tennis_winner	0.000
_has_part	0.100	dog_breed/color	0.158
_part_of	0.117	river/mouth	0.280
_member_holonym	0.048	cause_of_death	0.557
_synset_domain_topic_of	0.146	religion	0.593
_derivationally_related_form	0.138	category	0.605
_member_of_domain_topic	0.084	judge	0.805
_member_meronym	0.076	country_of_origin	0.852
average MRP	0.104	average MRP	0.386

Table 5: Effective rate of images information.

#Triples of Relation	Image Effective Rate
0-50	0.715
50-100	0.794
100-500	0.802
500-1000	0.839
1000-2000	0.906

From Table 4, we also find that different relations have different MRPs, which indicates that different relations have different sensitivities to visual information. There are some relations that can make good use of visual information, such as $\xrightarrow{\text{active\_moietie}}$ and $\xrightarrow{\text{river/mouth}}$. By analyzing the reasons, we find that $\xrightarrow{\text{active\_moietie}}$ is a reflexive relation, and most entities on $\xrightarrow{\text{river/mouth}}$ have visually checkable images. As a contrast, the accuracy of abstract relations such as $\xrightarrow{\text{judge}}$ is relatively low.

Moreover, we propose a new metric, the image efficient rate, which refers to the proportion of the number of relations whose MRP is less than 0.5 relative to the number of all relations. We segment the relations according to the number of triples contained in the relations and then calculate the image efficient rate of each segment. The results are shown in Table 5. We find that as the number of triples contained in the relationship increases, the image efficient rate also gradually increases. It may lead to the conclusion that the more complex the data is, the more benefit from the visual information.

4.4 Sensitivity of Image Encoder

In this subsection, we mainly discuss the impact of image encoders used in our model. We analyze the RSME's performance with different types of image encoders in the link prediction experiment. As a contrast, most of the current embedding models only use CNN

Table 6: The comparisons of image encoders.

Models		Mean Rank Rate	
		WN18-IMG-S	FB15K-IMG-S
pHash	phash_32	0.486	0.483
	phash_128	0.509	0.495
CNN	VGG16	0.166	0.429
	ResNet50	0.171	0.423
Vision Transformer	VIT	0.104	0.386

(e.g., VGG16 [26] and AlexNet [14]) to extract the visual features of the images, and ignore the analysis of the image encoder's role in embedding models.

Experimental Setup: We try three kinds of image encoders in this paper. The first one is CNN, the second one is the perceptual hash algorithm, and the third one is the vision transformer [7]. We choose VGG16 and Resnet50 for CNN based experiments. This is because VGG16 is the currently most widely used image encoder in multi-modal KG embedding models. As for Resnet50, it is the SOTA CNN model in recent years. In summary, we use 5 different image encoders to do the link prediction experiments on WN18-IMG-S and FB15K-IMG-S. We use the average MRP as the evaluation metric. The smaller the average MRP, the better the performance of the image encoder.

Result: The results are shown in Table 6. We find that the perceptual hash algorithm can not capture the characteristics of the images no matter in WN18-IMG-S or FB15K-IMG-S, even if the length of the perceptual hash algorithm is increased to 128. This indicates that the traditional Hamming distance-based metric is not suitable as the image encoders in multi-modal KG embedding. As for the CNN based models, we find that Resnet50 outperforms VGG16 on both datasets, which indicates that a good CNN can extract better visual features. Surprisingly, we find that the vision transformer significantly outperforms all other image encoders. This is due to the better global perception of vision transformer. The huge performance gap between image encoders suggests that the performance of the multi-modal KG embedding model does not only depend on the model itself, but also is sensitive to the used image encoders.

5 RELATED WORK

5.1 Unimodal KG Embedding Models

Translation-based KG embedding models are known for their simplicity and efficiency. TransE [4] is the first translation based model with the engine function satisfying $\|\mathbf{h}+\mathbf{r}-\mathbf{t}\| \approx \mathbf{0}$. Despite its simplicity and efficiency, TransE has flaws in dealing with 1-to-N, N-to-1, and N-to-N relations. TransH [33] and TransR [18] overcome the flaw by allowing entity to have relation-based embeddings. TransH projects the entity to a hyperplane of relation r. TransR projects the entity to a characteristic space by a relation-based matrix. TransD [15] simplifies TransR by decomposing the projection matrix into a unit matrix plus a product of two vectors, which reduces the number of parameters. TranSparse [16] simplifies TransR through a sparse matrix. TorusE [8] defines the translations on a compact Lie group. SOTA method RotatE [28] proposes a rotational model taking translation as a rotation in complex space. RESCAL [24] is the first bilinear model. RESCAL represents KG as a three-way tensor. The three ways are the entities and the relationship. DistMult [35] simplifies RESCAL by restricting M_r to diagonal matrices. However, the DistMult algorithm cannot handle asymmetric relations, HolE [23] combines the expressive power of RESCAL with the efficiency and simplicity of DistMult by using the circular correlation operation. ComplEx [29] entends HolE to the complex space so as to better model asymmetric relations. However, both translation-based models and bilinear models only focus on the structural information between triples and ignored the rich visual context in multi-modal KGs.

5.2 Multi-modal KG Embedding Models

To encode image features in KG embeddings, IKRL [34] learn visual information and structural information separately based on TransE [4]. On this basis, Mousselly et al [22] and TransAE model [32] jointly learn the visual and structural features into unified knowledge embeddings. Mousselly et al used three different methods, i.e., simple concatenation, DeViSE [10], and Imagined [6] to integrate multi-modal information, and TransAE utilized an auto-encoder to fuse them. Although existing multi-modal KG representation learning methods have shown promising performance, they believe that the learned embeddings are expected to be better since the visual modality intuitively contributes to rich information in contents. Despite of their achieved extent success, but it is not always clear to what extent truly visual context is required for KG embedding. Images may also introduce noise and lead to uncertainty about whether the visual context really improves embedding quality.

6 CONCLUSION

In this work, we attempt to probe the utility of the rich visual context in KG representation learning. We find that image resources in KG indeed help to improve the quality of learned KG embeddings. We also argue that visual information is not always useful. To validate our hypothesis, a relation sensitive multi-modal embedding model, i.e., RSME, is proposed to automatically encourage or filter the influence of additional visual context during the representation learning. We conduct extensive link prediction experiments on our approach compared with the state-of-the-art methods to show that leveraging the visual input to generating better KG embeddings is possible under appropriate circumstances. We also probe the effect of different visual feature encoders used in the proposed model. The experimental results validate the importance of visual feature encoder settings. We hope the results presented here are helpful for future multi-modal KG research work.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China with Grant Nos. 61906037 and 62072099; the Fundamental Research Funds for the Central Universities with Grant No. 4309002159 and 2242021k10011.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia-A crystallization point for the Web of Data. *Journal of web semantics* 7, 3 (2009), 154–165.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *2008 ACM SIGMOD International Conference on Management of Data*. 1247–1250.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*. 1–9.

[5] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 134–147.

[6] Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *31st AAAI Conference on Artificial Intelligence*. 4378–4384.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Takuma Ebisu and Ryutaro Ichise. 2018. Toruse: Knowledge graph embedding on a lie group. In *32rd the AAAI Conference on Artificial Intelligence*. 1819–1826.

[9] Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *2018 Conference on Empirical Methods in Natural Language Processing*. 2974–2978.

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. *Advances in Neural Information Processing Systems* (2013), 2121–2129.

[11] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[13] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *12th ACM International Conference on Web Search and Data Mining*. 105–113.

[14] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016).

[15] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 687–696.

[16] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2016. Knowledge graph completion with adaptive sparse transfer matrix. In *30th AAAI Conference on Artificial Intelligence*. 985–991.

[17] Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1227–1235.

[18] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *29th AAAI Conference on Artificial Intelligence*. 2181–2187.

[19] Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual Pivoting for (Unsupervised) Entity Alignment. In *35th AAAI Conference on Artificial Intelligence*. 4257–4266.

[20] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *European Semantic Web Conference*. Springer, 459–474.

[21] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[22] Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *7th Joint Conference on Lexical and Computational Semantics*. 225–234.

[23] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *30th AAAI Conference on Artificial Intelligence*. 1955–1961.

[24] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A Three-Way Model for Collective Learning on Multi-Relational Data.. In *International Conference on Machine Learning*. 809–816.

[25] Xia-mu Niu and Yu-hua Jiao. 2008. An overview of perceptual hashing. *Acta Electronica Sinica* 36, 7 (2008), 1405–1411.

[26] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[27] 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 *29th ACM International Conference on Information and Knowledge Management*. 1405–1414.

[28] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*. 1–18.

[29] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*. 2071–2080.

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

[31] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.

[32] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks*. IEEE, 1–8.

[33] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *28th AAAI Conference on Artificial Intelligence*. 1112–1119.

[34] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *26th International Joint Conference on Artificial Intelligence*. 3140–3146.

[35] 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 *International Conference on Learning Representations*.