Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion

Derong Xu
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
derongxu@mail.ustc.edu.cn

Tong Xu*
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
tongxu@ustc.edu.cn

Shiwei Wu
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
dwustc@mail.ustc.edu.cn

Jingbo Zhou
Business Intelligence Lab,
Baidu Research
zhoujingbo@baidu.com

Enhong Chen
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
cheneh@ustc.edu.cn

ABSTRACT

Knowledge Graph Completion (KGC), aiming to infer the missing part of Knowledge Graphs (KGs), has long been treated as a crucial task to support downstream applications of KGs, especially for the multimodal KGs (MKGs) which suffer the incomplete relations due to the insufficient accumulation of multimodal corpus. Though a few research attentions have been paid to the completion task of MKGs, there is still a lack of specially designed negative sampling strategies tailored to MKGs. Meanwhile, though effective negative sampling strategies have been widely regarded as a crucial solution for KGC to alleviate the vanishing gradient problem, we realize that, there is a unique challenge for negative sampling in MKGs about how to model the effect of KG relations during learning the complementary semantics among multiple modalities as an extra context. In this case, traditional negative sampling techniques which only consider the structural knowledge may fail to deal with the multimodal KGC task. To that end, in this paper, we propose a MultiModal Relation-enhanced Negative Sampling (MMRNS) framework for multimodal KGC task. Especially, we design a novel knowledge-guided cross-modal attention (KCA) mechanism, which provides bi-directional attention for visual & textual features via integrating relation embedding. Then, an effective contrastive semantic sampler is devised after consolidating the KCA mechanism with contrastive learning. In this way, a more similar representation of semantic features between positive samples, as well as a more diverse representation between negative samples under different relations could be learned. Afterwards, a masked gumbel-softmax optimization mechanism is utilized for solving the non-differentiability of sampling process, which provides effective parameter optimization compared with traditional sample strategies. Extensive experiments on three multimodal KGs demonstrate that our MMRNS framework could significantly outperform the state-of-the-art baseline methods, which validates the effectiveness of relation guides in multimodal KGC task.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Information systems** → *Multimedia information systems.*

KEYWORDS

Multi-modal, Knowledge graph completion, Negative sampling

ACM Reference Format:
Derong Xu, Tong Xu*, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548388

1 INTRODUCTION

Recent years have witnessed the booming of multimodal knowledge graphs (KGs). Multimodal KGs usually extend traditional KGs by supplementing multimodal data, like visual and audio attributes, to provide a physical world meaning to the symbols of traditional KGs [7, 15, 38, 49]. Along this line, various downstream applications, e.g., multimodal NER [26], visual question answering [16] and recommender system [31, 36] are widely supported. Unfortunately, due to the insufficient accumulation of multimodal corpus, existing multimodal KGs may suffer even more severe incompleteness compared with traditional KGs, which fatally impairs their usability and effectiveness. In this case, knowledge graph completion (KGC) solutions to multimodal scenario, which targets at automatically inferring missing facts, have attracted wide attention. Specifically, previous KGC methods mainly try to construct negative samples by uniform sampling, which suffer from the vanishing gradient problem in the later phase of training [40, 48]. Therefore, a special-designed negative sampling strategy tailored to multimodal KGs are urgently required.

[*] Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548388

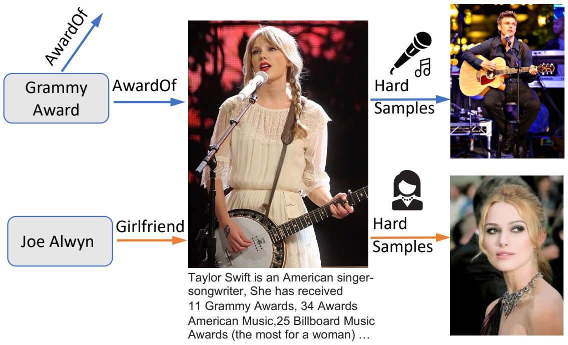

AwardOf
Grammy Award
AwardOf
Hard Samples
Joe Alwyn
Girlfriend
Hard Samples
Taylor Swift is an American singer-songwriter, She has received 11 Grammy Awards, 34 Awards American Music,25 Billboard Music Awards (the most for a woman) …

Figure 1: Examples of Hard Samples with Different Relations.

Recently, some efforts have been made for identifying hard negative sampling of traditional KGs [20]. For instance, IGAN [40] and KBGAN [5] both introduced a generative adversarial network to find high-quality negative samples, while NScaching [48] utilized extra memory to cache negative samples with a large score of KGC model. Besides, SANS [1] captured negative samples following the assumption that neighborhoods of positive entity are more likely hard samples. Though these techniques have achieved competitive performance on traditional KGs, however, they may fail to be applied to multimodal KGs, since current techniques mainly focus on the structural knowledge, while the rich multimodal cues are not fully utilized, which severely degrades the effectiveness.

Moreover, when jointly learning the multimodal attributes, relations in KGs may play an important role, as they could guide the learning of complementary semantics among multiple modalities as an extra context. Let's illustrate the impact of KG relations with a toy example shown in Figure 1. Usually, following the basic rules, negative samples having similar attributes and semantic information with positive samples are more likely to be hard samples. Thus, when selecting the hard negative examples for the entity *Taylor Swift*, given the attributes as *female* and *singer*, we should attempt to highlight features that reflect all of these attributes in their visual and textual information. However, as we all know, different attributes should be highlighted with considering different relations. For instance, with regard to the relation *AwardOf*, we expect to capture the multimodal cues which are mainly related to *singers* and *music*, e.g., a singer (no matter male or female) who is playing guitar in a concert. Correspondingly, for the relation *Girlfriend*, examples with *female* attribute may be a better choice. In this case, a more comprehensive solution which summarizes the multimodal cues deeply coupled with KG relations is required.

To address this problem, we propose a novel Knowledge-guided Cross-modal Attention (KCA) mechanism, which integrates multiple relations of the same entity to estimate the bi-directional attention weights of multimodal semantic features. Specifically, two parts are designed, in which one part summarizes the multimodal cues via mutual attention for relation-irrelevant features, while another part jointly reasons the multimodal attention in a bi-directional manner with embedding relations for relation-guided features, e.g., *singer, music* and related visual factors under the relation *AwardOf*. Moreover, the widely-seen *1-to-Many* relations in KGs, e.g., the relation *AwardOf* may connect *Grammy Award* and quite a lot of famous singers as winners of this award, naturally produces some positive triples in KGs, i.e., two similar entities could be both positive. This phenomenon motivates us to capture the similarity of semantic features between positive samples, as well as the diversity between negative samples under *1-to-Many* relations. Thus, based on the KCA mechanism, we introduce contrastive loss to build a contrastive semantic sampler, which aims to further learn multimodal semantic similarity/difference representation between positive and negative samples to estimate the sampling distribution.

Along this line, in this paper, we design a MultiModal Relation-enhanced Negative Sampling (MMRNS) framework to figure out hard negative samples by jointly utilizing the multimodal data and complex KG relations to enhance the semantic representation of entities, with the KCA mechanism enhanced by the contrastive semantic sampler. Afterwards, considering that non-differentiable sampling process may lead to the difficulty of refining sampling network parameters end-to-end via the optimization of KGC model, we further adapt the masked gumbel-softmax tool to achieve a differentiable solution for the sampling network. To be specific, we integrate mask operation on the basis of gumbel-softmax [17] to ensure that positive samples can be filtered out during forward propagation, and gradient can be returned during backward propagation. In addition, a variable factor varying with the number of iterations is utilized to dynamically tackle the exploration-exploitation tradeoff in early and later training phases. Technical contribution of this paper could be summarized as follows:

- To the best of our knowledge, we are the first to discuss the negative sampling strategy issue for the KGC task of multimodal KGs.
- A novel knowledge-guided attention mechanism is proposed, enhanced by a contrastive semantic sampler, to carry out cross-modal semantic learning under the guidance of complex KG relations.
- A masked gumbel-softmax tool is adapted to achieve back-propagation of gradient for optimizing the network parameters by KGC model loss.
- Extensive evaluations have proved the effectiveness and robustness of our negative sampling method by summarizing multimodal cues and revealing complex relations.

2 RELATED WORK

In this section, prior arts of KGC task, as well as related efforts on negative sampling strategy, will be summarized below.

2.1 Knowledge Graph Completion

Knowledge Graph Completion (KGC), aiming to predict the missing part of Knowledge Graphs, has been extensively studied in recent years [41]. Traditionally, distance-based models such as TransE [4], TransD [18], and TransR [23] have been proposed to model relation as a distance between head and tail entity for learning entity and relation embeddings. Another group of techniques is Semantic matching models including RESCAL [28], DistMult [46], ComplEx [37] and so on. There are also several recent state-of-the-art models such as GC-OTE [35] and PairRE [6]. However, most prior arts focus on designing a better scoring function for KGC, yet ignored the importance of negative sampling strategy, which may limit the further improvement in performance of these methods.

2.2 Negative sampling for KGC

Rather than applying uniform distribution sampling like most KGC models, several effective sampling strategies have been proposed in recent years [22, 32]. For instance, TransH [43] defined a Bernoulli distribution to replace head or tail by considering the complex relations such as 1-to-Many, but it is still a fixed sampling distribution, thus lacking flexibility. Also, IGAN [40] and KBGAN [5] both introduced a generative adversarial network(GAN) to get high-quality negative samples, in which the generator produces sampling distribution and the discriminator produces reward to optimize the generator by policy gradient [34]. However, such GAN-based methods are harder to train [13]. NScaching [48] proposed an efficient sampling scheme, which uses extra memory to cache negative samples with large scores, and samples the negative triples from it. Besides, SANS [1] considered using structural knowledge in KGs, which treated a subset of entities restricted to the entity's k-hop neighborhoods as hard samples. In summary, the previous approaches have demonstrated their effectiveness. The core points to find the hard samples are either using the structural knowledge of KGs or trying to use the negative sample scores. However, they still suffer from two problems: 1) The models trained with structural knowledge can only provide limited negative score information due to the incompleteness of KGs; 2) A more effective parameter optimization strategy is required to utilize the negative score of KGC model.

2.3 Multimodal Knowledge Graph

In recent years, increasing efforts have been made on the MKGs related tasks [8, 9, 27, 29, 47, 49]. For example, IKRL [44] and RSME [39] tried to combine image with structured knowledge for KG embedding. Also, TransAE [42] integrated visual and textual information by extending TransE to a multimodal score function. Besides, KBLRN [11] learned knowledge base representations from latent, relational, and numerical features. Though they all get competitive performance, we observe that there is still a lack of special-designed negative sampling strategies tailored to MKGs. In this paper, for identifying hard negative samples, we proposed a novel knowledge-guided cross-modal attention and construct a contrastive semantic sampler to enhance the semantic representation of multimodal entities with the guide of relation. At the same time, a new optimization strategy is adapted to effectively update the parameters of the multimodal sampling network.

3 METHODOLOGY

3.1 Preliminaries & Problem Definition

Given a knowledge graph $\mathcal{G} = \{(h, r, t)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, we denote $\mathcal{E}$ by entity set and $\mathcal{R}$ by relation set. Each triple in KG is represented as (h, r, t), which means head entity $h \in \mathcal{E}$ and tail entity $t \in \mathcal{E}$ are connected by a directed relation $r \in \mathcal{R}$. Also, we represent the embedding of tail entity and relation by $\boldsymbol{t} \in \mathbb{R}^{d_{emb}}$ and $\boldsymbol{r} \in \mathbb{R}^{d_{emb}}$, respectively. Besides, we denote $\boldsymbol{e_i} \in \mathbb{R}^{d_i \times d_N}$ as visual features and $\boldsymbol{e_t} \in \mathbb{R}^{d_t \times d_M}$ as textual features to describe the multimodal cues.

In this way, the KGC task can be modeled as a ranking problem, i.e., given a positive triple (h, r, t^+) and several negative triples (h, r, t^-), the KGC model aims to improve the score of positive triple and lower the score of negative triples with an effective scoring function. Along this line, the goal of our sampling strategy is to utilize the positive triple and the corresponding multimodal data to maximize the sampling probability of hard negative samples t^-, which are semantically similar to the positive one, to improve the discriminating ability of the model.

3.2 Knowledge-guided Cross-modal Attention

Then, we turn to introduce the detail of Knowledge-guided Cross-modal Attention (KCA) mechanism to learn cross-modal bi-directional attention weights by integrating multiple relations.

Specifically, KCA first tries to capture the interactions between different modalities, i.e., image and text, which aims to highlight the same semantic features between cross-modal data simultaneously to learn relation-irrelevant features. We denote relation-irrelevant features by cross-modal features that are all important under different relations to identify hard samples. For instance, in Figure 1, the negative samples of *Taylor Swift* are expected to be a person-related entity that contains more attributes related to the human body or face, rather than other irrelevant entities like locations, regardless of the relation being *AwardOf* or *Girlfriend*.

Meanwhile, KCA further integrates relational information after describing the multimodal interaction to guide the model in determining which multimodal semantic features should be highlighted to learn relation-guided features. For example, when the relation is *AwardOf*, KCA aims to enhance the cross-modal attention of attributes such as *singers* and *music*. When the relation is *Girlfriend*, KCA aims to enhance the cross-modal attention of attributes such as *female*. It is worth noting that the relation, as a kind of categorical data, contains limited and coarse-grained tag information and usually has no semantic similarities or correlations with image and text. Therefore, when introducing relation for guidance, we begin by modeling the interaction of textual and visual features, and then introduce relation embedding to guide the cross-modal attention weights of image and text, respectively.

Along this line, given the visual features $\boldsymbol{e_i}$ and textual features $\boldsymbol{e_t}$, they are first fed into a full connection network for nonlinear mapping as well as dimensional unification:

$$\hat{\boldsymbol{e}}_{\boldsymbol{i}} = R(\boldsymbol{e_i W_i} + \boldsymbol{b_i}) \in \mathbb{R}^{d_i \times d_{att}} \quad \hat{\boldsymbol{e}}_{\boldsymbol{t}} = R(\boldsymbol{e_t W_t} + \boldsymbol{b_t}) \in \mathbb{R}^{d_t \times d_{att}} \tag{1}$$

where $R(\cdot)$ is the active function LeakyRELU [45], $\boldsymbol{W}$ and $\boldsymbol{b}$ mentioned in this paper all represent the trainable weights and bias, respectively. The cross-modal matrix $\boldsymbol{M} \in \mathbb{R}^{d_i \times d_t}$ is calculated by

$$\boldsymbol{M} = \hat{\boldsymbol{e}}_{\boldsymbol{i}} \cdot \hat{\boldsymbol{e}}_{\boldsymbol{t}}^{T} \tag{2}$$

where $\boldsymbol{M}$ aims to capture and highlight the same semantic features among image and text. Here, the module is divided into four branches as shown in Figure 3: ① text-guided visual attention, ② relation-text-guided visual attention, ③ relation-image-guided textual attention, and ④ image-guided textual attention.

In the branch ①, KCA normalizes $\boldsymbol{M}$ to produce the attention weights on visual regions conditioned by each sentence of text. The attention weights are multiplied by image features $\hat{\boldsymbol{e}}_{\boldsymbol{i}}$ to derive attended relation-irrelevant visual representation $\boldsymbol{e}_{ir}^{i}$, which is

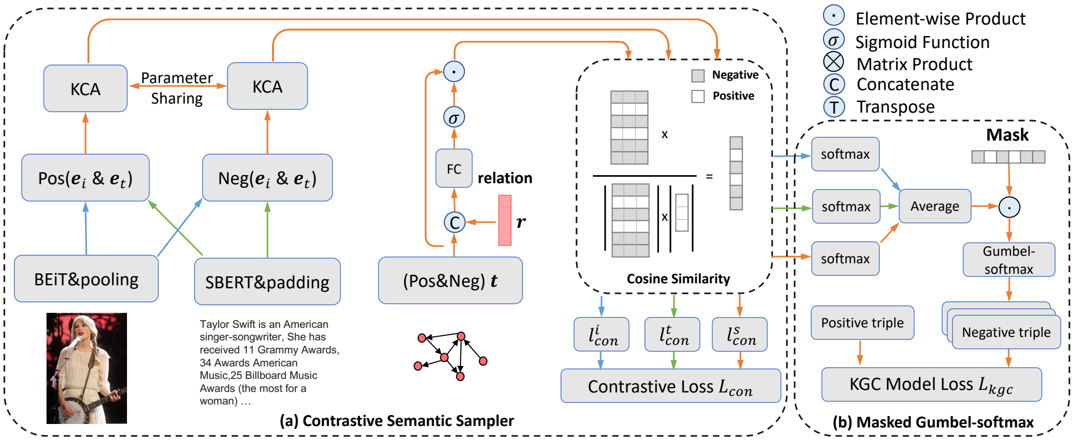


Figure 2: The Overall Framework of MMRNS.

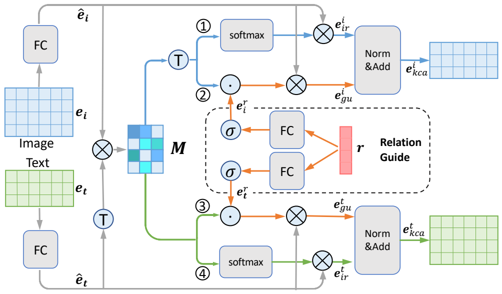


Figure 3: Knowledge-guided Cross-modal Attention (KCA).

general for any relation types:

$$e^i_{ir} = \hat{e}_i \cdot softmax(M^T, dim = -1) \tag{3}$$

In the branch ②, KCA aims to further integrate relation embedding to guided cross-modal semantic information. The difference between ② and ① is that ② employs KG relations to guide the normalized attention weight. Under this situation, the attention weights are also multiplied by image feature $\hat{e}_i$ to derive attended relation-guided visual representation e^i_{gu}:

$$e^i_{gu} = \hat{e}_i \cdot (M^T \odot e^r_i) \tag{4}$$

Here e^r_i and e^r_t are produced by feeding the embedding of relation r into two different full connection, which guides the bi-directional generation of visual and textual attention, respectively.

$$e^r_i = \sigma(r \cdot W^r_i + b^r_i) \quad e^r_t = \sigma(r \cdot W^r_t + b^r_t) \tag{5}$$

Correspondingly, the branches ③ and ④ attempt to learn attended textual representation guided by image and relation, whose motivations are similar to those of branches ② and ①. Both cross-modal and relation guided representations e^i_{ir} and e^i_{gu} are fed into a layer normalization [2] to unify the distribution, and then added to get knowledge-guided visual representation:

$$e^i_{kca} = Norm(e^i_{ir}) + Norm(e^i_{gu}) \tag{6}$$

Similarly, we produce the knowledge-guided textual representation e^t_{kca} as the same procedure.

3.3 Contrastive Semantic Sampler

Afterwards, we further construct a contrastive semantic sampler to calculate the sampling distribution of negative samples. The sampler firstly applies pretrained models to extractor semantic features and then uses KCA mechanism to model multimodal interaction with the guide of relation. The core point of our sampler is to further learn multimodal semantic representation by mining the similarities and differences between positive and negative samples.

3.3.1 ***Feature Preprocessing.*** We first extract preliminary visual features by BEiT [3], which can be used to learn semantic regions and object boundaries. We apply average pooling to the semantic visual representation to reduce computational complexity. We then extract the preliminary textual features by SBERT [30], which has a significant improvement on common semantic textual similarity tasks. Also, we apply cutting and padding to make the representation tensors with the same dimension. Since entities are also structural embedding as the relations, we just concatenate and feed them into a full connection network to integrate relation information.

$$e_s = t \cdot \sigma(concat(r, t) \cdot W_s + b_s) \tag{7}$$

3.3.2 ***Cosine Similarity.*** The preliminary features of image-text pairs for positive and negative samples are both fed into KCA respectively. The KCA for positive and negative samples shares the parameters. The visual features similarity between visual representation of two entities z_i and z_j is measured using cosine similarity. δ is a small number to ensure that the denominator is not zero.

$$sim^i(z_i, z_j) = \frac{{z_i}^T z_j}{\|z_i\|\|z_j\| + \delta} \tag{8}$$

3.3.3 ***Contrastive Loss.*** Finally, we build a contrastive loss function similar with [10], which uses similarity as the input, but only has one positive sample pair. However, we have multiple positive samples due to the 1-to-Many relations in KGs as mentioned above.

Therefore, the goal of this module is to narrow the gap between positive samples while widening the gap between positive and negative samples. In addition, We integrate the self-adversarial technique [32] in our framework to further improve the model performance. The loss weights $p(h_i, r, t_i)$ for i_{th} triple are calculated from the score of KGC model. We set the weight of those triples that are not sampled to $1/|\mathcal{E}|$:

$$p(h_i, r, t_i) = \begin{cases} \frac{exp(\alpha \cdot KGC(h_i, r, t_i))}{\sum\limits_{j \in S} exp(\alpha \cdot KGC(h_j, r, t_j))} & i \in S \\ 1/|\mathcal{E}| & otherwise \end{cases} \tag{9}$$

where S is the set of sampling triples, α is the temperature of sampling. In this way, the final contrastive loss function for visual features similarity is as follows:

$$l^i_{con} = -log \frac{\sum\limits_{j \in P} p(h_j, r, t_j) exp(\boldsymbol{sim}^i(\boldsymbol{z}, \boldsymbol{z}_j))}{\sum\limits_{n \in N} p(h_n, r, t_n) exp(\boldsymbol{sim}^i(\boldsymbol{z}, \boldsymbol{z}_n))} \tag{10}$$

where P is the positive sample set, N is the negative sample set. We simultaneously calculate the similarity for textual and structural features by Equation 8, as well as the contrastive loss by Equation 10, which are denoted by l^t_{con} and l^s_{con}, respectively. The total contrastive loss is obtained by averaging them:

$$L_{con} = (l^i_{con} + l^t_{con} + l^s_{con})/3 \tag{11}$$

3.4 Masked Gumbel-Softmax

In this section, we will explain how to use the proposed differentiable sampling approach, which integrates mask operation with gumbel-softmax [17], to ensure effective gradient back propagation. The mask operation aims to overcome the problem of introducing gumbel-softmax into the sampling process of KGC.

3.4.1 ***Gumbel-Softmax.*** Since the sampling process of categorical distribution is independent of the optimization process, the gradient of KGC model is unable to be back-propagated to the sampling network. Therefore, the trainable parameters of the contrastive semantic sampler can not be optimized in an end-to-end manner with the training phase of KGC model. To achieve gradient back propagation, we introduce the gumbel-softmax re-parameterization trick [17], which produces a continuous distribution that can approximate samples from a discrete probability distribution $\boldsymbol{p}$ by using the softmax function as a differentiable approximation to argmax: $\boldsymbol{y} = softmax((log(\boldsymbol{p}) + \boldsymbol{g})/\tau)$. Where, each element g_i in $\boldsymbol{g}$ is i.i.d samples drawn from standard Gumbel distribution [14, 25].

3.4.2 ***Masked vector.*** Considering that the semantic similarities of positive and negative samples in image, text and structure are individually used to compute probability distribution $\boldsymbol{p}$, we utilize softmax to transform the similarity into sampling probability:

$$\boldsymbol{p} = (SF(\boldsymbol{sim}^i/\epsilon) + SF(\boldsymbol{sim}^t/\epsilon) + SF(\boldsymbol{sim}^s/\epsilon))/3 \tag{12}$$

where $SF(\cdot)$ denotes softmax function, ϵ is a factor to balance exploration and exploitation and will be elaborated below. However, $\boldsymbol{p}$ is not the final sampling probability distribution. Since 1-to-Many relations are very common in KGs, not all the entities can be treated as negative samples. Most common methods will filter out positive samples [4, 32]. A common way to accomplish it is to set positive positions in the sampling probability distribution $\boldsymbol{p}$ to zero. But it will make gumbel-softmax non-differentiable, which contradicts our purpose. Therefore, we propose a non-differentiable masked vector, where the values of negative positions are set to 1.0, and the values of positive positions are set to a number very close to zero. The probability distribution $\boldsymbol{p}$ is element-wise multiplied by the masked vector. We noticed that multiplication can be replaced by addition due to the log function to reduce computational complexity. Following is the masked gumbel-softmax:

$$\boldsymbol{y_m} = SF((log(\boldsymbol{p}) + log(\boldsymbol{mask}) + \boldsymbol{g})/\tau) \tag{13}$$

here $\boldsymbol{y_m}$ is the sampling result. It is worth noting that the masked vector also benefits achieving sampling without replacement. The total loss L is obtained by adding loss of KGC model L_{kgc} and loss of sampler L_{con}. The effect of loss rate β is analysed in Section 4.5.

$$L = L_{kgc} + \beta L_{con} \tag{14}$$

3.4.3 ***Exploration and Exploitation.*** Here, considering to make the sampling strategy adaptive in different training periods, we further define an exploration and exploitation factor ϵ. The motivation is to learn both hard and simple samples in the early training phase, and pay more attention to the utilization of hard samples in the later training phase. The value of ϵ decreases as the number of iterations increases. The effect of ϵ_o will be discussed in detail in Section 4.5.

$$\epsilon = \epsilon_o/(1 + log(iter)) \tag{15}$$

4 EXPERIMENTS

In this section, we will conduct extensive experiments on three real-world datasets to validate our framework, and then reveal some interesting findings on the impacts of complex relations.

4.1 Multimodal datasets

All experiments are conducted on three multimodal KGs, including MMKB-DB15K [24] and our two self-constructed datasets, i.e. MKG-W and MKG-Y, as shown in Table 1. Specifically, MMKB-DB15K [24] is an open-source multimodal knowledge graph, whose structured knowledge is a subset of DBpedia [21], and images are crawled from image search engines. The search queries are entity name, entity notable type and Wikipedia URIs. Considering that each entity lacks textual description, we supplemented the textual information of entities from the database of DBpedia [21].

At the same time, our two self-constructed datasets, namely Multimodal KG-Wikipedia and Multimodal KG-YAGO, i.e. MKG-W and MKG-Y, are constructed based on [33], which extracted the structured knowledge from Wikipedia and YAGO. We further extended images of entities through web search engines, by asking human experts to manually screen out the appropriate images, and meanwhile ensuring that each entity has five to thirty images. For textual descriptions, we crawled them from DBpedia and then aligned them with corresponding entities through extra *sameAs* link provided by [33]. We randomly divided all KGs datasets by the ratio 8:1:1 for training, validation and testing sets respectively.

Dataset	Ent	Rel	Train	Valid	Test	Image	Text
MKG-W	15000	169	34196	4276	4274	14463	14123
MKG-Y	15000	28	21310	2665	2663	14244	12305
MMKB-DB15K	12842	279	79222	9902	9904	12818	9078

Table 1: Statistics of Datasets.

4.2 Baseline Methods

To evaluate the performance of our MMRNS framework, we compare our method with following strategies for negative sampling, including Uniform, Bernoulli [43], NSCaching [48], KBGAN [5], SANS [1], NS-KGE[22]. The details are described as follows:

- **Uniform**, which samples from a uniform distribution.
- **Bernoulli** [43], which defines a Bernoulli distribution to replace entities, whose parameters depend on the mapping property of the relation.
- **KBGAN** [5], which introduces generative adversarial networks to get high-quality negative samples.
- **Nscaching** [48], which utilizes extra memory to cache negative samples with large scores, and samples by weight.
- **SANS** [1], which assumes the entities that are mutually close are more likely to be related, and treats them as hard samples.
- **NS-KGE** [22], which considers all of the negative instances for training, to provide better efficiency.

The experiments are conducted on the sampling strategies based on several state-of-the-art KGC models, including TransE [4], DistMult [46], ComplEx [37], RotatE [32], PairRE [6] and GC-OTE [35]. The self-adversarial technique is integrated into RotatE, PairRE and GC-OTE. We evaluate all approaches with the following most-used metrics in KGC: mean rank (MR), mean reciprocal rank (MRR), and Hits@n (n = 1, 3, 10).

4.3 Experimental Settings

4.3.1 ***Preprocessing.*** We fed the visual data into BEiT [3]. We chose the pretained model with a base-sized architecture, patch resolution of 16x16, and fine-tuning resolution of 224x224 to get one randomly selected image representation with a size of 197x768. Then we apply a 2D average pooling with kernel size of (9, 3) and stride of (8, 2) over it. We fed the textual data into SBERT [30], and the used pretrained model is 'multi-qa-MiniLM-L6-cos-v1', which was tuned for semantic search. We cut and pad the sentences to ensure dimension consistency. Both hyperparameters for BEiT and SBERT are set by default values.

4.3.2 ***Implementation Details.*** We utilized grid search to discover the optimal hyperparameters for each approach on the valid set and evaluate them on the test set. Since the parameters set for KGC model and sampling strategy is too large to afford grid search simultaneously, we fix the critical KGC model hyperparameters and tune the sampling strategies hyperparameters. Following other stat-of-the-art sampling strategies and KGC model, we chose to fix the number of negative sampling to 20, embedding dim to 200 for all KGC models in three MKGs. In MKG-W and MKG-Y, the margin for TransE, RotatE and PairRE are fixed to 3.0, and the regularization of DistMult and ComplEx are fixed to $5 \times e^{-5}$. In MMKB-DB15K, the margin and regularization are fixed to 6.0 and $1 \times e^{-5}$, respectively. The temperature of self-adversarial is set to 0.5 in three MKGs. After fixing the hyperparameters of KGC models, the hyperparameters of all sampling strategies are searched based on their papers. We use Adam [19] to optimize our model. The trainable parameters of sampling network are initialized by Xavier normalization [12]. The learning rate of sampling network is searched from $\{1 \times e^{-4}, 5 \times e^{-5}\}$. The factor ϵ_o is searched from $\{1, 3, 5, 10\}$. The loss rate β is searched from $\{0.1, 0.01, 0.005, 0.003\}$. In the training process, rather than using all entities as the candidate negative samples to learn the hard samples, we randomly sample a subset of entities first. The size of pre-samples can be set to $|\mathcal{E}|/10$. Such simplified operation reduce the computational cost significantly while maintaining performance.

4.4 Experimental Results

The overall comparison is summarized in Table 2. We observe that our MMRNS strategy achieves the best performance in most of the evaluation metrics. Especially for the Hits@10, our approach gets the best results on all datasets and all KGC models. What's more, MMRNS obtains a significant improvement of 6.4%, 2.0%, and 7.5% on TransE, DistMult, and ComplEx on the MKG-W dataset over its best competitors, and also obtains an improvement of 2.6%, 6.5%, and 1.6% on TransE, DistMult, and ComplEx on the MKG-Y dataset.

The uniform sampling strategy randomly chooses negative samples at all phases of training. However, some negative samples are simple to learn, while others are more difficult. Under the ranking evaluation metrics of KGC task, sampling and training samples with a low rank are not very helpful to performance. Meanwhile, the number of negative samples can not be too large due to the class-imbalance problem. Thus uniform strategy performs worse than other strategies. We also observe that our approach is better than uniform sampling in almost all metrics, which indicates MMRNS is able to effectively figure out the hard negative samples. For example, MMRNS obtains an improvement of 6.5%, 23.8%, and 12.3% on TransE, DistMult, and ComplEx on the MKG-W dataset over uniform sampling. Although Bernoulli and SANS do not take advantage of negative sample scores learned by KGC models, they utilize the structural information of KGs, which provides prior knowledge for their negative sampling. Thus they get better performance than the fixed uniform sampling. However, the structural information is still limited. Our method takes into account the rich semantic information in multimodalities including text and image, thus achieving better performance. NS-KGE may have better computational efficiency and space efficiency, but its performance is lower than our model in most situation. RotatE, PairRE and GC-OTE both utilized the self-adversarial technique, and Table 2 demonstrates MMRNS get better results in most of metrics, which indicates our model can be well combined with this technique.

4.5 Analysis on Hyperparameters

Moreover, we empirically evaluate the influence of hyperparameters setting of KGC models and MMRNS.

4.5.1 ***Parameters of KGC models.*** In order to show the consistent improvement of the performance of our approach under

KGC Models	Strategies	MKG-W					MKG-Y					MMKB-DB15K				
		MRR↑	MR↓	H1↑	H3↑	H10↑	MRR↑	MR↓	H1↑	H3↑	H10↑	MRR↑	MR↓	H1↑	H3↑	H10↑
TransE[4]	Uniform	29.19	1067	21.06	33.20	44.23	30.73	1987	23.45	35.18	43.37	24.86	675	12.78	31.48	47.07
	Bernoulli	28.98	1184	22.99	31.48	40.11	31.12	1896	25.69	35.09	39.20	30.11	736	19.04	36.74	49.86
	KBGAN	24.21	1450	14.51	30.74	40.45	26.92	2220	19.90	32.02	37.69	24.00	729	5.47	**39.72**	50.36
	Nscaching	28.90	1106	22.24	32.22	40.97	29.91	1897	24.78	33.01	38.27	**33.03**	813	**23.31**	38.74	50.29
	SANS	30.22	**489**	23.64	32.78	42.64	27.51	**846**	21.65	30.46	38.44	26.33	**602**	13.87	33.65	48.80
	NS-KGE	28.50	911	19.77	32.80	44.27	31.58	1441	24.76	35.73	43.63	24.18	722	13.21	29.44	44.28
	Ours	**33.51**	926	**26.25**	**36.79**	**47.11**	**34.81**	1791	**28.59**	**38.71**	**44.76**	26.61	663	13.40	34.58	**50.60**
DistMult[46]	Uniform	20.99	1147	15.93	22.28	30.86	25.04	**2646**	19.33	27.80	35.95	23.03	**814**	14.78	26.28	39.59
	Bernoulli	25.76	1663	**21.02**	27.42	34.77	32.47	4069	29.63	33.81	37.10	**27.05**	1040	**20.04**	**30.13**	40.37
	KBGAN	19.11	2141	7.69	27.23	37.44	14.82	3507	0.51	27.41	34.55	23.36	985	16.43	25.34	36.30
	Nscaching	24.70	1631	20.34	26.01	32.74	32.55	4160	29.80	34.62	37.31	26.38	1182	19.33	29.63	39.74
	SANS	23.21	1334	18.46	24.41	32.20	27.52	3328	22.34	30.76	37.24	24.72	1257	16.82	28.23	39.76
	NS-KGE	25.67	1938	20.85	27.01	35.00	4.05	3867	**31.35**	35.58	38.62	23.62	1259	16.64	26.23	37.34
	Ours	**25.92**	**1084**	19.73	**27.99**	**38.20**	**32.78**	3481	27.99	**36.06**	**41.13**	22.80	864	14.06	26.32	**40.42**
ComplEx[37]	Uniform	24.93	1114	19.09	26.69	36.73	28.71	**2645**	22.26	32.12	40.93	27.48	**757**	18.37	31.57	45.37
	Bernoulli	27.14	1547	22.50	28.51	36.16	33.21	4167	**30.42**	35.00	38.11	**30.68**	1062	**24.20**	**33.50**	43.15
	KBGAN	20.78	2054	9.92	28.44	38.13	16.99	3094	3.09	28.93	36.17	19.53	1292	9.25	32.02	42.39
	Nscaching	27.12	1521	22.55	28.52	35.66	31.94	4107	29.03	33.91	37.01	30.17	1149	23.57	33.13	42.62
	SANS	26.73	1485	21.81	28.11	36.23	27.30	3818	22.45	30.35	36.27	28.94	1440	21.10	32.80	43.57
	NS-KGE	28.65	1405	23.40	30.58	38.37	**34.43**	3447	30.24	37.32	41.62	27.14	991	20.42	29.67	40.24
	Ours	**28.98**	**1112**	22.69	**31.32**	**41.26**	31.98	3245	25.63	35.29	**42.28**	27.25	943	17.25	31.93	**45.50**
RotatE[32]	Uniform	33.67	1169	26.80	36.68	46.73	34.95	2427	29.10	38.35	45.30	29.28	714	17.87	36.12	49.66
	SANS	33.32	**768**	27.35	35.66	44.67	35.28	**1404**	29.77	38.45	44.93	**30.51**	**641**	**19.13**	**37.19**	50.72
	Ours	**34.13**	1144	**27.37**	**37.48**	**46.82**	**35.93**	2504	**30.53**	**39.07**	**45.47**	29.67	708	17.89	36.66	**51.01**
PairRE[6]	Uniform	34.40	**823**	28.24	36.71	46.04	32.01	1821	25.53	35.84	43.89	31.12	**597**	21.62	35.91	49.30
	SANS	32.73	936	27.02	34.54	43.47	32.79	**1650**	26.71	36.37	43.54	31.16	621	21.48	36.23	49.32
	Ours	**35.03**	843	**28.59**	**37.49**	**47.47**	**34.96**	1870	**29.18**	**38.34**	**44.87**	**31.92**	614	**22.45**	**36.84**	**50.01**
GC-OTE[35]	Uniform	33.92	1057	26.55	35.96	46.05	32.95	**1938**	26.77	36.44	**44.08**	31.85	**620**	22.11	36.52	**51.18**
	Ours	**34.32**	**1006**	**27.14**	**36.82**	**47.01**	**33.38**	2381	**27.65**	**36.78**	43.44	**32.68**	641	**23.01**	**37.86**	50.96

Table 2: Results of five KGC models compared with other SOTA sampling strategies on three MKGs. H1, H3 and H10 mean Hits@1, Hits@3 and Hits@10, respectively. Smaller MR means the better result, other metrics are the larger the better. The best results are highlighted in bold, and the second results are highlighted with an underline. Bernoulli is implemented by the code of Nscaching [48], other methods are all implemented by their own open-source codes.

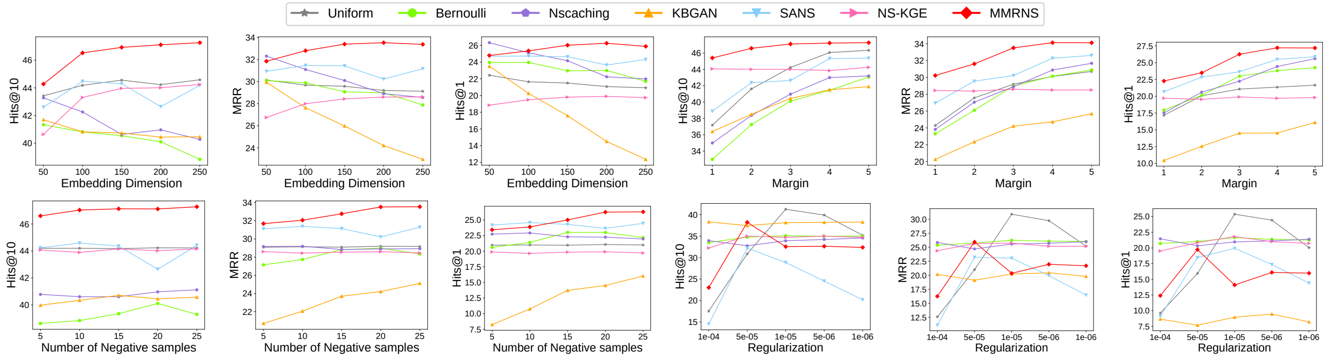


Figure 4: Parameter sensitivity analysis with varying values of KGC model. When one parameter is evaluated, other parameters are set by default values as described in Section 4.5.

different KGC models hyperparameters, we tuned the hyperparameters of TransE and DistMult to observe the change curve of performance on MKG-W dataset, as shown in Figure 4. We observed that MMRNS achieves the best performance under most KGC parameters and metrics, which demonstrates our approach is robust enough to adapt to different situations. We also realize that the performance of other baselines decreases when increasing the embedding dimension, while our approach can avoid the overfitting caused by high-dimensional embedding and achieve better performance.

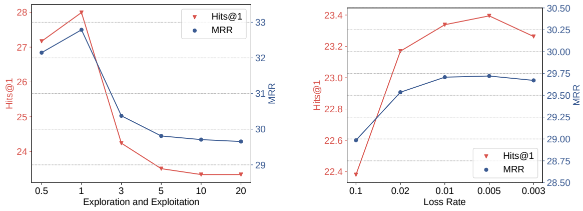


Figure 5: Parameter evaluation of MMRNS.

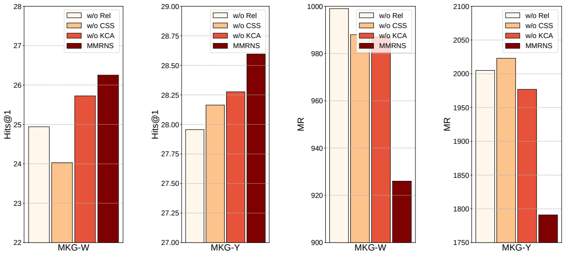


Figure 6: Ablation Study on MKG-W and MKG-Y with TransE.

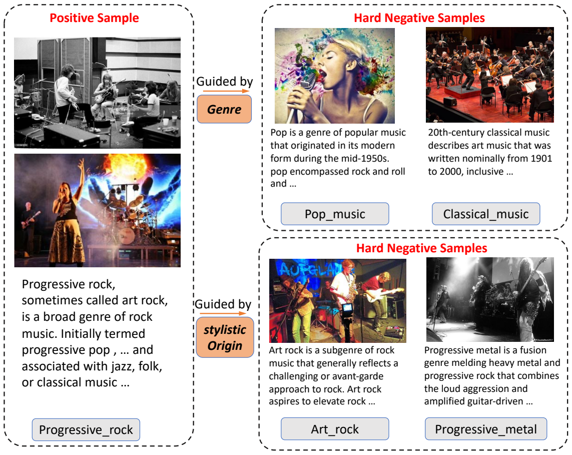
*4.5.2 **Parameters of MMRNS**.* We further evaluate the influence of parameters of MMRNS including exploration-exploitation factor ϵ and loss rate β as shown in Figure 5. The ϵ indicates the degree of utilization for sampling distribution given by the sampling network. We find that the best performance is obtained when ϵ_o equals 1, and the performance significantly decreases when ϵ_o dropped to 3, which demonstrates our sampling approach is helpful for achieving better performance. What's more, Higher values of ϵ tend to result in a smoother curve. The sampling distribution with a very large value of ϵ will approximate the uniform distribution. Loss rate β is used to adjust the effect ratio of loss L_{kgc} and loss L_{con} to the trainable parameters of the sampling network. We observed that the best performance is achieved when loss rate equals 0.005.

4.6 Ablation Study

To further demonstrate the effect of each proposed component, we conducted ablation study on MKG-W and MKG-Y based on TransE by designing different variants of MMRNS.

- **w/o KCA**: MMRNS without KCA, we directly calculate the cosine similarity of positive and negative image-text pairs after a full connection.
- **w/o CSS**: MMRNS without Contrastive Semantic Sampler, we simply cut off the optimization of contrastive loss on parameters of the sampler.
- **w/o Rel**: MMRNS without relation, we remove the part of the relation embedding integrated into each module of MMRNS.

According to the results shown in Figure 6, we observe a drop in the performance for both w/o KCA, w/o CSS and w/o Rel on MKG-W and MKG-Y datasets. These reductions indicate that complex relation information is an essential part and can not be ignored, and also demonstrate the effectiveness of our proposed modules to utilize multimodal data and integrate knowledge-guided relation embedding for finding hard samples.



Progressive rock, sometimes called art rock, is a broad genre of rock music. Initially termed progressive pop , ... and associated with jazz, folk, or classical music ...

Progressive_rock

Pop is a genre of popular music that originated in its modern form during the mid-1950s. pop encompassed rock and roll and ...

Pop_music

20th-century classical music describes art music that was written nominally from 1901 to 2000, inclusive ...

Classical_music

Art rock is a subgenre of rock music that generally reflects a challenging or avant-garde approach to rock. Art rock aspires to elevate rock ...

Art_rock

Progressive metal is a fusion genre melding heavy metal and progressive rock that combines the loud aggression and amplified guitar-driven ...

Progressive_metal

Figure 7: Case Study on MMKB-DB15K with TransE.

4.7 Case Study

Finally, we present a case study to show that our MMRNS does well in integrating multimodal cues and complex relations information to figure out hard negative samples. Specifically, we choose the intersection set among the top 10 negative samples score of KGC model, and the top 10 of negative sampling results. As shown in Figure 7, We can observe that MMRNS is able to summarize similar multimodal cues to find hard samples, such as *music* and *person* related semantic attributes. We also find the difference between negative samples under the guidance of relations. When the relation is *Genre*, negative samples are the same kind of genres like *Classical music* different from *rock music*. The highlighted multimodal semantic attributes are such as *music*. Yet, when the relation is *stylisticOrigin*, negative samples are more likely to be the same rock music like *Art rock*. The highlighted attributes are more related to *rock group* and *electric guitar*.

5 CONCLUSION

In this paper, we have proposed a MultiModal Relation-enhanced Negative Sampling framework to figure out hard negative samples for KGC. In order to address the unique challenge in negative sampling of MKGs, we first designed a novel knowledge-guided cross-modal attention, which combines multiple relations to guided cross-modal attention weights between semantic features of image and text. Then, a contrastive semantic sampler was built to learn the semantic similarity between positive samples, and the diversity between negative samples under multiple relations. In addition, we proposed a masked gumbel-softmax optimization technique, which enables the sampling process differentiable for training. We evaluated our approach on three data sets and six KGC models by comparing with several state-of-the-art negative sampling techniques. Extensive evaluations had demonstrated the effectiveness of MMRNS framework to identify hard negative samples.

6 ACKNOWLEDGMENTS

This work was supported by the grants from the National Natural Science Foundation of China (No.62072423).

REFERENCES

[1] Kian Ahrabian, Aarash Feizi, Yasmin Salehi, William L Hamilton, and Avishek Joey Bose. 2020. Structure Aware Negative Sampling in Knowledge Graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6093–6101.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[5] Liwei Cai and William Yang Wang. 2018. KBGAN: Adversarial Learning for Knowledge Graph Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1470–1480.

[6] Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. PairRE: Knowledge Graph Embeddings via Paired Relation Vectors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4360–4369.

[7] Liyi Chen, Zhi Li, Weidong He, Gong Cheng, Tong Xu, Nicholas Jing Yuan, and Enhong Chen. 2022. Entity Summarization via Exploiting Description Complementarity and Salience. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[8] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: entity alignment for multi-modal knowledge graph. In *Proc. of KSEM*.

[9] Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multi-modal Siamese Network for Entity Alignment. In *Proc. of KDD*.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[11] Alberto García-Durán and Mathias Niepert. 2018. KBlrn: End-to-End Learning of Knowledge Base Representations with Latent, Relational, and Numerical Features. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, Amir Globerson and Ricardo Silva (Eds.). 372–381.

[12] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[14] Emil Julius Gumbel. 1954. *Statistical theory of extreme values and some practical applications: a series of lectures*. Vol. 33. US Government Printing Office.

[15] Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2281–2290.

[16] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rkE3y85ee

[18] 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 (volume 1: Long papers)*. 687–696.

[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *The third International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[20] Bhushan Kotnis and Vivi Nastase. 2017. Analysis of the impact of negative sampling on link prediction in knowledge graphs. *arXiv preprint arXiv:1708.06816* (2017).

[21] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web* 6, 2 (2015), 167–195.

[22] Zelong Li, Jianchao Ji, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Chong Chen, and Yongfeng Zhang. 2021. Efficient non-sampling knowledge graph embedding. In *Proceedings of the Web Conference 2021*. 1727–1736.

[23] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

[24] 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.

[25] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).

[26] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2000–2008.

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

[28] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Icml*.

[29] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3208–3218.

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[31] 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 *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1405–1414.

[32] 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*.

[33] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2326–2340.

[34] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* 12 (1999).

[35] Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2713–2722.

[36] Shaohua Tao, Runhe Qiu, Yuan Ping, and Hui Ma. 2021. Multi-modal Knowledge-aware Reinforcement Learning Network for Explainable Recommendation. *Knowledge-Based Systems* 227 (2021), 107217.

[37] 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.

[38] Meng Wang, Guilin Qi, HaoFen Wang, and Qiushuo Zheng. 2019. Richpedia: A Comprehensive Multi-Modal Knowledge Graph. In *Semantic Technology: 9th Joint International Conference, JIST 2019, Hangzhou, China, November 25–27, 2019, Proceedings* (Hangzhou, China). Springer-Verlag, Berlin, Heidelberg, 130–145. https://doi.org/10.1007/978-3-030-41407-8_9

[39] 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 International Conference on Multimedia*. 2735–2743.

[40] Peifeng Wang, Shuangyin Li, and Rong Pan. 2018. Incorporating gan for negative sampling in knowledge representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[41] 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.

[42] 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.

[43] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

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

[45] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).

[46] 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 *The third International Conference on Learning Representations*.

[47] Yingying Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2020. Multimodal multi-relational feature aggregation network for medical knowledge representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3956–3965.

[48] Yongqi Zhang, Quanming Yao, Yingxia Shao, and Lei Chen. 2019. NSCaching: simple and efficient negative sampling for knowledge graph embedding. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 614–625.

[49] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *arXiv preprint arXiv:2202.05786* (2022).