

Modality-Aware Negative Sampling for Multi-modal Knowledge Graph Embedding

Yichi Zhang, Mingyang Chen, Wen Zhang[†]
 Zhejiang University, Hangzhou, China
 {zhangyichi2022, mingyangchen, zhang.wen}@zju.edu.cn

Abstract—Negative sampling (NS) is widely used in knowledge graph embedding (KGE), which aims to generate negative triples to make a positive-negative contrast during training. However, existing NS methods are unsuitable when multi-modal information is considered in KGE models. They are also inefficient due to their complex design. In this paper, we propose Modality-Aware Negative Sampling (MANS) for multi-modal knowledge graph embedding (MMKGE) to address the mentioned problems. MANS could align structural and visual embeddings for entities in KGs and learn meaningful embeddings to perform better in multi-modal KGE while keeping lightweight and efficient. Empirical results on two benchmarks demonstrate that MANS outperforms existing NS methods. Meanwhile, we make further explorations about MANS to confirm its effectiveness.

I. INTRODUCTION

Knowledge graphs (KGs) [1], [2] represent real-world knowledge in the form of triple (h, r, t) , which indicates the entity h and the entity t are connected by the relation r . Multi-modal KGs (MMKGs) are the KGs that consist of rich modal information such as images and text. Nowadays, KGs and MMKGs have been widely used in AI-related tasks like question answering [3], recommendation systems [4], language modeling [5] and telecom fault analysis [6].

Meanwhile, KGs as well as MMKGs are usually far from complete and comprehensive because many triples are unobserved, which restricts the application of KGs and makes knowledge graph completion (KGC) a significant task. Knowledge graph embedding (KGE) [7]–[10] is a popular and universal approach for KGC, which represents entities and relations of KGs in a continuous low-dimension vector space. In the usual paradigm, KGE models would design a score function to estimate the plausibility of triples with entity and relation embeddings. These embeddings are structural embeddings since they can encode information about triple structures. As for MMKGs, embedding-based methods can still work by utilizing multi-modal information. Nevertheless, existing multi-modal KGE (MMKGE) [11]–[13] methods design additional embeddings to represent the modal information, which would also participate in the score function.

Negative sampling (NS) [7] is a widely used technology for training KGE models, which aims to generate manual negative triples by randomly replacing entities for positive-negative contrast. NS would guide the KGE model to give higher

scores for the positive triples. An outstanding NS strategy would obviously improve the performance of KGE models to discriminate the triple plausibility.

Though existing NS methods [14]–[19] have tried different ways to obtain high-quality negative samples, they have one drawback that cannot be ignored: they are designed for general KGE models and **underperform in MMKGE**. As for MMKGE, entities may have multiple heterogeneous embeddings such as visual and structural embeddings. However, NS for the general KGE models will treat multiple embeddings of an entity as a whole and replace them together with embeddings of another entity, which we think is entity-level. Such design implicitly assumes that different embeddings of an entity have been aligned and model could distinguish the two embeddings of each entity, which weakens the model’s capability of aligning different embeddings and results in less semantic information being learned by the embeddings. Besides, we should also take the efficiency of the method into account while considering the multi-modal scenario, as those existing approaches design many complex modules (e.g. GAN [14], large-scale caches [15], manual rules [18], entity clustering [19]) to sample high-quality negative samples. We think they are over-designed and make the NS method computationally expensive.

To address the mentioned challenges, we propose Modality-Aware Negative Sampling (MANS for short) strategy for MMKGE. MANS is a lightweight but effective NS strategy designed for MMKGE. We first propose visual NS (MANS-V for short), a modal-level sampling strategy that would sample only negative visual features for contrast. We employ MANS-V to achieve modality alignment for multiple entity embeddings and guide the model to learn more semantic information from different perspectives by utilizing multi-modal information. We further extend MANS-V to three combined strategies, called two-stage, hybrid, and adaptive negative sampling respectively. All of the NS methods make up MANS together. Our Contribution could be summarized as follows:

- To the best of our knowledge, MANS is the first work focusing on the negative sampling strategy for multi-modal knowledge graph embedding.
- In MANS, we propose MANS-V to align different modal information. Furthermore, we extend it to three combined NS strategies with different settings.

[†]Corresponding Author.

- We conduct comprehensive experiments on two knowledge graph completion tasks with two MMKG datasets. Experiment results illustrate that MANS could outperform the baseline methods in various tasks.
- We further carry out extensive analysis to explore several research questions about MANS to demonstrate the details of MANS.

II. RELATED WORKS

A. Knowledge Graph Embedding

Knowledge Graph Embedding (KGE) [20] is an important research topic for knowledge graphs, which focuses on embedding the entities and relations of KGs into low-dimensional continuous vector space.

General KGE methods utilize the triple structure to embed entities and relations and follow the research paradigm that defines a score function to measure the plausibility of triples in the given KG. Negative sampling (NS) is a significant technology widely used when training KGE models. During training, positive triples should get higher scores than those negative triples, which are generated by NS.

Previous KGE methods can be cursorily divided into several categories. Translation-based methods like TransE [7] and TransH [21] modeling the triples as the translation from head to tail entities with a distance-based scoring function. Semantic-based methods like DistMult [9] and ComplEx [8] use similarity-based scoring functions. Neural network-based methods [22], [23] employ neural networks to capture features from entities and relations and score the triples. Several KGE methods modeling triples with various mathematical structures, such as RotatE [10], ConE [24]. Some recent methods [25], [26] combine rule learning / analogical inference and KGE together to enhance the interpretability of KGE models.

B. Multi-modal Knowledge Graph Embedding

The KGE methods mentioned before are unimodal approaches as they only utilize the structure information from KGs. For multi-modal Knowledge Graphs (MMKGs), the modal information like images and text should also be highly concerned as another embedding for each entity and relation. Existing methods usually extract modal information using pre-trained models and project the modal information into the same representation space as structural information. IKRL [11] apply VGG [27] to extract visual information of entities' images and scoring a triple with both visual information and structure information using TransE [7]. TransAE [13] also employs TransE as the score function and exact modal information with a multi-modal auto-encoder. Mosselley et al [28] and Pezeshkpour et al [12] use VGG [27] and GloVe [29] to separately extract visual and textual information and then fused them into multi-modal information. Recently, RSME [30] focused on preserving truly valuable images and discarding the useless ones with three gates.

C. Negative Sampling in Knowledge Graph Embedding

Negative sampling (NS) aims to generate negative triples which don't appear in existing KGs. Those negative triples will participate in the training process of KGE models by contrasting them with positive triples. Therefore, many NS methods are proposed to generate high-quality negative samples. Normal NS [7] randomly replaces the head or tail entity with another entity with the same probabilities. KBGAN [31] and IGAN [14] apply Generative Adversarial Networks (GANs) [32] to select harder negative samples. NSCaching [15] store the high-quality negative triples with cache during training to achieve efficient sampling. NS-KGE [17] employs a unified square loss to avoid NS during training. It is called no-sampling but all-sampling. SANS [16] utilize the graph structure to sample high-quality negative samples. CAKE [18] construct commonsense from KGs to guide NS. EANS [19] propose a clustering-based negative sampling strategy with an auxiliary loss function. VBKGC [33] propose a twins negative sampling method for different parts of the score function.

However, many of the NS methods have their shortcomings which leads to the dilemma of NS for MMKGE. On the one hand, they are not lightweight enough as extra modules are introduced in the models. On the other hand, they are designed for unimodal knowledge graph embedding. Such a strategy performs well in general KGE because each entity has only one structural embedding. As many MMKGE models define multiple embeddings for each entity, the alignment between different embeddings is also significant but ignored by existing methods.

III. PROBLEM FORMULATION

In this section, we would introduce the basic pipeline of multi-modal knowledge graph embedding (MMKGE) in a three-step format. We first formally describe what a MMKG is and the embeddings we design for the MMKGE task. Then we detailedly introduce the modules of the MMKGE model. Eventually, we would show the training objective of MMKGE model and emphasis the process of negative sampling.

A. Basic Definition

A MMKG can be denoted as $\mathcal{G}_M = (\mathcal{E}, \mathcal{R}, \mathcal{I}, \mathcal{T})$, where $\mathcal{E}, \mathcal{R}, \mathcal{I}, \mathcal{T}$ are the entity set, relation set, image set and triple set. Entities in \mathcal{E} may have 0 to any number of images in \mathcal{I} , and the image set of entity e is denoted as I_e .

We denote \mathbf{e}_s and \mathbf{e}_v as the structural embedding and visual embedding for an entity e , respectively. Therefore, the entity e can be represented by two embedding vectors $\mathbf{e}_s, \mathbf{e}_v$. Besides, we denote \mathbf{r} as the structural embedding of relation r .

B. MMKGE Framework

In this paper, we employ a general MMKGE framework as the backbone model. The model architecture is shown in Figure 1, which consists of a visual encoder and a score function.

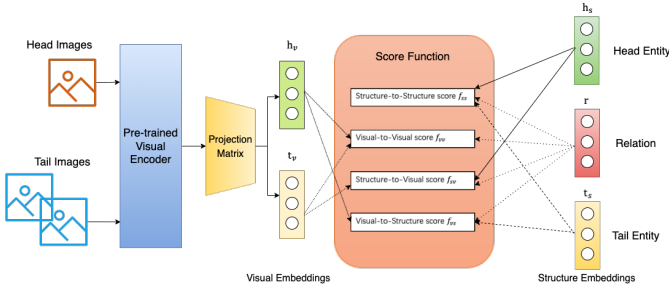


Fig. 1. Our Multi-modal KGE model architecture

1) *Visual Encoder*: Visual encoder, which is denoted as E_{img} , aims to capture the visual feature of entities and project them into the same representation space of structural embeddings. For those entities with more than one image, we use mean pooling to aggregate the visual feature. The visual embedding \mathbf{e}_v of entity e can be denoted as:

$$\mathbf{e}_v = \mathbf{W} \times \frac{1}{|I_e|} \sum_{I_e^k \in I_e} E_{img}(I_e^k) \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{d \times d_v}$ is the projection matrix, d is the dimension of both structural and visual embedding and d_v is the dimension of the output dimension of the visual encoder. In this paper, we employ pre-trained VGG-16 [27] as the visual encoder.

2) *Score Function*: The score function is denoted as $\mathcal{F}(h, r, t)$. Both the structural embeddings \mathbf{e}_s and visual embeddings \mathbf{e}_v will be considered in the score function. The overall score function consists of four parts, aiming to learn the embeddings in the same vector space, which can be denoted as: $\mathcal{F}(h, r, t) = f(\mathbf{h}_s, \mathbf{r}, \mathbf{t}_s) + f(\mathbf{h}_v, \mathbf{r}, \mathbf{t}_v) + f(\mathbf{h}_s, \mathbf{r}, \mathbf{t}_v) + f(\mathbf{h}_v, \mathbf{r}, \mathbf{t}_s)$, where f is the TransE score [7].

Besides, the overall score function $\mathcal{F}(h, r, t)$ can be divided into two parts, unimodal scores, and multi-modal scores. The unimodal scores only consider single-modal embedding of entities while multi-modal scores use both structural embeddings and visual embeddings. Under such criteria, $f(\mathbf{h}_s, \mathbf{r}, \mathbf{t}_s), f(\mathbf{h}_v, \mathbf{r}, \mathbf{t}_v)$ are unimodal scores and $f(\mathbf{h}_s, \mathbf{r}, \mathbf{t}_v), f(\mathbf{h}_v, \mathbf{r}, \mathbf{t}_s)$ are multi-modal scores. Such a distinction of scores will play an important role in adaptive NS.

C. Sampling and Training

The general target of a MMKGE model is to give higher scores for the positive triples and lower scores for the negative triples. In another word, the MMKGE model would discriminate the plausibility of a given triple by its score, which is widely used in KGC to predict the missing triples. Margin-rank loss is a general training objective extensively used in the MMKGE model [11], [12]. It could be denoted as:

$$\mathcal{L} = \sum_{(h, r, t) \in \mathcal{T}} \sum_{(h', r', t') \in \mathcal{T}'} \max(\gamma - \mathcal{F}(h, r, t) + \mathcal{F}(h', r', t')) \quad (2)$$

where γ is the margin, (h, r, t) is the positive triple in the KG and (h', r', t') is the negative triples.

Besides, a given KG usually consists of the observed facts, which are all positive triples. We need to generate the negative triple (h', r', t') manually. Such a process is what we call negative sampling (NS). In normal NS, one of the head and tail entities is randomly replaced. In this setting, h', t' are still the entities in \mathcal{E} . This also means that normal NS is an entity-level sampling strategy as it samples negative entities for a given positive triple. As we have analyzed in the previous section, normal NS is suitable for general KGE models but fails when it comes to the MMKGE. In the next section, we will introduce our NS methods to sample better negative triples.

IV. METHODOLOGY

Normal NS is an entity-level strategy, as all the embeddings of the selected entity are replaced by the negative ones. However, our approach differs. In this section, we would briefly introduce our **Modality-Aware Negative Sampling (MANS)**. MANS is based on visual negative sampling (MANS-V for short), which is a modal-level NS strategy and would sample negative visual embeddings for a finer contrast. We further combine MANS-V and normal NS with a sampling proportion β and propose three more comprehensive NS settings. They are two-stage negative sampling (MANS-T), hybrid negative sampling (MANS-H), and adaptive negative sampling (MANS-A).

A. Visual Negative Sampling (MANS-V)

MANS-V aims to sample the negative visual embeddings that do not belong to the current entity to teach the model to identify the visual features corresponding to each entity, which could achieve the modality alignment between structural and visual embeddings. In our context, modality alignment means that the model could identify the relations between the two modal embeddings, which we think is of great importance in MMKGE.

MANS-V is a modal-level method that would sample negative visual embeddings. The negative triple (h', r', t') generated by MANS-V preserves the original structural embeddings but the visual embedding of the replaced entity is changed. For example, if we replace head entity h with another entity h' , the embeddings of h' used during training is $\mathbf{h}_s, \mathbf{h}'_v$. For tail entity, the embeddings of t' is $\mathbf{t}_s, \mathbf{t}'_v$. In MANS-V, the replaced entity is a virtual negative entity that doesn't exist in \mathcal{E} . An intuitive example of MANS-V is shown in Figure 2.

Thus, MANS-V is a more fine-grained strategy compared with normal sampling. It changes the granularity of NS from the whole entity to the single modal embedding of the entity. By sampling only negative visual embeddings, MANS-V could achieve alignment between different modal embeddings for an entity.

KGE models would learn to align the two embeddings for each entity by MANS-V. However, learning to discriminate the plausibility of triples is still significant, which could be achieved by normal NS. Hence, we consider that MANS-V could play an important role as the auxiliary to enhance the

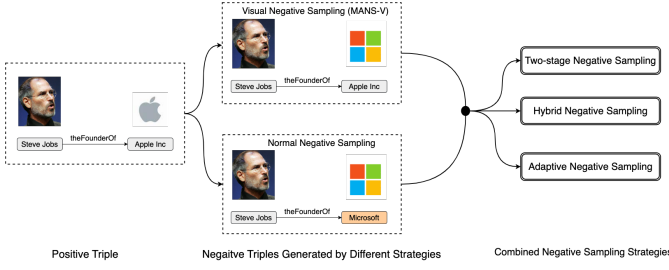


Fig. 2. An example of MANS-V. Only negative visual feature is sampled compared with normal negative sampling. We further combine MANS-V with normal NS to get three more NS strategies.

normal NS and we propose three combination strategies for comprehensive training.

B. Two-Stage Negative Sampling (MANS-T)

MANS-T divides the training process into two different stages:

- In **Stage1**, MANS-V is applied to train the model. The model would learn to align different modal embeddings in this stage.
- In **Stage2**, we employ normal sampling and train the model to discriminate the plausibility. As the structural and visual embeddings are aligned inside each entity, the model would learn better in this stage.

We assume that the total training epoch is M and the proportion of MANS-V is β_1 , then the turning point for stage switching is:

$$M_0 = \beta_1 \times M \quad (3)$$

which means training epoch $[0, M_0]$ is **Stage1** and $[M_0 + 1, M]$ is **Stage2**. It's not difficult to find that normal NS and MANS-V are two special cases of MANS-T when $\beta_1 = 0$ (normal) or $\beta_1 = 1$ (image).

C. Hybrid Negative Sampling (MANS-H)

As MANS-T divides the NS from the view of training epochs, MANS-H would apply two sampling strategies in each training epoch. Compared with the two-stage setting, MANS-H is more progressive.

In each mini-batch of one training epoch, we assume that the batch size is N and the MANS-V proportion is β_2 , for each triple, we sample k negative samples, then the total number of negative samples is kN and the total negative triples generated by MANS-V is:

$$N_0 = \beta_2 \times kN \quad (4)$$

which means that N_0 negative samples are randomly generated by MANS-V in a mini-batch and others are generated by normal NS. During the whole training process, MANS-H will be applied and the negative samples are blended from multiple sampling strategies. In MANS-H, the sampling proportion β_2 is a tunable hyper-parameter. The same as the two-stage setting, MANS-H becomes normal NS when $\beta_2 = 0$ and MANS-V when $\beta_2 = 1$.

D. Adaptive Negative Sampling (MANS-A)

MANS-A is an improved version of MANS-H, which no longer needs to tune the sampling proportion anymore. MANS-A will change the proportion β_3 adaptively. The adaptive sampling proportion β_3 would be determined by different scores of the training data.

As mentioned before, the overall score function $\mathcal{F}(h, r, t)$ can be divided into unimodal scores and multi-modal scores. We could denote the two parts as:

$$\mathcal{F}_{unimodal}(h, r, t) = f(\mathbf{h}_s, \mathbf{r}, \mathbf{t}_s) + f(\mathbf{h}_v, \mathbf{r}, \mathbf{t}_v) \quad (5)$$

$$\mathcal{F}_{multimodal}(h, r, t) = f(\mathbf{h}_s, \mathbf{r}, \mathbf{t}_v) + f(\mathbf{h}_v, \mathbf{r}, \mathbf{t}_s) \quad (6)$$

We define a function $\Phi(h_i, r_i, t_i)$ to discriminate whether the triple (h_i, r_i, t_i) need MANS-V. The function $\Phi(h_i, r_i, t_i)$ is defined as:

$$\Phi(h_i, r_i, t_i) = \begin{cases} 0 & \mathcal{F}_{multimodal} \geq \mathcal{F}_{unimodal} \\ 1 & \mathcal{F}_{multimodal} < \mathcal{F}_{unimodal} \end{cases} \quad (7)$$

which means that, when multi-modal score $\mathcal{F}_{multimodal}(h_i, r_i, t_i)$ is higher than the unimodal score, MANS-V will be applied. As MANS-V would align different modal embeddings and achieve higher multi-modal scores. Hence, the adaptive proportion β_3 for each batch is defined as:

$$\beta_3 = \frac{1}{N} \sum_{i=1}^N \Phi(h_i, r_i, t_i) \quad (8)$$

where $(h_i, r_i, t_i) (i = 1, 2, \dots, N)$ is the batch data. With sampling proportion β_3 , the MANS-H would be applied during the training of this batch. The biggest difference between adaptive and MANS-H is that we define an adaptive sampling proportion β_3 and no longer need to tune it anymore, which could reduce the workload for searching better hyper-parameters.

V. EXPERIMENTS

In this section, we will present the detailed experiment settings and the experimental results to show the advantages of MANS. We design several experiments to answer the following research questions (RQs):

- **RQ1:** Could MANS outperform the baseline methods and achieve new state-of-the-art (SOTA) results in various KGC tasks?
- **RQ2:** As a new hyper-parameter β is introduced in our method, how to select better sampling proportion $\beta_i (i = 1, 2)$ for MANS-T and MANS-H?
- **RQ3:** Is MANS-A a reasonable and effective design? What is the trend of the sampling proportion β_3 in MANS-A during training?
- **RQ4:** Is MANS efficient and lightweight compared with existing NS methods?
- **RQ5:** Could MANS learn better embeddings with more semantic information compared with normal NS?

TABLE I
STATISTICS OF DATASETS

Dataset	Entities	Relations	Images	Triples
FB15K	14951	1345	13444	592213
DB15K	14777	279	12841	99028

A. Datasets

In our experiments, we use two well-known MMKG datasets (FB15K, DB15K with extra images of entities) proposed in [34], the statistical information of the datasets is shown in Table I.

B. Evaluation and Implementation Details

1) *Tasks and Evaluation Protocol*: We evaluate our method on two tasks, link prediction, and triple classification [7]. The link prediction task aims to predict the missing entity for a given query $(h, r, ?)$ or $(?, r, t)$ with the KGE model. We evaluate the link prediction task by mean rank (MR) [7], mean reciprocal rank (MRR) [10] and Hit@K ($K=1,3,10$) [7]. Besides, we follow the filter setting [7] which would remove candidate triples that have already appeared in the datasets.

Triple classification task would predict the given triple (h, r, t) is true or not. Thus, we evaluate the task with accuracy (Acc), precision (P), recall (R), and F1-score (F1), which are the common metrics for the binary classification task.

2) *Baselines*: For the link prediction task, we employ the normal NS [7] and several recent SOTA NS methods as the baselines. They are No-Samp [17], NSCaching [15], SANS [16], CAKE [18], and EANS [19], which enhance the normal NS from their different perspectives. We utilize their official code to conduct baseline results. For the triple classification task, we compare the performance of MANS with normal NS, as other NS methods do not focus on this task and give the corresponding implementations.

3) *Experiments Settings*: For experiments, we set both structural embedding and visual embedding size $d_e = 128$ for each model. The dimension of visual features captured by a pre-trained VGG-16 model is $d_v = 4096$. For those entities which have no image, we employ Xavier initialization [35] for their visual features. We set the number of negative triples to 1 and train each model with 1000 epochs.

During training, we divide each dataset into 400 batches and apply IKRL [11] as the MMKGE model. We use the default Adam optimizer for optimization and tune the hyper-parameters of our model with grid search. The margin γ is tuned in $\{4.0, 6.0, 8.0, 12.0\}$ and learning rate η is tuned in $\{0.001, 0.01, 0.1, 1\}$. Besides, for two-stage and MANS-H, we tuned the sampling proportion β_1, β_2 from 0.1 to 1.0.

For baselines, we have taken full account of the parameter settings in the original paper [15], [17]–[19]. All the experiments are conducted on one Nvidia GeForce 3090 GPU. Our code of MANS is released in <https://github.com/zjukg/MANS>.

C. RQ1: Main Results

To answer RQ1, we conduct experiments on two KGC tasks. The evaluation results of the link prediction task are shown in Table II and the triple classification results are in Table III. From the experimental results, We can conclude the following points:

Poor performance of the baselines. We could observe that existing NS methods have poor performance and they are even worse than the normal NS. According to our previous analysis, these NS methods are designed for general KGE models and are unsuitable for the multi-modal scenario where modal information is carefully considered. They could not align different embeddings of each entity and get bad performance in MMKGE.

The outperformance of MANS. MANS could achieve better link prediction results compared with baselines. For example, MANS-A achieves much better Hit@1 on FB15K compared with baselines (from 0.318 to 0.353, a relative improvement of 9.9%). Besides, MANS performs particularly well in Hit@1 and MRR, which are sensitive to high-rank results [15]. This means that MANS can largely improve the accurate discriminatory ability of the model by aligning structural and visual embeddings.

Necessity and effectiveness of MANS-V. According to the previous section, MANS-V is designed to align different modal information. Though it does not perform better than baseline methods, MANS-V is the fundamental component of the other three settings of MANS. Besides, we could prove with such a result that both modal alignment and positive-negative discrimination are important for MMKGE, which could be achieved by MANS-V and normal NS respectively. MANS-T, MANS-H, and MANS-A could perform better because they combine the advantages of both. In summary, MANS-V is a necessary design for MMKGE.

Comparison of different MANS settings. As we propose three different settings of MANS, we could observe from Table II that all of the three settings (MANS-T, MANS-H, MANS-A) outperform the baseline methods. Experiment results demonstrate that MANS-H and MANS-A would perform better than MANS-T. Meanwhile, MANS-H and MANS-A have their advantages on different datasets and metrics, but the overall difference of link prediction performance between MANS-H and MANS-A is not notable. Nevertheless, the proportion β_2 of MANS-H needs to be tuned several times to find the best choice while MANS-A could adaptively change the proportion β_3 during training and get good performance without hyper-parameter tuning. For the mentioned reasons, we believe that the overall performance of MANS-A is better than MANS-T and MANS-H. MANS-A is free of proportion tuning and could achieve outstanding results.

Universality of MANS. From Table III, we could see that three settings of MANS could achieve better triple classification results on four metrics compared with normal NS. Besides, MANS-A outperforms MANS-T and MANS-H on accuracy and F1-score. In summary, the results show that our

TABLE II

EVALUATION RESULTS FOR LINK PREDICTION. THE BEST RESULTS OF EACH METRIC ARE IN BOLD AND THE SECOND BEST RESULTS ARE UNDERLINED.

Model	FB15K					DB15K				
	MRR↑	MR↓	Hit@10↑	Hit@3↑	Hit@1↑	MRR↑	MR↓	Hit@10↑	Hit@3↑	Hit@1↑
Normal [7]	0.479	95	<u>0.755</u>	0.604	0.314	0.303	685	0.542	0.410	0.167
No-Samp [17]	0.109	1594	0.212	0.130	0.051	0.151	456	0.271	0.171	0.087
NSCaching [15]	0.329	121	0.526	0.374	0.224	0.291	835	0.471	0.344	0.192
SANS [16]	0.394	109	0.635	0.466	0.264	0.276	703	0.413	0.387	0.127
CAKE [18]	0.395	68	0.647	0.467	0.262	-	-	-	-	-
EANS [19]	0.483	111	0.739	0.597	0.327	0.269	1036	0.489	0.353	0.141
MANS-V	0.454	103	0.713	0.552	0.305	0.274	506	0.525	0.333	0.165
MANS-T	0.485	93	0.748	0.591	0.333	0.307	615	<u>0.546</u>	0.411	0.178
MANS-H	<u>0.493</u>	92	0.756	0.606	<u>0.351</u>	<u>0.329</u>	553	0.541	<u>0.414</u>	<u>0.204</u>
MANS-A	0.499	<u>88</u>	0.749	<u>0.601</u>	0.353	0.332	<u>549</u>	0.550	0.420	0.204

TABLE III

EVALUATION RESULTS FOR TRIPLE CLASSIFICATION

Dataset	Model	Accuracy	Precision	Recall	F1-score
FB15K	Normal	95.2	94.7	95.7	95.2
	MANS-V	96.5	95.8	97.2	96.5
	MANS-T	96.2	95.7	96.8	96.2
	MANS-H	96.5	95.9	97.3	96.5
	MANS-A	96.6	96.1	97.2	96.7
DB15K	Normal	86.6	88.1	84.7	86.4
	MANS-V	85.6	85.3	85.9	85.7
	MANS-T	87.4	88.1	86.4	87.3
	MANS-H	87.9	87.3	88.9	88.1
	MANS-A	88.0	87.1	89.2	88.1

design of MANS could benefit the MMKGE model in various KGC tasks such as link prediction and triple classification, which means that MANS is a universal approach for better KGC.

D. RQ2: Proportion Selection

TABLE IV

BEST SAMPLING PROPORTION (β_1 FOR MANS-T AND β_2 FOR MANS-H) FOR LINK PREDICTION TASK.

	FB15K	DB15K
MANS-T	0.4	0.3
MANS-H	0.3	0.3

Though MANS achieved good performance on link prediction and other tasks, a fact cannot be ignored is that MANS might require more effort to tune the sampling proportion (β_1, β_2 for MANS-T and MANS-H respectively). The optimal proportions for MANS-T and MANS-H are shown in Table IV, we further explore the impact of sampling proportion on the link prediction task. It would answer RQ2 and guide us in choosing the best sampling proportion.

It is worth mentioning that when $\beta_i = 0.0 (i = 1, 2)$, both MANS-T and MANS-H degrade to normal negative sampling. When $\beta_i = 1.0 (i = 1, 2)$, both of them become MANS-I. Thus, they can be baselines for comparison.

We could observe that the trends of MANS-T and MANS-H are almost identical. For MANS-T, the best proportion $\beta_1 =$

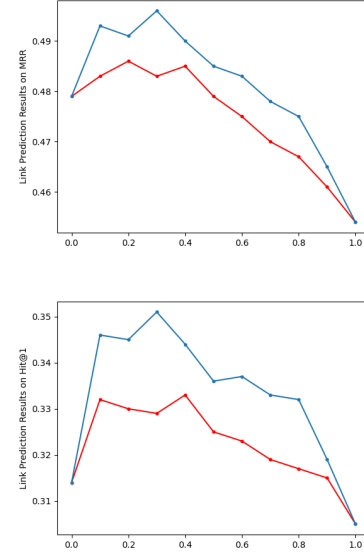


Fig. 3. Impact of sampling proportion β_1, β_2 for two-stage (MANS-T, the red line) and hybrid (MANS-H, the blue line) negative sampling. The experiments are based on FB15K dataset and TransE base score function.

0.3, and for MANS-H the best proportion $\beta_2 = 0.4$. In the range of 0.1 to 0.4, MMKGE models trained with MANS-T and MANS-H perform better. Meanwhile, we could find that as the proportion of image negative sampling increases (when $\beta_1, \beta_2 \geq 0.5$), the model performance would get down and might be worse than normal negative sampling. In the range of 0.1 to 0.4, the performance of each strategy has just little changes most of the time. Therefore, the best choice for sampling proportion should most likely be in this range.

E. RQ3: Adaptive Setting

From the previous experiments, we could observe that the performance of MANS-A is close to and slightly better than MANS-H most of the time. In this section, we will dive into MANS-A and make further exploration to illustrate the rationality of its design and answer RQ3.

We record the adaptive proportion β_3 for each batch of data in each training epoch and then calculate the average

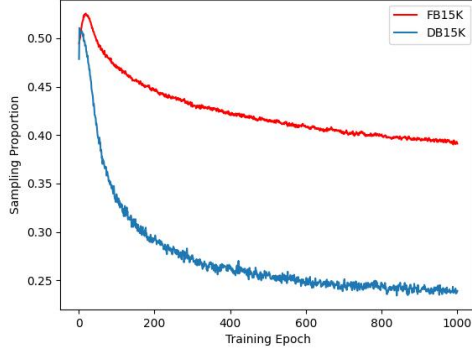


Fig. 4. Trend of adaptive sampling proportion β_3 during whole training process on two datasets.

adaptive proportion of all the batches for each epoch. The trends of adaptive sampling proportion β_3 for different models and datasets in each training epoch are shown in Figure 4.

According to Figure 4, the adaptive proportion β_3 usually becomes stable during the training process. We could pay attention to the stable part of each curve. Compared with the optimal proportions for MANS-H (which can be found in the previous section), we could find that design of MANS-A is reasonable as the adaptive proportion β_3 in MANS-A is close to the optimal settings in MANS-H. For example, the stable sampling proportions on FB15K and DB15K are nearly 0.4 and 0.3. They are close to the optimal or sub-optimal β_2 of MANS-H. This suggests that the adaptive setting MANS-A would find the suitable proportion β_3 which is consistent with MANS-H but free of tuning. In summary, the design of MANS-A is reasonable and effective.

F. RQ4: Efficiency

As we mentioned earlier, MANS is more lightweight and efficient compared with existing methods because it is free of over-designed. Therefore, we evaluate the training speed of each NS method and list the results in Table V, aiming to answer RQ4. The experiments are conducted on a single Nvidia GeForce RTX 3090 GPU.

From the table, we could find that the training speed of MANS is closer to the normal NS. Even the most complicated MANS-A is more efficient than several baselines. Though No-Samp [17] is very fast, it fails to perform well in MMKGE according to the link prediction results in Table II.

We also list the extra modules proposed by each method. Unlike random walks in SANS [16] and entity clustering in EANS [19], our visual NS is not computationally intensive, which is the reason why MANS is lightweight enough. Besides, we have found in practice that NSCaching [15] and No-Samp [17] would consume lots of memory and GPU resources, which is $1.13\times$ (NSCaching [15]) and $6.65\times$ (No-Samp [17]) than MANS-A. In summary, MANS is lightweight and efficient enough and could make the training process faster compared with other NS methods. We have achieved

TABLE V
THE TRAINING SPEED OF DIFFERENT NS METHODS AND THE EXTRA MODULES PROPOSED BY THEM

Method	Traning Speed(s/epoch)	Extra Module
Normal [7]	14.3	-
No-Samp [17]	0.2	Full-batch Traning
NSCaching [15]	16.7	Entity Caching
SANS [16]	16.6	Random Walks
EANS [19]	60.9	Entity Clustering
MANS-I	14.8	Visual NS
MANS-T	14.5	
MANS-H	15.1	
MANS-A	16.1	

a significant improvement in two tasks of KGC with our lightweight design.

G. RQ5: Embedding Visualization

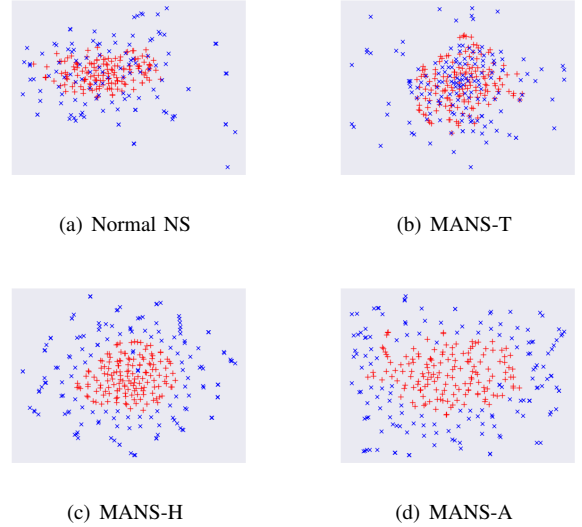


Fig. 5. Visualization using t-SNE for entity embeddings trained with different NS methods. The marker $+$ and \times represent the structural and visual embeddings.

To evaluate the effectiveness of MANS and answer RQ5 in a straightforward view, we apply t-SNE to visualize the structural and visual embeddings of entities. We select the entities with type */award/award_winner* in FB15K and the results are shown in Figure 5.

From the visualization results, we could observe that the distribution of structural and visual embeddings of normal NS is close to each other. This means that the semantic information they express is relatively similar. However, the embedding distribution of MANS-H and MANS-A shows a more clear boundary between the two kinds of embeddings compared with normal NS, which means the learned embeddings have more semantic information and the MMKGE model can clearly distinguish them to enhance the model performance, which is consistent with the link prediction performance in Table II. Thus, RQ5 is solved and we could

conclude that MANS could guide the MMKGE model to learn meaningful and semantic-rich embeddings.

VI. CONCLUSION

In this paper, we propose MANS, a modality-aware negative sampling method for MMKGE, which focuses on the alignment between different modal embeddings of a MMKGE model. MANS is the first NS method designed especially for MMKGE while achieving efficiency and effectiveness to solve the problems of existing NS methods. We first propose visual negative sampling (MANS-V) and extend MANS-V to three different settings called MANS-T, MANS-H, and MANS-A. Besides, we conduct comprehensive experiments on two public benchmarks and two classic tasks to demonstrate the performance of MANS compared with several state-of-the-art NS methods. In the future, we plan to conduct more in-depth research about MMKGE from two perspectives: (1) developing more robust solutions to achieve modal alignment and fusion of MMKG, (2) attempting to make co-design of the MMKGE model and NS method for better performance.

ACKNOWLEDGEMENT

This work is funded by Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020017) and Yongjiang Talent Introduction Programme (No. 2022A-238-G).

REFERENCES

- [1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *Proc. of ISWC*, 2007.
- [2] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proc. of SIGMOD*, 2008.
- [3] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "QA-GNN: reasoning with language models and knowledge graphs for question answering," in *Proc. of NAACL*, 2021.
- [4] W. Zhang, C. M. Wong, G. Ye, B. Wen, W. Zhang, and H. Chen, "Billion-scale pre-trained e-commerce product knowledge graph model," in *Proc. of ICDE*, 2021.
- [5] W. Liu, P. Zhou, Z. Zhang, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-BERT: enabling language representation with knowledge graph," in *Proc. of AAAI*, 2020.
- [6] Z. Chen, W. Zhang, Y. Huang, M. Chen, Y. Geng, H. Yu, Z. Bi, Y. Zhang, Z. Yao, W. Song *et al.*, "Tele-knowledge pre-training for fault analysis," *arXiv preprint arXiv:2210.11298*, 2022.
- [7] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. of NeurIPS*, 2013.
- [8] T. Trouillon, C. R. Dance, É. Gaussier, J. Welbl, S. Riedel, and G. Bouchard, "Knowledge graph completion via complex tensor factorization," *J. Mach. Learn. Res.*, 2017.
- [9] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proc. of ICLR*, 2015.
- [10] Z. Sun, Z. Deng, J. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *Proc. of ICLR*, 2019.
- [11] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *Proc. of IJCAI*, 2017.
- [12] P. Pezeshkpour, L. Chen, and S. Singh, "Embedding multimodal relational data for knowledge base completion," in *Proc. of EMNLP*, 2018.
- [13] Z. Wang, L. Li, Q. Li, and D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," in *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 2019, pp. 1–8.
- [14] P. Wang, S. Li, and R. Pan, "Incorporating GAN for negative sampling in knowledge representation learning," in *Proc. of AAAI*, 2018.
- [15] Y. Zhang, Q. Yao, Y. Shao, and L. Chen, "Nscaching: Simple and efficient negative sampling for knowledge graph embedding," in *Proc. of ICDE*, 2019.
- [16] K. Ahrabian, A. Feizi, Y. Salehi, W. L. Hamilton, and A. J. Bose, "Structure aware negative sampling in knowledge graphs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 6093–6101.
- [17] Z. Li, J. Ji, Z. Fu, Y. Ge, S. Xu, C. Chen, and Y. Zhang, "Efficient non-sampling knowledge graph embedding," in *Proc. of WWW*, 2021.
- [18] G. Niu, B. Li, Y. Zhang, and S. Pu, "CAKE: A scalable commonsense-aware framework for multi-view knowledge graph completion," in *Proc. of ACL*, 2022.
- [19] S. Je, "Entity aware negative sampling with auxiliary loss of false negative prediction for knowledge graph embedding," *CoRR*, vol. abs/2210.06242, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2210.06242>
- [20] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, 2017.
- [21] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. of AAAI*, 2014.
- [22] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *Proc. of AAAI*, 2018.
- [23] X. Jiang, Q. Wang, and B. Wang, "Adaptive convolution for multi-relational learning," in *Proc. of NAACL*, 2019.
- [24] Z. Zhang, J. Wang, J. Chen, S. Ji, and F. Wu, "Cone: Cone embeddings for multi-hop reasoning over knowledge graphs," in *Proc. of NeurIPS*, 2021.
- [25] W. Zhang, B. Paudel, L. Wang, J. Chen, H. Zhu, W. Zhang, A. Bernstein, and H. Chen, "Iteratively learning embeddings and rules for knowledge graph reasoning," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. ACM, 2019, pp. 2366–2377. [Online]. Available: <https://doi.org/10.1145/3308558.3313612>
- [26] Y. Zhen, Z. Wen, C. Mingyang, H. Yufeng, Y. Yi, and C. Huajun, "Analogical inference enhanced knowledge graph embedding," *arXiv preprint arXiv:2301.00982*, 2023.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of ICLR*, 2015.
- [28] H. M. Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proc. of AACL*, 2018.
- [29] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. of EMNLP*, 2014.
- [30] M. Wang, S. Wang, H. Yang, Z. Zhang, X. Chen, and G. Qi, "Is visual context really helpful for knowledge graph? A representation learning perspective," in *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 2021.
- [31] L. Cai and W. Y. Wang, "KBGAN: adversarial learning for knowledge graph embeddings," in *Proc. of NAACL*, 2018.
- [32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of NeurIPS*, 2014.
- [33] Y. Zhang and W. Zhang, "Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling," *arXiv preprint arXiv:2209.07084*, 2022.
- [34] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, and D. S. Rosenblum, "MMKG: multi-modal knowledge graphs," in *The Semantic Web - 16th International Conference, ESWC 2019, Portoro, Slovenia, June 2-6, 2019, Proceedings*, 2019.
- [35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, ser. JMLR Proceedings, Y. W. Teh and D. M. Titterton, Eds., vol. 9. JMLR.org, 2010, pp. 249–256.