Cross-modal Knowledge Graph Contrastive Learning for Machine Learning Method Recommendation

Xianshuai Cao
School of Software, Shandong University, Jinan, China
xianshuaicao@mail.sdu.edu.cn

Yuliang Shi*
School of Software, Shandong University; Dareway Software Co., Ltd, Jinan, China
shiyuliang@sdu.edu.cn

Jihu Wang
School of Software, Shandong University, Jinan, China
jihu_wang@mail.sdu.edu.cn

Han Yu*
School of Computer Science and Engineering, Nanyang Technological University, Singapore
han.yu@ntu.edu.sg

Xinjun Wang
School of Software, Shandong University; Dareway Software Co., Ltd, Jinan, China
wxj@sdu.edu.cn

Zhongmin Yan
School of Software, Shandong University, Jinan, China
yzm@sdu.edu.cn

ABSTRACT

The explosive growth of machine learning (ML) methods is overloading users with choices for learning tasks. Method recommendation aims to alleviate this problem by selecting the most appropriate ML methods for given learning tasks. Recent research shows that the descriptive and structural information of the knowledge graphs (KGs) can significantly enhance the performance of ML method recommendation. However, existing studies have not fully explored the descriptive information in KGs, nor have they effectively exploited the descriptive and structural information to provide the necessary supervision. To address these limitations, we distinguish descriptive attributes from the traditional relationships in KGs with the rest as structural connections to expand the scope of KG descriptive information. Based on this insight, we propose the Cross-modal Knowledge Graph Contrastive learning (CKGC) approach, which regards information from descriptive attributes and structural connections as two modalities, learning informative node representations by maximizing the agreement between the descriptive view and the structural view. Through extensive experiments, we demonstrate that CKGC significantly outperforms the state-of-the-art baselines, achieving around 2% higher accurate click-through-rate (CTR) prediction, over 30% more accurate top-10 recommendation, and over 50% more accurate top-20 recommendation compared to the best performing existing approach.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Machine learning**.

*Yuliang Shi and Han Yu are the corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548273

KEYWORDS

knowledge graph, recommender system, cross modalities, contrastive learning

ACM Reference Format:
Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. 2022. Cross-modal Knowledge Graph Contrastive Learning for Machine Learning Method Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548273

1 INTRODUCTION

The rapid evolution of scientific research has generated an overwhelming volume of scholarly information, which has brought great pressure on information retrieval by researchers. The scholarly recommendation is a proven efficient approach to tackle this problem. Typical researches of it include paper recommendation, collaboration recommendation and venue recommendation [28].

As an important scholarly area, machine learning (ML) has been experiencing rapid development in recent years. In each subfield of machine learning (e.g., computer vision, natural language processing, federated learning [10]), numerous diverse methods emerge every year. However, the prosperous boom has resulted in serious information overload. Selecting the most suitable ML method for a given task scenario has thus become a significant challenge for researchers and practitioners. Thus, there is a critical demand for targeted recommender systems in the machine learning field to solve the above problems. To be specific, given a machine learning instance (e.g., a task or a dataset, here we treat dataset as the task scenario), suitable ML methods need to be accurately recommended.

Moreover, the rich connections between entities in machine learning constitute the knowledge graph (KG) of this domain. Knowledge graphs have been shown to be able to greatly improve the efficiency of recommendations with the rich semantic information and diverse types of relationships. Specifically, knowledge graphs in the machine learning domain can benefit method recommendations from two aspects. Firstly, in machine learning knowledge graphs (MLKGs), the original interaction information between datasets

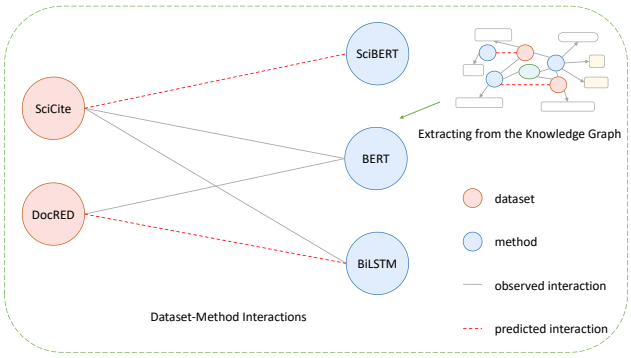

Figure 1: Schema of the machine learning method recommendation, the dataset-method interaction information is extracted from the machine learning knowledge graph.

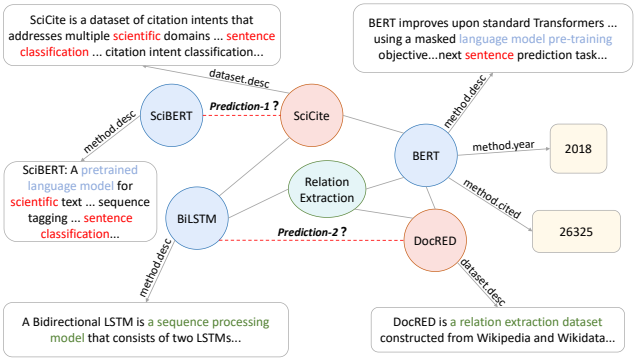

Figure 2: An illustration of the descriptive and structural information used for the dataset-method interaction prediction in a machine learning knowledge graph.

and methods can be obtained for training ML method recommendation models. A schematic example is shown in Figure 1. Secondly, as shown in Figure 2, MLKGs can provide both structural and descriptive information for ML method recommendations, which can improve performance while enhancing explainability.

However, when employing knowledge graphs to support recommendations, traditional models [22, 26] usually focus on structural information, but ignore descriptive information. When entities are less connected, it is difficult to learn a high-quality representation. Actually, in the machine learning method recommendation scenario, the emerging entities are less connected. Thus, their associated descriptive information should play a more important role.

To exploit multi-modal information, previous studies have incorporated text descriptions [27, 29], relationship types [12, 31], and visual information [30] into knowledge graphs to obtain more effective knowledge representations. However, these approaches are not designed for ML method recommendation tasks. Recently, several research works [2, 19] have introduced multi-modal information into knowledge graphs for enhancing recommendations and achieved promising results. Despite its effectiveness, we argue that there are limitations in the following two aspects.

(1) Limited Descriptive Information. Although recent studies have fused multi-modal descriptive information such as descriptive text and even visual information into knowledge graphs, they have not distinguished the descriptive attributes of entities from the traditional knowledge graph relationships. In general, for a given node in a knowledge graph, there are two main sources of information: 1) the descriptive attribute information that characterizes itself, and 2) the structural connection information with other nodes. As shown in Figure 2, in a machine learning knowledge graph (taking the *BERT* method as an example), its descriptive information contains not only the textual description, but also the time the method was proposed (i.e., *2018*) and the number of citations (i.e., *26325*). In addition, its structural information includes its directly connected first-order neighbors (e.g., task *Relation Extraction*) and indirectly connected higher-order neighbors (e.g., method *SciBERT*). Current research tends to treat some important descriptive attributes (e.g., (*BERT*, *method.year*, 2018)) as structural relationships. In this case, information that reflects the uniqueness of nodes has suffered seriously. Meanwhile, descriptive attributes can hardly contribute to modeling the neighborhood structure of entities, since they are usually sparsely connected.

(2) Mutual Effect in Information Aggregation. Current research works tend to learn independently before integrating the information from different modalities in the knowledge graph, without considering the supervisory effect among them. Specifically, descriptive and structural information alone can provide important supporting information when performing method recommendation (i.e., recommending the appropriate method for a given dataset). For example, in Figure 2, based on the similarity of the descriptive information between the dataset entity *SciCite* and the method entity *SciBERT*, the first interaction prediction (i.e., `Prediction-1`) can predict a higher probability of interaction between the two. Additionally, for the second interaction prediction (i.e., `Prediction-2`), the closeness in the structural information between the dataset entity *DocRED* and the method entity *BiLSTM* can be utilized as they have very similar neighborhood information, which is a strong reason to recommend the method to this dataset. Naturally, descriptive information is more direct than the structural ones, and can better represent the uniqueness and differences of nodes. In contrast, structural information can provide neighborhood insight. Thus, both types of information can unveil additional insights by mutually providing supervision signals. However, current research works tend to independently learn the representation of multiple modalities such as structural, textual or visual, without analyzing and leveraging the supervisory effects between them.

To address these limitations, in this paper, we propose the *Cross-modal Knowledge Graph Contrastive learning (CKGC)* method, which learns node representations in the knowledge graph from descriptive and structural views respectively to obtain a generalized representation by maximizing the mutual information between these two modalities. We summarize our contributions as follows.

- We categorize traditional knowledge graph relationships into descriptive attributes and structural connections, to consider more descriptive information that reflects the distinctiveness of entities.

- We propose a cross-modal knowledge graph contrastive learning approach, which captures entity features from different modalities and supervises each other to obtain more effective entity representations.
- Through extensive experiments on a real dataset, we demonstrate the effectiveness of the proposed method, which significantly outperforms the state of the art.

2 RELATED WORK

In this section, we introduce existing work related to our research, including scholarly recommendation and knowledge graph-based recommendation.

2.1 Scholarly Recommendation.

Scholarly recommendation is a research topic which focuses on recommending the most appropriate scholarly works from a massive number of candidates by mining the potential relationships in scholarly big data. The main representative studies of in this field include paper recommendation [6, 16, 21], collaboration recommendation [11, 13, 20] and venue recommendation [4, 14, 15]. Paper recommendation [6] or citation recommendation [16] aims to recommend relevant papers for study or citation, the collaboration recommendation [13] helps researchers find scholars with similar interests in collaboration while the venue recommendation [15] could build a bridge between researchers and conferences.

However, scholarly recommendation or knowledge mining has not been sufficiently studied for the machine learning domain, where serious information overload has been observed. In this paper, we investigate the method recommendation problem which recommends the most suitable machine learning method for a specified task scenario. Besides, we fully exploit and utilize the structural and descriptive information in the knowledge graph to enhance the performance of method recommendation.

2.2 Knowledge Graph-based Recommendation.

In general, existing knowledge graph-based recommendations can be classified into three categories, embedding-based methods, path-based methods, and unified methods. The embedding-based methods [23, 32] first utilize knowledge graph embedding techniques to obtain the entity representation and then fuse it into the recommendation framework. However, these methods only focus on the first-order connectivity in the knowledge graph without capturing higher-order connectivity. The path-based methods [18, 33] provide interpretable guidance for recommendations by exploring the similarity of connecting paths or patterns. The disadvantage of these approaches is that they usually rely on domain knowledge to predefine meta-paths. The unified methods [25, 26] integrate the semantic representation and path connectivity of entities, which are mainly based on the idea of embedding propagation, enhancing the representation of central nodes by aggregating neighborhood information.

However, most of these methods above only utilize the structural information in the knowledge graph while ignoring the descriptive information. Recent studies [2, 19] have further improved the performance of recommendation by introducing multimodal descriptive information, such as descriptive text or visual signals, into the knowledge graph. Although recent works consider multimodal descriptive information, they fail to distinguish descriptive attributes in traditional relations nor do they take into account that different modalities can provide meaningful supervisory signals to each other. In this paper, we distinguish traditional relations into two categories, descriptive attributes and structural connections, together with cross-modal contrastive learning to obtain a more effective node representation.

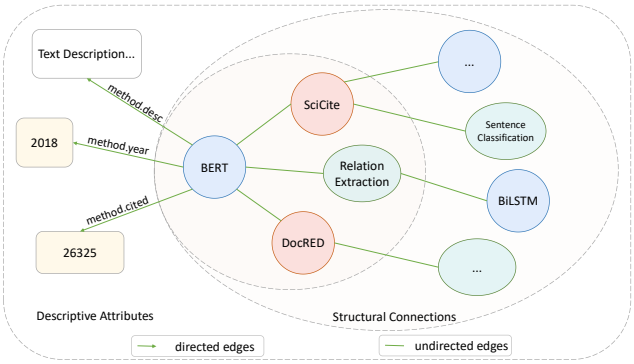

Figure 3: Distinction between descriptive attributes and structural connections.

3 PROBLEM FORMULATION

The focus of this paper is on the ML method recommendation problem, which predicts the probability of a given ML method to be suitable for a given dataset. In this scenario, we have a set of M datasets $\mathcal{D} = \{d_1, d_2, ..., d_M\}$ and a set of N ML methods $\mathcal{M} = \{m_1, m_2, ..., m_N\}$. The dataset-method interaction matrix $Y = \{y_{dm} \mid d \in \mathcal{D}, m \in \mathcal{M}\}$ is defined according to the observed implicit feedback, where $y_{dm} = 1$ indicates that ML method d has interacted with m before; otherwise, $y_{dm} = 0$.

Additionally, $\mathcal{G} = \{(h, r, t) \mid h \in \mathcal{E}, r \in \mathcal{R}, t \in \mathcal{E}\}$ represents a machine learning knowledge graph (MLKG), where h, r, t denote the "head entity-relation-tail entity" in a knowledge triple. For example, (*DocRED, dataset.task, Relation Extraction*) represents that the *DocRED* dataset can be used for the *Relation Extraction* task. $\mathcal{E}$ and $\mathcal{R}$ are the sets of entities and relations in the knowledge graph, respectively.

Moreover, we distinguish the descriptive attributes from the structural connections by verifying the existence of the inverse relation for each relation, which can be defined as follows:

Definition 1 (Relationship Distinction in the Knowledge Graph). For any triple $h_i \xrightarrow{r_i} t_i$, relation r_i is regarded as a structural connection if there exists an inverse relation $-r_i$ of it to enable the existence of triple $h_i \xleftarrow{-r_i} t_i$. For instance, in the triple (*BERT, method.dataset, SciCite*), the inverse relation *dataset.method* of the relation *method.dataset* states that there exists a triple (*SciCite, dataset.method, BERT*). Relationships other than structural connections are treated as descriptive attributes, such as the relation *method.year*. Specifically, the descriptive attributes are denoted as attribute-type directed edges by $r^a \in \mathcal{R}_A$ while the structural

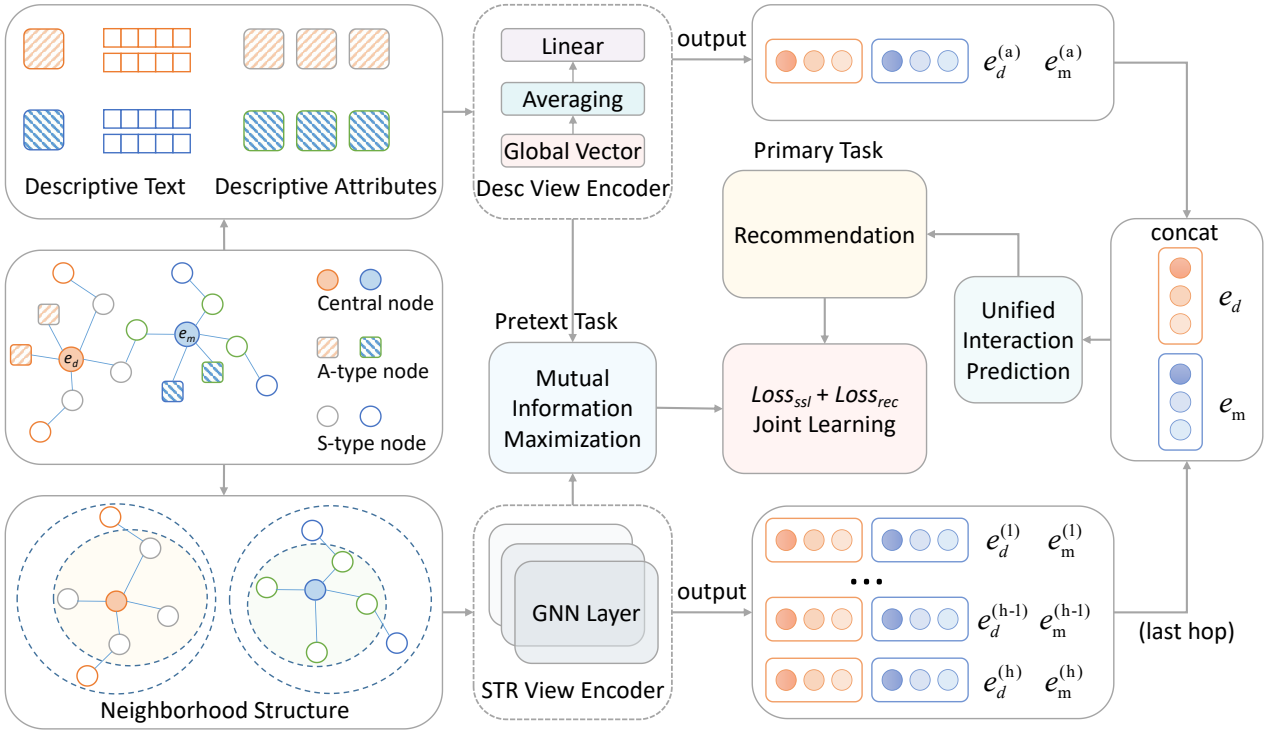

Figure 4: The framework of CKGC. The left part illustrates the discrimination between descriptive attributes and structural connections of given nodes. The right side represents the cross-modal contrastive learning of descriptive and structural views.

connections are represented as structure-type undirected edges by $r^s \in \mathcal{R}_S$. Here, $\mathcal{R} = \mathcal{R}_A \cup \mathcal{R}_S$.

For example, in Figure 3, the *BERT* entity has descriptive attributes such as text description, year of publication, and number of citations. Besides, *BERT* has the dataset entity *SciCite* and the task entity *Relation Extraction* as the first-order connected neighbors, through which it can reach higher-order neighbors.

By mapping the datasets and ML methods to the corresponding entities in the MLKG, respectively ($d, m \in \mathcal{E}$), the MLKG is able to provide descriptive attributes and structural connections for the interaction prediction. For example, dataset instance *DocRED* and ML method instance *BERT* correspond to the entities with the same name in the knowledge graph. Given the interaction matrix Y and the knowledge graph $\mathcal{G}$, we aim to predict whether the dataset will interact with candidate ML methods that are not in the interaction record of the dataset. Our goal is to learn a prediction function $\hat{y}_{dm} = \mathcal{F}(d, m \mid Y, \mathcal{G}, \Theta)$, where $\hat{y}_{dm}$ denotes the probability that a given dataset d adopts ML method m, and Θ denotes the parameters of the prediction function $\mathcal{F}$.

4 THE PROPOSED APPROACH

The overall framework of CKGC [1] is shown in Figure 4. Firstly, for a given dataset and ML method pair, the descriptive and structural information in the knowledge graph is retrieved. Then, the descriptive view features and structural view features are captured by the *Descriptive View Encoder* and *Structural View Encoder*, respectively. Finally, by maximizing the consistency of node representations from views of the two different modalities, a more generalizable representation can be obtained to improve the downstream ML method recommendation task.

4.1 The Descriptive View Encoder

The descriptive information can directly illustrate the features of the node and highlight its uniqueness. For a node e in the MLKG, its relevant descriptive information set $\mathcal{A}(e)$ can be accessed by the attribute-type edges, where the set includes descriptive text and other attributes. Here, $\mathcal{A}(e) = \{t \mid (h, r, t) \in \mathcal{G} \text{ and } r \in \mathcal{R}_A\}$, which is denoted as *A-type node* in Figure 4.

Since the descriptive values are often numerical or textual, they are treated as natural language. Then, the pre-trained word vectors are used to obtain the description text embedding $\mathbf{s}_t$ and other descriptive feature embeddings $\mathbf{s}_o$. However, the pre-trained vector dictionary does not contain all numbers when obtaining embedding representations of the numerical values. For example, the number of citations for method *BERT* is 26325, which has no corresponding pre-trained vector. Therefore, we do a simple preprocessing on the numeric values. Specifically, we first normalize them, then multiply them by 100 followed by a rounding operation so that all values are in the range of 0 to 100. This approach is similar to the idea

[1] The approach can be seen as a further extension and optimization based on the DEKR [2]. High-quality entity representations can be yielded by considering more descriptive attributes and exploiting the supervisory effect between two different modalities.

of bucketing, which can be regarded as distinguishing different numeric values into multiple categories.

After that, the initial embeddings $\mathbf{s}_t$ and $\mathbf{s}_o$ are projected into a low-dimensional vector space and concatenated to derive the descriptive view representation $\mathbf{v}_a$ of the node:

$$\begin{aligned} \mathbf{v}_t = \mathbf{W}_t\mathbf{s}_t,\ \mathbf{v}_o = \mathbf{W}_o\mathbf{s}_o \\ \mathbf{v}_a = \mathbf{v}_t \parallel \mathbf{v}_o. \end{aligned} \tag{1}$$

Note that there are several alternative ways to obtain the natural language embeddings (e.g., SIF [1], BERT [5]). Since our main focus here is on the identification and introduction of the descriptive attributes in knowledge graphs, the simple yet effective approach of [2] will suffice.

4.2 The Structural View Encoder

The structural information describes the neighborhood environment and the topological structure of the node within the graph, offering useful contextual information about the node. Given a node e in the MLKG, its first-order neighbors $\mathcal{N}(e)$ can be obtained via its structure-type edges, where $\mathcal{N}(e) = \{t \mid (h, r, t) \in \mathcal{G}\ and\ r \in \mathcal{R}_S\}$ (denoted as *S-type node* in Figure 4).

Intuitively, the representation of node e can be derived by aggregating the information from connected neighbors into its own representation, which can be formulated as:

$$\mathbf{e}^{(1)} = \text{LeakyReLU}(\mathbf{W}_0(\mathbf{e}^{(0)} + \mathbf{e}^{(0)}_{\mathcal{N}(e)})). \tag{2}$$

Here, $\mathbf{e}^{(1)}$ is the representation of node e after aggregating the first-order (i.e., one-hop) neighbor information. LeakyRelu is the activation function, $\mathbf{W}_0$ denotes a learnable transformation matrix. In addition, $\mathbf{e}^{(0)}_{\mathcal{N}(e)}$ is the linear integration of node e's neighborhood, which can be expressed as:

$$\mathbf{e}_{\mathcal{N}(e)} = \sum_{e' \in \mathcal{N}(e)} \pi(e')\mathbf{e}', \tag{3}$$

where $\pi(e')$ controls the number of messages propagated from the neighborhood to discriminate the contribution of different neighbors. It can be computed as:

$$\pi(e') = \frac{\exp(\pi(e'))}{\sum_{v \in \mathcal{N}(e)} \exp(\pi(v))}. \tag{4}$$

Similarly, node e can leverage information from its multi-hop neighbors through higher-order connections. The representation of node e with the aggregated h-hop neighbors is expressed as:

$$\mathbf{e}^{(h)} = \text{LeakyReLU}(\mathbf{W}_{h-1}(\mathbf{e}^{(h-1)} + \mathbf{e}^{(h-1)}_{\mathcal{N}(e)})), \tag{5}$$

where $\mathbf{e}^{(h-1)}$ and $\mathbf{e}^{(h-1)}_{\mathcal{N}(e)}$ denote the representation of node e and its neighbors after (h-1)-hop aggregation, respectively.

It is worth noticing that we could have employed more advanced techniques to learn the graph representations here, such as removing the nonlinear layers in the graph neural network [8] or connecting the representations of each layer with residuals [3]. By adopting the classical knowledge graph neural network structure, fair experimental comparisons with CKGC can be made to investigate the specific designs proposed in this paper.

4.3 Cross-modal Contrastive Learning

After propagation and information aggregation on the MLKG, the corresponding descriptive feature representation $\mathbf{e}^{(a)}$ (i.e.,$\mathbf{v}_a$) and the structural feature representation $\mathbf{e}^{(s)}$ (i.e.,$\mathbf{e}^{(h)}$) of node e can be obtained. In particular, unlike the structure-type neighbors, the attribute-type neighbors which provide descriptive information are only linked to the central node which they modify. Thus, their high-order connections do not need to be considered.

In order to take full advantage of the supervision signals between the views of different modalities, the mutual information between the descriptive and structural views is maximized, which can be regarded as a contrastive self-supervised strategy to encourage the consistency of node representation under views of different modalities. Specifically, we adopt InfoNCE [7] to model the mutual information for efficiency, which can be formulated as:

$$\begin{aligned} \mathcal{L}_{ssl} = \sum_{e \in \mathcal{E}} (-\log \left(\frac{\exp(s(e^{(a)}, e^{(s)})/\tau))}{\sum_{v \in \mathcal{E}} \exp(s(e^{(a)}, v^{(s)})/\tau)} \right) \\ -\log \left(\frac{\exp(s(e^{(s)}, e^{(a)})/\tau))}{\sum_{v \in \mathcal{E}} \exp(s(e^{(s)}, v^{(a)})/\tau)} \right)), \end{aligned} \tag{6}$$

where $s(\cdot)$ is the function for measuring the correlation between two representations with cosine similarity. τ is the temperature parameter in the softmax function. By doing so, the agreement between the descriptive and structural representations of a target node e in the knowledge graph can be maximized.

4.4 Model Prediction and Optimization

Based on the learned node representations, for a given dataset-ML method-pair e_d and e_m, we can obtain its final multimodal representation as follow:

$$\mathbf{e}_d = \mathbf{e}^{(a)}_d \parallel \mathbf{e}^{(s)}_d, \quad \mathbf{e}_m = \mathbf{e}^{(a)}_m \parallel \mathbf{e}^{(s)}_m, \tag{7}$$

where $\parallel$ represents the concatenation operation. Thereafter, we employ a unified interaction prediction function to calculate the interaction probabilities of the dataset-ML method-pair:

$$\hat{y}_{dm} = \sigma(\Psi\ (\mathbf{e}_d \parallel \mathbf{e}_m)). \tag{8}$$

The function $\Psi(\cdot)$ is set as a NeuMF layer [9], which can model both linear and nonlinear interactions under both views uniformly. In addition, $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Moreover, benefiting from the node representation consistency, which is encouraged via cross-modal contrastive learning, the concatenation of different views has not led to a degradation in prediction performance (We present the detailed experimental results in subsubsection 5.4.1).

In order to strengthen the performance of ML method recommendation capability of CKGC, we leverage multi-task learning to jointly optimize the recommendation task and the pretextual self-supervised task (i.e., cross-modal contrastive learning) with the following combined loss function:

$$\mathcal{L} = \sum_{d \in \mathcal{D}, m \in \mathcal{M}} \mathcal{J}(\hat{y}_{dm}, y_{dm}) + \lambda_1 \mathcal{L}_{ssl} + \lambda_2 \|\Theta\|_2^2. \tag{9}$$

$\mathcal{J}$ is the cross-entropy function. λ_1 and λ_2 are hyperparameters controlling the relative strength of the self-supervised task and the L_2 regularization. Θ is the parameter set.

Table 1: Key statistics of the dataset.

Knowledge Graph		Interaction Dataset	
# Entities	17,483	# Datasets	2,092
# Descriptive Relations	12	# Methods	6,239
# Structural Relations	11	# Interactions	13,732
# Triples	117,245	Density	0.00105

5 EXPERIMENTAL EVALUATION

In this section, we conduct extensive experiments on a real-world dataset to evaluate the effectiveness of the proposed CKGC. We aim to answer the following research questions.

- **Q1**: How does CKGC perform in comparison to state-of-the-art knowledge graph-based recommendation models?
- **Q2**: What is the impact of key designs (e.g., descriptive attributes, cross-modal contrastive learning) on the CKGC?
- **Q3**: Can CKGC provide an in-depth analysis of cross-modal knowledge graph contrastive learning?

5.1 Real-World Dataset

Our experiments are based on the machine learning dataset in [2][2], which covers 19 machine learning domains (e.g., computer vision, natural language processing, reinforcement learning, graphs). It includes 2,093 ML task datasets, 7,644 ML methods, 517 ML tasks, 4,338 academic papers and 2,872 open source repositories, which together with their relationships constitute the machine learning knowledge graph. Following the practice in [2], after data cleaning and pre-processing, we retain 2,092 ML datasets and 6,239 ML methods to construct the interaction data. More detailed information about the dataset is shown in Table 1.

5.2 Experiment Settings

5.2.1 ***Comparison Baselines.*** We compare CKGC with seven state-of-the-art models, including traditional KG-free method (BPR), unified knowledge graph-based methods (KGCN, KGNN-LS, and KGAT), and methods that introduce multimodal information into the knowledge graph (CKE, MKGAT, and DEKR).

- **BPR** [17]: This method is based on the traditional matrix factorization model and optimized using Bayesian analysis.
- **KGCN** [25]:This method extends the non-spectral GCN approaches to the knowledge graph, which derives the representation of entities by selectively and biasedly aggregating their neighborhood information.
- **KGNN-LS** [24]: This method converts knowledge graphs into user-specific graphs and utilizes label smoothing regularization during information aggregation to generate personalized item embeddings.
- **KGAT** [26]: This method combines the user-item graph into the knowledge graph to obtain a collaborative knowledge graph, on which an attentive neighborhood aggregation mechanism is applied to generate the representation of users and items.
- **CKE** [32]: This method is a typical embedding-based approach, which combines the collaborative filtering with structural knowledge, textual knowledge, and visual knowledge in a unified framework for recommendation.
- **MKGAT** [19]: This method introduces multimodal knowledge graphs to the recommendation , in which a multimodal graph attention mechanism is performed for information propagation to derive aggregated embedding representations.
- **DEKR** [2]: This method constructs a description-enhanced knowledge graph, which captures the neighborhood structure and text description information of entities for the recommendation.

5.2.2 ***Evaluation Metrics.*** We conduct experiments on two tasks: 1) click-through rate (CTR) prediction, and 2) top-K recommendation. In the CTR prediction task, the commonly used metrics AUC, accuracy and F1-score are adopted. To evaluate the top-K recommendations task, we adopt precision@K, recall@K and ndcg@K metrics, where the value of K is set to 10 and 20 in the experiments.

5.2.3 ***Parameter Settings.*** We implement the proposed CKGC approach on Pytorch. We set the descriptive and structural embedding dimensions to 64, and the number of graph convolution iterations to 3. All methods adopt the Adam optimizer with batch size set to 128 for fair comparison. For hyperparameters such as learning rate, regularization coefficient, we perform a grid search to obtain the optimal settings of each comparison approach. In our proposed CKGC model, the learning rate searches in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ while λ_1 and λ_2 are tuned in $\{10^{-2}, 5\times10^{-2}, 10^{-1}, 5\times10^{-1}\}$ and $\{10^{-6}, 5\times10^{-6}, ..., 10^{-1}, 5\times10^{-1}\}$ respectively.

5.3 Performance Comparison (Q1)

The performance comparison results of all methods in CTR prediction and top-K recommendation are reported in Table 2. The main observations are as follows:

- CKGC consistently outperforms all baselines for all evaluation metrics in both the CTR prediction task and the top-K recommendation task. Specifically, CKGC improves over the best performing baseline DEKR by a large margin, especially on the top-K recommended tasks. The "Outperformance" row in Table 2 represents the percentage improvement of CKGC over the optimal baseline DEKR.
- In comparison with the traditional models, the performance of recommendations can be significantly improved by leveraging the knowledge graph (KG) as side information.
- The unified KG-based recommendation models outperform CKE on the CTR prediction task, which can be attributed to the capability of modeling higher-order connectivity. However, the CKE model achieves comparable performance with some unified models (e.g., KGCN, KGAT) on the top-K recommendation task. This is because in the CTR scenario, the rich information provided by higher-order connectivity helps to improve the efficiency of prediction. However, in the top-K recommendation scenario, the noise introduced by the higher-order connectivity can negatively impact the recommended results.

[2]https://github.com/cxsss/DEKR.

Table 2: Performance comparison results.

	CTR Prediction			Top-10 Recommendation			Top-20 Recommendation		
Models	AUC	Accuracy	F1-score	Precision	Recall	NDCG	Precision	Recall	NDCG
BPR	0.7518	0.6466	0.6862	0.0360	0.1132	0.1107	0.0205	0.1552	0.1227
KGCN	0.8112	0.7366	0.7435	0.0298	0.0896	0.0736	0.0232	0.1364	0.0904
KGNN-LS	0.8231	0.7221	0.7506	0.0309	0.0526	0.0374	0.0225	0.0724	0.0469
KGAT	0.8394	0.7396	0.7598	0.0591	0.1431	0.1209	0.0373	0.1932	0.1381
CKE	0.7684	0.6535	0.6966	0.0511	0.1178	0.0995	0.0342	0.1438	0.1135
MKGAT	0.8829	0.7638	0.7897	0.0615	0.1786	0.1317	0.0416	0.2601	0.1672
DEKR	0.9687	0.9172	0.9197	0.0642	0.2155	0.1598	0.0462	0.3268	0.1946
CKGC	**0.9851**	**0.9472**	**0.9453**	**0.0860**	**0.3603**	**0.2527**	**0.0715**	**0.5267**	**0.3097**
Outperformance	1.69%	3.27%	2.78%	34.38%	67.19%	57.51%	55.43%	61.17%	59.15%

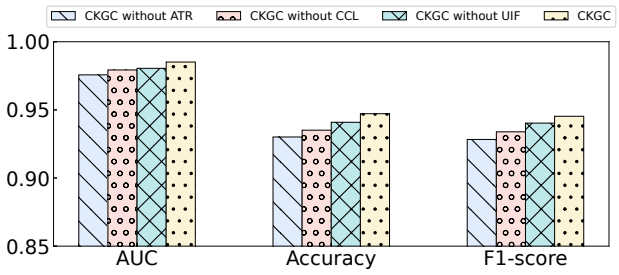

Figure 5: Performance comparison between CKGC and its variants.

- Compared with KGCN, KGNN-LS, and KGAT, MKGAT and DEKR achieve better performance by introducing multi-modal knowledge. Moreover, MKGAT performs weaker than DEKR because it merges different modalities into one representation, which may reduce the quality of entity representation since the representation spaces of various modalities tend to be different. Although the CKE model considers multimodal information as well, its simple utilization of knowledge representation limits the model performance.
- CKGC learns effective entity representations by considering more descriptive attributes and encouraging different modalities to provide supervisory signals to each other, thus further improving performance over DEKR.

5.4 Study of CKGC (Q2)

5.4.1 ***Ablation Study.*** To further explore the effectiveness of key designs in CKGC, three variants of CKGC are investigated in the ablation study:

(1) **CKGC without ATR**, which excludes the distinction of attribute-type relations (ATR);
(2) **CKGC without CCL**, which removes the cross-modal contrast learning (CCL) component;
(3) **CKGC without UIF**, which does not use the uniform interaction prediction function (UIF).

Table 3: Effect of relationship distinction.

Models	Precision@20	Recall@20	NDCG@20
CKGC with 25% attributes distinction	0.0310	0.2885	0.1628
CKGC with 50% attributes distinction	0.0355	0.3362	0.1685
CKGC with 75% attributes distinction	0.0525	0.5023	0.2721
CKGC regard all relations as attributes	0.0294	0.1635	0.1053
CKGC	**0.0715**	**0.5267**	**0.3097**

Here, only one key design is changed for each variant compared to the overall model. For the variant without UIF, descriptive and structural features are passed through MLP separately.

The results are shown in Figure 5. It is clear that removing any of these key design components results performance degradation. This shows the need for distinguishing descriptive attributes from structural connections as well as the positive effect of cross-modal contrast learning. In addition, it is worth noting that no performance degradation has occurred as a result of merging the learned representations of the two modal views into a unified framework for prediction. This suggests that the consistency of the node representations has been increased by mutual supervision between the two views.

5.4.2 ***Effect of Attribute Discrimination.*** To evaluate the effectiveness of distinguishing traditional relationships into descriptive attributes and structural connections, we performed experiments with different distinction ratios. For all descriptive attributes finally distinguished in the CKGC model, we set the distinction to 25%, 50%, and 75%, respectively, and we also set a case to treat all traditional relations as descriptive attributes. Considering that different attributes may have varying degrees of contribution, for each proportion, we randomly selected the corresponding number of attributes in five times and calculated the average of the results.

It can be observed in Table 3 that the recommended performance is enhanced as the distinction ratio increases from 25% to 50% to 75%, and the optimal performance is achieved when all the descriptive attributes are distinguished by CKGC. This indicates that treating descriptive attributes as structural connections is not conducive to learning higher-quality node representations. However, if all the traditional relations are treated as descriptive attributes, the

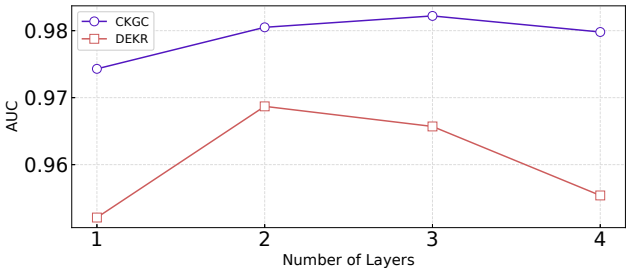

Figure 6: AUC performance under different layer settings.

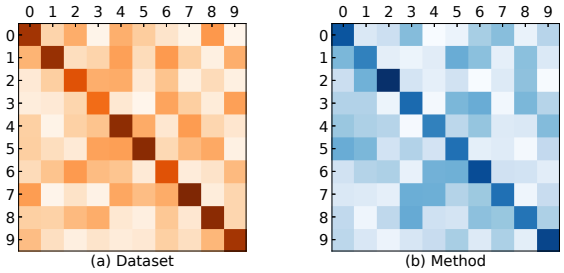

Figure 7: Visualization of cosine similarity between the embedding representations from two modal views.

performance degrades severely, due to the fact that the entitys' structural information is ignored in this case.

5.5 In-depth Analysis (Q3)

5.5.1 ***Effect of Model Depth.*** Learning high-quality node representations is the key of knowledge graph-based recommendations. To explore the impact on node representation learning while using descriptive and structural representations of nodes for cross-modal contrastive learning, we analyze the performance of CKGC when the graph neural network is stacked with different numbers of layers (equal to the order of the furthest neighbors), and compare it with the best performing existing approach DEKR.

As shown in Figure 6, CKGC outperforms DEKR when the number of layers is set to 1, which can be attributed to its richer descriptive information. When the number of layers increases from 1 to 2, the performance of both models improves. This reflects the facilitative effect of higher-order connectivity. However, the performance of DEKR decreases when the layer number increases to 3. A similar trend can be observed for CKGC with layers of 4. The reason might be that more noise has been brought in by farther neighbors. Nevertheless, CKGC has only a small drop in performance when stacking more layers, which demonstrates its ability to preserve the distinctiveness of the nodes due to the supervision provided by the descriptive modal view for the structural modal view.

5.5.2 ***Case Study.*** To intuitively demonstrate the influence of cross-modal contrastive learning on entity representation learning, we randomly selected 10 dataset entities and 10 method entities. Specifically, we computed the cosine similarity between the final descriptive view representation and the structural view representation pair by pair. The resulting matrix is visualized in Figure 7, where darker shades of color indicate larger values.

From Figure 7 can find that, in most cases, the diagonal values (i.e., the cosine similarity computed from two views of the same entity) are the highest. In contrast, the majority of the non-diagonal values (i.e., the cosine similarity calculated from two views of different entities) are lower. Obviously, in general, the values on the diagonal are significantly higher than those on the non-diagonal. This indicates that cross-modal contrastive learning encourages the improvement of the consistency from different view representations for a given method or dataset through the supervisory role between different modalities. It is worth pointing out that, the two view representations of the same entity do not reach the maximum cosine similarity taking value, that is, the case where the two representations are identical. This suggests that the complementary information of the respective parts is also preserved while encouraging the consistency of both.

6 CONCLUSION AND FUTURE WORK

In this paper, we focus on the information overload in the field of machine learning method recommendation and proposed the Cross-modal Knowledge Graph Contrastive learning (CKGC) approach to address this challenge. We offer a unique insight by classifying the relationships in machine learning knowledge graphs into descriptive attributes and structural connections, so that the information reflecting the distinctiveness of nodes is fully preserved together with the node topology. Based on this insight, we incorporate two encoders to learn descriptive view representations and structural view representations into CKGC. The mutual information between them is then maximized to achieve cross-modal knowledge graph contrastive learning. In this way, more generalized node representations can be learned, which in turn, can enhance ML method recommendation performance. Extensive experiments on a real-world dataset have shown significant advantages of CKGC over the state-of-the-art approaches.

This work explores multimodal information such as descriptive attributes and text as well as considers the role of cross-modal supervision. In future work, we plan to consider more modal information and focus more on taking into account the type of descriptive information together with fusing the different modalities in a more deliberate manner. In addition, our approach can not only be applied to the field of machine learning, but can also be extended to the scenario of general domain recommendation.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their rigorous comments. This work is supported by the Key Research and Development Plan of Shandong Province (Major Scientific and Technological Innovation Project) (2021CXGC010103). Han Yu is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-019); the Joint NTU-WeBank Research Centre on Fintech (Award No: NWJ-2020-008); the Nanyang Assistant Professorship (NAP); the RIE 2020 Advanced Manufacturing and Engineering (AME) Programmatic Fund (No. A20G8b0102), Singapore; and Future Communications Research & Development Programme (FCP-NTU-RG-2021-014).

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR*.

[2] Xianshuai Cao, Yuliang Shi, Han Yu, Jihu Wang, Xinjun Wang, Zhongmin Yan, and Zhiyong Chen. 2021. DEKR: Description Enhanced Knowledge Graph for Machine Learning Method Recommendation. *SIGIR* (2021), 203–212.

[3] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2020. Revisiting Graph based Collaborative Filtering: A Linear Residual Graph Convolutional Network Approach. *ArXiv* abs/2001.10167 (2020).

[4] Zhen Chen, Feng Xia, Huizhen Jiang, Haifeng Liu, and Jun Zhang. 2015. AVER: Random Walk Based Academic Venue Recommendation. *WWW* (2015).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

[6] Travis Ebesu and Yi Fang. 2017. Neural Citation Network for Context-Aware Citation Recommendation. *SIGIR* (2017).

[7] Michael U. Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*.

[8] Xiangnan He, Kuan Deng, Xiang Wang, Yaliang Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *SIGIR* (2020).

[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. *Proceedings of the 26th International Conference on World Wide Web* (2017).

[10] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning* 14, 1-2 (2021), 1–210.

[11] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. 2014. ACRec: a co-authorship based random walk model for academic collaboration recommendation. *WWW* (2014).

[12] Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2016. Knowledge Representation Learning with Entities, Attributes and Relations. In *IJCAI*.

[13] Zheng Liu, Xing Xie, and Lei Chen. 2018. Context-aware Academic Collaborator Recommendation. *KDD* (2018).

[14] Jarana Manotumruksa, Craig MacDonald, and Iadh Ounis. 2017. A Deep Recurrent Collaborative Filtering Framework for Venue Recommendation. *CIKM* (2017).

[15] Jarana Manotumruksa, Craig MacDonald, and Iadh Ounis. 2018. A Contextual Attention Recurrent Architecture for Context-Aware Venue Recommendation. *SIGIR* (2018).

[16] Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 2014. ClusCite: effective citation recommendation by information network-based clustering. *KDD* (2014).

[17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*.

[18] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and Philip S. Yu. 2019. Heterogeneous Information Network Embedding for Recommendation. *IEEE Transactions on Knowledge and Data Engineering* 31 (2019), 357–370.

[19] Rui Sun, Xuezhi Cao, Yan Zhao, Juncheng Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal Knowledge Graphs for Recommender Systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020).

[20] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. 2012. Cross-domain collaboration recommendation. In *KDD*.

[21] Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*.

[22] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. *CIKM* (2018).

[23] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep Knowledge-Aware Network for News Recommendation. *Proceedings of the 2018 World Wide Web Conference* (2018).

[24] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. 2019. Knowledge-aware Graph Neural Networks with Label Smoothness Regularization for Recommender Systems. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).

[25] Hongwei Wang, Miao Zhao, Xing Xie, Wenjie Li, and Minyi Guo. 2019. Knowledge Graph Convolutional Networks for Recommender Systems. *WWW* (2019).

[26] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).

[27] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph and Text Jointly Embedding. In *EMNLP*.

[28] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. 2017. Big Scholarly Data: A Survey. *IEEE Transactions on Big Data* 3 (2017), 18–35.

[29] Han Xiao, Minlie Huang, Lian Meng, and Xiaoyan Zhu. 2017. SSP: Semantic Space Projection for Knowledge Graph Embedding with Text Descriptions. In *AAAI*.

[30] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *IJCAI*.

[31] Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2016. Representation Learning of Knowledge Graphs with Hierarchical Types. In *IJCAI*.

[32] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. *KDD* (2016).

[33] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Lee. 2017. Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017).