A Mutual Information Perspective on Knowledge Graph Embedding

Jiang Li, Xiangdong Su*, Zehua Duo, Tian Lan, Xiaotao Guo, Guanglai Gao
[1] College of Computer Science, Inner Mongolia University, China
[2] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China
[3] Inner Mongolia Key Laboratory of Multilingual Artiffcial Intelligence Technology, China
lijiangimu@gmail.com,cssxd@imu.edu.cn

Abstract

Knowledge graph embedding techniques have emerged as a critical approach for addressing the issue of missing relations in knowledge graphs. However, existing methods often suffer from limitations, including high intra-group similarity, loss of semantic information, and insufficient inference capability, particularly in complex relation patterns such as 1-N and N-1 relations. To address these challenges, we introduce a novel KGE framework that leverages mutual information maximization to improve the semantic representation of entities and relations. By maximizing the mutual information between different components of triples, such as (h, r) and t, or (r, t) and h, the proposed method improves the model's ability to preserve semantic dependencies while maintaining the relational structure of the knowledge graph. Extensive experiments on benchmark datasets demonstrate the effectiveness of our approach, with consistent performance improvements across various baseline models. Additionally, visualization analyses and case studies demonstrate the improved ability of the MI framework to capture complex relation patterns.

1 Introduction

Knowledge graphs (KGs) represent a structured form of storing and organizing knowledge, which has demonstrated immense potential in various applications such as search engines (Yang et al., 2019), intelligent question-answering systems (Saxena et al., 2022), and recommendation systems (Wang et al., 2024). KGs are typically represented as triples (h, r, t), where h denotes the head entity, r represents the relation, and t denotes the tail entity. However, real-world KGs often contain a significant number of missing relations, which has led to extensive research on knowledge graph embedding (KGE).

KGE techniques aim to map entities and relations into a low-dimensional vector space, enabling more effective inference and prediction, thereby addressing the challenge of missing relations in KGs. This typically involves predicting missing links within a KG. For instance, given a triple in the form of $(h, r, ?)$, the model predicts the correct tail entity t based on the head entity h and relation r. Similarly, for $(?, r, t)$, the model predicts the correct head entity h that based on the relation r and tail entity t. Existing KGE methods perform geometric operations in the embedding space, transforming entities and relations into actionable vector representations. For instance, TransE (Bordes et al., 2013) uses translation operations in real vector space to represent entities and relations. ComplEx (Trouillon et al., 2016) introduces a complex vector space. QuatE (Zhang et al., 2019) further employs rotation operations to encode KGs in the quaternion space.

Despite significant progress made by existing methods, several key limitations remain. **Firstly**, high intra-group similarity is a major concern. For 1-N or N-1 relation patterns, these methods often embed multiple entities related to the same head or tail entity in a similar manner, thereby obscuring potential semantic differences between tail or head entities and limiting the model's reasoning capability. **Secondly**, loss of semantic information is prevalent. These methods typically focus on minimizing the distance between (h, r) and t or between (t, r) and h in the embedding space, emphasizing the modeling of structural information in KGs by employing translation or rotation operations, while neglecting the preservation of semantic information. This bias can lead to the omission of important semantic characteristics of individual entities, thus weakening the model's performance in capturing complex semantic relations.

To address these challenges, we propose a novel KGE approach based on a mutual information

* Corresponding Author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22152–22166

framework. Mutual information (MI), a measure of dependency between two variables, enables the model to capture deep semantic connections between entities and relations. The motivation for our method is drawn from its (MI) proven success in various deep learning tasks. For instance, Do et al. (2021) demonstrated that maximizing mutual information between different views facilitates efficient clustering, while Peng et al. (2020) introduced Graphical Mutual Information (GMI) for graph representation learning, showing that this technique effectively retains rich structural information in unsupervised learning. Furthermore, Wu et al. (2023) applied mutual information to align themes across languages in cross-lingual topic modeling, preventing representation degradation and enabling accurate multilingual topic alignment. Inspired by these tasks, we leverage mutual information maximization to enhance the representation of semantic information in KGs. By maximizing the mutual information between different components of triples (e.g., (h, r) and t, or (r, t) and h), our method not only preserves geometric structural information but also improves the model's ability to represent complex semantic relationships. Experimental results indicate that the proposed KGE method based on mutual information achieves significant performance improvements in KGE tasks. In summary, our study provides a novel theoretical foundation and practical approach for KGE, introducing a mutual information maximization strategy that enhances the model's ability to capture complex semantic relationships and improve inference performance.

2 Related Work

2.1 Knowledge Graph Embedding Methods

Knowledge Graph Embedding (KGE) techniques transform the entities and relationships in knowledge graphs into low-dimensional vector spaces. These embeddings are optimized to preserve the structural and semantic information of the graph. Early methods, such as TransE (Bordes et al., 2013), TransH (Wang et al., 2014), TransR (Lin et al., 2015), PaiRE (Chao et al., 2021) and CompoundE (Ge et al., 2023) encode KG embedding in the Euclidean space. Recent work, ExpressivE (Pavlovic and Sallinger, 2023) enhances Euclidean models by modeling relations as geometric regions, enabling the capture of a wider range of relational patterns within Euclidean space.

To better capture complex relation patterns, KGE methods have explored alternative embedding spaces. RotatE (Sun et al., 2019), ComplEx (Trouillon et al., 2016) model KGs in the complex vector space. QuatE (Zhang et al., 2019), QuatRE (Nguyen et al., 2022), QuatSE (Li et al., 2022), TransERR (Li et al., 2024) DCNE (Dong et al., 2024) and DaBR (Wang et al., 2025) extend this by leveraging quaternion spaces to represent interactions between entities and relations. These methods demonstrate significant improvements in encoding diverse relationship patterns. However, they often suffer from high intra-group similarity and limited semantic preservation.

In recent years, contrastive learning-based KGE methods (Luo et al., 2021; Hu et al., 2024) have gained attention for learning more discriminative representations by contrasting positive and negative samples. While our work also draws on contrastive learning ideas, it differs by maximizing the mutual information between the query (h, r) and the target entity t, combined with minimizing conditional entropy, thus presenting an information-theory-driven optimization objective.

Another key line of KGE research leverages non-Euclidean embedding spaces, including hyperbolic, spherical, and mixed-curvature geometries, to better capture hierarchy and cyclicity structural patterns. Poincaré Embeddings (Nickel and Kiela, 2017) first demonstrated that hyperbolic space effectively models hierarchical structures due to its exponential capacity. Building on this, RefH, RotH, and AttH (Chami et al., 2020) extend hyperbolic embeddings with reflection- and rotation-based operators combined with attention mechanisms to model diverse relation types. HGCN (Chami et al., 2019) further generalizes graph convolutional networks to hyperbolic space using Riemannian geometry, enabling inductive reasoning over hierarchical KGs. To model mixed structural patterns, DGS (Iyer et al., 2022) embeds cyclic relations in spherical space and hierarchical ones in hyperbolic space, linking them via a shared bridge space. Similarly, NMM (Iyer et al., 2024) combines spherical and hyperbolic embeddings to capture homophily and influence, respectively. Its extension, NMM-GNN, introduces a non-Euclidean variational autoencoder with a space unification loss, achieving strong results on KG prediction tasks.

Our proposed method builds upon these approaches by integrating mutual information maximization, which enhances the discriminative ca-

pacity of entity embeddings by aligning relational contexts more effectively. This work specifically focuses on models defined in real, complex, and quaternion spaces, while non-Euclidean embeddings are left for future investigation.

2.2 Mutual Information Maximization

Mutual Information (MI) is a fundamental concept in information theory that quantifies the dependency between two random variables. It has been widely applied across domains to model variable interactions and enhance representation learning (Bachman et al., 2019; Hjelm et al., 2018; Kong et al., 2019; Chi et al., 2020; Wu et al., 2023). For instance, in clustering tasks, maximizing MI across different data views has been shown to improve cluster quality by encouraging compact intra-cluster representations and distinct inter-cluster separations (Do et al., 2021). Similarly, in topic modeling, methods like InfoCTM (Wu et al., 2023) leverage MI to align cross-lingual topic representations, thereby reducing topic redundancy and improving coherence. In graph representation learning, Graphical Mutual Information (GMI) maximization has been used to preserve both local and global graph structure. Furthermore, in cross-lingual scenarios, maximizing MI prevents degenerate representations, ensuring robust alignment across languages. Song et al. (2024) integrate global and local knowledge constraints to enhance the pre-trained language model's ability to comprehend query contexts. These studies illustrate that MI maximization not only enhances traditional representation learning tasks but also plays an essential role in cross-domain applications, especially in multilingual and graph-structured data contexts. While prior KGE methods focus on improving relational modeling through enhanced score functions or embedding spaces, they often neglect the underlying semantic dependencies among entities and relations. Our work introduces a novel perspective by incorporating MI maximization into the KGE framework.

3 Methodology

3.1 Problem Setting and Notations

A knowledge graph is typically denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$, $\mathcal{R}$ and $\mathcal{T}$ represent the sets of entities, relations and triples (h, r, t), respectively. The goal of KGE is to compute a score function $f_r(h, t)$ that assigns higher scores to valid triples than to invalid ones. Specifically, KGE aims to learn low-dimensional representations of entities and relations to facilitate missing link prediction in KGs. It involves predicting the tail entity t given a tuple $(h, r, ?)$, or conversely, predicting the head entity h for a tuple $(?, r, t)$.

In traditional KGE methods, such as TransE, entity and relation vector representations are learned by minimizing an score function. These methods tend to embed multiple entities associated with the same head or tail entity into nearby spatial positions, leading to a high intra-group similarity problem. This phenomenon is particularly pronounced when dealing with 1-N and N-1 relations, hindering the model's ability to distinguish between different tail or head entities.

Specifically, for 1-N relations, given a head entity h and a relation r, there exist multiple tail entities $\{t_1, t_2, \ldots, t_N\}$ associated with them. The objective function of traditional embedding methods usually takes the form:

$$E(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|, \tag{1}$$

where $\mathbf{h}$, $\mathbf{r}$, and $\mathbf{t}$ represent the embedding vectors of the head entity, relation, and tail entity, respectively. To minimize the score function, the model adjusts all tail entity embeddings $\{\mathbf{t}_i\}$ to be as close as possible to $\mathbf{h} + \mathbf{r}$. Consequently, these tail entity embeddings are also very close to each other, i.e.,

$$\|\mathbf{t}_i - \mathbf{t}_j\| \approx 0, \quad \forall i, j \in \{1, 2, \ldots, N\}, \tag{2}$$

which leads to high intra-group similarity, making it difficult for the model to distinguish between different tail entities. A similar situation applies to N-1 relations, resulting in high intra-group similarity among head entities. Similarly, models like RotatE, ComplEx, QuatE, etc. also face these potential issues.

3.2 Mutual Information Maximization for Reducing Intra-Group Similarity

To address this issue, we introduce a mutual information maximization strategy. MI measures the dependency between two random variables. By maximizing the mutual information between the conditional variables and the target variable, the model can capture richer dependency structures, thereby enhancing its discriminative ability.

Proposition 1. *Maximizing the mutual information $I((h,r);t)$ between (h,r) and t for 1-N relations and $I((r,t);h)$ between (r,t) and h for N-1 relations reduces the conditional entropy $H(t|h,r)$ and $H(h|r,t)$, thereby improving the model's ability to distinguish between entities.*

Proof. Let $I((h,r);t)$ denote the mutual information between (h,r) and t. Mutual information is formally defined as:

$$I((h,r);t) = H(t) - H(t|h,r), \tag{3}$$

where $H(t)$ is the entropy of the tail entity, and $H(t|h,r)$ is the conditional entropy of t given (h,r).

Maximizing $I((h,r);t)$ is equivalent to minimizing $H(t|h,r)$, as $H(t)$ is independent of the model parameters because it reflects the marginal distribution of t determined by the dataset and does not depend on the conditional relationships that the model learns.

The conditional entropy $H(t|h,r)$ is defined as:

$$H(t|h,r) = -\sum_{t\in\mathcal{E}_t} p(t|h,r)\log p(t|h,r), \tag{4}$$

where $\mathcal{E}_t$ is the set of all possible tail entities, and $p(t|h,r)$ is the probability of selecting t given (h,r).

By minimizing $H(t|h,r)$, the model reduces the uncertainty in predicting the tail entity t given the head entity h and the relation r. This optimization encourages the model to assign higher probabilities to the correct t and lower probabilities to incorrect ones, making $p(t|h,r)$ more focused and distinctive. As a result, the embeddings of tail entities $\{\mathbf{t}_i\}$ are adjusted to reflect their unique relationships with (h,r), spreading them apart in the embedding space. This separation reduces overlap among similar tail entities, effectively lowering high intra-group similarity and improving the model's ability to distinguish between them.

Similarly, for N-1 relations, maximizing $I((r,t);h)$ is equivalent to minimizing $H(h|r,t)$, which is defined as:

$$H(h|r,t) = -\sum_{h\in\mathcal{E}_h} p(h|r,t)\log p(h|r,t), \tag{5}$$

where $\mathcal{E}_h$ is the set of all possible head entities. By minimizing $H(h|r,t)$, the model enhances the discriminability among head entities for a given (r,t).

In summary, by introducing a mutual information maximization strategy and minimizing conditional entropy, we enhance the model's ability to capture subtle differences between entities, alleviating the high intra-group similarity problem in 1-N and N-1 relations. This method is mathematically proven to be effective and improves the discriminative capability of entity representations in the embedding space, thereby promoting the reasoning performance of the model.

3.3 Mutual Information Lower Bound Optimization Using InfoNCE

In the previous section, we demonstrated that mutual information can effectively reduce uncertainty in entity predictions by minimizing conditional entropy, thereby enhancing the model's discriminative power. However, directly optimizing mutual information is generally intractable in neural network-based models. In our approach, we employ two lower bound to approximate mutual information maximization, including InfoNCE and Jensen-Shannon Divergence (JSD).

Since directly computing mutual information $I(U,\hat{U})$ is often infeasible (Song and Ermon, 2019), researchers typically optimize its variational lower bounds. Among these bounds, the InfoNCE lower bound (Oord et al., 2018; Logeswaran and Lee, 2018; Poole et al., 2019) has proven to be highly effective in practice. The InfoNCE bound is formally defined as:

$$\begin{aligned} &I(U,\hat{U}) \geq I_{\text{InfoNCE}}(U,\hat{U}) \triangleq \\ &\mathbb{E}_{p(u_{1:N})p(\hat{u}|u_1)}\left[\log\frac{\exp(f(\hat{u},u_1))}{\sum_{j=1}^{N}\exp(f(\hat{u},u_j))}\right] + \log N \\ &= -\mathcal{L}_{\text{contrast}} + \log N, \end{aligned} \tag{6}$$

where U and $\hat{U}$ represent random variables from two distinct views. $u_{1:N}$ are N samples drawn from the distribution p_U, and $\hat{u}$ is a positive sample associated with u_1 from the distribution $p_{\hat{U}}$. The pair $(\hat{u},u_1)$ is defined as a positive pair, while the pairs $(\hat{u},u_j)$ $(j=2,\ldots,N)$ are treated as negative pairs. The function $f(x,y)$, often referred to as a "critic," measures the similarity between two representations x and y. The term $\mathcal{L}_{\text{contrast}}$ is widely recognized as the contrastive loss in previous works (Tian et al., 2020; Chen et al., 2020). The bound arises from the fact that:

$$\log\frac{\exp(f(\hat{u},u_1))}{\sum_{j=1}^{N}\exp(f(\hat{u},u_j))} \leq 0. \tag{7}$$

Thus, the InfoNCE lower bound $I_{\text{InfoNCE}}(U, \hat{U})$ is upper-bounded by $\log N$. This implies two key properties: (i) The InfoNCE bound becomes loose when the true mutual information $I(U, \hat{U})$ significantly exceeds $\log N$. (ii) Increasing the number of negative samples N tightens the bound, providing a better approximation of the true mutual information.

Despite its bias, the InfoNCE lower bound has much lower variance compared to other unbiased mutual information estimators (Poole et al., 2019). This low-variance property ensures stable training, making InfoNCE a practical and reliable choice for mutual information maximization in neural network-based frameworks.

Specifically, we aim to maximize the mutual information $I((h, r); t)$ between (h, r) and t, as well as the mutual information $I((r, t); h)$ between (r, t) and h. To maximize the mutual information between (h, r) and t, we leverage a contrastive learning strategy to improve the model's representation of tail entities. Specifically, the positive pair (h, r, t^+) is sampled from real triples in the knowledge graph, while the negative pairs (h, r, t^-) are generated by randomly replacing the tail entity. The sample set S is defined as follows:

$$S = \{(h, r, t^+), (h, r, t_1^-), (h, r, t_2^-), \ldots, (h, r, t_n^-)\}, \tag{8}$$

where (h, r, t^+) is the positive sample, and $\{(h, r, t_1^-), (h, r, t_2^-), \ldots, (h, r, t_n^-)\}$ are n negative samples. To optimize the distinction between positive and negative samples, we define the following contrastive learning loss function:

$$\mathcal{L}_{\text{contrast}}^{(h,r)\to t} = -\mathbb{E}_S \left[\log \frac{\exp(f(h, r, t^+))}{\exp(f(h, r, t^+)) + \sum_{i=1}^{n} \exp(f(h, r, t_i^-))} \right], \tag{9}$$

where $f(h, r, t)$ represents the matching score between the head entity and relation for the tail entity. By minimizing this loss function, we effectively maximize a lower bound on the mutual information between (h, r) and t, thereby improving the model's ability to predict tail entities given the head entity and relation.

Similarly, for the relation and tail entity pair (r, t), we aim to maximize the mutual information between them and the head entity h, denoted as $I((r, t); h)$. We define the sample set as follows:

$$S = \{(r, t, h^+), (r, t, h_1^-), (r, t, h_2^-), \ldots, (r, t, h_n^-)\}, \tag{10}$$

and the corresponding contrastive learning loss function is:

$$\mathcal{L}_{\text{contrast}}^{(r,t)\to h} = -\mathbb{E}_S \left[\log \frac{\exp(f(r, t, h^+))}{\exp(f(r, t, h^+)) + \sum_{i=1}^{n} \exp(f(r, t, h_i^-))} \right]. \tag{11}$$

By minimizing this loss function, the model maximizes the mutual information between (r, t) and h, thus enhancing the ability to distinguish among head entities. The total contrastive learning objective is given by the weighted sum of the two parts:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{contrast}}^{(h,r)\to t} + \lambda_2 \mathcal{L}_{\text{contrast}}^{(r,t)\to h}, \tag{12}$$

where λ_1 and λ_2 are weight parameters used to balance the two mutual information maximization objectives.

The final optimization objective combines the original loss of the KGE models with the mutual information maximization strategies described above. Specifically, the total loss can be expressed as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{KGE}} + \mathcal{L}_{\text{total}}, \tag{13}$$

where $\mathcal{L}_{\text{KGE}}$ represents the original loss function of the baseline KGE models, and $\mathcal{L}_{\text{total}}$ is the mutual information-based InfoNCE defined in the previous section. By combining the structural loss $\mathcal{L}_{\text{orig}}$ and the mutual information maximization loss $\mathcal{L}_{\text{total}}$, the model simultaneously preserves the structural characteristics of the knowledge graph while enhancing the semantic distinctions between entities and relations.

3.4 Mutual Information Lower Bound Optimization Using JSD

To further enhance the model's ability to capture the complex dependencies between entities and relations in a knowledge graph, we propose an alternative optimization approach using Jensen-Shannon Divergence (JSD) to estimate the mutual information lower bound. JSD is a symmetric and effective metric that can measure the difference between two probability distributions. In my work, I

aim to estimate and maximize the mutual information between the joint representation of the head entity and relation ($\mathbf{H}_q$) and the representation of the tail entity ($\mathbf{H}_{[T]}$), using a JSD-based lower bound estimator. Specifically, the joint representation of the head entity and relation $\mathbf{H}_q$ and the representation of the tail entity $\mathbf{H}_{[T]}$ are treated as two random variables whose representations are drawn from potentially different distributions, and the goal is to capture their correlations by maximizing mutual information. According to the theoretical foundation provided by MINE (Mutual Information Neural Estimation) (Belghazi et al., 2018), the mutual information can be estimated through a lower bound:

$$I(\mathbf{H}_q; \mathbf{H}_{[T]}) \geq \hat{I}_{\text{JSD}}(\mathbf{H}_q; \mathbf{H}_{[T]}). \tag{14}$$

To this end, we introduce the Jensen-Shannon Mutual Information Estimator as follows:

$$\begin{aligned}\hat{I}_\theta^{(\text{JSD})}(\mathbf{H}_q, \mathbf{H}_{[T]}) := \mathbb{E}_P[-\text{sp}(T_\theta(\mathbf{H}_q, \mathbf{H}_{[T]}))] \\ -\mathbb{E}_{P'}[\text{sp}(T_\theta(\mathbf{H}'_q, \mathbf{H}_{[T]}))],\end{aligned} \tag{15}$$

Where $\mathbf{H}_q$ represents the joint representation of the head entity and relation, and $\mathbf{H}_{[T]}$ represents the tail entity's representation. $\mathbf{H}'_q$ is a representation sampled from other queries in the same mini-batch, serving as a negative sample. The function $T_\theta(\cdot)$, parameterized as a neural network or designed based on a scoring function in KGE models, serves as a discriminator to assign a compatibility score to the input representations. $\text{sp}(x) = \log(1 + e^x)$ is the softplus activation function, used to ensure non-negativity and smooth gradient calculation.

Our optimization goal is to maximize the above mutual information lower bound to enhance the model's representation learning. The optimization objective is given by:

$$\theta = \arg\max_\theta \frac{1}{|\mathcal{G}|} \sum_{b_j \in G} \hat{I}_\theta^{(\text{JSD})}(\mathbf{H}_q^{b_j}, \mathbf{H}_{[T]}^{b_j}), \tag{16}$$

where b_j is mini batch from training dadsets. To optimize the model parameters, we define the following loss function:

$$\mathcal{L}_{JSD} = \sum_{b_j \in \mathcal{G}} (E_\theta^{P'} - E_\theta^P), \tag{17}$$

where E_θ^P represents the expectation for the query and tail entity, and $E_\theta^{P'}$ denotes the expectation for negative distribution. The final optimization objective is obtained by integrating the JSD-based loss into the original KGE framework as follows:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{KGE}} + \lambda_3 \mathcal{L}_{\text{JSD}}, \tag{18}$$

where λ_3 is a weighting coefficient. By minimizing this loss function, the model effectively enhances its ability to distinguish between positive and negative samples and maximizes the mutual information between entities and relations, thereby improving the quality of knowledge embeddings.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed methods, we conduct experiments on two widely used benchmark datasets for knowledge graph embedding tasks: WN18RR (Dettmers et al., 2018) and FB15K-237 (Toutanova and Chen, 2015). A summary of these datasets is presented in Appendix A.

4.2 Evaluation Protocol

Following the standard link prediction evaluation protocol used in previous studies, we evaluate the proposed approaches by ranking each test triple against all possible substitutions of head and tail entities: (h', r, t) and (h, r, t'), $\forall h', t' \in \mathcal{E}$, where $\mathcal{E}$ is the set of all entities in the knowledge graph. We follow the standard evaluation protocol, including metrics such as Mean Reciprocal Rank (MRR) and Hits@K (K = 1, 3, 10), which are widely used in knowledge graph completion tasks. MRR measures the average reciprocal rank of the ground-truth entity in the ranked list of predictions. Higher MRR values indicate better ranking performance. Hits@K calculates the proportion of test triples for which the ground-truth entity appears in the top K predictions.

5 Results and Analysis

5.1 Overall Results

In this section, We select representative works from different representation spaces to validate the effectiveness of our proposed model: TransE (Bordes et al., 2013) in the real space, RotatE (Sun et al., 2019) and ComplEx (Trouillon et al., 2016) in the complex space, QuatE (Zhang et al., 2019) in the quaternion space, and CP (Lacroix et al.,

	WN18RR				FB15K-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE (Bordes et al., 2013)	19.46	**3.48**	32.39	46.28	28.71	19.89	31.63	46.68
TransE-MI	**20.50**	2.09	**36.07**	**48.61**	**30.28**	**21.56**	**33.38**	**47.61**
CP (Lacroix et al., 2018)	41.60	**39.34**	42.36	45.88	26.74	18.93	28.90	42.59
CP-MI	**42.12**	38.64	**42.96**	**49.27**	**29.69**	**21.77**	**32.28**	**45.46**
RESCAL (Nickel et al., 2011)	42.81	40.87	43.41	46.51	29.75	21.48	**32.67**	**46.16**
RESCAL-MI	**43.36**	**41.40**	**44.37**	**49.12**	**29.83**	**21.90**	32.40	45.43
ComplEx (Trouillon et al., 2016)	43.22	41.07	43.78	47.53	28.38	20.31	30.89	44.76
ComplEx-MI	**43.61**	**41.24**	**44.40**	**47.93**	**31.28**	**22.76**	**34.10**	**48.40**
RotatE (Sun et al., 2019)	43.51	**41.31**	44.24	48.01	27.61	19.68	29.78	43.88
RotatE-MI	**43.70**	41.19	**44.46**	**48.82**	**31.15**	**22.80**	**34.08**	**47.88**
QuatE (Zhang et al., 2019)	44.74	**42.39**	45.56	49.27	30.61	22.14	33.51	47.51
QuatE-MI	**45.08**	42.20	**45.90**	**50.65**	**32.21**	**23.44**	**35.20**	**50.07**

Table 1: Link prediction results on WN18RR and FB15K-237 datasets based on InfoNCE.

	WN18RR				FB15K-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE (Bordes et al., 2013)	**19.46**	**3.48**	**32.39**	**46.28**	28.71	19.89	31.63	46.68
TransE-MI	19.26	3.40	31.80	46.20	**29.91**	**21.24**	**32.99**	**47.08**
CP (Lacroix et al., 2018)	41.60	39.34	42.36	45.88	26.74	18.93	28.90	42.59
CP-MI	**42.46**	**40.11**	**43.30**	**47.00**	**28.87**	**21.17**	**31.37**	**44.21**
RESCAL (Nickel et al., 2011)	42.81	**40.87**	43.41	46.51	29.75	21.48	**32.67**	**46.16**
RESCAL-MI	**42.99**	40.83	**43.86**	**47.02**	**29.94**	**21.96**	32.56	45.77
ComplEx (Trouillon et al., 2016)	43.22	41.07	43.78	47.53	28.38	20.31	30.89	44.76
ComplEx-MI	**43.88**	**41.38**	**44.70**	**48.61**	**30.49**	**22.14**	**33.42**	**47.21**
RotatE (Sun et al., 2019)	43.51	41.31	44.24	48.01	27.61	19.68	29.78	43.88
RotatE-MI	**44.08**	**41.34**	**45.28**	**49.46**	**30.81**	**22.58**	**33.50**	**47.20**
QuatE (Zhang et al., 2019)	44.74	**42.39**	45.56	49.27	30.61	22.14	33.51	47.51
QuatE-MI	**45.02**	42.36	**45.90**	**50.24**	**32.01**	**23.33**	**34.90**	**49.56**

Table 2: Link prediction results on WN18RR and FB15K-237 datasets based on JSD.

2018) and RESCAL (Nickel et al., 2011) based on tensor decomposition methods. Table 1 and Table 2 present the link prediction results on the WN18RR and FB15K-237 datasets, showcasing the performance improvements of models when incorporating the MI maximization framework, both using the InfoNCE lower bound and JSD lower bound. All models were trained within a unified framework, with a dimension of 100, a maximum of 50 epochs, and no additional regularization functions. All experimental results are derived from our experiments. In most cases, the MI maximization approach consistently outperforms baseline models, with the "Model-MI" versions achieving higher MRR and Hits@K (K = 1, 3, 10) across all evaluated models. In both datasets, the consistent performance improvements can be attributed to the MI framework's ability to retain semantic dependencies between entities and relations, resulting in richer and more robust representations. The results validate the effectiveness of the MI maximization approach in enhancing the performance of KGE models.

5.2 Complex Relations Modeling

Table 3 presents the results of the MI-enhanced models on two types of complex relation patterns, 1-N and N-1, on the FB15K-237 dataset. For 1-N relations, where a single head entity is associated with multiple tail entities, the MI framework significantly improves performance. For example, TransE-MI achieves an MRR of 30.17, com-

	FB15K-237					
	1-N			N-1		
	MRR	H@1	H@10	MRR	H@1	H@10
TransE	26.28	16.07	47.84	32.23	23.47	50.20
TransE-MI	**30.17**	**19.87**	**51.32**	**34.77**	**25.73**	**52.63**
CP	25.30	15.94	44.89	31.34	23.03	48.10
CP-MI	**30.51**	**20.87**	**50.26**	**35.12**	**26.72**	**51.79**
RESCAL	28.29	18.41	48.23	33.84	25.26	50.56
RESCAL-MI	**30.51**	**20.88**	**50.12**	**35.31**	**26.93**	**51.76**
ComplEx	27.19	17.52	47.28	32.59	24.17	49.45
ComplEx-MI	**31.11**	**20.96**	**51.91**	**36.10**	**27.35**	**53.68**
RotatE	25.95	16.45	45.93	31.74	23.46	48.54
RotatE-MI	**31.50**	**21.43**	**52.34**	**36.15**	**27.48**	**53.54**
QuatE	29.69	19.46	50.61	34.98	26.15	52.38
QuatE-MI	**31.82**	**21.47**	**53.46**	**36.56**	**27.70**	**54.53**

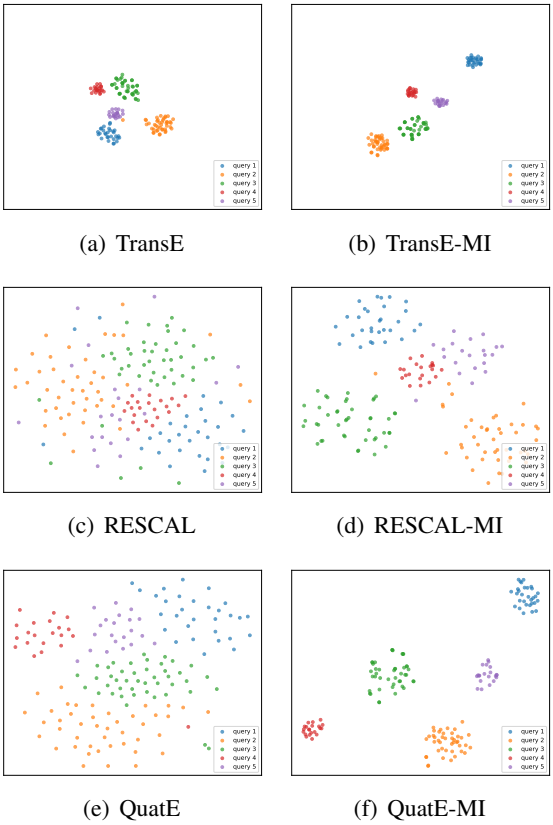
Table 3: Link prediction results on Different Relation Types (1-N and N-1) on FB15K-237.

pared to 26.28 for TransE, while Hits@10 increases from 47.84 to 51.32. Similarly, ComplEx-MI and RotatE-MI improve MRR from 27.19 to 31.11 and from 25.95 to 31.50, respectively. These improvements highlight the MI framework's ability to reduce high intra-group similarity among tail entities by better preserving semantic distinctions.

For N-1 relations, where multiple head entities share the same tail entity, the MI-enhanced models again show consistent improvements. TransE-MI increases MRR from 32.23 to 34.77, and Hits@1 rises from 23.47 to 25.73, demonstrating improved precision in ranking head entities. More advanced models, such as RESCAL-MI, show notable gains, with MRR increasing from 33.84 to 35.31 and Hits@1 improving from 25.26 to 26.93. These results demonstrate that the MI framework enhances the performance of knowledge graph embedding models by better capturing complex relational dependencies and improving the entity and relation representations.

5.3 Visualization Analysis

Figure 1 illustrates the T-SNE (Van der Maaten and Hinton, 2008) visualization of tail entity embeddings for various models before and after applying MI maximization. In this visualization, each point represents a tail entity, and points of the same color belong to tail entities that share the same (h, r) context (1-N relations). Across all baseline models, embeddings before MI enhancement display significant overlap among points of different colors, indicating high intra-group similarity and insufficient semantic distinction between tail entities. After incorporating the MI framework, the embeddings demonstrate a more distinct clustering structure for tail entities within the same (h, r) context, with reduced overlap between different clusters. For example, TransE-MI exhibits a clearer separation of clusters compared to TransE, while RESCAL-MI and QuatE-MI show tighter and more coherent clusters, reflecting their enhanced ability to capture semantic dependencies and mitigate high intra-group similarity. Overall, these results demonstrate that the MI framework effectively enhances the semantic relationships between entities and relations, resulting in more meaningful representations. Please refer to Appendix D for additional model visualizations.



(a) TransE (b) TransE-MI (c) RESCAL (d) RESCAL-MI (e) QuatE (f) QuatE-MI

Figure 1: Visualization of the embeddings of tail entities using T-SNE. A point represents a tail entity. Points in the same color represent tail entities that have the same (h, r) context (1-N).

5.4 A Case Study

To provide deeper insights into the impact of MI maximization on knowledge graph embedding models, we present a case study on the FB15K-237 dataset. Table 4 illustrates the result of RotatE and RotatE-MI in predicting the tail entity for the query

Query	(Dena Higley, people/person/profession, ?)	
Answer	writer	
Model	**RotatE**	**RotatE-MI**
Rank 1	actor	• **writer**
Rank 2	film director	actor
Rank 3	author	television producer
Rank 4	• **writer**	journalist
Rank 5	television producer	film director
Rank 6	journalist	author

Table 4: A Case Study of Tail Prediction on FB15K-237. The • refers to the correct answer.

(Dena Higley, people/person/profession, ?), where the correct answer is writer. For the baseline model, RotatE ranks writer in the fourth position, behind other incorrect predictions such as actor, film director, and author. In contrast, the MI-enhanced model, RotatE-MI, successfully ranks writer at the top position. The improvement can be attributed to the mutual information framework, which better preserves the semantic dependencies between entities and relations. By maximizing the mutual information between (h, r) and t, the MI framework enables the model to better capture the relational context of the query and reduce confusion among closely related professions such as actor and author. Overall, the results validate the effectiveness of mutual information maximization in improving the precision of tail entity predictions in knowledge graph completion tasks.

6 Conclusion

In this study, we proposed a mutual information maximization framework to address challenges in KGE, such as high intra-group similarity and semantic information loss. By enhancing the mutual information between relational components, our approach improves the discriminative power of entity and relation embeddings. Experimental results demonstrate that MI-enhanced models outperform their baselines in link prediction metrics like MRR and Hits@K. Improved clustering and semantic separation in the embedding space further validate the framework's effectiveness in handling complex 1-N and N-1 relation patterns. In conclusion, our MI framework provides a general and effective enhancement to KGE models, improving both semantic representation and reasoning capabilities.

Limitations

In this study, we explore the integration of mutual information maximization into knowledge graph embedding tasks to address challenges like high intra-group similarity and semantic information loss. Similar to most existing KGE models, our method cannot generalize to unseen entities that were absent during training, which remains an important direction for future research.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), National Education Science Planning Project (Grant No. BIX230343), The Central Government Fund for Promoting Local Scientiffc and Technological Development (Grant No. 2022ZY0198), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Inner Mongolia Autonomous Region Science and Technology Planning Project (Grant No. 2023YFSH0017), Hohhot Science and Technology Project (Grant No. 2023-Zhan-Zhong-1), Science and Technology Program of the Joint Fund of Scientiffc Research for the Public Hospitals of Inner Mongolia Academy of Medical Sciences (Grant No.2023GLLH0035), The Project of Innovation Research in Postgraduate at Inner Mongolia University (11200-5223737).

References

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. 2020. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545*.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32.

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)*, pages 4360–4369, Online. Association for Computational Linguistics.

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*, pages 1597–1607. PMLR.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press.

Kien Do, Truyen Tran, and Svetha Venkatesh. 2021. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9928–9938.

Yao Dong, Qingchao Kong, Lei Wang, and Yin Luo. 2024. Dual complex number knowledge graph embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 5391–5400. ELRA and ICCL.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12:2121–2159.

Xiou Ge, Yun Cheng Wang, Bin Wang, and C-C Jay Kuo. 2023. Compounding geometric operations for knowledge graph completion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6947–6965.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Jie Hu, Hongqun Yang, Fei Teng, Shengdong Du, and Tianrui Li. 2024. A knowledge graph completion model based on triple level interaction and contrastive learning. *Pattern Recognition*, 156:110783.

Roshni Iyer, Yewen Wang, Wei Wang, and Yizhou Sun. 2024. Non-euclidean mixture model for social network embedding. *Advances in Neural Information Processing Systems*, 37:111464–111488.

Roshni G Iyer, Yunsheng Bai, Wei Wang, and Yizhou Sun. 2022. Dual-geometric space embedding model for two-view knowledge graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 676–686.

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2019. A mutual information maximization perspective of language representation learning. *arXiv preprint arXiv:1910.08350*.

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *International Conference on Machine Learning*, pages 2863–2872. PMLR.

Jiang Li, Xiangdong Su, Xinlan Ma, and Guanglai Gao. 2022. Quatse: Spherical linear interpolation of quaternion for knowledge graph embeddings. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 209–220. Springer.

Jiang Li, Xiangdong Su, Fujun Zhang, and Guanglai Gao. 2024. TransERR: Translation-based knowledge graph embedding via efficient relation rotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16727–16737, Torino, Italia. ELRA and ICCL.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2181–2187. AAAI Press.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Zhiping Luo, Wentao Xu, Weiqing Liu, Jiang Bian, Jian Yin, and Tie-Yan Liu. 2021. Kge-cl: Contrastive learning of tensor decomposition based knowledge graph embeddings. *arXiv preprint arXiv:2112.04871*.

Dai Quoc Nguyen, Thanh Vu, Tu Dinh Nguyen, and Dinh Phung. 2022. Quatre: Relation-aware quaternions for knowledge graph embeddings. In *Companion Proceedings of the Web Conference 2022*, pages 189–192.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 809–816. Omnipress.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Aleksandar Pavlovic and Emanuel Sallinger. 2023. Expressive: A spatio-functional embedding for knowledge graph completion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. 2020. Graph representation learning via graphical mutual information maximization. In *Proceedings of The Web Conference 2020*, pages 259–270.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR.

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2814–2828. Association for Computational Linguistics.

Jiaming Song and Stefano Ermon. 2019. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*.

Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao Yu, and Jun Zhao. 2024. Multilingual knowledge graph completion from pretrained language models with knowledge constraints. *arXiv preprint arXiv:2406.18085*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 57–66.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Lingzhi Wang, Shafiq Joty, Wei Gao, Xingshan Zeng, and Kam-Fai Wong. 2024. Improving conversational recommender system via contextual and time-aware modeling with less domain-specific knowledge. *IEEE Transactions on Knowledge and Data Engineering*.

Weihua Wang, Qiuyu Liang, Feilong Bao, and Guanglai Gao. 2025. Distance-adaptive quaternion knowledge graph embedding with bidirectional rotation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4219–4231, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1112–1119. AAAI Press.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023. Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13763–13771.

Yueji Yang, Divykant Agrawal, HV Jagadish, Anthony KH Tung, and Shuang Wu. 2019. An efficient parallel keyword search engine on knowledge graphs. In *2019 IEEE 35th international conference on data engineering (ICDE)*, pages 338–349. IEEE.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2731–2741.

A Statistics of the Benchmark Datasets

We summarize the detailed information of these datasets in Table 5.

WN18RR WN18RR is a subset of WordNet, which is a lexical database of English. The dataset was designed to address issues of test set leakage in the original WN18 dataset. It contains 93,003 triples split into 86,835 for training, 3,034 for validation, and 3,134 for testing. WN18RR focuses on more challenging reasoning tasks by excluding reversible relations and ensuring that the evaluation requires compositional reasoning. Entities represent synsets, and relations describe lexical or semantic connections between synsets.

FB15K-237 FB15K-237 is derived from Freebase, a large-scale knowledge graph containing real-world facts. Similar to WN18RR, FB15K-237 addresses test set leakage by removing near-duplicate triples from the original FB15K dataset. It consists of 310,116 triples with 14,541 entities and 237 relations, split into 272,115 for training, 17,535 for validation, and 20,466 for testing.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#Train	#Valid	#Test
WN18RR	40,943	11	86,835	3,034	3,134
FB15K-237	14,541	237	272,115	17,535	20,466

Table 5: Statistics of the WN18RR and FB15K-237.

B Experimental Setup

We implemented our model using Python and PyTorch library. All experiments are trained on a single NVIDIA Tesla V100 with 32GB memory. We use Adagrad (Duchi et al., 2011) optimizer and the best hyperparameters based on the performance on the validation datasets. The learning rate is set in [0.1, 0.01] in all cases, the embedding dimension d is set in [100, 512, 1000, 2000], the batch size is set in [200, 1000]. The best models are selected by early stopping on the validation datasets, and the max epoch is 50. We search the weights λ_1, λ_2 and λ_3 in Eq. 12 and Eq. 18 in [1, 0.1, 0.01,0.001]. We search the temperature τ for InfoNCE in [0.2, 0.5, 0.8, 1.0].

In our implementation, we employ a self-supervised contrastive learning framework. By optimizing contrastive losses such as InfoNCE or JSD, this framework aims to maximize the mutual information lower bound $I(\mathbf{z}_{hr}; \mathbf{z}'_{hr})$. Here, $\mathbf{z}_{hr} = \phi(\mathbf{e}_h, \mathbf{e}_r)$ is the composed query representation derived from an input triplet (h, r, t_{key}), and $\mathbf{z}'_{hr} = \phi(\mathbf{e}_{h'}, \mathbf{e}_{r'})$ is its positive counterpart.

The theoretical rationale for this approach, connecting the optimization of $I(\mathbf{z}_{hr}; \mathbf{z}'_{hr})$ to the KGE objective related to t_{key}, is as follows. Maximizing $I(\mathbf{z}_{hr}; \mathbf{z}'_{hr})$ through InfoNCE forces the query representation $\mathbf{z}_{hr}$ to be invariant to augmentations. We hypothesize that these augmentations are specifically designed to perturb features that are irrelevant to t_{key}, allowing the model to focus on the core semantic features of $\mathbf{z}_{hr}$ that are crucial for predicting t_{key}. This process can be understood as the "purification" of $\mathbf{z}_{hr}$, which reduces the influence of noise and non-essential information, thus enhancing the model's ability to capture and represent the dependencies between (h, r) and t_{key}.

In this study, the KGE task aims to optimize $\mathbf{z}_{hr}$ for accurate prediction of t_{key}. From an information-theoretic perspective, this corresponds to maximizing $I(\mathbf{z}_{hr}; t_{\text{key}})$, which can be achieved by minimizing the conditional entropy $H(t_{\text{key}}|\mathbf{z}_{hr})$, thereby reducing prediction uncertainty. Hence, by refining $\mathbf{z}_{hr}$ through mutual information maximization, we enhance its ability to capture relevant dependencies, directly improving KGE performance. Thus, by optimizing $I(\mathbf{z}_{hr}; \mathbf{z}'_{hr})$, we indirectly enhance $I(\mathbf{z}_{hr}; t_{\text{key}})$ by "purifying" $\mathbf{z}_{hr}$, making it more robust and semantically focused. Our code is available at https://github.com/dellixx/KGE-MI.

C Additional Experimental Results

Table 6 and Table 7 report the link prediction results on the WN18RR and FB15K-237 datasets, where the MI maximization framework is implemented using both InfoNCE and JSD lower bounds. Across all baseline models, the "Model-MI" versions consistently outperform their original counterparts in terms of metrics. We further enhance the performance of our models by incorporating additional regularization functions, which help to better capture the dependencies between entities and relations. The results for baseline models are derived from the original papers. On WN18RR, ComplEx-MI and RotatE-MI achieve significant gains in both MRR and Hits@K. Similar trends are observed on FB15K-237, with notable improvements in MRR and Hits@10 for models such as TransE-MI and QuatE-MI. The JSD-based models also show consistent improvements across both datasets. For ex-

	WN18RR				FB15K-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE (Bordes et al., 2013)	.226	-	-	.501	.294	-	-	.465
TransE-MI	.244	.055	.407	.516	.308	.218	.340	.491
CP (Lacroix et al., 2018)	.438	.414	.445	.485	.333	.247	.363	.508
CP-MI	.466	.421	.480	.559	.367	.271	.404	.557
RESCAL (Nickel et al., 2011)	.455	.419	.461	.493	.353	.264	.385	.528
RESCAL-MI	.509	.466	.526	.590	.368	.272	.406	.550
ComplEx (Trouillon et al., 2016)	.460	.428	.473	.522	.346	.256	.386	.525
ComplEx-MI	.498	.453	.516	.583	.355	.261	.393	.542
RotatE (Sun et al., 2019)	.476	.428	.492	.571	.338	.241	.375	.533
RotatE-MI	.496	.453	.512	.577	.354	.262	.390	.539
QuatE (Zhang et al., 2019)	.481	.436	.500	.564	.311	.221	.342	.495
QuatE-MI	.490	.447	.504	.574	.348	.255	.381	.534

Table 6: Link prediction results on WN18RR and FB15K-237 datasets based on InfoNCE.

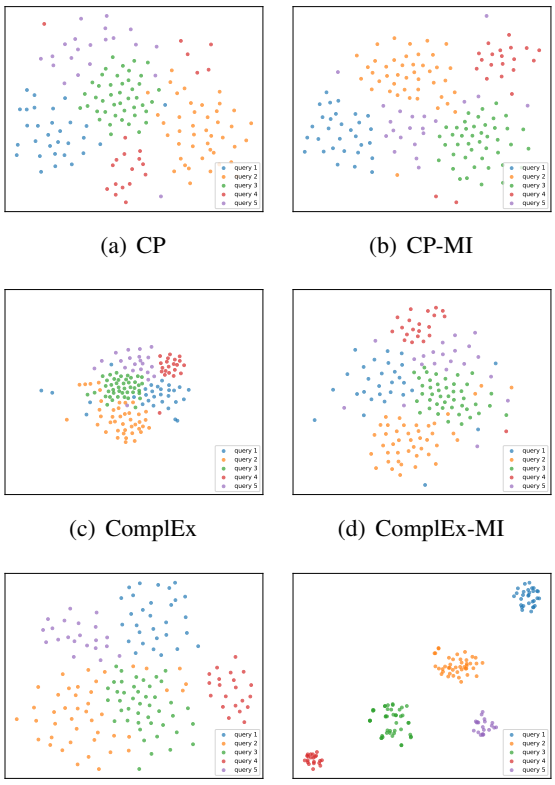
	WN18RR				FB15K-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE (Bordes et al., 2013)	.226	-	-	.501	.294	-	-	.465
TransE-MI	.240	.053	.398	.514	.303	.211	.336	.487
CP (Lacroix et al., 2018)	.438	.414	.445	.485	.333	.247	.363	.508
CP-MI	.462	.420	.478	.550	.362	.268	.401	.545
RESCAL (Nickel et al., 2011)	.455	.419	.461	.493	.353	.264	.385	.528
RESCAL-MI	.502	.463	.521	.588	.364	.270	.397	.548
ComplEx (Trouillon et al., 2016)	.460	.428	.473	.522	.346	.256	.386	.525
ComplEx-MI	.489	.444	.509	.580	.355	.261	.393	.542
RotatE (Sun et al., 2019)	.476	.428	.492	.571	.338	.241	.375	.533
RotatE-MI	.491	.450	.508	.566	.346	.260	.388	.534
QuatE (Zhang et al., 2019)	.481	.436	.500	.564	.311	.221	.342	.495
QuatE-MI	.488	.446	.503	.562	.347	.249	.376	.532

Table 7: Link prediction results on WN18RR and FB15K-237 datasets based on JSD.

ample, TransE-MI achieves an MRR of 0.240 on WN18RR, a 6.2% increase over TransE. JSD optimization benefits advanced models like ComplEx-MI and RotatE-MI, improving their ability to capture entity-relation dependencies more effectively. On FB15K-237, QuatE-MI achieves an MRR of 0.347, demonstrating its ability to improve tail entity prediction precision.

D Additional Visualization and Analysis

Figure 2 presents the visualization of tail entity embeddings for CP (MI), ComplEx (MI), and RotatE (MI) models before and after applying MI maximization. Each point represents a tail entity, and points of the same color correspond to tail entities sharing the same (h, r) context (1-N relations). For the baseline models (CP, ComplEx, and RotatE), the embeddings show substantial overlap among points of different colors, indicating high intra-group similarity and a lack of semantic distinction between tail entities. After incorporating the MI framework, the enhanced models (CP-MI, ComplEx-MI, and RotatE-MI) exhibit clearer and more distinct clustering structures. Specifically, tail entities within the same (h, r) context are more tightly grouped, while points from different groups are more effectively separated. For example, CP-MI and ComplEx-MI display better-defined cluster boundaries compared to their original versions, while RotatE-MI achieves a significant reduction in intra-group overlap, resulting in highly distinguishable clusters. These results highlight the ability of the MI framework to improve the semantic organization of embeddings, leading to more meaningful



(a) CP (b) CP-MI

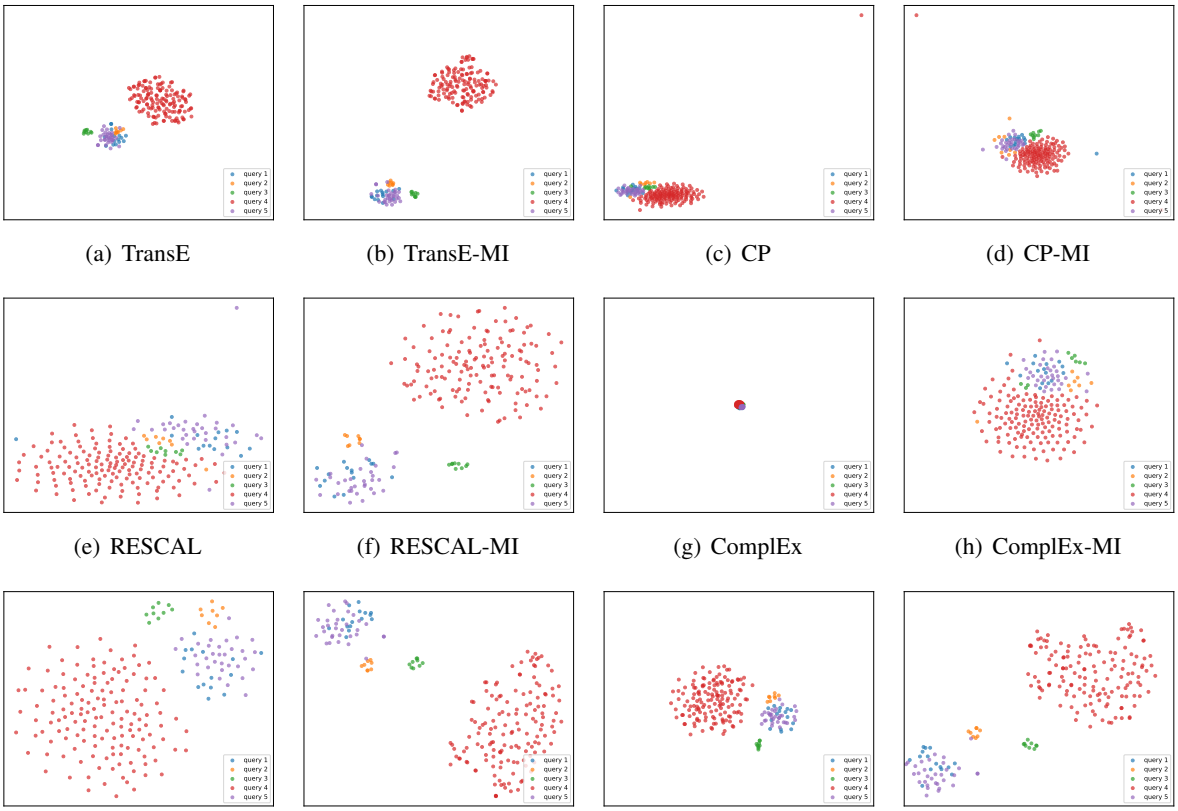
(c) ComplEx (d) ComplEx-MI

(e) RotatE (f) RotatE-MI

Figure 2: Visualization of the embeddings of tail entities using T-SNE for CP (MI), ComplEx (MI) and RotatE (MI). A point represents a tail entity. Points in the same color represent tail entities that have the same (h, r) context (1-N).

and discriminative representations.

Figure 3 illustrates the visualization of head entity embeddings for various models under the N-1 context before and after applying MI maximization. In the baseline models, embeddings show significant overlap among different groups (colors), indicating low discriminability of head entities. After incorporating the MI framework, the enhanced models (e.g., TransE-MI, RESCAL-MI, and RotatE-MI) exhibit better separation and clearer clustering, effectively reducing intra-group similarity and improving the differentiation of head entity representations.



(a) TransE (b) TransE-MI (c) CP (d) CP-MI

(e) RESCAL (f) RESCAL-MI (g) ComplEx (h) ComplEx-MI

(i) RotatE (j) RotatE-MI (k) QuatE (l) QuatE-MI

Figure 3: Visualization of the embeddings of head entities using T-SNE. A point represents a head entity. Points in the same color represent head entities that have the same (r, t) context (N-1).