MQuinE: a Cure for "Z-paradox" in Knowledge Graph Embedding

Yang Liu, Huang Fang
Cognitive Computing Lab, Baidu Research
Beijing, China
{liuyang173,fanghuang}@baidu.com

Yunfeng Cai, Mingming Sun
Beijing Institute of Mathematical Sciences and Applications
Beijing, China
{caiyunfeng,sunmingming}@bimsa.cn

Abstract

Knowledge graph embedding (KGE) models achieved state-of-the-art results on many knowledge graph tasks including link prediction and information retrieval. Despite the superior performance of KGE models in practice, we discover a deficiency in the expressiveness of some popular existing KGE models called *Z-paradox*. Motivated by the existence of Z-paradox, we propose a new KGE model called *MQuinE* that does not suffer from Z-paradox while preserves strong expressiveness to model various relation patterns including symmetric/asymmetric, inverse, 1-N/N-1/N-N, and composition relations with theoretical justification. Experiments on real-world knowledge bases indicate that Z-paradox indeed degrades the performance of existing KGE models, and can cause more than 20% accuracy drop on some challenging test samples. Our experiments further demonstrate that MQuinE can mitigate the negative impact of Z-paradox and outperform existing KGE models by a visible margin on link prediction tasks.

1 Introduction

Knowledge graphs (KGs) consist of many facts that connect real-world entities (e.g., humans, events, words, etc.) with various relations. Each fact in a knowledge graph is usually represented as a triplet (h, r, t), where h, t are respectively the head and tail entities and r is the relation; the triplet (h, r, t) indicates that the head entity h has the relation r to the tail entity t. Due to the prevalence of relational data in practice, KG has a wide range of applications including recommendation systems (Wang et al., 2018; Ma et al., 2019; Wang et al., 2019), natural language processing (NLP) (Sun et al., 2018), question answering (QA) (Huang et al., 2019) and querying (Chen et al., 2022).

Embedding-based models (Bengio et al., 2003; Blei et al., 2003) have revolutionized certain fields of machine learning including KG in the past two decades. Simply speaking, knowledge graph embedding (KGE) models map each entity and relation into a vector or matrix and calculate the probability of a fact triple through some score functions. KGE models are space and time efficient. More importantly, KGE models such as TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), OTE (Tang et al., 2020), etc., are quite expressive; it was shown that KGE models if designed carefully, can capture various relation patterns including symmetry/asymmetry, inversion, composition, injective and non-injective relations. Due to the efficiency and expressiveness of KGE models, embedding-based models have achieved state-of-the-art performance on many KG applications and are widely deployed in practice.

Despite the popularity of KGE models on various KG applications, in this work, we discover a bottleneck, termed as "Z-paradox", of the expressiveness of some existing KGE models. Below we give a short description and illustration of Z-paradox and present its formal definition and some related properties in Section 3.

Z-paradox. Though popular KGE models (e.g., TransE, RotatE, OTE) have already taken various relation patterns into account, there are still limitations. In what follows we introduce a limitation of popular KGE models. Specifically, in Figure 1.1, there are four entities e_1, e_2, e_3, e_4, with e_1 linking to e_2, e_3 linking to both e_2 and e_4. The task is to determine whether e_1 links to e_4 or not. A good KGE model should permit both

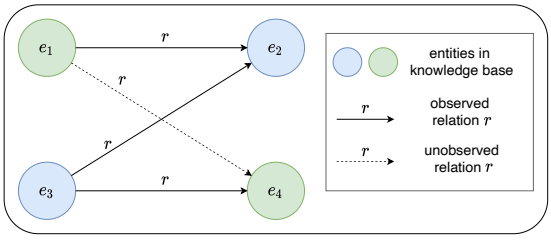

Figure 1.1: An illustration of Z-paradox

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9837–9850
November 12-16, 2024

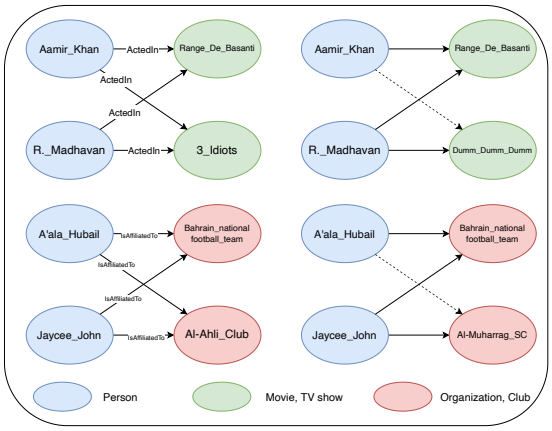

Figure 1.2: Illustration of Z-paradox in YAGO3-10.

scenarios. However, we find that many popular KGE models such as TransE (Bordes et al., 2013) and RotatE (Sun et al., 2019) would guarantee e_1 links to e_4 regardless of whether e_1 actually has relation r to e_4 or not. We term this phenomenon *Z-paradox* due to the graph structure in Figure 1.1. To be more concrete, let us take some examples from the YAGO3-10 knowledge base, where some actors and movies are connected by the relation *ActedIn*. We illustrate the phenomenon via Figure 1.2, where the dotted arrow means that popular KGE models infer that the arrow-tail links to the arrowhead, which contradicts with the true facts. In the upper left plot, we observe that *Aamir_Khan* has acted in *Range_De_Basanti*, *R._Madhaven* has acted in *Range_De_Basanti* and *3_Idiots*, then the KGE model would infer that *Aamir_Khan* has acted in *3_Idiots*, which is correct. However, in the upper right plot, we observe that *Aamir_Khan* has acted in *Range_De_Basanti*, *R._Madhaven* has acted in *Range_De_Basanti* and *Dumm_Dumm_Dumm*, then the KGE model would also infer that *Aamir_Khan* has acted in *Dumm_Dumm_Dumm*, which is incorrect. The same phenomenon occurs in the lower left and right plots.

We demonstrate that the Z-pattern is indeed a serious issue for standard KG benchmark datasets. For example, about 35% of the test facts in the FB15k-237 dataset are negatively affected by the Z-pattern, and KGE models such as TransE and RotatE can suffer more than 20% accuracy drop on these test facts; see Section 5.3 and Section 5.4 for details. To mitigate the negative impact of the Z-pattern, we propose a new KGE model to overcome the *Z-paradox*. Moreover, the new model can ensure both the robust expressiveness and the ability to model various relation patterns, i.e., preserves the good properties of existing KGE models. The new model embeds a triplet (h, r, t) by five matrices $(\mathbf{H}, \langle \mathbf{R}^h, \mathbf{R}^t, \mathbf{R}^c \rangle, \mathbf{T})$, where the matrices $\mathbf{H}, \mathbf{T}$ denote the embeddings of the head entity h and tail entity t respectively, the matrix triplet $\langle \mathbf{R}^h, \mathbf{R}^t, \mathbf{R}^c \rangle$ represents the embedding of the relation r. We term the new model **M**atrix **Quin**tuple **E**mbedding (MQuinE). We show that MQuinE enjoys good theoretical properties and can achieve promising empirical results. Compared with existing KGE models on the challenging FB15k-237 dataset, MQuinE obtains a 10% improvement of Hit@10 on test facts that are negatively impacted by the Z-pattern, and attains 7% and 4% overall improvement of Hit@1 and Hit@10 on all test facts.

Our contributions are summarized as follows:

1) A newly-defined phenomenon in the knowledge graph named *Z-paradox* has been discovered and we prove that existing translation-based KGE models all suffer from Z-paradox. Theoretically, we present a necessary condition for the occurrence of Z-paradox.

2) We propose MQuinE, a new KGE model that is free from Z-paradox meanwhile can still model complex relations including (a)symmetric, inverse, 1-N/N-1/N-N, and composition relations.

3) Experimental results of MQuinE on standard benchmark datasets validate that MQuinE can indeed overcome the negative impact of Z-paradox; MQuinE outperforms existing KGE methods by a large margin on most benchmark datasets.

2 Related works

We summarize some popular KGE models in Table C.2. We go through some existing KGE methods and discuss how they relate to our work.

Translation distance based methods. Translation distance based approaches evaluate the plausibility of fact triples by comparing the distances between the head and tail entity embeddings after some relation transformations. Inspired by *word2vec* (Mikolov et al., 2013), Bordes et al. (2013) first introduced the idea of translation invariance into the knowledge graph embedding domain and proposed the TransE model. Sun et al. (2019) proposed RotatE and characterized relations as rotations between the head and tail entities in complex space; it was shown that many desirable properties, such as symmetry/asymmetry, inversion, and Abelian composition, can be achieved by RotatE. Tang et al. (2020) extended RotatE from the 2-

dimensional complex domain to high-dimensional space. Lu and Hu (2020) proposed DensE to better model the complex composition relation patterns.

Bilinear semantic matching methods. Nickel et al. (2011) first introduced the idea of tensor decomposition to model triple-relational data. Yang et al. (2014) later proposed a simple and effective bilinear model called DisMult and achieved promising empirical results. Subsequent works such as ComplEX (Trouillon et al., 2016), TuckER (Balažević et al., 2019), DihEdral (Xu and Li, 2019), QuatE (Zhang et al., 2019) and SEEK (Xu et al., 2020) adopted more complicated bilinear operations to either improve the expressiveness of DisMult or decrease the model complexity.

Deep learning methods. Vashishth et al. (2020) proposed COMPGCN to incorporate multi-relational information into graph convolutional networks which leverages a variety of composition operations from knowledge graph embedding techniques to embed both nodes and relations in a graph jointly. Dettmers et al. (2018) proposed ConvE and used convolutional neural networks to model multi-relational data. Subsequent works (Nathani et al., 2019; Vashishth et al., 2020) brought more advanced neural network architectures such as graph convolutional networks and graph attention networks. More recently, Wang et al. (2021) proposed M^{2}GNN and embeds entities and relations into the mixed-curvature space with trainable heterogeneous curvatures. Zhou et al. (2022) proposed JointE and adopted both 1-dimensional and 2-dimensional convolution operations to capture the latent knowledge more carefully.

3 Z-paradox and its cure: MQuinE

In this section, we give a formal definition of Z-paradox and propose a new KGE model called MQuinE that can circumvent Z-paradox while having strong expressiveness.

3.1 Z-paradox

Definition 1 (Z-paradox)**.** Given a KGE model parameterized by $\{\boldsymbol{e}_i\}_{i=1}^{|\mathcal{E}|}, \{\boldsymbol{r}_i\}_{i=1}^{|\mathcal{R}|}$ and a score function $s(\cdot)$ such that $s^* := \inf s$. For any $e_1, e_2, e_3, e_4 \in \mathcal{E}, r \in \mathcal{R}$, if

$$s(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) = s(\boldsymbol{e}_3, \boldsymbol{r}, \boldsymbol{e}_2) = s(\boldsymbol{e}_3, \boldsymbol{r}, \boldsymbol{e}_4) = s^* \tag{1}$$

implies that $s(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_4) = s^*$ must hold, then we say the KGE model suffers from Z-paradox.

Consider a KGE model that suffers from Z-paradox. If $e_1 \to e_2, e_3 \to e_2, e_3 \to e_4$, i.e., (1) holds, then $e_1 \to e_4$ must holds regardless the fact is true or not. It is obvious that a KGE model that suffers from Z-paradox has an inherent deficiency in its expressiveness, and a good KGE model should be able to circumvent Z-paradox. Next, we show that a wide range of existing KGE models indeed suffer from Z-paradox and therefore have limited expressiveness.

Proposition 3.1. *Given a KGE model parameterized by $\{\boldsymbol{e}_i\}_{i=1}^{|\mathcal{E}|}, \{\boldsymbol{r}_i\}_{i=1}^{|\mathcal{R}|}$ and a score function $s(\cdot)$. If $s(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ can be expressed as*

$$s(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) = \|f(\boldsymbol{h}, \boldsymbol{r}) - g(\boldsymbol{t}, \boldsymbol{r})\|$$

for some functions $f(\cdot)$ and $g(\cdot)$, and $s^* := \inf s = 0$, then the KGE model suffers from Z-paradox.

Proof. First, we notice that $s(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) = s^* = 0$ implies $f(\boldsymbol{h}, \boldsymbol{r}) = g(\boldsymbol{t}, \boldsymbol{r})$. For four entities e_1, e_2, e_3, e_4 satisfying

$$s(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) = s(\boldsymbol{e}_3, \boldsymbol{r}, \boldsymbol{e}_2) = s(\boldsymbol{e}_3, \boldsymbol{r}, \boldsymbol{e}_4) = 0,$$

we have

$$\begin{aligned} f(\boldsymbol{e}_1, \boldsymbol{r}) &= g(\boldsymbol{e}_2, \boldsymbol{r}) \\ f(\boldsymbol{e}_3, \boldsymbol{r}) &= g(\boldsymbol{e}_2, \boldsymbol{r}) \\ f(\boldsymbol{e}_3, \boldsymbol{r}) &= g(\boldsymbol{e}_4, \boldsymbol{r}). \end{aligned}$$

Then it follows that

$$\begin{aligned} & s(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_4) \\ =& \|f(\boldsymbol{e}_1, \boldsymbol{r}) - g(\boldsymbol{e}_4, \boldsymbol{r})\| \\ =& \|[f(\boldsymbol{e}_1, \boldsymbol{r}) - g(\boldsymbol{e}_2, \boldsymbol{r})] - [f(\boldsymbol{e}_3, \boldsymbol{r}) - g(\boldsymbol{e}_2, \boldsymbol{r})] \\ & + [f(\boldsymbol{e}_3, \boldsymbol{r}) - g(\boldsymbol{e}_4, \boldsymbol{r})]\| = 0, \end{aligned}$$

i.e., (e_1, r, e_4) holds. This completes the proof. □

Remark 3.1. Proposition 3.1 indicates that all existing translation-based KGE models suffer from Z-paradox, including TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), OTE (Tang et al., 2020) and MQuadE (Yu et al., 2021), etc.

Remark 3.2. KGE models with bilinear score functions may also suffer from Z-paradox under certain conditions. This is because bilinear score functions can be transformed into distance-based score functions. For example, we have

$$2\mathbf{h}^T\mathbf{R}\mathbf{t} = -\|\mathbf{R}\mathbf{t} - \mathbf{h}\|^2 + \|\mathbf{h}\|^2 + \|\mathbf{R}\mathbf{t}\|^2.$$

When $\mathbf{h}, \mathbf{t}$ both have fixed norms and $\mathbf{R}$ is an orthogonal matrix, then maximizing $\mathbf{h}^T\mathbf{R}\mathbf{t}$ is equivalent to minimizing $\|\mathbf{R}\mathbf{t} - \mathbf{h}\|^2$. Setting $s(h,r,t) = \|\mathbf{R}\mathbf{t} - \mathbf{h}\|^2$, by Proposition 3.1, the model suffers from Z-paradox provided that $\inf s = 0$. The above procedure can be applied to other bilinear KGE models including DisMult (Yang et al., 2014), ComplexEX (Trouillon et al., 2016), DihEdral (Xu and Li, 2019), QuatE (Zhang et al., 2019), SEEK (Xu et al., 2020), Tucker (Balažević et al., 2019).

3.2 MQuinE

In this section, we introduce our model – MQuinE; before that, we review some fundamental relation patterns which need to be considered for KGE models.

Symmetric/Asymmetric relation. A relation r is symmetric iff the fact triple (h,r,t) holds $\Leftrightarrow$ the fact triple (t,r,h) holds. And a relation r is asymmetric iff the fact triples (h,r,t) and (t,r,h) do not hold simultaneously.

Inverse relation. A relation r_2 is the inversion of the relation r_1 iff the fact triple (h,r_1,t) holds $\Leftrightarrow$ the fact triple (t,r_2,h) holds.

Relation composition. A relation r_3 is the composition of relation r_1 and r_2 (denoted by $r_3 = r_1 \oplus r_2$) iff the facts (a,r_1,b) and (b,r_2,c) imply the fact (a,r_3,c).

Abelian (non-Abelian). If $r_1 \oplus r_2 = r_2 \oplus r_1$, the composition $r_1 \oplus r_2$ is Abelian; otherwise, it is non-Abelian.

1-N/N-1/N-N relation. A relation r is a 1-N / N-1 relation if there exist at least two distinct tail/head entities such that (h,r,t_1), (h,r,t_2) / $(h_1,r,t), (h_2,r,t)$ hold. A relation r is an N-N relation if it is both 1-N and N-1.

Next, we propose MQuinE, which preserves the aforementioned relation patterns, moreover, circumvents the Z-paradox. Specifically, for a fact triple (h,r,t), we use

$$s(h,r,t) = \|\mathbf{H}\mathbf{R}^h - \mathbf{R}^t\mathbf{T} + \mathbf{H}\mathbf{R}^c\mathbf{T}\|_F^2 \quad (2)$$

to measure the plausibility of the fact triple. Here $\mathbf{H}, \mathbf{T} \in \boldsymbol{R}^{d\times d}$ are symmetric matrices and denote the embeddings of the head entity and tail entity, respectively, and the matrix triple $\langle \mathbf{R}^h, \mathbf{R}^t, \mathbf{R}^c \rangle \in \boldsymbol{R}^{d\times d} \times \boldsymbol{R}^{d\times d} \times \boldsymbol{R}^{d\times d}$ denotes the embedding of the relation r. Specifically, for a true fact triple (h,r,t), we expect $s(h,r,t) \approx 0$, moreover, for a false triple, we hope the score is relatively large.

3.3 Expressiveness of MQuinE

In this section, we theoretically show that MQuinE is able to model symmetric/asymmetric, inverse, 1-N/N-1/N-N, Abelian/non-Abelian compositions relations, more importantly, MQuinE does not suffer from Z-paradox.

Theorem 3.2. *MQuinE can model the symmetry/asymmetry, inverse, 1-N/N-1/N-N relations.*

Theorem 3.3 (Composition)**.** *MQuinE can model the Abelian/non-Abelian compositions of relations.*

The proofs of Theorem 3.2 and Theorem 3.3 are provided in Appendix.

Theorem 3.4 (No Z-paradox)**.** *MQuinE does not suffer from Z-paradox.*

Proof. We show the result via two examples. Let

$$\mathbf{R}^h = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{R}^t = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix},$$
$$\mathbf{R}^c = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \mathbf{E}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{E}_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

On one hand, set $\mathbf{E}_4 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$, it holds that

$$s(e_1,r,e_2) = 0, \quad s(e_3,r,e_2) = 0,$$
$$s(e_3,r,e_4) = 0, \quad s(e_1,r,e_4) = 1.$$

In other words, given a Z-pattern, e_1 does not link to e_4. On the other hand, set $\mathbf{E}_4 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, it holds that

$$s(e_1,r,e_2) = 0, \quad s(e_3,r,e_2) = 0,$$
$$s(e_3,r,e_4) = 0, \quad s(e_1,r,e_4) = 0.$$

In other words, given a Z-pattern, e_1 may link to e_4. This completes the proof. □

Remark 3.3. Set $\mathbf{R}^c = 0$, MQuinE becomes MQuadE (Yu et al., 2021). So it is not surprising to draw the conclusion that MQuinE can model various relation patterns (Theorem F.1, Theorem F.2, Theorem F.3 and Theorem 3.3) since MQuadE can. The cross term $\mathbf{H}\mathbf{R}^c\mathbf{T}$ plays the central role in circumventing Z-paradox. Adding such a cross term to MQuadE to obtain MQuinE is nontrivial, the key insight is Proposition 3.1 — without a cross term between the head and tail entities, a distance-based model must suffer from Z-paradox.

Relation	Relation matrix property
1-N relation	$\text{rank}(\mathbf{R}^t) + \text{rank}(\mathbf{R}^c) < d$
N-1 relation	$\text{rank}(\mathbf{R}^h) + \text{rank}(\mathbf{R}^c) < d$
N-N relation	$\text{rank}(\mathbf{R}^t) + \text{rank}(\mathbf{R}^c) < d$ and $\text{rank}(\mathbf{R}^h) + \text{rank}(\mathbf{R}^c) < d$
Symmetric relation	$(\mathbf{R}^t)^T = \mp\mathbf{R}^h$, $(\mathbf{R}^c)^T = \pm\mathbf{R}^c$
r_2 inversion of r_1	$\mathbf{R}_1^h = (\mathbf{R}_2^t)^T, \mathbf{R}_1^t = (\mathbf{R}_2^h)^T, \mathbf{R}_1^c = -(\mathbf{R}_2^c)^T$
Compositions $r_3 = r_1 \oplus r_2$	$\mathbf{R}_3^h = \mathbf{R}_1^h\mathbf{R}_2^h, \mathbf{R}_3^t = \mathbf{R}_1^t\mathbf{R}_2^t, \mathbf{R}_3^c = \mathbf{R}_1^h\mathbf{R}_2^c + \mathbf{R}_1^c\mathbf{R}_2^t$
$r_1 \oplus r_2$ Abelian composition	$\mathbf{R}_1^h\mathbf{R}_2^h = \mathbf{R}_2^h\mathbf{R}_1^h, \mathbf{R}_1^t\mathbf{R}_2^t = \mathbf{R}_2^t\mathbf{R}_1^t, \mathbf{R}_1^h\mathbf{R}_2^c + \mathbf{R}_1^c\mathbf{R}_2^t = \mathbf{R}_2^h\mathbf{R}_1^c + \mathbf{R}_2^c\mathbf{R}_1^t$
$r_1 \oplus r_2$ Non-Abelian composition	$\mathbf{R}_1^h\mathbf{R}_2^h \neq \mathbf{R}_2^h\mathbf{R}_1^h$, or $\mathbf{R}_1^t\mathbf{R}_2^t \neq \mathbf{R}_2^t\mathbf{R}_1^t$, or $\mathbf{R}_1^h\mathbf{R}_2^c + \mathbf{R}_1^c\mathbf{R}_2^t \neq \mathbf{R}_2^h\mathbf{R}_1^c + \mathbf{R}_2^c\mathbf{R}_1^t$

Table 3.1: The various relations which can be modeled with different matrices.

Algorithm 1: Z-sampling

Input: the set of entities $\mathcal{E}$, the set of relations $\mathcal{R}$, the set of observed triplets $\mathcal{O}$, a triplet (h, r, t), number of negative samples $m \in \mathbb{N}$, number of Z-samples $k \in \mathbb{N}$.

Fix h, r, sample m tail entities $\{t_i\}_{i=1}^m$ s.t. $(h, r, t_i) \notin \mathcal{O}\ \forall i \in [m]$, set $S_{\text{neg}} = \{(h, r, t_i)\}_{i=1}^m$;

Set $S_Z = \emptyset$;

for $i = 1, 2, \ldots, m$ **do**

  Collect all Z-patterns associated with (h, r, t_i), i.e., $e_2, e_3 \in \mathcal{E}$ s.t. $(h, r, e_2), (e_3, r, e_2), (e_3, r, t) \in \mathcal{O}$;

  $S_Z = S_Z \cup \{(h, r, e_2), (e_3, r, e_2), (e_3, r, t)\}$;

end

Uniform randomly select k triplets in S_Z and remove other triplets from S_Z;

Output: S_{neg} and S_Z.

The comparison of MQuinE and some existing KGE models in expressiveness are given in Table C.1. To our knowledge, MQuinE is the only KGE model that does not suffer from Z-paradox while preserving the ability to capture all relation patterns.

4 Learning of MQuinE

Parameterization and regularization. We constrain the entity embedding matrices to be symmetric and parameterize each entity matrix $\mathbf{E}$ by a lower triangular matrix $\mathbf{A}$ and its transpose, i.e.,$\mathbf{E} = \mathbf{A} + \mathbf{A}^T$. We use the Frobenius norm of entity and relation embedding matrices as regularization.

Z-sampling. The negative sampling technique plays an important role in the training of KGE models. Classic negative sampling first sample a batch of observed facts. Then for each fact (h, r, t) in the batch, fix the head entity h and relation r and sample m tail entities $\{t_i\}_{i=1}^m$ such that $\{(h, r, t_i)\}_{i=1}^m$ are not observed. Lastly, perform a gradient step to decrease the score of positive samples and increase the score of negative samples. To mitigate the effect of Z-patterns more explicitly and fully exploit the benefit of MQuinE, we propose a new sampling technique called *Z-sampling*. Given a positive fact (h, r, t), Z-sampling first sample m negative samples $\{(h, r, t_i)\}_{i=1}^m$ following exactly the same procedure as the classic negative sampling, then it collects all Z-patterns from observed facts that are related to the sampled negative samples, i.e.,

$$S_{\text{Z}} = \cup_{i=1}^m \{(h, r, e_2), (e_3, r, e_2), (e_3, r, t_i) \\ |(h, r, e_2), (e_3, r, e_2), (e_3, r, t_i) \in \mathcal{O}\}.$$

Lastly, the Z-sampling uniform randomly samples k facts from S_Z and treats them as positive facts. The detailed algorithm of Z-sampling is given in Algorithm 1. Z-sampling explicitly tries to minimize the score of positive facts in the Z-pattern and maximize the score of negative facts simultaneously. Experiments on KG benchmark datasets demonstrate the effectiveness of Z-sampling; see Section 5.6 for details. Note that Algorithm 1 can also be applied to ranking the head entity h (fixing r, t) with minor modifications, we omit the details.

Objective function. The loss function $\mathcal{L}_{(h,r,t)}$ with respect to an observed fact (h, r, t) is

$$\begin{aligned}\mathcal{L}_{(h,r,t)} = &-\log\sigma(\gamma - s(h, r, t)) \\ &- \lambda_{\text{neg}}\mathbb{E}_{(h,r,t')\sim S_{\text{neg}}}\left[\log\sigma(s(h, r, t') - \gamma)\right] \\ &- \lambda_Z\mathbb{E}_{(h,r,t'')\in S_Z}\left[\log\sigma(\gamma - s(h, r, t''))\right],\end{aligned}$$

where $\gamma > 0$ is a pre-defined margin, σ is the sigmoid function, i.e., $\sigma(x) = 1/(1 + e^{-x})$, S_{neg} and S_Z are the negative samples and Z-samples as defined in Algorithm 1, and $\lambda_{\text{neg}}, \lambda_Z > 0$ are used to control the trade-off between positive and negative samples.

5 Experiments

We conduct experiments to demonstrate the impact of Z-paradox for existing KGE models on KG benchmark datasets and evaluate the performance of MQuinE. First, we introduce the experimental setup in Section 5.1, including the description of benchmark datasets, evaluation tasks, evaluation metrics and baseline methods. In Section 5.3, we propose a metric called *Z-value* to quantify the number of Z-patterns related to a given fact and gather statistics of Z-values to quantify the effect of Z-patterns for our experimental datasets. For each dataset, we divide their test samples into easy, neutral, and hard cases according to their Z-value. In Section 5.4, we evaluate the performance of MQuinE against other competitive baseline methods on the easy, neutral, and hard cases respectively; we also report the overall improvement of MQuinE in Section 5.5. Lastly, we conduct an ablation study to evaluate the effectiveness of Z-sampling with different KGE models in Section 5.6.

5.1 Experimental setup

Dataset. We conduct experiments on five large-scale benchmark datasets — FB15k-237 (Toutanova and Chen, 2015), WN18 (Bordes et al., 2013), WN18RR (Dettmers et al., 2018), YAGO3-10 (Mahdisoltani et al., 2014), and CoDEx (Safavi and Koutra, 2020) (CoDEx-L, CoDEx-M, CoDEx-S). The detailed statistics of these datasets are given in Table D.1. These datasets contain various relations including 1-N, N-1, N-N, and composition relations, and are suitable for evaluating complex KGE models. We follow the train/validation/test split from Sun et al. (2019) and divide the observed facts into training, validation, and testing by 8:1:1.

Evaluation task and metrics. We evaluate the performance of KGE models on the link prediction task. Given a query fact (h, r, t), the link prediction task requires one to fix h, r and rank t among all possible tail entities $t' \in \mathcal{E}$ except those t''s such that (h, r, t')'s appear in the training set. We use the mean reciprocal rank (MRR), mean rank (MR), Hits@N (N = 1,3,10) as our evaluation metrics.

Baselines. We compare MQuinE with KGE baselines including TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), DisMult (Yang et al., 2014), ComplEX (Trouillon et al., 2016), DihEdral (Xu and Li, 2019), QuatE (Zhang et al., 2019), TuckER (Balažević et al., 2019), ConvE (Dettmers et al., 2018), OTE (Tang et al., 2020), BoxE (Abboud et al., 2020), HAKE (Zhang et al., 2020), MQuadE (Yu et al., 2021), ExpressivE (Pavlović and Sallinger, 2023), DualE (Cao et al., 2021) and HousE (Li et al., 2022).

5.2 Implementation details

Our model. We initialized the element of entity matrices from the normal distribution $\mathcal{N}(0, 0.01)$ and the relation matrices $\mathbf{R}^h, \mathbf{R}^t, \mathbf{R}^c$ are initiated as the identity matrix. We apply grid search to find the best hyper-parameters of MQuinE. The tuning ranges of hyper-parameters are as follows: number of Z-samples $k \in \{10, 32, 64\}$, dimension of entity and relation matrices $d \in \{20, 25, 32, 35, 42\}$, batch size $b \in \{256, 512, 1024\}$, self-adversarial temperature $\alpha \in \{0.5, 1\}$, fixed margin $\gamma \in \{6, 9, 12, 15, 21, 24\}$, number of negative samples $m \in \{128, 256, 512, 1024\}$, initial learning rate $\eta \in \{10^{-3}, 10^{-4}, 10^{-5}\}$, regularization coefficient $\lambda_{\text{reg}} \in \{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, negative sampling coefficient $\lambda_{\text{neg}} \in \{0.5, 1.0, 2.0\}$. The best hyperparameters for each dataset are given in Table E.6.

Baselines. We follow the implementation of RotatE[1] and use the best configuration of TransE, RotatE, DisMult, and ComplEX reported[2]. We set the number of Z-sampling k to 32 for experiments in Section 5.6.

5.3 Statistics of Z-patterns

We define two statistics that characterize the Z-patterns of a fact. The first one is Z-value, which can be used to measure the number of Z-patterns associated with a fact (h, r, t). The second one is the rank of a fact based on its Z-value.

Definition 2 (Z-value)**.** Given a fact (h, r, t), define

$$n_Z(h, r, t) := \big|\{(e_2, e_3) \mid e_2 \neq e_3; (h, r, e_2), (e_3, r, e_2), (e_3, r, t) \in \mathcal{O}\}\big|,$$

which is the number of Z-patterns connected with (h, r, t)[3].

Definition 3. Define $\text{rank}_Z(h, r, t)$ as follows

$$\sum_{t' \in \mathcal{E},\ (h,r,t') \notin \mathcal{O}} \mathbf{1}_{\{n_Z(h,r,t') \geq n_Z(h,r,t)\}},$$

[1] https://github.com/DeepGraphLearning/KnowledgeGraphEmbedding

[2] https://github.com/DeepGraphLearning/KnowledgeGraphEmbedding/blob/master/best_config.sh

[3] Z-value differs from the concept of Z-score in statistics.

Models	FB15k-237			WN18RR			YAGO3-10		
	Easy	Neutral	Hard	Easy	Neutral	Hard	Easy	Neutral	Hard
ComplEX	67.3%	47.3%	48.9%	96.3%	50.5%	49.6%	0.678%	0.455%	0.518%
DisMult	63.2%	45.2%	37.2%	96.1%	48.6%	48.4%	0.627%	0.356%	0.526%
TransE	64.0%	48.0%	41.9%	90.9%	48.2%	53.7%	0.665%	0.404%	0.665%
RotatE	67.7%	48.1%	39.7%	84.7%	53.4%	49.2%	0.751%	0.457%	0.691%
MQuinE	69.7%	52.9%	**52.7%**	85.2%	56.8%	**62.2%**	0.783%	0.543%	**0.759%**

Table 5.1: Hits@10 on easy, neutral, and hard cases of FB15k-237, WN18RR and YAGO3-10.

Case name	Condition
Easy case	$\text{rank}_Z(h,r,t) < 10$.
Neutral case	$n_Z(h,r,t)$ is tied for the 10^{th} place.
Hard case	otherwise.

Table 5.2: Case splitting description.

Dataset	Easy case	Neutral case	Hard case
FB15k-237	6,681 (33%)	6,546 (32%)	7,239 (35%)
WN18	356 (7%)	4,331 (87%)	313 (6%)
WN18RR	314 (10%)	2,679 (86%)	123 (4%)
YAGO3-10	1,248 (25%)	1,074 (21%)	2,678 (54%)

Table 5.3: Statistics of Z-patterns in the testing set responding to the training set.

which is the rank of $n_Z(h,r,t)$ among unobserved candidate facts.

Intuitively, for two facts (h,r,t) and (h,r,t'), if $n_Z(h,r,t) \gg n_Z(h,r,t')$, then KGE models that suffer from Z-paradox would incline to assign (h,r,t) a lower score[4] compared with (h,r,t') and rank (h,r,t) higher than (h,r,t'). Therefore, for a test fact (h,r,t), if there are many facts (h,r,t') such that (h,r,t')'s do not appear in the training set and $n_Z(h,r,t') \gg n_Z(h,r,t)$, then for KGE models that suffer from Z-paradox, this test fact should be hard for them to predict. Motivated by the above logic, we use $\text{rank}_Z(h,r,t)$ to measure the level of difficulty for predicting (h,r,t). In Table 5.2, we divide the test facts into three categories, namely, easy, neutral and hard cases. For easy cases, we require $n_Z(h,r,t)$ to be top-9; for neural cases, we require $n_Z(h,r,t)$ is tied for the 10th place; other cases are categorzied into hard cases. In Table 5.3, we summarize the ratio of easy, neutral and hard cases in our experimental datasets. We can observe that both FB15k-237 and YAGO3-10 have a notable number of hard cases while most test facts of WN18 and WN18RR are neutral cases.

5.4 Case study of MQuinE

We evaluate the performance of MQuinE and other baseline methods on the easy/neutral/hard cases, respectively. The results on FB15k-237 and WNRR18 are shown in Table 5.1, respectively. We can observe that the Hits@10 on hard cases is significantly lower than the Hits@10 on easy cases when using CompLEX, DisMult, TransE and RotatE; the prediction accuracy on hard cases is about 20% lower than the accuracy on easy cases. This observation indicates that Z-paradox is indeed a serious issue and can significantly degrade the performance of existing KGE models. The results in Table 5.1 also show that MQuinE can significantly improve the performance on hard cases; the Hits@10 obtained by MQuinE is 13.0% higher than RotatE on FB15k-237. In the meanwhile, MQuinE does not sacrifice accuracy on easy and neutral cases; the Hits@10 of MQuinE is 2.0% and 4.8% higher than RotatE on easy and neutral cases, respectively. A similar conclusion can be drawn for the WN18RR dataset, in which the Hits@10 of MQuinE is about 13.0% higher than RotatE and 10.6% higher than TransE on WN18RR.

5.5 Overall evaluation on link prediction

The overall evaluation results of MQuinE and other KGE baseline methods on FB15k-237, WN18RR, YAGO3-10, and CoDEx are presented in Table 5.4 and some missing results are provided in Appendix E. The metric values of baseline methods are taken directly from their original papers. Overall, we observe that MQuinE outperforms all the existing KGE methods with a visible margin in all metrics on FB15k-237, where Hits@1 is 7% higher than other methods and Hits@10 reaches 58.8%. Similarly, MQuinE also exceeds other baseline methods in most metrics on WN18RR, YAGO3-10, WN18 and CoDEx. In addition, we evaluate MQuinE with the node classification task on CoDEx-S and CoDEx-M, and our results shown in Table E.5 (in Appendix E) outperforms the baselines metrics reported by Safavi and Koutra (2020).

5.6 Evaluation of Z-sampling

We examine the impact of Z-sampling on our model as while as other baseline methods, the results are given in Table 5.5. To fairly evaluation the

[4] A lower score means a more plausible fact.

Models	FB15k-237			WN18RR			YAGO3-10		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
DisMult	0.241	0.155	0.419	0.443	0.403	0.534	0.340	0.240	0.540
ComplEX	0.247	0.158	0.428	0.472	0.432	0.550	0.360	0.260	0.550
DihEdral	0.320	0.230	0.502	0.486	0.443	0.557	0.472	0.381	0.643
TuckER	0.353	0.260	0.536	0.470	0.443	0.526	0.527	0.446	0.676
ConvE	0.325	0.237	0.501	0.430	0.400	0.520	0.520	0.450	0.660
TransE	0.294	-	0.465	0.466	0.422	0.555	0.467	0.364	0.610
RotatE	0.336	0.241	0.530	0.476	0.428	0.571	0.495	0.402	0.670
ExpressivE	0.333	0.243	0.512	0.482	0.407	**0.619**	-	-	-
BoxE	0.337	0.238	0.538	0.451	0.400	0.541	**0.567**	**0.494**	0.699
HAKE	0.346	0.250	0.542	**0.497**	0.452	0.582	0.545	0.462	0.694
MQuadE	0.356	0.260	0.549	0.426	0.427	0.564	0.536	0.449	0.689
DualE	0.330	0.237	0.518	0.482	0.440	0.561	-	-	-
HousE	0.361	0.266	0.551	0.496	0.452	0.585	0.565	0.487	0.703
MQuinE	**0.420**	**0.332**	**0.588**	0.492	**0.454**	0.603	0.566	0.492	**0.711**

Models	CoDEx-S			CoDEx-M			CoDEx-L		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
RESCAL	0.404	0.293	0.623	0.317	0.244	0.456	0.304	0.242	0.419
TransE	0.354	0.219	0.634	0.303	0.223	0.454	0.187	0.116	0.317
ComplEx	**0.465**	0.372	0.646	**0.337**	0.262	**0.476**	0.294	0.237	0.400
ConvE	0.444	0.343	0.635	0.318	0.239	0.464	0.303	0.240	0.420
TuckER	0.444	0.339	0.638	0.328	0.259	0.458	0.309	0.244	0.430
MQuinE	0.443	**0.379**	**0.652**	0.335	**0.320**	**0.476**	**0.326**	**0.267**	**0.440**

Table 5.4: Overall evaluation results on the FB15k-237, WN18RR, and YAGO3-10, and CoDEx datasets.

Models	Without Z-sampling			With Z-sampling		
	Hits@N			Hits@N		
	1	3	10	1	3	10
DisMult	0.155	0.263	0.419	0.212 (↑ 5.7%)	0.326 (↑ 6.3%)	0.422 (↑ 0.3%)
ComplEX	0.158	0.275	0.428	0.231 (↑ 7.4%)	0.350 (↑ 7.5%)	0.454 (↑ 2.6%)
TransE	-	-	0.465	0.234	0.370	0.484 (↑ 1.9%)
RotatE	0.234	0.366	0.524	0.230 (↓ 0.4%)	0.356 (↓ 1.0%)	0.508 (↓ 1.6%)
MQuadE	0.248	0.377	0.529	0.269 (↑ 2.1%)	0.342 (↓ 3.5%)	0.504 (↓ 2.5%)
MQuinE	0.274	0.375	0.532	**0.332 (↑ 5.8%)**	**0.440 (↑ 6.5%)**	**0.588 (↑ 5.6%)**

Table 5.5: Effect of Z-sampling on the FB15k-237 dataset.

effect of Z-sampling, we rerun DisMult, ComplEX, TransE, RotatE and MQuadE with/without Z-sampling based on their original implementation. We can observe that the Z-sampling strategy can significantly improve the performance of DisMult, ComplEX, but degrades the performance of RotatE and MQuadE a little bit. This demonstrates that Z-sampling is a useful technique for KGE models that do not suffer from Z-paradox. Not surprisingly, the Z-sampling strategy improves the performance of MQuinE significantly; Z-sampling improves both Hits@1, Hits@3 and Hits@10 for more than 5%. Moreover, during our numerical experiments, we also observed that Z-sampling can stabilize the training of MQuinE and make MQuinE more robust to the different hyperparameter setups.

6 Conclusion

In this paper, we introduce a phenomenon called Z-paradox and show that many existing KGE models suffer from it both theoretically and empirically. To overcome Z-paradox, we propose a new KGE model *MQuinE* that can circumvent Z-paradox while maintaining strong expressiveness. Experiments on real-world knowledge bases suggests that the Z-paradox indeed degrades the performance of existing KGE models and strongly support the effectiveness of MQuinE.

Limitations. There are a few limitations for MQuinE. Z-paradox holds for translation-distance based score function and apply to bilinear and deep learning based KGE models under some certain conditions.

References

Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. Boxe: A box embedding model for knowledge base completion. *Advances in Neural Information Processing Systems*, 33:9649–9661.

Ivana Balažević, Carl Allen, and Timothy Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

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

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. 2021. Dual quaternion knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6894–6902.

Xuelu Chen, Ziniu Hu, and Yizhou Sun. 2022. Fuzzy logic based logical query answering on knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3939–3948.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 105–113.

Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, et al. 2022. House: Knowledge graph embedding with householder parameterization. In *International Conference on Machine Learning*, pages 13209–13224. PMLR.

Haonan Lu and Hailin Hu. 2020. Dense: An enhanced non-abelian group representation for knowledge graph embedding. *arXiv preprint arXiv:2008.04548*.

Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly learning explainable rules for recommendation with knowledge graph. In *The world wide web conference*, pages 1210–1221.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. 2014. Yago3: A knowledge base from multilingual wikipedias. In *7th biennial conference on innovative data systems research*. CIDR Conference.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. 2019. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4710–4723.

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

Aleksandar Pavlović and Emanuel Sallinger. 2023. Expressive: A spatio-functional embedding for knowledge graph completion. In *International Conference on Learning Representations*.

Tara Safavi and Danai Koutra. 2020. CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8328–8350, Online. Association for Computational Linguistics.

Mingming Sun, Xu Li, and Ping Li. 2018. Logician and orator: Learning from the duality between language and knowledge in open domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2119–2130.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2713–2722.

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 *International conference on machine learning*, pages 2071–2080. PMLR.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*, pages 1835–1844.

Shen Wang, Xiaokai Wei, Cicero Nogueira Nogueira dos Santos, Zhiguo Wang, Ramesh Nallapati, Andrew Arnold, Bing Xiang, Philip S Yu, and Isabel F Cruz. 2021. Mixed-curvature multi-relational graph neural network for knowledge graph completion. In *Proceedings of the Web Conference 2021*, pages 1761–1771.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958.

Canran Xu and Ruijiang Li. 2019. Relation embedding with dihedral group in knowledge graph. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 263–272.

Wentao Xu, Shun Zheng, Liang He, Bin Shao, Jian Yin, and Tie-Yan Liu. 2020. Seek: Segmented embedding of knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3897.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Jinxing Yu, Yunfeng Cai, Mingming Sun, and Ping Li. 2021. Mquade: a unified model for knowledge fact embedding. In *Proceedings of the Web Conference 2021*, pages 3442–3452.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 3065–3072. AAAI Press.

Zhehui Zhou, Can Wang, Yan Feng, and Defang Chen. 2022. Jointe: Jointly utilizing 1d and 2d convolution for knowledge graph embedding. *Knowledge-Based Systems*, 240:108100.

Appendix

A Organization

In Section 2, we give an overview of KGE models. We introduce the *Z-paradox* bottleneck along with its theoretical properties, and propose our new KGE model MQuinE in Section 3 and Section 4 respectively. In Section 5, we evaluate the effect of Z-paradox on standard KG benchmarks and empirically compare MQuinE with other competitive baselines. At last, we give a conclusion of our work and discuss the future direction in Section 6.

B Notation Description

We denote the set of entities as $\mathcal{E}$ and the set of relations as $\mathcal{R}$. Following the conventional notation, we represent a knowledge graph as a set of triplets $\mathcal{O} = \{(h_i, r_i, t_i) \mid h_i, t_i \in \mathcal{E}, r_i \in \mathcal{R}\}_{i=1}^{n}$, where n is the number of observed facts. For each entity e and relation r, we use their bold version $\boldsymbol{e}$ and $\boldsymbol{r}$ to denote their embeddings. A KGE model is associated with a score function $s(\cdot) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$. Given a fact (h, r, t), the KGE model tends to predict it to be true if $s(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$ is small and false otherwise. We use bold capital letters. e.g., $\mathbf{A}, \mathbf{B}, \mathbf{H}, \mathbf{T}$ to denote matrices, and use $\|\cdot\|$ to denote the Euclidean norm of vectors or the Frobenius norm of matrices.

C Missing summary table of score functions and properties for knowledge graph embedding models.

We provide a summary Table C.2 of score functions and their mathematical forms for different KGE models.

D Dataset details

Statistics of the benchmark datasets are summarized in Table D.1. We give a brief overview of them in the following:

FB15k-237. FB15k-237 is a subset of the Freebase (Bollacker et al., 2008) knowledge graph which contains 237 relations. The FB15k (Bordes et al., 2013) dataset, a subset of Free-base, was used to build the dataset by (Toutanova and Chen, 2015) in order to study the combined embedding of text and knowledge networks. FB15k-237 is more challenging than the FB15k dataset because FB15k-237 strips out the inverse relations.

Model	Sym/Asym	Inversion	Composition	Injective	Non-injective	Z-paradox
TransE	✗/✔	✔	✔	✔	✗	✗
TransX	✔/✔	✗	✗	✔	✗	✗
DisMult	✔/✗	✗	✗	✔	✗	✔
ComplEX	✔/✔	✔	✗	✔	✗	✔
RotatE	✔/✔	✔	✔	✔	✗	✗
OTE	✔/✔	✔	✔	✔	✔	✗
BoxE	✔/✔	✔	✗	✔	✔	✔
ExpressivE	✔/✔	✔	✔	✔	✔	✗
MQuadE	✔/✔	✔	✔	✔	✔	✗
MQuinE	✔/✔	✔	✔	✔	✔	✔

Table C.1: The pattern modeling and inference abilities of several models.

Model Category	Model	Score function $s(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$	Representation of parameters
Bilinear	DisMult	$-\langle \mathbf{h}, \mathbf{r}, \mathbf{t} \rangle$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^k$
	ComplEX	$-\mathrm{Re}(\langle \mathbf{h}, \mathbf{r}, \bar{\mathbf{t}} \rangle)$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^k$
	DihEdral	$-\mathbf{h}^T \mathbf{R} \mathbf{t}$	$\mathbf{h}, \mathbf{t} \in \mathbb{R}^{2k}, \mathbf{R} \in \mathbb{D}_K^k$
	QuatE	$-\left\langle \mathbf{h} \otimes \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{t} \right\rangle$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{H}^k$
	SEEK	$\sum_{x,y} \langle \mathbf{r}_x, \mathbf{h}, \mathbf{t}_{w_{x,y}} \rangle$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^k$
	TuckER	$-\mathcal{W} \times_1 \mathbf{h} \times_2 \mathbf{r} \times_3 \mathbf{t}$	$\mathbf{h}, \mathbf{t} \in \mathbb{R}^k, \mathbf{r} \in \mathbb{R}^l, \mathcal{W} \in \mathbb{R}^{k \times l \times k}$
Deep learning	ConvE	$g(\mathrm{vec}(g([\mathbf{h}, \mathbf{r}] * \mathbf{w}))W)\mathbf{t}$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^k, \mathbf{w} \in \mathbb{R}^{m_1}, W \in \mathbb{R}^{m_2}$
Translation-based	TransE	$\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^k$
	RotatE	$\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^k, \|\mathbf{r}_i\| = 1$
	OTE	$\|\mathbf{h}\Phi(\mathbf{R}) - \mathbf{t}\|$	$\mathbf{h}, \mathbf{t} \in \mathbb{R}^k, \mathbf{R} = diag\{\mathbf{R}_1, \cdots, \mathbf{R}_s\}, \mathbf{R}_i \in \mathbb{R}^{k/s \times k/s}$
	MQuadE	$\|\mathbf{HR} - \widehat{\mathbf{R}}\mathbf{T}\|_F$	$\mathbf{H}, \mathbf{T}, \mathbf{R}, \widehat{\mathbf{R}} \in \mathbb{R}^{p \times p}$, $\mathbf{H}, \mathbf{T}$ are symmetric
Ours	**MQuinE**	$\|\mathbf{H}\mathbf{R}^h - \mathbf{R}^t\mathbf{T} + \mathbf{H}\mathbf{R}^c\mathbf{T}\|_F$	$\mathbf{H}, \mathbf{T}, \mathbf{R}^h, \mathbf{R}^t, \mathbf{R}^c \in \mathbb{R}^{p \times p}$ $\mathbf{H}, \mathbf{T}$ are symmetric

Table C.2: The score functions of different KGE models.

WN18. WN18 is a subset of the WordNet (Fellbaum, 1998), a lexical database for the English language that groups synonymous words into synsets. WN18 contains relations between words such as *hypernym* and *similar_to*.

WN18RR. WN18RR is a subset of WN18 that removes symmetry/asymmetry and inverse relations to resolve the test set leakage problem. WN18RR is suitable for the examination of relation composition modeling ability.

YAGO3-10. YAGO3-10 is a subset of the YAGO knowledge base (Mahdisoltani et al., 2014) whose entities have at least 10 relations. The dataset contains descriptive relations between persons, movies, places, etc.

CoDEx. CoDEx (Safavi and Koutra, 2020) is a set of knowledge graph completion datasets extracted from Wikidata and Wikipedia that improve upon existing knowledge graph completion benchmarks in scope and level of difficulty.

E More experimental results and hyperparameters

The results of link prediction on the WN18 dataset and some missing results on FB15k-237, WN18RR, and YAGO3-10 are shown in Table E.4, Table E.1, Table E.2 and Table E.3. The overall results with CoDEx dataset (Safavi and Koutra, 2020) are presented in Table E.5.

F Proofs for Theorem 3.2 and Theorem 3.3

For better illustration, we restate Theorem 3.2 as Theorem F.1, Theorem F.2, Theorem F.3.

Theorem F.1. ***MQuinE*** *can model the symmetry/asymmetry relations.*

Proof. A relation r is symmetric iff $(h, r, t) \Leftrightarrow (t, r, h)$. In MQuinE, it requires

$$\mathbf{H}\mathbf{R}^h - \mathbf{R}^t\mathbf{T} + \mathbf{H}\mathbf{R}^c\mathbf{T} = 0$$
$$\Leftrightarrow \mathbf{T}\mathbf{R}^h - \mathbf{R}^t\mathbf{H} + \mathbf{T}\mathbf{R}^c\mathbf{H} = 0.$$

Dataset	#Entity	# Relation	# Train	# Valid	# Test
FB15k	14,951	1,345	483,142	50,000	59,071
FB15k-237	14,541	237	272,115	17,535	20,466
WN18	40,943	18	141,442	5,000	5,000
WN18RR	40,943	11	86,835	3,034	3,134
YAGO3-10	123,182	37	1,079,040	5,000	5,000
CoDEX-L	77,951	69	551,193	30,622	30,622
CoDEX-M	17,050	51	185,584	10,310	10,310
CoDEX-S	2,034	42	32,888	1,827	1,828

Table D.1: Statistics of the datasets.

Models	Metrics				
	MRR	MR	Hits@N 1	Hits@N 3	Hits@N 10
DisMult	0.241	254	0.155	0.263	0.419
ComplEX	0.247	339	0.158	0.275	0.428
DihEdral	0.320	-	0.230	0.353	0.502
QuatE	0.311	176	0.221	0.342	0.495
TuckER	0.353	162	0.260	0.387	0.536
ConvE	0.325	224	0.237	0.356	0.501
TransE	0.294	357	-	-	0.465
RotatE	0.336	177	0.241	0.373	0.530
BoxE	0.337	163	0.238	0.374	0.538
OTE	0.351	-	0.258	0.388	0.537
MQuadE	0.356	174	0.260	0.392	0.549
MQuinE	**0.420**	**109**	**0.332**	**0.440**	**0.588**

Table E.1: Overall evaluation results on the FB15k-237 dataset.

Note that

$$\begin{aligned} & \mathbf{T}\mathbf{R}^h - \mathbf{R}^t\mathbf{H} + \mathbf{T}\mathbf{R}^c\mathbf{H} = 0 \\ \Leftrightarrow \quad & -\mathbf{H}^T(\mathbf{R}^t)^T + (\mathbf{R}^h)^T\mathbf{T}^T + \mathbf{H}^T(\mathbf{R}^c)^T\mathbf{T}^T = 0 \\ \overset{(a)}{\Leftrightarrow} \quad & -\mathbf{H}(\mathbf{R}^t)^T + (\mathbf{R}^h)^T\mathbf{T} + \mathbf{H}(\mathbf{R}^c)^T\mathbf{T} = 0, \end{aligned}$$

where (a) uses the fact that the entity matrix in MQuinE is symmetric. Hence, if

$$(\mathbf{R}^t)^T = \mp\mathbf{R}^h, \quad (\mathbf{R}^c)^T = \pm\mathbf{R}^c,$$

the relation is symmetric, otherwise, asymmetric. □

Theorem F.2. ***MQuinE*** *can model the inverse relations.*

Proof. A relation r_1 is the inverse relation of r_2 iif $(h, r_1, t) \Leftrightarrow (t, r_2, h)$. In MQuinE, it requires

$$\begin{aligned} & \mathbf{H}\mathbf{R}_1^h - \mathbf{R}_1^t\mathbf{T} + \mathbf{H}\mathbf{R}_1^c\mathbf{T} = 0 \\ \Leftrightarrow & \mathbf{T}\mathbf{R}_2^h - \mathbf{R}_2^t\mathbf{H} + \mathbf{T}\mathbf{R}_2^c\mathbf{H} = 0. \end{aligned}$$

Using the fact that the entity matrix is symmetric, we have

$$\begin{aligned} & \mathbf{T}\mathbf{R}_2^h - \mathbf{R}_2^t\mathbf{H} + \mathbf{T}\mathbf{R}_2^c\mathbf{H} = 0 \\ \Leftrightarrow & \mathbf{H}(\mathbf{R}_2^t)^T - (\mathbf{R}_2^h)^T\mathbf{T} - \mathbf{T}(\mathbf{R}_2^c)^T\mathbf{H} = 0. \end{aligned}$$

Models	Metrics				
	MRR	MR	Hits@N 1	Hits@N 3	Hits@N 10
DisMult	0.443	4999	0.403	0.453	0.534
ComplEX	0.472	5702	0.432	0.488	0.550
DihEdral	0.486	-	0.443	0.505	0.557
TuckER	0.470	-	0.443	0.482	0.526
ConvE	0.430	-	0.400	0.440	0.520
TransE	0.466	-	0.422	-	0.555
RotatE	0.476	3340	0.428	0.492	0.571
BoxE	0.451	3207	0.400	0.472	0.541
MQuadE	0.426	6114	0.427	0.462	0.564
MQuinE	**0.492**	**2599**	**0.454**	**0.518**	**0.603**

Table E.2: Overall evaluation results on the WN18RR dataset.

Models	Metrics				
	MRR	MR	Hits@N 1	Hits@N 3	Hits@N 10
DisMult	0.340	5926	0.240	0.380	0.540
ComplEX	0.360	6351	0.260	0.400	0.550
DihEdral	0.472	-	0.381	0.523	0.643
TuckER	0.527	3306	0.446	0.576	0.676
ConvE	0.520	2792	0.450	0.560	0.660
RotatE	0.495	1767	0.402	0.550	0.670
BoxE	**0.567**	1164	**0.494**	0.611	0.699
MQuadE	0.536	1337	0.449	0.582	0.689
MQuinE	0.566	**992**	0.492	**0.629**	**0.711**

Table E.3: Overall evaluation results on the YAGO3-10 dataset.

Therefore, simply set

$$\mathbf{R}_1^h = (\mathbf{R}_2^t)^T, \quad \mathbf{R}_1^t = (\mathbf{R}_2^h)^T, \quad \mathbf{R}_1^c = -(\mathbf{R}_2^c)^T,$$

the conclusion follows. □

Theorem F.3. ***MQuinE*** *can model the 1-N/N-1/N-N relations.*

Proof. **1-N relations.** A relation r is 1-N iff there exist two distinct fact triples (h, r, t_1) and (h, r, t_2). Set $s(h, r, t_1) = s(h, r, t_2) = 0$, we get

$$\begin{aligned} & \mathbf{H}\mathbf{R}^h - \mathbf{R}^t\mathbf{T_1} + \mathbf{H}\mathbf{R}^c\mathbf{T_1} = 0, \\ & \mathbf{H}\mathbf{R}^h - \mathbf{R}^t\mathbf{T_2} + \mathbf{H}\mathbf{R}^c\mathbf{T_2} = 0, \end{aligned}$$

Models	Metrics				
	MRR	MR	Hits@N 1	3	10
ComplEX	0.941	-	0.936	0.945	0.947
DihEdral	0.946	-	0.942	0.949	0.954
TuckER	0.953	-	**0.949**	0.955	0.958
ConvE	0.943	374	0.935	0.946	0.956
TransE	-	263	-	-	0.754
RotatE	0.949	**09**	0.944	0.952	0.959
MQuadE	0.897	268	0.893	0.926	0.941
MQuinE	**0.958**	189	0.937	**0.957**	**0.975**

Table E.4: Results of link prediction on the WN18 dataset.

Models	CoDEx-S		CoDEx-M	
	Acc	F1-score	Acc	F1-score
RESCAL	0.843	0.852	0.818	0.815
TransE	0.829	0.837	0.797	0.803
ComplEx	0.836	0.846	0.824	0.818
ConvE	0.841	0.846	0.826	**0.829**
TuckER	0.840	0.846	0.823	0.816
MQuinE	**0.876**	**0.883**	**0.831**	0.828

Table E.5: Overall evaluation results on the CoDEx datasets for triple classification.

where $\mathbf{H}, \mathbf{T_1}, \mathbf{T_2}$ are the embedding matrices of h, t_1, t_2, and $\langle \mathbf{R}^h, \mathbf{R}^t, \mathbf{R}^c \rangle$ is the matrix triple of the relation r.

By simple calculations, we have

$$(-\mathbf{R}^t + \mathbf{H}\mathbf{R}^c)(\mathbf{T_1} - \mathbf{T_2}) = 0.$$

Now let us set $\mathrm{rank}(\mathbf{R}^t) + \mathrm{rank}(\mathbf{R}^c) < d$, then

$$\mathrm{rank}(-\mathbf{R}^t + \mathbf{H}\mathbf{R}^c) \leq \mathrm{rank}(\mathbf{R}^t) + \mathrm{rank}(\mathbf{R}^c) < d,$$

i.e., $-\mathbf{R}^t + \mathbf{H}\mathbf{R}^c$ is low rank. Then $\mathbf{T}_1$ and $\mathbf{T}_2$ can be distinct, **MQuinE** can model 1-N relations.

N-1 relations. A relation r is N-1 iff there exist two distinct fact triples (h_1, r, t) and (h_2, r, t). Similar to the proof for 1-N relations, we have

$$\mathbf{H_1}\mathbf{R}^h - \mathbf{R}^t\mathbf{T} + \mathbf{H_1}\mathbf{R}^c\mathbf{T} = 0,$$
$$\mathbf{H_2}\mathbf{R}^h - \mathbf{R}^t\mathbf{T} + \mathbf{H_2}\mathbf{R}^c\mathbf{T} = 0,$$

where $\mathbf{H_1}, \mathbf{H_2}, \mathbf{T}$ are the embedding matrices of h_1, h_2, t. Then it follows that

$$(\mathbf{H_1} - \mathbf{H_2})(\mathbf{R}^h + \mathbf{R}^c\mathbf{T}) = 0.$$

Set $\mathrm{rank}(\mathbf{R}^h) + \mathrm{rank}(\mathbf{R}^c) < d$, **MQuinE** can model N-1 relations.

N-N relations. Set $\mathrm{rank}(\mathbf{R}^t) + \mathrm{rank}(\mathbf{R}^c) < d$ and $\mathrm{rank}(\mathbf{R}^h) + \mathrm{rank}(\mathbf{R}^c) < d$. The conclusion follows. □

Theorem 3.2 (Compositions). **MQuinE** can model the Abelian/non-Abelian compositions of relations.

Proof. A relation r_3 is a composition of r_1 and r_2 iff we have $(e_1, r_1, e_2), (e_2, r_2, e_3) \rightarrow (e_1, r_3, e_3)$. In **MQuinE**, it requires

$$\mathbf{E_1}\mathbf{R}_1^h - \mathbf{R}_1^t\mathbf{E_2} + \mathbf{E_1}\mathbf{R}_1^c\mathbf{E_2} = 0,$$
$$\mathbf{E_2}\mathbf{R}_2^h - \mathbf{R}_2^t\mathbf{E_3} + \mathbf{E_2}\mathbf{R}_2^c\mathbf{E_3} = 0.$$

Rewrite the above two equalities as

$$\mathbf{E_1}\mathbf{R}_1^h - (\mathbf{R}_1^t - \mathbf{E_1}\mathbf{R}_1^c)\mathbf{E_2} = 0,$$
$$\mathbf{E_2}(\mathbf{R}_2^h + \mathbf{R}_2^c\mathbf{E_3}) - \mathbf{R}_2^t\mathbf{E_3} = 0.$$

Then it follows that

$$\begin{aligned}&\mathbf{E_1}\mathbf{R}_1^h(\mathbf{R}_2^h + \mathbf{R}_2^c\mathbf{E_3})\\ =&(\mathbf{R}_1^t - \mathbf{E_1}\mathbf{R}_1^c)\mathbf{E_2}(\mathbf{R}_2^h + \mathbf{R}_2^c\mathbf{E_3})\\ =&(\mathbf{R}_1^t - \mathbf{E_1}\mathbf{R}_1^c)\mathbf{R}_2^t\mathbf{E_3},\end{aligned}$$

where is equivalent to

$$\mathbf{E_1}\mathbf{R}_1^h\mathbf{R}_2^h = \mathbf{R}_1^t\mathbf{R}_2^t\mathbf{E_3} - \mathbf{E_1}(\mathbf{R}_1^h\mathbf{R}_2^c + \mathbf{R}_1^c\mathbf{R}_2^t)\mathbf{E_3}.$$

Let

$$\mathbf{R}_3^h = \mathbf{R}_1^h\mathbf{R}_2^h, \quad \mathbf{R}_3^t = \mathbf{R}_1^t\mathbf{R}_2^t,$$
$$\mathbf{R}_3^c = \mathbf{R}_1^h\mathbf{R}_2^c + \mathbf{R}_1^c\mathbf{R}_2^t.$$

Then we know that (e_1, r_3, e_3) holds.

Abelian/Non-Abelian compositions By definition, if

$$\mathbf{R}_1^h\mathbf{R}_2^h = \mathbf{R}_2^h\mathbf{R}_1^h, \quad \mathbf{R}_1^t\mathbf{R}_2^t = \mathbf{R}_2^t\mathbf{R}_1^t,$$
$$\mathbf{R}_1^h\mathbf{R}_2^c + \mathbf{R}_1^c\mathbf{R}_2^t = \mathbf{R}_2^h\mathbf{R}_1^c + \mathbf{R}_2^c\mathbf{R}_1^t,$$

r_3 is an Abelian composition, otherwise, non-Abelian. □

Dataset	k	d	b	α	γ	m	λ_{reg}	λ_{neg}
FB15k-237	32	38	1024	0.5	12.0	256	0.01	1.0
WN18	64	35	1024	1.0	9.0	256	5e-3	1.0
WN18RR	64	35	1024	0.5	12.0	512	0.01	1.0
YAGO3-10	64	18	4096	1.0	32.0	512	5e-3	1.0
CoDEx-S	32	32	1024	0.5	12.0	256	0.01	1.0
CoDEx-M	32	32	1024	0.5	12.0	256	0.01	1.0
CoDEx-L	32	32	1024	0.5	12.0	256	5e-3	1.0

Table E.6: The best hyperparameters of MQuinE on four datasets in our experiments.