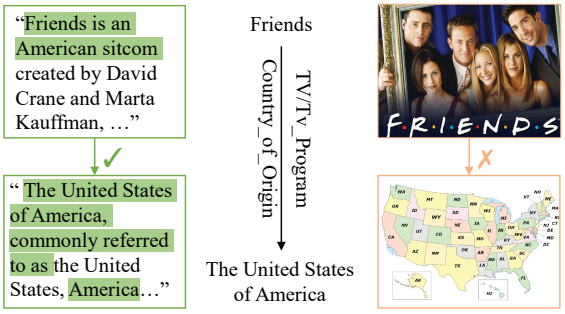
MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion

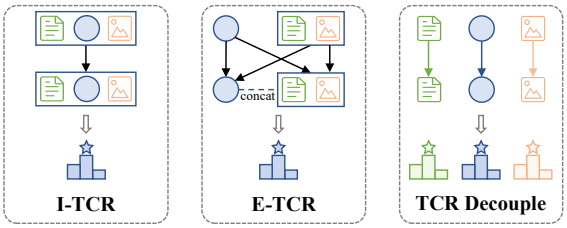
Yu Zhao[1] **Xiangrui Cai**[1]* **Yike Wu**[2] **Haiwei Zhang**[1]
Ying Zhang[3] **Guoqing Zhao**[4] **Ning Jiang**[4]
[1] College of Cyber Science, TKLNDST, Nankai University, Tianjin, China
[2] School of Journalism and Communication, Nankai University, Tianjin, China
[3] College of Computer Science, Nankai University, Tianjin, China
[4] Mashang Consumer Finance Co, Ltd
zhaoyu@dbis.nankai.edu.cn, {caixr,wuyike,zhhaiwei,yingzhang}@nankai.edu.cn

Abstract

Multimodal knowledge graph completion (MKGC) aims to predict missing entities in MKGs. Previous works usually share relation representation across modalities. This results in mutual interference between modalities during training, since for a pair of entities, the relation from one modality probably contradicts that from another modality. Furthermore, making a unified prediction based on the shared relation representation treats the input in different modalities equally, while their importance to the MKGC task should be different. In this paper, we propose **MoSE**, a **Mo**dality **S**plit representation learning and **E**nsemble inference framework for MKGC. Specifically, in the training phase, we learn modality-split relation embeddings for each modality instead of a single modality-shared one, which alleviates the modality interference. Based on these embeddings, in the inference phase, we first make modality-split predictions and then exploit various ensemble methods to combine the predictions with different weights, which models the modality importance dynamically. Experimental results on three KG datasets show that MoSE outperforms state-of-the-art MKGC methods. Codes are available at https://github.com/OreOZhao/MoSE4MKGC.



(a) An example of contradictory relations between modalities.



(b) Implicit (I-) and Explicit (E-) Tight-Coupling Relation (TCR), and TCR Decouple (Ours). Note that the arrows with the same color represent *the same* relation embedding.

Figure 1: (a) displays an example of multimodal triples, in which text modality demonstrates relevant relation ✓ and visual modality demonstrates irrelevant relation ✗. (b) visualizes existing approaches with implicit and explicit tight-coupling relation (TCR), and our approach with decoupled relations.

1 Introduction

Multimodal knowledge graphs (MKGs) organize multimodal facts in the form of entities and relations, and have been successfully applied to various knowledge-driven tasks (Marino et al., 2019; Sun et al., 2020; Zhang et al., 2018; Zhu et al., 2022). To address the inherent incomplete problems in MKGs, multimodal knowledge graph completion (MKGC) has been proposed (Xie et al., 2016, 2017), which utilizes auxiliary visual or text information to help predict missing entities. As shown in Figure 1a, given the head entity Friends and the relation $\xrightarrow{country}$, MKGC is required to predict the tail entity The United States of America. It can be observed that the descriptions attached to entities provide supplementary information for entity prediction.

Existing MKGC methods usually share a common relation embedding across all modalities for a pair of entities, which tightly couples multiple relations from different modalities. We define this paradigm of MKGC as **Tight-Coupling Relation (TCR)**. As shown in Figure 1b, according to the way that the relations from different modalities are coupled, existing methods can be roughly divided into two categories: Implicit TCR (I-TCR) meth-

*Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10527 - 10536

ods and Explicit TCR (E-TCR) methods. I-TCR methods (Wang et al., 2021, 2019) usually first fuse multimodal information of an entity into a single embedding, and then learn a unified relation representation based on the embedding. E-TCR methods (Mousselly-Sergieh et al., 2018; Xie et al., 2016, 2017) directly model the relationship between separate multimodal information of entities without fusion. They usually learn a single relation embedding to simultaneously represent all intra-modal and inter-modal relations.

Although existing MKGC methods have achieved promising results, they are limited by TCR in two folds: (1) **Modality relation contradiction.** The TCR usually simultaneously represents multiple relations from different modalities only with a single embedding. However, for a pair of entities, the relation from one modality probably contradicts that from another modality. For example, as shown in Figure 1a, the description *"American sitcom"* of entity Friends demonstrates the relation $\overset{country}{\longrightarrow}$ to entity The United States of America, while the images do not. The inherent contradiction of TCR results in modality interference during representation learning in MKGs. (2) **Modality difference ignorance.** Based on TCR, existing methods usually treat the input in different modalities equally and make a unified prediction, which ignores the difference of modality importance. However, different modalities vary in data quality and entity coverage, and should contribute to the final prediction in varying degrees.

To overcome the above limitations, we propose a **Mo**dality **S**plit learning and **E**nsemble inference framework, MoSE. As shown in Figure 1b, in the training phase, MoSE decouples TCR and learns multiple modality-split relation embeddings instead of a single modality-shared one, which alleviates mutual interference between modalities. In the inference phase, MoSE first makes predictions for each modality separately based on the modality-split embeddings, and then merges them into the final prediction. We explore the best combination of modality predictions with various ensemble methods, and model the modality importance by modulating the modality weights dynamically. Experimental results and analysis on three widely-used datasets show that MoSE outperforms state-of-the-art methods for MKGC task.

Overall, the contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to deal with the modality contradiction of relation representation and discuss modality importance in MKGC task.
- We propose a modality-split learning and ensemble inference framework MoSE for MKGC, which decouples the tight-coupling relation embedding into modality-split ones in the training phase, and modulate modality importance adaptively in the inference phase.
- Experiment results on three datasets demonstrate that MoSE outperforms 9 baselines and obtain the state-of-the-art performance in MKGC task. The results also show that text modality is a useful complement for MKGC rather than visual modality.

2 Related Work

Existing researches on MKGC mainly focus on extending unimodal knowledge graph embedding (KGE) models to further exploit multimodal information. We notice that for a pair of entities in an MKG, existing multimodal KGE methods all exploit a modality-shared relation embedding which tightly couples multiple relations from different modalities, which we call **Tight-Coupling Relation (TCR)**. We divide existing methods to two categories: implicit tight-coupling relation (I-TCR) methods and explicit tight-coupling relation (E-TCR) methods.

2.1 Implicit TCR Methods

I-TCR methods (Wang et al., 2021, 2019) fuse multiple modalities into a unified entity embedding and utilize a shared relation representation as shown in Figure 1b. Thus the learned relation implicitly fuses multimodal relations. TransAE (Wang et al., 2019) extends TransE with auto-encoder fusing visual and text information into entity representation. Recently, RSME (Wang et al., 2021) notices the noise in visual modality and propose a forget gate to adjust the fusion rate of image to entity embeddings and reaches state-of-the-art (SOTA) performance. Though I-TCR methods show promising improvements, they neglect modality contradictions in modality-shared relation representation. Moreover, they make unified predictions without assessing whether the modality information is relevant to final predictions. In terms of SOTA RSME,

the fusion ratio of visual information is determined by image information itself, i.e. similarity, regardless of modality importance to final prediction.

2.2 Explicit TCR Methods

E-TCR methods (Mousselly-Sergieh et al., 2018; Xie et al., 2016, 2017) utilize a shared relation embedding which tightly couples multiple relations between intra-modal and inter-modal entities. E-TCR methods learn representations with an overall score across all modalities: structure-structure, structure-visual/text, visual/text-structure, visual/text-visual/text, all connected by a single modality-shared relation embedding as shown in Figure 1b, which explicitly tightly couples multiple relations. DKRL (Xie et al., 2016) and IKRL (Xie et al., 2017) extend TransE with text and visual modality respectively. MKB (Mousselly-Sergieh et al., 2018) extends IKRL (Xie et al., 2017) from visual modality to visual-text multimodalities. Although E-TCR methods project multimodal features to a common latent space, the inherent semantic contradiction of relations between different modalities is not eliminated. Moreover, they utilize weighted concatenation of multimodal entities to make a unified prediction and does not consider modality importance either.

3 Methodology

3.1 Preliminaries

In this section, we introduce the notation used in this paper and formulate the MKGC task.

KGC task. Knowledge graph is a collection of factual triples $\mathcal{G} = \{(h, r, t)\}$, where head entity and tail entity $h, t \in \mathcal{E}$ and relation $r \in \mathcal{R}$. The KGE model (1) represents entities and relations to vectors $\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}$, (2) utilizes a score function $f(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$ to decode the plausibility of a triple to scores. For a particular query $q = (h, r, ?)$, the KGC task aims at ranking all possible entities and obtaining prediction preference.

MKGC task. In MKGs, each entity $e \in \mathcal{E}$ has multimodal embeddings $\boldsymbol{e}_m, m \in \mathcal{M} = \{\mathcal{S}, \mathcal{V}, \mathcal{T}\}$, which denotes structure, visual, and text modality respectively. We use $\boldsymbol{e}_s, \boldsymbol{e}_v, \boldsymbol{e}_t$ to denote corresponding entity embedding, where visual and text embedding is projection of extracted features $\boldsymbol{e}_v = W_v f_v, \boldsymbol{e}_t = W_t f_t$.

TCR methods. I-TCR methods design a fusion mechanism $\Phi(\{\boldsymbol{e}_m\}), m \in \mathcal{M}$ to get a fused embedding of multimodal entities and extends the score function as $f(h, r, t) = f(\Phi(\{\boldsymbol{h}_m\}), \boldsymbol{r}, \Phi(\{\boldsymbol{t}_m\}))$. E-TCR methods utilize score function across all modalities, where $f(h, r, t) = \sum_{i=1}^{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} f(\boldsymbol{h}_i, \boldsymbol{r}, \boldsymbol{t}_j)$. It is worth noting that both ways utilize a modality-shared relation representation.

3.2 Overview

Figure 2 shows our **Mo**dality **S**plit learning and **E**nsemble inference framework, `MoSE`, for multimodal knowledge graph completion task. We first decouple TCR to modality-split relation embeddings corresponding to each modality. With the decoupled TCR, we construct modality-split triple representations for each modality to prevent modality interference in representations. Through the KGE score function, the modality-split representations are decoded to corresponding score distributions. In the training phase, we train modality-split entity and relation representations with intra-modal scores simultaneously. Considering visual and text modalities usually embody more contradictory and uncertain noise than structure modality, we apply confidence-constraint training objectives for the two modalities. In the inference phase, we exploit ensemble inference to combine the modality-split predictions and obtain the final predictions. We explored three kinds of ensemble inference methods aiming at modeling modality importance.

3.3 Modality-Split MKG Construction

In our paper, we assume that the TCR embedding used in existing methods represents multiple contradictory relations simultaneously and results in modality interference. Thus we propose to decouple the TCR and construct a modality-split MKG. With decoupling the contradictory relations from a modality-shared embedding to multiple modality-split relation embeddings, `MoSE` alleviates modality interference in relation representations. Formally, we construct decoupled modality-split relation embeddings for each relation type. In this paper, we construct structure, visual and text modality relation embedding $\boldsymbol{r}_s$ $\boldsymbol{r}_v$ and $\boldsymbol{r}_t$ for relation r. Since relation representation is decoupled, we also avoid modality fusion in entities, which also prevents interference within entity representations. Together with the modality-split entity and relations, we form a modality-split KG of multiple unimodal KGs which have identical topology but different entity and relation representations.

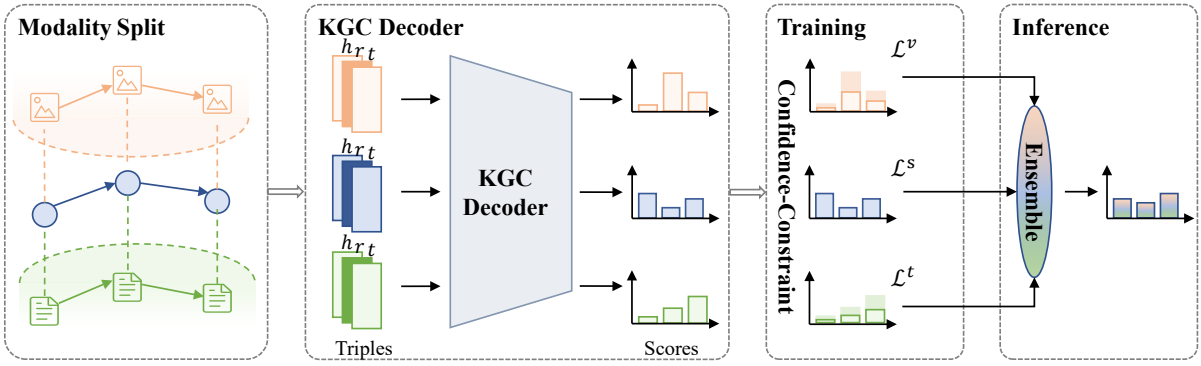


Figure 2: The framework of **Mo**dality-**S**plit learning and **E**nsemble Inference, MoSE, for multimodal knowledge graph completion.

3.4 KGC Decoder

With the modality-split KG construction, the score function is also separated to multiple scores denoted as $f_m(h,r,t) = f(\boldsymbol{h}_m, \boldsymbol{r}_m, \boldsymbol{t}_m), m \in \mathcal{M}$. With the modality-split architecture, MoSE is able to present score distribution for each modality. For each query triple (h,r,t), the decoder gives different scores depending on the learned representation, which intuitively reflects the strengths and limitations of each modality for entity predictions.

3.5 Training

We utilize multi-class cross-entropy (CE) loss for training following Lacroix et al. (2018). Given a query $(h,r,?)$, we construct corrupted triples by replacing the tail entity with randomly selected entities in $\mathcal{E}$. We also construct reverse triple $(?,r^{-1},h)$ for each triple in the training set and apply the same setting. For all triples, KGC decoder provides corresponding probability of truth $p_m(t|(h,r)) = softmax(f_m(h,r,t))$ computed with a $softmax$ applied to the output of the score function. We denote $p_m(t|(h,r))$ as probability obtained in modality m. The CE loss of modality m can be calculated as Equation (1).

$$\begin{aligned}\mathcal{L}_m &= -\sum_{t}^{|\mathcal{E}|} log(p_m(t|(h,r)) \\ &= CE(p_m(t|(h,r)))\,.\end{aligned} \tag{1}$$

Confidence-constraint Training. Visual and text modalities usually embody contradictory information due to data complexity and diversity. We notice that the contradiction lies in the fact that the modality information of *entity* is not always relevant to the *knowledge* of factual triple. In consequence, visual and text modality usually present uncertainty. To ease the uncertainty, we train the visual or text modality KGC in a confidence-constraint manner with a temperature-scaling technique (Guo et al., 2017). Since the predicted probability can approximately represent the confidence score of predictions, we simply compact the probability distribution by adding a temperature parameter $\mathcal{T}$ to the output of the KGC decoder as Equation (2) and extend CE loss to a confidence-constraint form as Equation (3). In this way, the confidence of predictions is constrained and the distribution is softened while prediction results remain the same. Formally, we use the confidence-constraint loss for visual and text modality as $\mathcal{L}_v^{cc}$ and $\mathcal{L}_t^{cc}$ respectively.

$$p_m^{cc}(h,r,t) = softmax(\frac{f_m(h,r,t)}{\mathcal{T}})\,. \tag{2}$$

$$\mathcal{L}_m^{cc} = CE(p_m^{cc}(t|(h,r)))\,. \tag{3}$$

Overall Objective. In the training phase, we simultaneously train three modalities to learn intra-modal representations separately after the overall objective $\mathcal{L}_{KGC}$. The overall KGC objective is the sum of modality losses as Equation (4).

$$\mathcal{L}_{KGC} = \mathcal{L}_s + \mathcal{L}_v^{cc} + \mathcal{L}_t^{cc}\,. \tag{4}$$

3.6 Inference

In the inference phase, we explore combination mechanisms of modality predictions by modeling modality weights. Modalities have strengths and limitations due to data quality and entity coverage, which are always complementary to each other. Appropriately adjusting modality weights to fully exploit the complementary strengths would lead to better prediction performance.

Ensemble Inference. Inspired by Chen et al. (2020), we exploit ensemble inference to obtain final predictions. We propose to directly combine scores instead of ranks since information from score distributions may get lost during the ranking process. For each query, we obtain three score distributions $f_m(h,r,t), m \in \mathcal{M} = \{\mathcal{S}, \mathcal{V}, \mathcal{T}\}$ from three modalities, which could directly reflect the strengths and limitations of modality for entity predictions. The scores are combined as Equation (5).

$$\mathcal{F}(h,r,t) = \sum_{m \in \mathcal{M}} w_m f_m(h,r,t)\,. \quad (5)$$

Next, we propose three variants of MoSE: MoSE-AI, MoSE-BI and MoSE-MI, which varies in the modality weight w_m calculation. We utilize a small amount of unbiased meta-set to learn modality weights that can be finely transferred to test-set. We choose validation-set as meta-set.

Equal-importance Average Inference. We utilize modality average weight without considering modality importance as a baseline MoSE-AI. For all modalities, we average the scores to obtain final prediction as Equation (6).

$$\mathcal{F}^{AI}(h,r,t) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} f_m(h,r,t)\,. \quad (6)$$

Relation-aware Boosting Inference. We find that entity-relevant triples are sparse and thus hard to capture the accurate correlation between entity and modality importance. In this paper, we assume that the relation of each modality varies in relevance level. So we propose to learn modality weight in relation-level to adjust modality importance to final predictions. We divide meta-set by relation type and upgrade RankBoost (Freund et al., 2003) mechanism to generate modality weights $w_m(r)$ corresponding to relation r and combine modality scores as Equation (7). The MoSE-BI Algorithm is illustrated in Appendix A.

$$\mathcal{F}^{BI}(h,r,t) = \sum_{m \in \mathcal{M}} w_m(r) f_m(h,r,t)\,. \quad (7)$$

Instance-specific Meta-Learner Inference. However, for KGs with fewer relations, such as WN9 (Xie et al., 2017) with only 9 relations, MoSE-BI is limited by coarse-grained relation-level weight learning. Thus we propose to train a meta-learner to find optimal weight functions for each triple instance. Following Shu et al. (2019), we exploit an MLP (Multilayer Perceptron) with only one hidden layer as a meta-learner to combine the scores and approximate true predictions. For a triple (h,r,t), we use the concatenation of three scores $F(h,r,t) = [f_m(h,r,t)], m \in \mathcal{M}$ as input, and train the weighted scores to fit the final predictions. The final prediction is obtained as Equation (8) where weight parameter $\mathbf{\Theta}$ is trained in meta-set and transferred to test-set.

$$\mathcal{F}^{MI}((h,r,t);\mathbf{\Theta}) = w(\mathbf{\Theta})F(h,r,t)\,. \quad (8)$$

The optimal weight $\mathbf{\Theta}$ is obtained with CE loss as Equation (9).

$$\mathcal{L}_{MI} = CE[softmax(\mathcal{F}^{MI}((h,r,t);\mathbf{\Theta}))]\,. \quad (9)$$

4 Experiments

4.1 Experimental Setting

Datasets. To evaluate the proposed model, we conduct experiments on three widely used KGC datasets: FB15K-237 (Toutanova et al., 2015), WN18 (Bordes et al., 2013), and WN9-IMG (Xie et al., 2017). The former two are unimodal KGC datasets with only structure modal, and the latter one contains both structure and visual modalities. We follow previous studies (Wang et al., 2021; Xie et al., 2016; Yao et al., 2019) to augment the text and visual modality information of each dataset. The dataset statistics are shown in Table 1.

Implementation details. To evaluate MoSE, four metrics are used, i.e., Hits@K, K=1, 3, 10, representing accuracy in top K predictions, and Mean Rank (MR). Higher Hits@K and lower MR indicate better performance. We use Pytorch 1.11.0 to implement MoSE. The operating system is Ubuntu 18.04.5. We use a single NVIDIA A6000 GPU with 48GB of RAM.

We report the results of three MoSE variants which vary in the inference methods. MoSE-AI refers to MoSE with average inference. MoSE-BI refers to MoSE with boosting inference. MoSE-MI refers to MoSE with meta-learner inference.

We follow the widely-used filtered setting (Bordes et al., 2013), i.e., excluding other true entities

Dataset	#Rel.	#Ent.	#Train	#Valid	#Test
FB15K-237	237	14,541	272,115	17,535	20466
WN18	18	40,943	141,442	5,000	5,000
WN9	9	6,555	11,741	1,337	1,319

Table 1: Datasets statistics for MKGC.

Model	FB15K-237				WN18				WN9			
	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓
Unimodal KGE methods												
TransE	0.198	0.376	0.441	323	0.040	0.745	0.923	357	0.864	0.901	0.917	146
DistMult	0.199	0.301	0.466	512	0.335	0.876	0.940	655	0.531	0.871	0.911	241
ComplEx	0.194	0.297	0.450	546	0.936	0.945	0.947	-	0.901	0.913	0.922	256
RotatE	0.241	0.375	0.533	177	0.942	0.950	0.957	254	0.889	0.906	0.922	175
Multimodal KGE methods												
IKRL (UNION)	0.194	0.284	0.458	298	0.127	0.796	0.928	596	-	-	0.938	21
TransAE	0.199	0.317	0.463	431	0.323	0.835	0.934	352	-	-	0.942	17
RSME	0.242	0.344	0.467	417	0.943	0.951	0.957	223	0.878	0.912	0.923	55
MoSE-AI	0.255	0.376	0.518	135	0.929	0.946	0.962	23	0.840	0.932	0.963	**4**
MoSE-BI	**0.281**	**0.411**	**0.565**	**117**	0.884	0.953	0.972	8	0.831	0.923	0.964	**4**
MoSE-MI	0.268	0.394	0.540	127	**0.948**	**0.962**	**0.974**	**7**	**0.909**	**0.937**	**0.967**	**4**
Pre-trained Language Model methods												
KG-BERT	-	-	0.420	153	0.117	0.689	0.926	58	-	-	-	-
MKGformer	0.256	0.367	0.504	221	0.944	0.961	0.972	28	-	-	-	-

Table 2: Knowledge graph completion performance on FB15K-237, WN18, and WN9. We highlight the **best** and the second best results of each column. MoSE-BI performs the best on FB15k-237, and MoSE-MI achieves the best performance on WN18 and WN9. We can also observe that existing multimodal KGE methods do not perform as well as RotatE, which is a unimodal KGE method.

when evaluating. We exploit ComplEx (Lacroix et al., 2018) as KGC Decoder. In this paper, we mainly focus on contradiction in relation embeddings. Thus, we employ SOTA pretrained encoder to extract visual and text features of entities, i.e., ViT (Dosovitskiy et al., 2020) following RSME (Wang et al., 2021) for visual modality and BERT (Kenton and Toutanova, 2019) for text modality. We use Adagrad (Duchi et al., 2011) to optimize the model. The hyperparameters are selected with the best Hits@10 on the validation set.

Baselines. We compare MoSE with several baselines to demonstrate the advantage of our framework. We mainly compare MoSE with KGE methods, which can be grouped into two categories: (1) the unimodal KGE methods, including TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2018); (2) the multimodal KGE methods, including a) E-TCR methods: IKRL (Xie et al., 2017), b) I-TCR methods: TransAE (Wang et al., 2019) and RSME (Wang et al., 2021). We also list the results of pre-trained language models (PLMs) for KGC, i.e., KG-BERT (Yao et al., 2019) and MKGformer (Chen et al., 2022).

4.2 Comparison to the Baselines

The experimental results in Table 2 show that MoSE obtains the best performance compared to all 9 baselines, which demonstrates the superiority of MoSE. Compared to unimodal KGE methods, MoSE outperforms the best unimodal method RotatE, while other multimodal methods do not. Compared to multimodal KGE methods, MoSE achieves 2% - 10% improvements in Hits@10 and 13 - 216 improvements in MR over the best existing methods. It is worth noting that even compared to the pre-trained language model methods, MoSE outperforms KG-BERT and MKGformer in all metrics on FB15K-237 and WN18 datasets.

Q1: Does MoSE *succeed in avoiding modality interference?* Compared with the corresponding base model, while other multimodal methods face a certain level of performance decline, MoSE achieves consistent improvements in all metrics. For example, the Hits@1 and Hits@3 of IKRL drop compared to those of TransE, and the Hits@3 of TransAE drops compared to that of TransE on the FB15K-237 dataset. Even SOTA RSME faces a slight drop on the WN9 dataset in terms of Hits@1 and Hits@3 compared to ComplEx. It reveals that MoSE can steadily enhance unimodal models with auxiliary modality information and successfully avoid modality interference to structure modality.

Q2: Is it necessary to assess modality importance? We explored three inference methods with modality importance in different aspects. MoSE-AI treats each modality equally and does not consider modality importance at all, while MoSE-BI considers modality importance in relation-level and MoSE-MI in instance-level. As shown in Table 2, MoSE-BI performs the best on FB15K-237 and

Model	FB15K-237				WN18				WN9			
	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓
I-TCR	0.192	0.303	0.439	439	0.945	0.953	0.958	298	0.588	0.755	0.847	126
E-TCR-AI	0.248	0.367	0.511	140	0.910	0.945	0.960	27	0.779	0.916	0.958	5
MoSE-AI	0.255	0.376	0.518	135	0.929	0.946	0.962	23	0.840	0.932	0.963	**4**
E-TCR-BI	0.271	0.402	0.554	121	0.858	0.945	0.968	10	0.756	0.914	0.959	**4**
MoSE-BI	**0.281**	**0.411**	**0.565**	**117**	0.884	0.953	0.972	8	0.831	0.923	0.964	**4**
E-TCR-MI	0.247	0.367	0.510	135	0.924	0.956	0.971	12	0.878	0.930	0.958	5
MoSE-MI	0.268	0.394	0.540	127	**0.948**	**0.962**	**0.974**	**7**	**0.909**	**0.937**	**0.967**	**4**

Table 3: Effectiveness of relation decoupling. For the I-TCR variation, we fuse the multimodal entities and exploit a shared relation representation, which yields a unified prediction. For the E-TCR variation, we replace the modality-split relation embeddings of MoSE with a shared relation embedding, which yields three prediction scores as well.

MoSE-MI performs the best on WN18 and WN9. All the best inference methods on the three datasets outperform MoSE-AI. It demonstrates the necessity of assessing modality importance for MKGC.

Q3: How to choose the suitable inference methods? As we can observe, different inference methods expert in different KG characteristics. The relation-aware inference MoSE-BI performs better in complex KGs with extensive relation types such as FB15K-237 (237 relations) and fails in KGs with fewer relation types such as WN18 and WN9 (18 and 9 relations respectively) while instance-specific inference MoSE-MI performs the opposite. The possible reason is that the inference methods are with different capabilities to approximate the optimal combination of modalities. MoSE-BI is easy to scale to KGs with more relations and able to achieve relatively better performance. Though MoSE-MI performs the best in two datasets, we believe that the single layer MLP may still limit the fitting capability of MoSE-MI.

4.3 Effectiveness of Relation Decoupling

Since the TCR baselines in Table 2 vary in KGC decoder and modality types, we further investigated different TCR variations of MoSE under the same setting to demonstrate the effectiveness of relation decoupling. The results are presented in Table 3. For I-TCR and E-TCR variation, we replace the modality-split relation embeddings in MoSE with a single modality-shared relation embedding. For I-TCR variation, we further fuse the multimodal entities with weighted concatenation, which yields a unified prediction.

As shown in Table 3, MoSE outperforms all its E-TCR variations under the same inference method. As for I-TCR method, the best performance of MoSE exceeds I-TCR in all metrics. It demonstrates the necessity of modality relation decoupling. We also notice that I-TCR exceeds MoSE-AI in Hits@1/3 and MoSE-BI in Hits@1 on WN18. The possible reason is that the modality information of WN18 has many mutual semantics. So modality fusion brings accurate entity representations. However, I-TCR obtains a large MR score, indicating it is not stable as MoSE for MKGC.

4.4 Modality Ablation

To demonstrate how each modality supports final predictions, we conduct modality ablation. Table 4 shows the experimental results obtained by (1) ensemble inference of three structure unimodal models Str-Str-Str-AI/BI/MI, (2) modality-split predictions obtained by KGC decoder Mose-Str/Vis/Text.

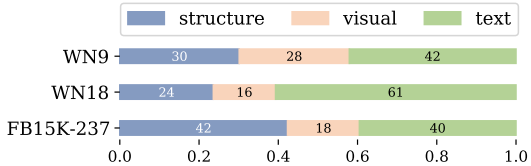
The improvements of Str-Str-Str over MoSE-Str is insignificant compared to that of MoSE-BEST over MoSE-Str. It reveals that MoSE improves the base unimodal model via effectively utilizing modality information instead of performing ensemble inference. For modality-split predictions MoSE-Str/Vis/Text, no single one of three prediction performances exceeds MoSE-BEST. It demonstrates that modalities in MoSE effectively enhance each other and successfully avoid modality mutual interference. The modality-split predictions also indicate the modality quality for assisting MKGC. The structure modality, which is directly learned from KGs, remains the best performance on all datasets, while visual modality has erratic performance and text modality consistently provides the best MR metric.

4.5 Case Study

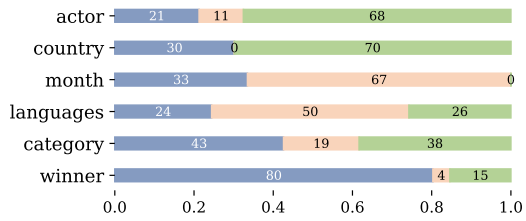
To demonstrate the intuitive ability of MoSE to assess modality importance, we conduct case studies with MoSE-BI, which provides modality weights

Model	FB15K-237				WN18				WN9			
	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓
MoSE-BEST	**0.281**	**0.411**	**0.565**	**117**	**0.948**	**0.962**	**0.974**	**7**	**0.909**	**0.937**	**0.967**	**4**
Str-Str-Str-AI	0.256	0.386	0.542	166	0.945	0.954	0.960	247	0.909	0.916	0.923	201
Str-Str-Str-BI	0.262	0.392	0.547	162	0.946	0.954	0.960	247	0.909	0.916	0.923	200
Str-Str-Str-MI	0.256	0.386	0.541	206	0.945	0.954	0.960	322	0.909	0.916	0.923	263
MoSE-Str	0.264	0.392	0.545	168	0.946	0.954	0.960	264	0.908	0.914	0.922	193
MoSE-Vis	0.167	0.242	0.329	890	0.527	0.611	0.685	2017	0.092	0.231	0.392	243
MoSE-Text	0.245	0.364	0.500	161	0.255	0.442	0.618	96	0.262	0.487	0.709	27

Table 4: The experimental results of modality ablation. MoSE-BEST refers to results obtained by the best variant of MoSE for each dataset, i.e., MoSE-BI for FB15K-237 and MoSE-MI for WN18 and WN9. Str, Vis, Text refer to structure, visual and text modality respectively. Str-Str-Str-AI/BI/MI refers to results obtained by replacing both visual and text modalities in MoSE with structure modality. MoSE-Str/Vis/Text refers to the modality-split prediction performances of each modality.



(a) Average modality weights (%) of datasets.



(b) Examples of modality weights (%) of FB15K-237 relations. Relations are abbreviated. See Appendix B for full relation names.

Figure 3: Modality weights (%) learned by MoSE-BI.

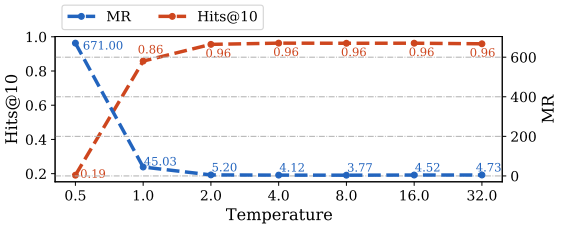


Figure 4: Temperature Parameter Analysis conducted by MoSE-AI on WN9.

for each modality corresponding to each relation. Figure 3 shows modality weights in Equation (7) to combine predictions from multiple sources.

Modality Importance. Figure 3a shows average modality weights on each dataset obtained by MoSE-BI. We can observe that text modality provides the greatest contributions on WN18 and WN9, while visual modality provides the minimum on all datasets. It demonstrates that text modality provides valuable information supporting knowledge predictions while visual modality in the opposite. The possible reason is that descriptions often mention relevant entities, while images are only highly related to entity itself.

Relation Cases. Figure 3b presents some examples to show how much each modality contributes to relation learning on FB15K-237. The higher level of modality importance often stems from more relation-relevant modality information. For example, for relation country_of_origin (abbr. country) shown in Figure 1a, the text modality provides more relevance information than visual modality. As shown in Figure 3b, text modality presents importance up to 70% while visual modality presents 0%. The results also demonstrate that MoSE-BI is able to identify which modality is more credible and then assign a higher weight in a fine-grained relation level.

4.6 Uncertainty in MKGs

To investigate the uncertainty of MKG predictions, we adjust the temperature parameter as shown in Figure 4. We use MoSE-AI to rule out the impact of ensemble inference. We vary the temperature $\mathcal{T}$ in Equation (2) from 2^{-1} to 2^5 with exponential growth. As the temperature increases, the performance tends to grow and converge to stable. When $\mathcal{T} = 2^{-1}$, the confidence to visual and text modality is enlarged and MoSE faces great performance decline. We can also observe that MoSE with larger $\mathcal{T}$ always outperforms $\mathcal{T} = 2^0 = 1$ in which the confidence is not constrained. It proves our assumption about the uncertainty of visual and text modalities.

5 Conclusion

In this paper, we propose a novel modality split learning and ensemble inference framework for multimodal knowledge graph completion called MoSE. MoSE first decouples modality-shared relation embedding to modality-split relation embeddings and performs modality-split representation learning in the training phase, aiming at overcoming modality relation contradiction. Then, MoSE exploits three ensemble inference techniques to combine the modality-split predictions by assessing modality importance. Experiment results demonstrate that MoSE outperforms state-of-the-art methods for MKGC task on three widely-used datasets.

Limitations

Despite that MoSE achieves some gains by modality-split learning and ensemble inference, MoSE still has the following limitations:

First, MoSE does not fully exploit visual modality. Since the image of the entity is highly *self-relevant* and covers less information about *other* related entities, we reduce the visual modality importance during ensemble inference to cater to the MKGC task, which heavily relies on the relationship between entities. Nevertheless, we believe there are other ways to exploit visual modality suitably.

Second, for a fair comparison, we follow SOTA method RSME (Wang et al., 2021) and utilize a single-image setting. We believe that under the multiple-image setting, the problem of modality relation contradiction still holds. Intuitively, even with more images, the image of the entity "The United States of America" in Figure 1a is unlikely to involve the entity "Friends". Quantitatively, the similarity of multiple images from the same entity is up to 99.250% on FB15K-237 and 99.255% on WN18 respectively. Therefore, there is little difference between single-image and multiple-image settings in our work. However, more images may introduce more side information, such as related entities, from which MKGC model may benefit.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. U1936206, 62272250, 62002178, U1903128), and Science and Technology on Communication Networks Laboratory, Shijiazhuang, Hebei, P. R. China (No. FFX22641X005).

References

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.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. *arXiv preprint arXiv:2205.02357*.

Xuelu Chen, Muhao Chen, Changjun Fan, Ankith Uppunda, Yizhou Sun, and Carlo Zaniolo. 2020. Multilingual knowledge graph completion via ensemble knowledge transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3227–3238.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204.

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

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32.

Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *CIKM*, pages 1405–1414.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. Rotate: Knowledge graph embedding by relational rotation in complex space. In *ICLR*.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, pages 1499–1509.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *ICML*.

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is visual context really helpful for knowledge graph? a representation learning perspective. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2735–2743.

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, volume 30.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *AAAI*.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *arXiv preprint arXiv:2202.05786*.

Appendix

A MoSE Boosting Inference Algorithm

In this section, we present the algorithm detail of relation-aware boosting inference method MoSE-BI in Algorithm 1. We first divide the meta-set $\mathcal{D}$ to relation-aware sets $\mathcal{D}_r$ by relation type. At each set, we exploit RankBoost (Freund et al., 2003) algorithm to model modality importance and combine modality scores to obtain final predictions.

The main idea of RankBoost is to turn a ranking problem into a classification problem. The score of corrupted entity $h_m(h,r,e)$ less than that of true tail entity $h_m(h,r,t)$ is seen as True prediction while False prediction in the opposite. It is worth noting that we select the best modality in each round to eliminate the impact of modality order.

B Relation names

Table 5 shows original relation names in Figure 3b.

Algorithm 1 MoSE Boosting Inference Algorithm

Input: The meta-set $\mathcal{D} = \{(h,r,t)\}, h,t \in \mathcal{E}, r \in \mathcal{R}$; Modality set $\mathcal{M}$; Modality-split scores $f_m(h,r,t), m \in \mathcal{M}$;
Output: $\mathcal{F}^{BI}(h,r,t) = \sum_{m\in\mathcal{M}} w_m(r) f_m(h,r,t)$
divide meta-set $\mathcal{D}$ by relation $r \in \mathcal{R}$ to $\mathcal{D} = \{\mathcal{D}_r\}$
for each relation set $\mathcal{D}_r \in \mathcal{D}$ **do**
  init $D_1(e) = \frac{1}{|\mathcal{M}|}, w_m(r) = 0, m \in \mathcal{M}$
  for each modality $m = 1, ..., |\mathcal{M}|$ **do**
    $h_m(e) = \begin{cases} 1, f_m(h,r,e) < f_m(h,r,t) \\ -1, f_m(h,r,e) >= f_m(h,r,t) \end{cases}$
    for $i = 1, ..., |\mathcal{M}|$ **do**
      $w_m^i = \frac{1}{2} ln(\frac{\sum_{e\in\mathcal{E}, h_m(e)=1} D_m(e)}{\sum_{e\in\mathcal{E}, h_m(e)=-1} D_m(e)})$
    select the best and unchosen modality with max weight $w_m = \max\{w_m^1, ..., w_m^{|\mathcal{M}|}\}$
    $Z_m = \sum_{e\in\mathcal{E}} D_m(e) exp(-w_m h_m(e))$
    $D_{m+1}(e) = \frac{D_m(e) exp(-w_m h_m(e))}{Z_m}$
    $w_m(r) = w_m(r) + w_m$
return $\mathcal{F}^{BI}(h,r,t) = \sum_{m\in\mathcal{M}} w_m(r) f_m(h,r,t)$

relation abbv.	relation name
actor	/tv/tv_program/regular_cast./tv/ regular_tv_appearance/actor
country	/tv/tv_program/country_of_origin
month	/travel/travel_destination/climate./ travel/travel_destination_monthly_climate/month
languages	/tv/tv_program/languages
category	/award/award_category/category_of
winner	/award/award_ceremony/awards_presented./ award/award_honor/award_winner

Table 5: Relation abbreviations and full names.