Multimodal Biological Knowledge Graph Completion via Triple Co-attention Mechanism

Derong Xu
State Key Laboratory of Cognitive Intelligence
University of Science and Technology of China
derongxu@mail.ustc.edu.cn

Jingbo Zhou*
Business Intelligence Lab
Baidu Research
zhoujingbo@baidu.com

Tong Xu*
State Key Laboratory of Cognitive Intelligence
University of Science and Technology of China
tongxu@ustc.edu.cn

Yuan Xia
Intelligent Healthcare Unit
Baidu
xiayuan@baidu.com

Ji Liu
Big Data Lab
Baidu Research
liuji04@baidu.com

Enhong Chen
State Key Laboratory of Cognitive Intelligence
University of Science and Technology of China
cheneh@ustc.edu.cn

Dejing Dou
BCG X
dejingdou@gmail.com

***Abstract*—Biological Knowledge Graphs (BKGs) can help to model complex biological systems in a structural way to support various tasks. Nevertheless, the incompleteness problem may limit the performance of existing BKGs, which still deserves new methods to reveal the missing relations. Though great efforts have been made to knowledge graph completion, existing methods are not easy to be adapted to the multimodal biological information such as molecular structures and textual descriptions. To this end, we propose a novel co-attention-based multimodal embedding framework, named CamE, for the multimodal BKG completion task. Specifically, we design a Triple Co-Attention (TCA) operator to capture and highlight the same semantic features among different modalities. Based on TCA, we further propose two components to handle multimodal fusion and multimodal entity-relation interaction, respectively. One is the multimodal TCA fusion module to achieve a multimodal joint representation for each entity in the BKG. It aims to project different modal information into a common space by capturing the same semantic features and overcoming the modality gap. The other is the relation-aware interactive TCA module to learn interactive representation by modelling the deep interaction between multimodal entities and relations. Extensive experiments on two real-world multimodal BKG datasets demonstrate that our method significantly outperforms several state-of-the-art baselines, including 10.3% and 16.2% improvement w.r.t MRR and Hits@1 metrics over its best competitors on public DRKG-MM dataset.**

***Index Terms*—Multi-Modal, Biological Knowledge Graph, Knowledge Graph Completion, Co-attention**

I. Introduction

Biological Knowledge Graphs (BKGs) are an emerging type of Knowledge Graph whose nodes represent biological entities (mainly including genes, disease, and compounds) and edges represent the relations among the entities [1]–[4]. With modelling the complex biological systems in such a structural way, BKGs are able to assist in a wide range of biological applications, such as protein targets identification, drug repurposing, and polypharmacy side-effects prediction [2], [5]–[7].

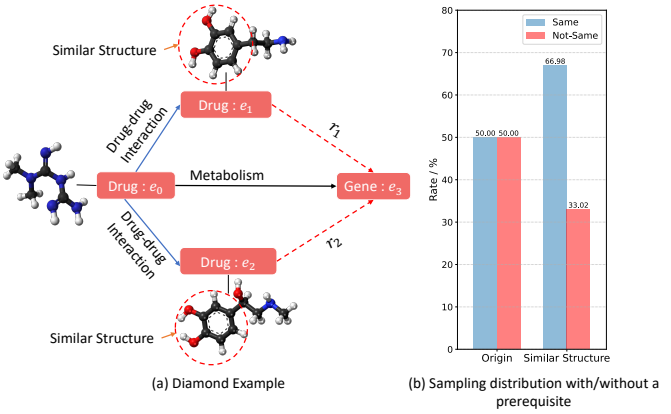


Fig. 1. An example to demonstrate that molecular structure (a kind of multimodal information of BKGs) is helpful for identifying relations of BKGs. Given a diamond from DRKG-MM data (a real-life BKG dataset used in the experiment section) that entities are $< e_0, e_1, e_2, e_3 >$ with e_0, e_1, e_2 being drugs, and e_3 being gene (as shown in Fig. 1(a)), we randomly select 10,000 such diamonds whose rate of "Same" ($r_1 = r_2$) and "Not-Same" ($r_1 \neq r_2$) is 50% vs 50%. Then we samples some diamonds from the 10,000 diamonds with a prerequisite that e_1 and e_2 have similar molecular structures. The results in Fig. 1(b) show that the average accuracy is 66.98%. While, theoretically, a random sampling should only yield 50% accuracy. It demonstrated that the introduction of structural similarity of molecule (e.g. whether the diamond is "Same" structure) could help to judge whether r_1 is equal to r_2, and thus enhance prediction performance. More details are elaborated in experimental analysis Section V-H1.

However, just like KGs in other domains, BKGs also suffer from the KG completion problem since the incompleteness of BKGs leads to performance degradation in many different downstream applications.

Meanwhile, we observe that rich semantic features hidden in multimodal information of BKGs like molecular structure could benefit the completion task. To demonstrate our claim, we show a toy example in Fig. 1, where the "Same" and "Not-Same" distribution of diamond relations is changed from "50.00% vs 50.00%" to "66.98% vs 32.02%" after considering the molecular structure similarity as a prerequisite to sampling.

*Jingbo Zhou and Tong Xu are corresponding authors. This work was done when Derong Xu was an intern at the Baidu Research under the supervision of Jingbo Zhou. The code is available at https://github.com/PaddlePaddle/PaddleHelix/tree/dev/research/CamE.

Meanwhile, the correlation between textual and molecular data is of much importance. For example, phenolic compounds comprise one or more aromatic rings with attached hydroxyl groups in their structures [8]; whereas they usually have the suffix "-phrine" in their names (it is also a kind of textual description) which reveals its medical properties and possibly related gene and disease, and benefits to predict the link in knowledge graphs. Therefore, it is valuable to jointly utilize the structured knowledge and multimodal information of BKGs as prior knowledge to improve the accuracy of the completion task.

Yet, it remains an outstanding challenge to integrate such biological multimodal information. Though graph embedding methods [9]–[11] have proven to be effective for KG completion, they mainly focus on the structured knowledge in KGs and lack the flexibility to consider the other modal data in BKGs like molecular structures and textual descriptions of entities. There are a few studies [12]–[16] to investigate the entity images and textual descriptions to extract knowledge representation. Nevertheless, since these methods are designed for images and texts, they cannot capture the underlying common semantic features between molecule, textual description, and structured knowledge in BKGs, resulting in poor performance as shown in our experimental evaluation.

To this end, we propose a novel triple Co-attention multimodal Embedding (CamE) framework specially designed for multimodal BKG completion. The intuition behind CamE is to capture the commonly-repeated information in different modalities of BKGs, which also serves as the anchor point to help other parts to align and promote modal fusion. For this purpose, we first introduce a dedicated Triple Co-Attention (TCA) operator. Then upon the TCA operator, CamE can be divided into two main components which are: 1) MultiModal TCA Fusion module (MMF); 2) Relation-aware Interactive TCA module (RIC).

Specifically, the TCA operator is expected to capture and highlight the same semantic features among modalities (i.e. molecular structure, textual description, and structured knowledge) via learning co-affinity and intra-affinity matrices. As far as we know, we are the first to adopt co-attention mechanism to handle the BKG completion task. Whereas, the original co-attention mechanism [17] cannot directly handle BKG completion task since it is designed for visual question answering tasks whose inputs are feature matrices of image and question. In our proposed TCA operator, we use one co-affinity matrix and two intra-affinity matrices to achieve the co-attention mechanism. The co-affinity matrix defines the mutual attention between two modality inputs. For each of the two inputs, the intra-affinity matrix defines its own attention, which shares parts of the mapping parameters with the co-affinity matrix to restrict the representation to the same subspace. Furthermore, we introduce multi-head TCA and design a learnable temperature sequence with fixed interval to increase the diversity of the multi-head.

Moreover, the MMF module is designed to fuse heterogeneous data from multimodal representations to a joint representation. In MMF, a pairwise TCA matching step conducts TCA operator between each pair of modalities to extract the common semantic features of each other. After that, an exchanging fusion step is further introduced to exchange the unimportant information of each individual modality measured by attention map, which achieves to bridge modality gap and promotes cross-modal fusion.

Besides, the RIC module aims to learn multimodal entity-relation interactive representation. The motivation of this module is that the interactions between entities and relations are particularly important and effective for KG completion tasks [11], [18], [19]. The RIC module can build the deep interaction between multimodal entities and relations using the TCA operator. In this way, all elements in multimodal representation of entities are possible to establish contact with all elements in relation embedding by multiplication. Finally, we construct a multi-channel feature map by stack modality joint and interactive representations as a multi-view of data, which are then fed into the convolutional neural network to infer the missing links.

We conduct extensive experimental evaluations on two real-world multimodal BKG datasets. Experimental results show our approach can achieve significant improvement over other state-of-the-art methods. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to integrate multimodal information including textual description and molecular structure for BKG completion task.
- We propose CamE, a novel co-attention-based multimodal embedding framework, which takes full advantage of multimodal information to tackle the BKG completion task. The core of CamE includes a specially designed TCA operator and two components to deal with multimodal fusion and multimodal entity-relation interaction respectively.
- Experimental results on two real-world multimodal biological datasets show that our framework significantly outperforms existing state-of-the-art methods by a large margin.

II. Related Work

Our work is closely related to KG completion and BKG applications. Here we briefly discuss the related works from these two perspectives as follow.

A. Knowledge Graph Completion

In the past decade, many research efforts have been devoted to the KG completion/embedding task. For example, TransE [9] models the relation as a translation between the head and tail entity. TransH [20], TransR [21], and TransD [22] extend this idea by using different projection strategies to deal with the complex relations such as the Many-to-1 relation. DistMult [18] proposes a simplified bilinear formulation to capture productive interactions in relational data and compute efficiently. There are also other recent studies, e.g., RotatE [10], ConvE [11], ComplEx [23], DualE [24] and PairRE [25]. ComplEx

[23] further extends the DistMult [23] in the complex vector space to capture both symmetric and antisymmetric relations. RotatE [10] models the relation as rotation from the head entity to the tail entity also in complex space. All these methods learn the embedding solely based on the structured knowledge in KG and ignore the rich multimodal information, which may limit the further improvement in performance.

There are a few studies using extra multimodal data to improve performance [26] [27] [28] [29]. IKRL [12] is the first work to combine the image with structured knowledge for KG embedding, which uses an attention-based method to choose the best image instance of each entity. Its score function of a triple is designed to utilize structured knowledge information as well as visual information following the framework of TransE. The authors in [14] try to integrate external text information and defines the energy of a triple as the sum of sub-energy functions of each modality. KBLRN [30] utilizes the probabilistic product of experts to integrate the relational and numerical features. Pezeshkpour et al. [31] further propose to learn multimodal embedding by utilizing different neural encoders for each modal data and integrate them into existing KG embedding methods. TransAE [13] combines multimodal autoencoder and TransE to get the entity representation. There is also a recent study [26] to leverage the transformer architecture to fuse the visual and text representation for KG completion. However, to the best of our knowledge, there is no previous study utilizing the graph data like molecule structure for KG completion [26].

B. Biological Knowledge Graph

Benefiting from the ability of KG to model relational data in complex biological systems, BKG has attracted a lot of research attention in recent years. DRKG [1] introduces a comprehensive BKG, which includes data from seven open data sources. The authors in [2] integrate various open biomedical resources in a unified format and creates a BKG from them. There is also an extensive study about how KG embedding models can be applied to different biological applications [32]. In this paper, we design a novel framework tailored to integrate the multimodal biological information to improve the performance of BKG completion task.

III. Preliminaries

We denote the multimodal BKG $\mathcal{G} = \{(h, r, t)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where $\mathcal{E}$ and $\mathcal{R}$ are the set of entities and relations. Each triplet (h, r, t) means a relationship of type $r \in \mathcal{R}$ between the head entity $h \in \mathcal{E}$ and tail entity $t \in \mathcal{E}$. We define the molecular embedding as $\boldsymbol{h_m} \in \mathbb{R}^{d_m}$, textual embedding as $\boldsymbol{h_t} \in \mathbb{R}^{d_t}$, and structured embedding as $\boldsymbol{h_s} \in \mathbb{R}^{d_s}$. Also, d_m, d_t and d_s are the corresponding numbers of dimensions. Moreover, we denote relation embedding as $\boldsymbol{r} \in \mathbb{R}^{d_r}$, tail entity embedding as $\boldsymbol{t}_s \in \mathbb{R}^{d_e}$, d_e and d_r are the dimensions of entity and relation embedding vector respectively.

The initial vector of textual description and molecular structure are obtained by pre-trained models before inputting into our model. For the textual information, we use CharacterBERT [33] to extract the textual feature vector of entities in DRKG-MM. For OMAHA-MM, we employ the BERT [34] trained on cased Chinese text to embed the textual data, which is implemented by the code of Transformers package [35]. We chose the output of the last layer and averaged all word vectors to get a final embedding. Specifically, the embedding of Gene is also generated by CharacterBERT [33]. As for the molecular information, we utilize a pre-training Graph Neural Networks [36] to get the molecular embedding. To be specific, we use the pre-trained GIN model, whose training strategy is to predict randomly masked node and edge attributes, to extract the molecular feature. What's more, the structural embedding is learned by CompGCN [37] with their official codes.

TABLE I
Explanations of key Notations

Notations	Explanations
$\boldsymbol{h_m}$	Molecular embedding
$\boldsymbol{h_t}$	Textual embedding
$\boldsymbol{h_s}$	Structured embedding
$\boldsymbol{r}$	Relation embedding
$\boldsymbol{t}_s$	Tail entity embedding
$\boldsymbol{h}_f$	Multimodal joint embedding
$\boldsymbol{W}$	Trainable projected matrix

The BKG completion task is to infer missing relations in a multimodal BKG, which can be modeled as a ranking problem for link prediction. Our ultimate goal is to develop an effective KG embedding model to make full use of the multimodal information, then improve the score of positive triplets (h, r, t^+) and reduce the score of negative triplets (h, r, t^-).

IV. Methodology

In this section, we elaborate on how to implement CamE in Fig. 2. which mainly consists of Triple Co-Attention (TCA) operator, MultiModal TCA Fusion module (MMF), and Relation-aware Interactive TCA module (RIC). TCA operator is expected to capture the mutually reinforcing co-attention semantic features among modalities via learning co-affinity and intra-affinity matrices. Based on the TCA operator, MMF is designed to fuse heterogeneous data from multimodal representations to a joint representation, and RIC is designed to learn multimodal entity-relation interactive representation. Through the two modules, CamE excepts to fully mines the same semantic information among modalities, and models the deep interaction between multimodal entity and relation. In following subsections, we will introduce these modules in detail step-by-step.

A. TCA Operator

We first present a novel TCA operator, which is designed to capture mutual influence and attention between two modality inputs like molecule graph and textual description. Originally, co-attention mechanism [17] is designed for visual question answering tasks whose inputs are image and question. To the best of our knowledge, we are the first to use the co-attention

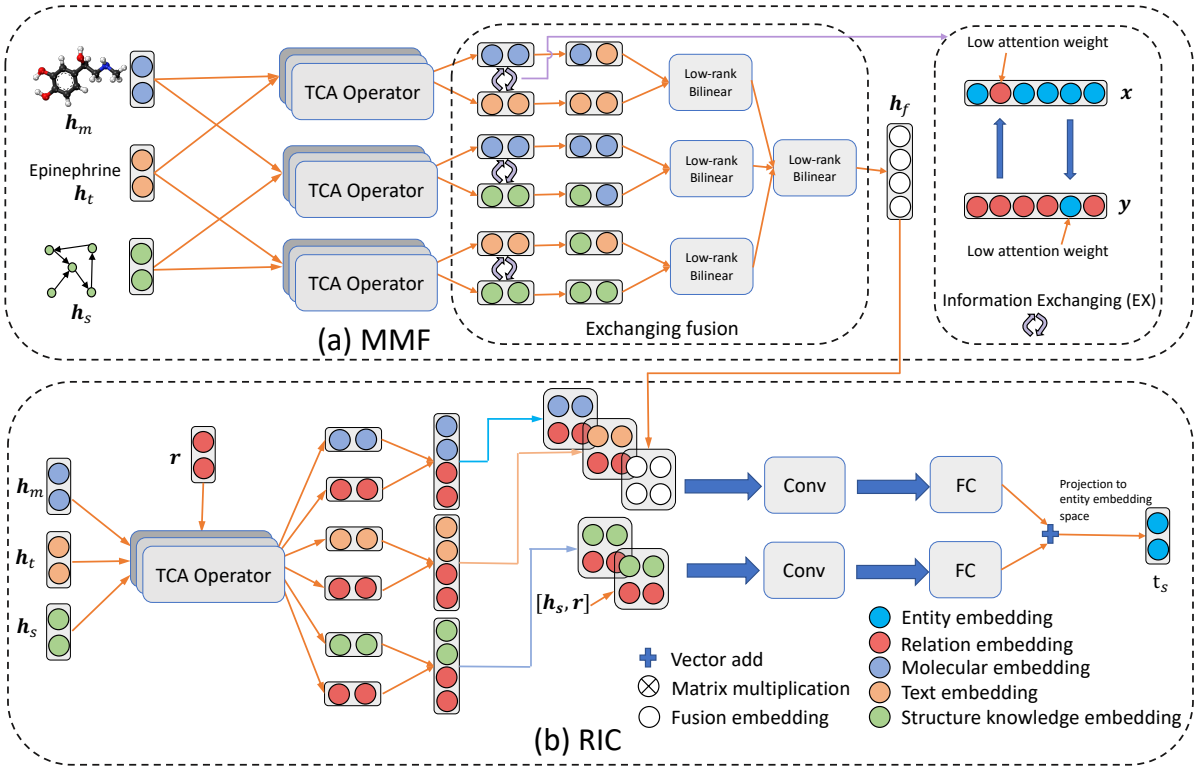


Fig. 2. The overall framework of CamE, which includes two modules, (a) MMF: Multimodal TCA Fusion Module and (b) RIC: Relation-aware Interactive TCA Module. Several TCA operators are stacked to represent the multi-head transformation, at which multiple outputs of TCA are fused together. Please refer to Section IV for details. The process of Information Exchanging is to exchanging each features of low attention, whose positions in a vector are replaced by the features of the other.

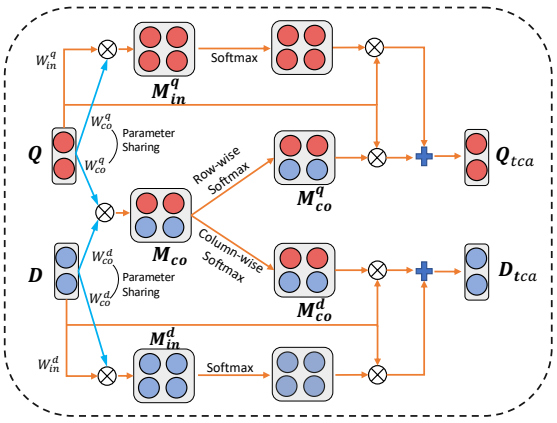


Fig. 3. Overview of triple Co-Attention (TCA) operator. the paired inputs of $\boldsymbol{Q}$ and $\boldsymbol{D}$ represent the feature vectors of two different modalities. The paired inputs are multiplied to triple affinity matrices to jointly learn the important information.

mechanism to tackle the BKG completion task. Since the input of CamE in our task is embedding vectors of entities (instead of matrices with sequence and patch information of questions and images), the existing co-attention mechanism is inappropriate to be applied on the multimodal data including molecule graph and textual description directly. Therefore, we specifically devise a new form of co-attention method, named as TCA operator, which takes the feature vectors from different modalities as inputs and projects them into a fused vector representation.

In general, TCA operator adopts triple affinity matrices to jointly learn the important information between two modality inputs, which aims to extract the mutually reinforcing co-attention features between the two input modality vectors, meanwhile extracts their own intra-attention features. Taking case 1 in Fig. 7 as an example, the entities with suffix "-cillin" in the text description also usually have a penicillin-type substructure in the molecule. This common semantic information reveal that the entities are a kind of penicillin, which is effective against many bacterial infections [38], [39]. From this information, we can know what diseases these entities are more likely to be associated with. TCA is designed to identify such features for modal fusion. We also apply scaling transformation and multi-head attention to increase the modelling ability of TCA.

One of triple affinity matrices is a co-affinity matrix with two modality data as inputs. The other two of them are intra-

affinity matrices with input from each modality but different non-linear mapping functions. The flow chart of TCA operator is shown in Fig. 3. In detail, We first present how to construct the co-affinity matrix. Specifically, given two input vectors $\boldsymbol{Q} \in \mathbb{R}^{d_1}$ and $\boldsymbol{D} \in \mathbb{R}^{d_2}$. $\boldsymbol{Q}$ and $\boldsymbol{D}$ serve as the feature vectors of two different modalities, and represent the pair inputs of Triple Co-Attention (TCA) operator. We expect the model to automatically highlight important parts between them. We first project them to a new vector space, which are then multiplied to get a co-affinity matrix $\boldsymbol{M_{co}} \in \mathbb{R}^{d_1 \times d_2}$ as follows:

$$\boldsymbol{M_{co}} = \sigma(\boldsymbol{Q}\boldsymbol{W_{co}^{q}}) \cdot \sigma(\boldsymbol{D^{T}}\boldsymbol{W_{co}^{d}}) \tag{1}$$

where σ is the sigmoid function, $\boldsymbol{W_{co}^{q}} \in \mathbb{R}^{d_1 \times d_1}$ and $\boldsymbol{W_{co}^{d}} \in \mathbb{R}^{d_2 \times d_2}$ are the learnable projection weights of $\boldsymbol{Q}$ and $\boldsymbol{D}$ respectively. Afterward, the obtained co-affinity matrix $\boldsymbol{M_{co}}$ is scaled though dividing by temperature τ. Then, we apply row-wise softmax operation and column-wise softmax operation to the scaled affinity matrix separately as follows:

$$\begin{aligned} \boldsymbol{M_{co}^{q}} &= softmax(\frac{\boldsymbol{M_{co}}}{\tau}, dim = 0) \in \mathbb{R}^{d_1 \times d_2} \\ \boldsymbol{M_{co}^{d}} &= softmax(\frac{\boldsymbol{M_{co}}}{\tau}, dim = 1) \in \mathbb{R}^{d_1 \times d_2} \end{aligned} \tag{2}$$

Finally, the original two vectors $\boldsymbol{Q}$ and $\boldsymbol{D}$ are multiplied by scaled affinity matrices respectively:

$$\begin{aligned} \boldsymbol{Q_{co}} &= \boldsymbol{Q^{T}} \cdot \boldsymbol{M_{co}^{q}} \\ \boldsymbol{D_{co}} &= \boldsymbol{M_{co}^{d}} \cdot \boldsymbol{D} \end{aligned} \tag{3}$$

Where $Q_{co} \in \mathbb{R}^{d_2}$ and $D_{co} \in \mathbb{R}^{d_1}$ represent the co-attention of two modalities learned by the co-affinity matrix. On the other hand, the procedure for obtaining the two intra-affinity matrices $\boldsymbol{M_{in}^{q}} \in \mathbb{R}^{d_1 \times d_2}$ and $\boldsymbol{M_{in}^{d}} \in \mathbb{R}^{d_1 \times d_2}$ is similar to that of obtaining the co-affinity matrix. They can be learned by:

$$\begin{aligned} \boldsymbol{M_{in}^{q}} &= \sigma(\boldsymbol{Q}\boldsymbol{W_{co}^{q}}) \cdot \sigma(\boldsymbol{Q}\boldsymbol{W_{in}^{q}}) \\ \boldsymbol{M_{in}^{d}} &= \sigma(\boldsymbol{D}\boldsymbol{W_{co}^{d}}) \cdot \sigma(\boldsymbol{D}\boldsymbol{W_{in}^{d}}) \end{aligned} \tag{4}$$

where $\boldsymbol{W_{co}^{q}} \in \mathbb{R}^{d_1 \times d_1}$ and $\boldsymbol{W_{co}^{d}} \in \mathbb{R}^{d_2 \times d_2}$ are the shared parameters of co-affinity and intra-affinity matrices to restrict the representation to the same subspace. $\boldsymbol{W_{in}^{q}} \in \mathbb{R}^{d_1 \times d_1}$ and $\boldsymbol{W_{in}^{d}} \in \mathbb{R}^{d_2 \times d_2}$ are learnable projection weights for intra-affinity matrices. The intra-attention of each inputs are obtained by multiplying intra-affinity matrices respectively:

$$\begin{aligned} \boldsymbol{Q_{in}} &= \boldsymbol{Q^{T}} \cdot softmax(\frac{\boldsymbol{M_{in}^{q}}}{\tau}, dim = 0) \\ \boldsymbol{D_{in}} &= \boldsymbol{D^{T}} \cdot softmax(\frac{\boldsymbol{M_{in}^{q}}}{\tau}, dim = 0) \end{aligned} \tag{5}$$

All the intra-attention and co-attention are added to get the final result:

$$\begin{aligned} \boldsymbol{Q_{tca}} &= \boldsymbol{Q_{co}} + \boldsymbol{Q_{in}} \\ \boldsymbol{D_{tca}} &= \boldsymbol{D_{co}} + \boldsymbol{D_{in}} \end{aligned} \tag{6}$$

where $\boldsymbol{Q_{tca}} \in \mathbb{R}^{d_2}$ and $\boldsymbol{D_{tca}} \in \mathbb{R}^{d_1}$ represent the output pair. In this way, we can extract the co-attention features and intra-attention features of multimodal data. We define the operator as $TCA(\cdot)$, which are served as core structure in the following two modules (MMF and RIC).

Furthermore, we introduce multi-head to our TCA to increase the modelling ability. The output of all heads will be concatenated together and then projected to the original dimension:

$$\begin{aligned} \boldsymbol{Q_{tca}} &= \boldsymbol{W}_{head}^{q}[\boldsymbol{Q_{tca}^{1}}; ...; \boldsymbol{Q_{tca}^{m}}] \\ \boldsymbol{D_{tca}} &= \boldsymbol{W}_{head}^{d}[\boldsymbol{D_{tca}^{1}}; ...; \boldsymbol{D_{tca}^{m}}] \end{aligned} \tag{7}$$

where $[;]$ represents concatenating the given vectors in the given dimension, m is the number of heads, and $\boldsymbol{W}_{head}^{d} \in \mathbb{R}^{d_2 \times md_2}$, $\boldsymbol{W}_{head}^{q} \in \mathbb{R}^{d_1 \times md_1}$ are trainable parameters.

In order to further improve the diversity of multi-head TCA, the triple affinity matrices $\boldsymbol{M_{co}}$, $\boldsymbol{M_{in}^{q}}$ and $\boldsymbol{M_{in}^{d}}$ are scaled through dividing by a learnable temperature sequence τ_i with fixed interval before the softmax operator, where $i \in \{1, ..., m\}$. The temperature of i-th head in the multi-head TCA is defined as:

$$\tau_i = \tau_{\circ} \cdot (\lambda \cdot i) \tag{8}$$

where $\tau_{\circ}$ is a learnable parameter and λ is a preset hyper-parameter. The effect of λ is also evaluated in Section V-E. In this way, We make the multi-head have flexible diversity, and the diversity is also learnable.

B. Multimodal TCA Fusion Module

The MMF aims to integrate heterogeneous data from multiple unimodal representations into a joint representation. Given molecular embedding $\boldsymbol{h_m}$, textual embedding $\boldsymbol{h_t}$ and structured embedding $\boldsymbol{h_s}$, MMF aims to enhance the semantic information appearing in different modalities. In the BKG, it is a common phenomenon that the same information may contain in different modalities. For instance, the name of phenolic compounds usually has a suffix "-phine" [8], and piperazine-derived compounds have the suffix "-azine" (the name is just a kind of textual description for this compound). The same information allows the model to comprehend the entity's true properties and meaning, thus MMF should pay special attention to such highlighted information. MMF mainly consists of two steps including 1) pairwise TCA matching; and 2) information exchanging fusion.

1) Pairwise TCA Matching: In this step, we expect to capture the common semantic features between two modalities by using $TCA(\cdot)$ for each pair inputs. During pairwise TCA matching, all the multimodal information of the entity including structured embedding $\boldsymbol{h}_s$, molecular embedding $\boldsymbol{h}_m$ and textual embedding $\boldsymbol{h}_t$ are fed into TCA operator. Each modality forms a pair as shown in Eqn. 9:

$$\begin{aligned} \hat{\boldsymbol{h}}_{x_1}, \hat{\boldsymbol{h}}_{y_1} &= TCA(\boldsymbol{W}_1\boldsymbol{h}_m, \boldsymbol{W}_2\boldsymbol{h}_t) \\ \hat{\boldsymbol{h}}_{x_2}, \hat{\boldsymbol{h}}_{y_2} &= TCA(\boldsymbol{W}_1\boldsymbol{h}_m, \boldsymbol{W}_3\boldsymbol{h}_s) \\ \hat{\boldsymbol{h}}_{x_3}, \hat{\boldsymbol{h}}_{y_3} &= TCA(\boldsymbol{W}_2\boldsymbol{h}_t, \boldsymbol{W}_3\boldsymbol{h}_s) \end{aligned} \tag{9}$$

where d_f is the fusion dimension, $\boldsymbol{W}_1 \in \mathbb{R}^{d_f \times d_m}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d_f \times d_t}$ and $\boldsymbol{W}_3 \in \mathbb{R}^{d_f \times d_s}$ are learnable parameters and aim to project the multi-modal vectors to the fusion dimension. In

this way, we obtain mutually reinforcing information between the two modalities, from both explicit and implicit semantic features.

2) Exchanging Fusion: In this step, we propose to exchange the information between modalities via the value of attention weight to further bridge the modality gap after pairwise TCA matching. The motivation is based on a model pruning method [40] which proposes an assumption of smaller-norm-less-information. It has been proved effective for multimodal fusion [41]. However, rather than utilizing the scaling factor of Batch-Normalization as the measurement of importance, we argue that the smaller attention weight provides less information, and has less influence on the final result. Thus we exchange the unimportant features for each modality measured by co-attention weight to other modal features, which alleviates the difference and heterogeneity between modalities.

Fig. 2(a) shows the process of information exchanging. For the unimportant part of the features (i.e. the position with small attention weight), it will be replaced with the information of another modality. An exchanging factor θ and layer Normalization [42] are introduced to determine which features need to be exchanged between modalities. Assuming that two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ need to be exchanged, the formula is defined as follows:

$$\begin{aligned} index &= where(ln(\boldsymbol{x}) < \theta) \\ \boldsymbol{x}[index] &= \boldsymbol{y}[index] \end{aligned} \tag{10}$$

$$\begin{aligned} index &= where(ln(\boldsymbol{y}) < \theta) \\ \boldsymbol{y}[index] &= \boldsymbol{x}[index] \end{aligned} \tag{11}$$

where $ln(\cdot)$ is the layer normalization, which is applied to each element. For $ln(x)$ or $ln(y)$, the features whose attention weight is less than the threshold will be identified and replaced with another modal information. We denote $EX(\boldsymbol{x}, \boldsymbol{y})$ as the whole exchanging procedure. How to set the value of θ is discussed in Section V-E. Each output pair information of TCA is exchanged by EX operation to further promote multimodal fusion:

$$\tilde{\boldsymbol{h}}_{x_i}, \tilde{\boldsymbol{h}}_{y_i} = EX(\hat{\boldsymbol{h}}_{x_i}, \hat{\boldsymbol{h}}_{y_i}) \qquad i \in \{1, 2, 3\} \tag{12}$$

Finally, we consider a low-rank bilinear fusion method, which is proposed to reduce the ranks of the bilinear weight matrix without losing representation capacity [43], [44]. This solution is to factor a 3D weight tensor into 2D weight matrices, assuming the weight tensor to be low-rank. However, existing low-rank bilinear fusion method cannot exploit complex interaction between modalities with raw inputs [45]. Whereas, in our method, by using the TCA and EX operation, the important part of modal features are fully extracted, and the modality gap is also bridged as much as possible. Thus, we fuse the fine processing data got from Eqn. 7, to a multimodal joint representation $\boldsymbol{h}_f \in \mathbb{R}^{d_f}$ by low-rank bilinear function as follows:

$$\boldsymbol{h}_f = \underset{i \in \{1,2,3\}}{\Omega} (\boldsymbol{P}^T (\sigma(\boldsymbol{U}_{x_i}^T \tilde{\boldsymbol{h}}_{x_i}) \circ \sigma(\boldsymbol{V}_{y_i}^T \tilde{\boldsymbol{h}}_{y_i})) + \boldsymbol{b}) \tag{13}$$

where $\circ$ means Hadamard product (element-wise product), and Ω denotes Hadamard product over a sequence of vector. $\boldsymbol{U}_{x_i} \in \mathbb{R}^{d_f \times d_f}$ and $\boldsymbol{V}_{y_i} \in \mathbb{R}^{d_f \times d_f}$ denote 2D matrices decomposed from a 3D tensor, and $\boldsymbol{b} \in \mathbb{R}^{d_f}$ is a bias vector.

C. Relation-aware Interactive TCA Module

We further propose a RIC module to build the interaction between multimodal entities and relation. Previous studies, such as ConvE [11], try to model the interaction between entity and relations by a convolution framework. In ConvE, the embedding of head entity and relation are concatenated and reshaped to a 2D matrix and fed into the convolutional network. We extend this idea to model the interaction between multimodal entities and relations. Rather than using concatenated interaction of ConvE, our RIC expects to build a deep connection between relations and all modal entities. We establish the mutual influence between entity vector and relation vector by TCA operator, which ensures all elements in entity embedding are possible to contact with all elements in relation embedding. The feature permutations are also proved to be effective by [46]. For each modal entity, the same operations are separately carried out with the relation $\boldsymbol{r}$, to get the multimodal entity-relation interactive representations $\boldsymbol{v}_i$:

$$\begin{cases} \boldsymbol{h}'_\omega, \boldsymbol{r}'_\omega = TCA(\boldsymbol{h}_\omega, \boldsymbol{r}) \\ \boldsymbol{v}_\omega = [\boldsymbol{h}'_\omega, \boldsymbol{r}'_\omega] \end{cases} \tag{14}$$

where $\omega \in \{t, m, s\}$. Finally, we consider stacking the modality joint representation $\boldsymbol{h}_f$ and interactive representation $\boldsymbol{v}_\omega$ to form a multi-channel feature map. The feature map represents multi-view from different modalities, which then are fed into the convolutional neural network. This multi-channel convolution paradigm further fuses multimodal information and increases the interaction. The output of convolution is flattened and projected to latent entity embedding space by a fully connected layer. In summary, the scoring function of triplet (h, r, t) is defined as follows:

$$\begin{aligned} \Phi = & f(\boldsymbol{h}_f \star (\boldsymbol{v}_t \boldsymbol{W}_t) \star (\boldsymbol{v}_m \boldsymbol{W}_m)) \boldsymbol{W}_1 \boldsymbol{h}_s \\ & + f(\boldsymbol{v}_s \star \boldsymbol{v}_0) \boldsymbol{W}_2 \boldsymbol{t}_s \end{aligned} \tag{15}$$

where $\star$ represents to reshape vector to 2D matrix and concatenate a sequence of matrices along a new dimension, $\boldsymbol{v}_0$ denotes $[\boldsymbol{h}; \boldsymbol{r}]$, $f(\cdot)$ means single layer convolution and full connection operator. $\boldsymbol{W}_t \in \mathbb{R}^{d_t \times d_f}$ and $\boldsymbol{W}_m \in \mathbb{R}^{d_w \times d_f}$ are learnable weights to project embedding $\boldsymbol{v}_t$ and embedding $\boldsymbol{v}_m$ to the dimension of fusion embedding $\boldsymbol{h}_f$, respectively. $\boldsymbol{W}_1 \in \mathbb{R}^{L_1 \times d_e}$, $\boldsymbol{W}_2 \in \mathbb{R}^{L_2 \times d_e}$ are also trainable weights for the full-connection network.

D. Model Optimization

We use 1-to-many scoring [11] to optimize our model, which calculates the scores of multiple negative samples simultaneously in a forward propagation and takes less computation time during the evaluation phase. We generate an inverse triplet (t, r^{-1}, h) from each triplet (h, r, t) in the whole dataset. All the original and inverse triplets are trained at

TABLE II
STATISTICS OF THE DATASET INFORMATION

Dataset	#Ent	#Rel	#Train	#Valid	#Test
DRKG-MM	97,238	107	4,699,408	587,424	587,426
OMAHA-MM	74,061	17	406,773	50,846	50,846

the same time, and ranked with whole entities. We minimize a Bernoulli negative log-likelihood loss function as defined below:

$$L = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{q}^{(i)} log(\boldsymbol{p}^{(i)}) + (1 - \boldsymbol{q}^{(i)}) log(1 - \boldsymbol{p}^{(i)}) \quad (16)$$

where n is the number of negative samples, $\boldsymbol{q} \in \mathbb{R}^n$ is the true label, and $\boldsymbol{p} \in \mathbb{R}^n$ is the predicted probabilities generated by applying the sigmoid function to Φ.

V. EXPERIMENTS AND RESULTS

In this section, we present experimental results on two BKG datasets to investigate the following research questions:

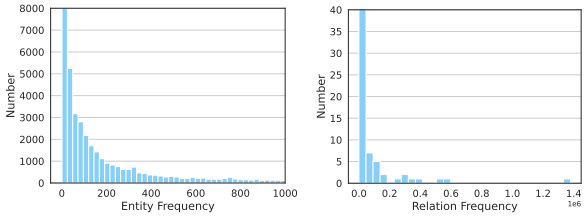
- **RQ1.** How does the proposed CamE model perform compared against the state-of-the-art methods?
- **RQ2.** How do the parameter settings affect the inference result?
- **RQ3.** Do the proposed modules and each modality benefit the performance of CamE?
- **RQ4.** How does CamE perform on different relation types?
- **RQ5.** Does CamE have the ability to learn the semantic relationship between multimodal data?
- **RQ6.** How does the convergence of CamE as compared with the state-of-the-art baselines?
- **RQ7.** How scalable is each module of CamE?

A. Datasets

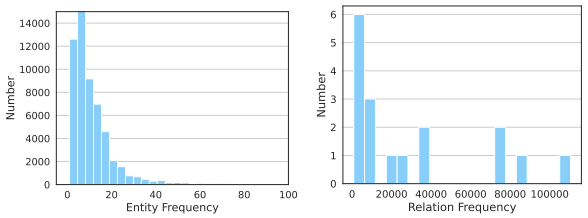
We use two real-world BKG datasets to evaluate our proposed method. The statistical results of these two datasets are shown in Table II. A large portion of KG entities and relations are actually long-tail as shown in Fig. 4. All datasets are split randomly according to 8:1:1 ratio for training, validation and test dataset.

1) DRKG-MM: DRKG-MM is based on a public Drug Repurposing Knowledge Graph (DRKG) [1], which is a comprehensive biological knowledge graph containing different types of entities such as genes, compounds, and diseases, etc. DRKG is constructed from six biological databases and more details about this dataset can be found in [1]. Compared with the original DRKG, we extended the molecular structure information to each drug, and added the textual descriptions for entities from a variety of open data sources including DrugBank[1], Hetionet[2], GNBR[3], String[4], IntAct[5] and DGIdb[6].

[1]https://go.drugbank.com
[2]https://het.io
[3]https://zenodo.org/record/1134693#.YTxrbZMzY_Z
[4]https://string-db.org
[5]https://www.ebi.ac.uk/intact
[6]http://www.dgidb.org



(a) Distribution of DRKG-MM



(b) Distribution of OMAHA-MM

Fig. 4. The histogram of entity and relation frequency in the real-world knowledge graph.

They are also the source of the structural knowledge of DRKG. The full name of Gene is from HUGO Gene Nomenclature Committee, HGNC[7]. Additional statistics are available in the open-source data[8].

2) OMAHA-MM: OMAHA-MM is extracted from a knowledge graph provided by Open Medical and Healthcare Alliance (OMAHA) [9]. Similar to DRKG-MM, OMAHA is also a BKG that has many entities and entities information, like diseases, symptoms, genes and gene mutations. It should be noted that the entities of compounds in OMAHA-MM don't contain the molecular information, which is not included in our experiments of overall comparison and ablation studies on OMAHA-MM. We extracted triplets from OMAHA to construct OMAHA-MM by three rules:

1) Following DRKG [1] and [6], we extract all triplets with the crucial elements in a biological KG, which includes the entity types of genes, diseases, and drugs.
2) We further exclude several entity types with less than 10,000 triplets and relation types with less than 1,000 triplets in OMAHA, since they didn't help to understand the crucial elements, and will add some noise to the overall KG.
3) We delete entities and corresponding triplets with degrees less than five to form our final OMAHA-MM. The reason for this process is to avoid the KG being too sparse to be able to learn graph embeddings. The original OMAHA data is extremely sparse which has 200k entities for 600k triplets. Refined knowledge graph have the same configuration with the most-used knowledge graphs such as FB15K [9], which has very few entities with a degree less than five (including indegree and outdegree).

[7]https://www.genenames.org
[8]https://github.com/gnn4dr/DRKG
[9]http://kg.omaha.org.cn

B. Experimental Settings

We utilize grid search on the valid set to get the best hyperparameters. All the learnable parameters are initialized by Xavier normalization [47]. We use Adam [48] to optimize our model. We conduct all experiments with the filtered setting following [9] for negative sampling. The model is trained on a NVIDIA RTX 3090 GPU. We fix the fusion dim d_f to 200, number of filter to 128, filter size to 9x9. The max training epoch is set to 500. We save the model parameters with best hits@10 on the valid set, and evaluate on the test set to get the final result. We get the reported results on DRKG-MM with the hyperparameters of learning rate: 0.001, embedding dim d_e and d_r: 500, number of negative sampling: 1-to-N (where N is the number of entities), exchanging factor θ: -0.5, number of multi-head m: 2, interval λ: 5. The best combination of parameters for OMAHA-MM is learning rate: 0.0005, embedding dim d_e and d_r: 100, number of negative sampling: 1-to-1000, exchanging factor θ: -2, number of multi-head m: 3, interval λ: 10.

C. Metrics and Baselines

Our model and the state-of-the-art baselines are evaluated under the three most-used metrics for knowledge graph completion tasks which are mean rank(MR), mean reciprocal rank (MRR), and Hits@n (n = 1, 3, 10). All the evaluated baselines can be divided into two groups: 1) unimodal KG completion methods which use only structure knowledge for this task including TransE [9], DistMult [18], ComplEx [23], ConvE [11], CompGCN [37], RotatE, a-RotatE [10], DualE [24] and PairRE [25]; and 2) multimodal KG completion methods which are designed for KG completion on other general multimodal KGs. Details of these unimodal baselines are presented below:

- TransE [9] models the relation r as a translation between the head entity h and tail entity t. It represents all entities and relations to the uniform continuous feature space, and learns low-dimensional and dense vectors for every entity and relation type by minimizing the score function $h + r - t$.
- DistMult [18] is a simplified bilinear model and restricts the relation matrix to be a diagonal matrix.
- ComplEx [23] extend the basic method of DistMult in the complex vector space to capture both symmetric and anti-symmetric relations.
- ConvE [11] is a neural network method, which stacks the head entity and relation, reshapes them to a 2D matrix, then uses 2D convolution and multi-layer perceptron over the matrix to model the interactions between entities and relations.
- CompGCN [37] is a Graph Convolutional framework, which leverages a variety of entity-relation composition operations to jointly embed both nodes and relations.
- RotatE and a-RotatE [10]. RotatE models the relation as rotation from the head entity to the tail entity in complex space. The scoring function is $-||h \circ r - t||^2$. These two models are both from [10] and have the same scoring function. The difference is that a-RotatE is trained with self-adversarial negative sampling yet RotatE is not.
- DualE [24] introduces dual quaternions into knowledge graph embeddings and designs a specific dual-quaternion-based multiplication to model relations as the compositions of a series of translation and rotation operations.
- PairRE [25] uses two vectors for relation representation to encode complex relations and multiple relation patterns simultaneously.

Details of these multimodal baselines are presented below:

- IKRL [12] is the first work to combine image with structured knowledge for KG embedding, which uses an attention-based method to choose the best image instance of each entity. Its scoring function is designed to utilize structured knowledge and visual information following the framework of TransE.
- MTAKGR [14] integrates external text information and defines the energy of a triple as the sum of sub-energy functions of each modality.
- TransAE [13] combines multimodal autoencoder and TransE to get the entity representation.
- MKGformer [26] is a hybrid transformer architecture with multi-level fusion which has coarse-grained prefix-guided interaction and fine-grained correlation-aware fusion modules to integrate visual and text representation.

TransE, DistMult, and ComplEx are implemented based on the code of RotatE. IKRL is implemented through the code of [14]. Since the pre-training model ViT and pre-training corpus used in MKGformer are highly coupled with visual data, MKGformer can not directly apply to our biological KGs. We reproduced its core structure "M-Encoder", including a Prefix-guided Interaction Module and Correlation-aware Fusion Module. "M-Encoder" is incorporated into our framework to replace our multimodal fusion and relation interaction. Other methods are all implemented by their official code.

D. Overall Comparison (RQ1)

In this section, we show the performance evaluation with the baselines. As we can see from Table III, CamE significantly outperforms all the baselines in almost all metrics. For instance, CamE got 10.3% (of MMR) and 16.2% (of Hits@1) improvement over its best competitors on the DRKG-MM dataset, and got 4.8% (of MRR) and 7.0% (of Hits@1) improvement over its best competitors on the OMAHA-MM dataset. The diverse improvement margin on different datasets of CamE is due to the fact that DRKG-MM is a more dense KG than OMAHA-MM. Compared with baselines, the only exception is that a-RotatE achieves the best performance on MR metric on OMAHA-MM dataset. a-RotatE (as well as PairRE) uses self-adversarial negative sampling technique, which automatically balances the updated degree of negative samples, making the model averaged in some situations. This is also the reason why they achieve relatively good performance on Hits@10 but not so well on Hits@1. Note that

TABLE III

LINK PREDICTION RESULTS COMPARED WITH OTHER STATE-OF-THE-ART METHODS ON THE DRKG-MM AND OMAHA-MM. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND RESULTS ARE HIGHLIGHTED WITH AN UNDERLINE. SMALLER *MR* MEANS THE BETTER RESULT, OTHER METRICS ARE THE LARGER THE BETTER.

Models	DRKG-MM					OMAHA-MM				
	MRR↑	MR↓	Hits@1↑	Hits@3↑	Hits@10↑	MRR↑	MR↓	Hits@1↑	Hits@3↑	Hits@10↑
Unimodal approach										
TransE [9]	15.6	822	4.0	21.1	35.3	19.1	867	10.5	22.2	35.4
DistMult [18]	19.2	1864	6.1	28.3	38.8	13.6	3637	7.9	14.7	25.2
ComplEx [23]	30.2	1857	22.4	33.3	43.9	25.0	1122	17.1	27.5	40.5
ConvE [11]	44.1	499	33.3	52.8	64.3	19.1	1979	12.8	20.9	31.7
CompGCN [37]	42.2	542	30.3	50.0	61.5	22.7	1588	13.6	22.4	39.0
RotatE [10]	25.3	699	9.5	35.6	50.3	20.0	858	11.5	23.2	36.5
a-RotatE [10]	39.2	653	19.0	51.6	64.2	22.2	**811**	13.3	25.5	39.7
DualE [24]	45.7	602	34.6	52.1	64.9	19.9	1951	11.5	22.9	36.5
PairRE [25]	36.8	612	17.9	51.1	65.5	24.6	1581	16.2	28.3	40.8
Multimodal approach										
IKRL [12]	12.7	680	6.1	12.5	24.0	16.5	1312	12.4	17.2	29.2
MTAKGR [14]	14.5	491	8.0	15.3	27.4	19.6	868	12.5	21.4	33.2
TransAE [13]	6.8	-	1.3	3.54	10.9	7.2	-	3.2	7.4	15.2
MKGformer [26]	45.4	428	34.6	54.7	64.4	24.8	880	17.2	26.8	38.9
CamE(ours)	**50.4**	**412**	**40.2**	**57.1**	**67.7**	**26.2**	871	**18.4**	**29.3**	**42.1**

TABLE IV

EVALUATION RESULTS OF MRR, H1(HITS@1) AND H10(HITS@10) ON DIFFERENT RELATION TYPES.

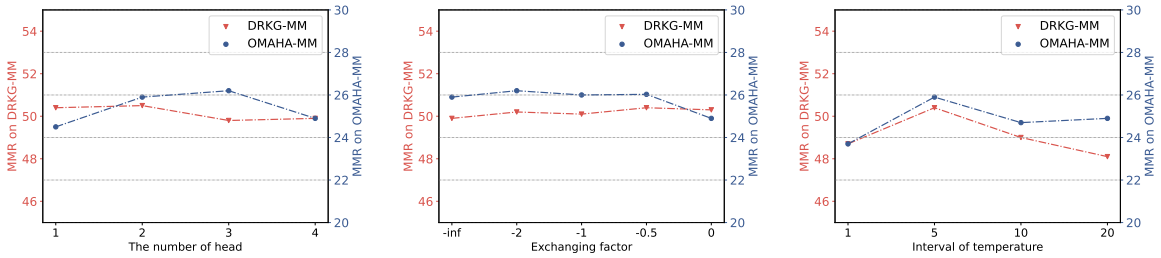
Relations	ConvE			a-RotatE			PairRE			DualE			CamE		
	MRR	H1	H10	MRR	H1	H10	MRR	H1	H10	MRR	H1	H10	MRR	H1	H10
Disease-Gene	9.1	4.0	19.7	9.5	4.5	19.9	8.2	3.4	18.2	5.8	2.5	12.4	**10.3**	**5.1**	**21.1**
Gene-Gene	49.6	37.9	66.0	32.9	4.6	67.3	36.2	9.0	69.0	**56.9**	**47.5**	**71.1**	52.0	40.2	68.1
Compound-Compound	59.0	44.3	87.2	55.6	38.9	85.4	44.1	20.3	83.8	48.7	29.9	82.5	**68.3**	**56.1**	**90.6**
Compound-Side-Effect	13.5	6.9	26.4	14.0	7.1	27.7	**16.3**	**9.1**	**31.3**	10.7	5.1	21.6	15.0	8.0	28.9
Compound-Gene	26.9	19.6	41.2	26.1	18.5	41.3	27.5	20.1	41.8	25.8	18.9	38.9	**29.0**	**21.4**	**43.7**
Compound-Disease	8.9	4.3	17.4	9.9	4.9	19.5	10.6	5.5	20.4	7.5	3.8	14.4	**11.0**	**5.8**	**21.2**

DualE and PairRE show inconsistent performance on DRKG-MM and OMAHA-MM. The reason might be that DRKG-MM is a dense KG, while OMAHA-MM is a sparse one. DualE and PairRE just adapt to correspond to these two types of KGs. However, CamE still can outperform both DualE and a-RotatE. ConvE utilizes the interaction between relations and entities to improve the prediction performance, which is some kind of similar to the RIC component of CamE (see discussion in Section IV-C. However, our method fully considered the multimodal data and modeled the interaction using extra multimodal entities, thus achieving much better performance.

CamE also can outperform all multimodal KG completion baselines. IKRL has an image encoder module and an aggregated image-based representation for ten image instances of each entity. On the DRKG-MM and OMAHA-MM dataset, since each entity has only one instance for each entity, we do not enable the aggregated representation. We input the feature vectors generated by the pre-training models, similar as our method, into all multimodal baselines. MTAKGR [14] is designed to use several crossed sub-scoring functions for each modality to conduct the completion task. TransAE [13] designs a multimodal autoencoder to learn entity embedding, but essentially it still adopts the score function of TransE and is difficult to handle complex interactions. As discussed before, CamE can fully utilize the correlation among different modalities which cannot be done by MTAKGR, IKRL, or TransAE, thus significantly outperforming them on both datasets. Note that existing multimodal baselines are designed for general KGs, such as Freebase, whose multi-modalities mainly include only visual, textual, and numerical information. For example, MKGformer [26] relies heavily on the pre-trained model parameters on visual and textual data, and fine-tunes on other tasks. Therefore they cannot work well on the biological data with molecule information.

E. Parameter Evaluation (RQ2)

In this section, we empirically evaluate the effect of three key factors: the number of heads m, the interval λ and exchanging factor θ for two purposes: 1) to study how to set the values and 2) to demonstrate the necessity to introduce the parameters. The number of heads indicates how many semantic features CamE can extract, the interval means the degree of diversity between multi-heads, and exchange factor indicates degree of information interaction. As we can see from Fig. 5(a), for DRKG-MM and OMAHA-MM datasets, the peak values of MRR for CamE appear when the number of head is equal to 2 and 3, respectively. The MRR of our model stably raises from 24.5% to 26.2% when changing number of head from 1 to 3 on OMAHA-MM. These results demonstrate that multi-head method is helpful for achieving



(a) (b) (c)

Fig. 5. Parameter evaluation with varying values. (a) The effect of the number of head; (b) The effect of exchanging factor; (c) The effect of interval of temperature with the number of head equals 2.

better performance. But too many heads may cause over fitting, and lead to performance degradation. As illustrated in Fig. 5(b), the best performance of DRKG-MM is achieved when $\theta = -0.5$, and the best performance of OMAHA-MM is $\theta = -2.0$. Note that since we adopt a layer normalization for the attention, the value of θ is $[-\infty, +\infty]$. These verify that introducing information exchanging strategy is helpful for our task. As shown in Fig. 5(c), we get the best results when interval $\lambda = 5$ with number of head equals 2. This not only indicates that it is important to increase the diversity between heads, but also indicates that multi-heads and intervals can promote each other.

F. Ablation Study (RQ3)

To further evaluate how each component affects the performance of the proposed model, here we conduct the ablation study on the datasets with designed different variants of CamE.

- **w/o EX:** CamE without the information exchanging process in the fusion module.
- **w/o TCA:** CamE without the triple co-attention operator.
- **w/o MMF:** CamE without multi-modal fusion component. MMF is replaced by simple multiplication.
- **w/o RIC:** CamE without interaction between multimodal entity and relation.
- **w/o M and R:** CamE without MMF as well as RIC. The rest structure simply stacks extra multimodal information.
- **w/o TD:** CamE without the information of textual description. Noticed that, our method can expand or compress with the number of modalities.
- **w/o MS:** CamE without the molecular structure information of compounds in the KG.

As shown in Fig. 6, the prediction performance of both datasets is reduced after removing the important components separately. If removing both MMF and RIC (i.e. w/o M and R), the performance is significantly reduced, which is even worse than ComplEx on OMAHA-MM. It indicates that simply combining multimodal information will introduce noise to the model which cannot be well utilized. TCA has a strong capability to extract important information between modalities, thus we can see significant improvement in two datasets, especially for OMAHA-MM, whose MRR raises from 23.8% to 26.4%. Although the EX module does have the ability to improve the prediction result, It is still limited, since other modules such as TCA have an enough capability to extract important information. We also show that after removing the information of each modality, the performance decreases to a certain extent. At the same time, the molecular structure seems to be more important than the textual description on DRKG-MM. In summary, the ablation study demonstrates that our method is effective to make the best of multimodal information on the BKG data.

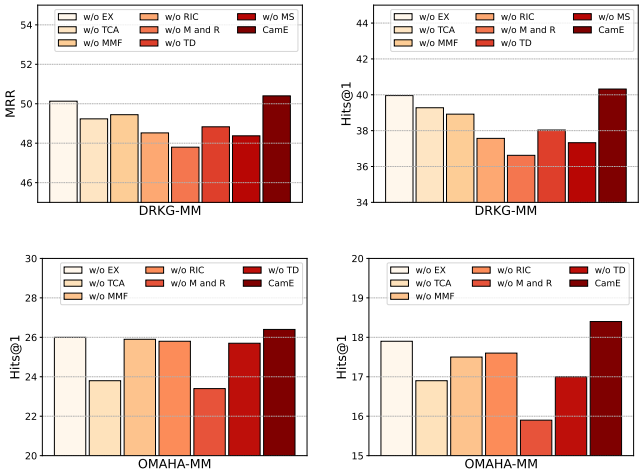


Fig. 6. Results of ablation study experiments.

TABLE V
STATISTICS OF DIFFERENT RELATION TYPES

Relations	Disease-Gene	Gene-Gene
Number of Triples	12,316	234,353
Relations	Compound-Side-Effect	Compound-Gene
Number of Triples	13,964	21,086
Relations	Compound-Compound	Compound-Disease
Number of Triples	138,754	8,451

G. Experiment on Different Relations (RQ4)

In this section, we profile the performance of CamE and baselines on different relation types on the datasets in Table IV. These relation types include Gene-Gene,

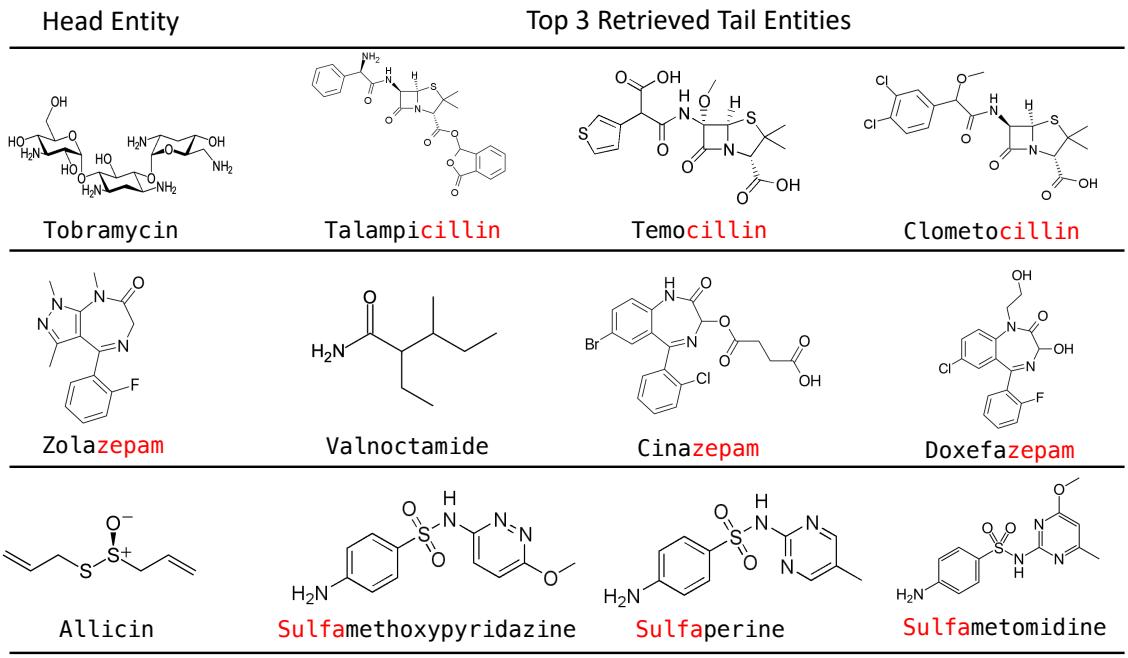


Fig. 7. Case study. The relation between head and tail entities is *Drug-drug Interaction*. The red font of the textual description indicates the prefix or suffix of the corresponding molecular structure.

Compound-Compound, Compound-Gene, Compound-Disease, Compound-SideEffect, and Disease-Gene. The number of triples of different relations types is shown in Table V. We can see the importance of Compounds and Genes. Most of them are very useful tasks for biological and biomedical applications. For example, Compound-Disease relation is relevant to drug repurposing, Compound-Compound relation is relevant to drug-drug interaction and Gene-Gene relation is related to the research of gene expression [6]. The results in Table IV are got by training on the whole KG, and testing on different relationship triples. As we can see, CamE performs better than baselines on most of the relation types. The ability of our model to fully utilize the molecular structure allows the compound-related triples to perform better. This experiment result also demonstrates that, with predicting missing links and supplementing data, the BKG completion potentially has a wide range of applications, such as identifying potential relations of drug-disease, drug-drug, and disease-gene.

H. Case Study (RQ5)

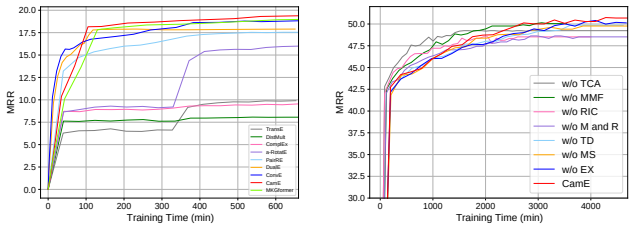
In Fig. 7 we present a case study to show the top 3 tail entities reasoned by CamE. As we can see from Fig. 7, the top 3 tail entities have some interesting semantic similarities, such as suffix "-cillin" and prefix "Sulfa-" in the textual description, whose corresponding molecular structures are the penicillin-type and Sulfonamides-type respectively. The same type of drug usually has a similar medical usage. For example, penicillin-type drugs are effective against many bacterial infections [38], which can be reflected in triplet "(Temocillin, Therapy, infections)". Therefore, this semantic information of entities can help us to reason whether there are some relations between entities. These also indicate that similar information does exist in the biological multimodal data (whether inside or between modalities), and CamE can well capture such common semantic features to improve the performance.

1) Diamond Example: We show a diamond example and the distribution of relations in Fig. 1, which is sampled from DRKG-MM. We restricted the e_0, e_1 and e_2 to be Drug and e_3 to be Gene, and randomly select 5,000 "Same" and 5,000 "Not-Same" diamond structures from all the diamonds that meet this requirement. Then, we randomly search 100 pairs of entities (e_1, e_2) with similarities greater than a certain threshold. We use the value of the vector inner product as a measure of similarity. The feature vectors of the molecule are generated by pre-trained Graph neural network [36]. The values of similarity among the top 100 are judged to be similar, which means "Same" structure. We repeat 100 times for the operation of searching 100 pairs of entities with different random seeds, and report the final average result. It can be seen from Fig. 1, that the molecular structure can help the relation completion task in a certain degree.

I. Evaluation of Efficiency (RQ6, RQ7)

In this section, we evaluate the execution time on DRKG-MM to explore the convergence of CamE compared with other baselines and the scalability of each module of CamE.

1) Evaluation of Convergence (RQ6): In this subsection, we report the training convergence of baselines and our



(a) (b)

Fig. 8. Testing MRR performance v.s. training time. (a) Comparison with other baselines. (b) Comparison with ablation models.

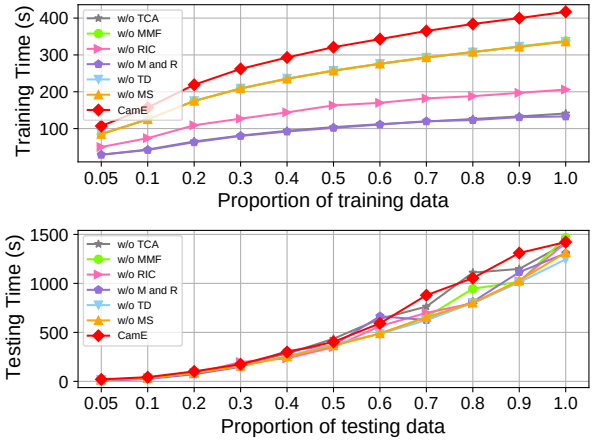


Fig. 9. Training and testing time on a single epoch with various KG sizes.

proposed various modules on DRKG-MM in Fig. 8. Since the inference phase of the KGC task takes a long time, even more than one day for some methods, we randomly selected 10,000 triples from the testing set for evaluation, which resulted in a different performance with Table III. We ignore some baselines, such as MTAKGR, because of a much longer convergence time. In Fig. 8(a), we do not utilize the pretrained structural embedding in CamE here for fair comparison. The experimental results have demonstrated that ConvE, DualE and PairRE require less time to achieve a comparable accuracy compared with the others. Our method converges more slowly in the early stage because of the introduction of multimodal data, yet it can achieve the best performance by fully modeling the multimodal data as the training time increases. Other methods, such as DistMult, converge at an early stage, but don't perform well due to the lack of ability to learn dense large-scale KG.In Fig. 8(b), the experimental results have also exhibited that, in the first few epochs, w/o TCA require less time to achieve a convergent result compared. However, It saw a great decrease in performance without TCA, which shows importance to trade off the performance against the efficiency.

2) Evaluation of Scalablity (RQ7): In this subsection, the scalability study is carried out in terms of training and testing time on DRKG-MM with various KG sizes, as shown in Fig. 9. The Y-axis represents the average execution time of an epoch. The X-axis represents the proportion of the original training and testing triples. The training time has a nearly linear scalability with the increase of the KG sizes, and the testing time also has almost linear scalability but with a larger slope. This is because the test phase requires ranking test triples against all other candidate triples, and a larger entity set will produce more candidate triples. When separately removing the MMF, TD, and MS from CamE, the training time changes with a similar trend. At the same time, CamE without TCA and CamE without MMF and RIC have similar training time costs, and their training cost is the smallest. The underlying reason is that the number of TCA utilized by each module is the same. It means that the most time-consuming training phase is actually the TCA operator. However, the TCA operator brings significant benefit for improving the prediction performance as shown in the ablation study. By contrast, different modules have similar testing time, which demonstrates that the inference time mainly depends on the task requirements.

VI. Conclusion

In this paper, we introduce a novel co-attention-based framework (CamE) for the multimodal BKG completion task. The major motivation is to capture commonly-repeated semantic features among multimodal information including molecule, textual description, and structured knowledge in BKG to predict missing links. To achieve this, we propose a novel triple co-attention operator (TCA) and construct a multimodal knowledge graph embedding framework. Specifically, we fuse the multimodal data with a multimodal TCA fusion module to achieve modality-joint representation. The module includes steps of pairwise TCA matching and exchanging fusion respectively. Furthermore, we propose to model the full interaction between multimodal entities and relation by a relation-aware interactive TCA module, which helps to obtain entity-relation interactive representation. Finally, modality-joint and interactive representation are integrated into a multi-channel feature map to infer missing links. Extensive experimental evaluation on two real-world multimodal BKGs demonstrated the effectiveness of our proposed method.

VII. Acknowledgements

This work was supported in part by the grants from National Natural Science Foundation of China (No.62222213, U22B2059, 62072423), and the USTC Research Funds of the Double First-Class Initiative (No.YD2150002009).

References

[1] V. N. Ioannidis, X. Song, S. Manchanda, M. Li, X. Pan, D. Zheng, X. Ning, X. Zeng, and G. Karypis, "Drkg - drug repurposing knowledge graph for covid-19," https://github.com/gnn4dr/DRKG/, 2020.

[2] B. Walsh, S. K. Mohamed, and V. Nováček, "Biokg: A knowledge graph for relational learning on biological data," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 3173–3180.

[3] J. Hao, C. J.-T. Ju, M. Chen, Y. Sun, C. Zaniolo, and W. Wang, "Bio-joie: Joint representation learning of biological knowledge bases," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2020, pp. 1–10.

[4] F. Z. Smaili, X. Gao, and R. Hoehndorf, "Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations," *Bioinformatics*, vol. 34, no. 13, pp. i52–i60, 2018.

[5] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[6] S. Bonner, I. P. Barrett, C. Ye, R. Swiers, O. Engkvist, A. Bender, C. T. Hoyt, and W. Hamilton, "A review of biomedical datasets relating to drug discovery: A knowledge graph perspective," *arXiv preprint arXiv:2102.10062*, 2021.

[7] Y. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal multi-relational feature aggregation network for medical knowledge representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3956–3965.

[8] H. Fiege, H.-W. Voges, T. Hamamoto, S. Umemura, T. Iwata, H. Miki, Y. Fujita, H.-J. Buysch, D. Garbe, and W. Paulus, *Phenol Derivatives*. John Wiley Sons, Ltd.

[9] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[10] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *International Conference on Learning Representations*, 2018.

[11] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2018.

[12] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3140–3146.

[13] Z. Wang, L. Li, Q. Li, and D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," in *2019 International Joint Conference on Neural Networks*. IEEE, 2019, pp. 1–8.

[14] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018, pp. 225–234.

[15] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[16] D. Xu, T. Xu, S. Wu, J. Zhou, and E. Chen, "Relation-enhanced negative sampling for multimodal knowledge graph completion," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3857–3866.

[17] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Advances in neural information processing systems*, vol. 29, pp. 289–297, 2016.

[18] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *The third International Conference on Learning Representations*, 2015.

[19] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[20] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.

[21] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2015.

[22] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 687–696.

[23] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International conference on machine learning*, 2016, pp. 2071–2080.

[24] Z. Cao, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "Dual quaternion knowledge graph embeddings," in *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6894–6902.

[25] L. Chao, J. He, T. Wang, and W. Chu, "PairRE: Knowledge graph embeddings via paired relation vectors," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4360–4369.

[26] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2022, p. 904–915. [Online]. Available: https://doi.org/10.1145/3477495.3531992

[27] M. Wang, S. Wang, H. Yang, Z. Zhang, X. Chen, and G. Qi, "Is visual context really helpful for knowledge graph? a representation learning perspective," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2735–2743.

[28] L. Chen, Z. Li, T. Xu, H. Wu, Z. Wang, N. J. Yuan, and E. Chen, "Multi-modal siamese network for entity alignment," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 118–126. [Online]. Available: https://doi.org/10.1145/3534678.3539244

[29] L. Chen, Z. Li, Y. Wang, T. Xu, Z. Wang, and E. Chen, "Mmea: entity alignment for multi-modal knowledge graph," in *Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020, Hangzhou, China, August 28–30, 2020, Proceedings, Part I 13*. Springer, 2020, pp. 134–147.

[30] A. García-Durán and M. Niepert, "Kblrn: End-to-end learning of knowledge base representations with latent, relational, and numerical features," in *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, 2018, Monterey, California, USA, August 6-10, 2018*, A. Globerson and R. Silva, Eds., 2018, pp. 372–381.

[31] P. Pezeshkpour, L. Chen, and S. Singh, "Embedding multimodal relational data for knowledge base completion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3208–3218. [Online]. Available: https://aclanthology.org/D18-1359

[32] S. K. Mohamed, A. Nounu, and V. Nováček, "Biological applications of knowledge graph embedding models," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 1679–1693, 2021.

[33] H. El Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii, "Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6903–6915.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[36] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. S. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," in *8th International Conference on Learning Representations*, 2020.

[37] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," *arXiv preprint arXiv:1911.03082*, 2019.

[38] N. Kardos and A. L. Demain, "Penicillin: the medicine with the greatest impact on therapeutic outcomes," *Applied microbiology and biotechnology*, vol. 92, no. 4, pp. 677–687, 2011.

[39] D. Luque Paz, I. Lakbar, and P. Tattevin, "A review of current treatment strategies for infective endocarditis," *Expert Review of Anti-infective Therapy*, vol. 19, no. 3, pp. 297–307, 2021.

[40] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2736–2744.
[41] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
[42] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
[43] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2247–2256.
[44] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.
[45] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5947–5959, 2018.
[46] S. Vashishth, S. Sanyal, V. Nitin, N. Agrawal, and P. Talukdar, "Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions," in *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 3009–3016.
[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *The third International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015.