IMF: Interactive Multimodal Fusion Model for Link Prediction

Xinhang Li
Department of Computer Science and Technology, Tsinghua Univerisity
Beijing, China
xh-li20@mails.tsinghua.edu.cn

Xiangyu Zhao*
School of Data Science, City Univerisity of Hong Kong
Hong Kong
xianzhao@cityu.edu.hk

Jiaxing Xu
School of Computer Science and Engineering, Nanyang Technological Univerisity
Singapore
jiaxing.xu@ntu.edu.sg

Yong Zhang*
Department of Computer Science and Technology, Tsinghua Univerisity
Beijing, China
zhangyong05@tsinghua.edu.cn

Chunxiao Xing
Department of Computer Science and Technology, Tsinghua Univerisity
Beijing, China
xingcx@tsinghua.edu.cn

ABSTRACT

Link prediction aims to identify potential missing triples in knowledge graphs. To get better results, some recent studies have introduced multimodal information to link prediction. However, these methods utilize multimodal information separately and neglect the complicated interaction between different modalities. In this paper, we aim at better modeling the inter-modality information and thus introduce a novel **I**nteractive **M**ultimodal **F**usion (**IMF**) model to integrate knowledge from different modalities. To this end, we propose a two-stage multimodal fusion framework to preserve modality-specific knowledge as well as take advantage of the complementarity between different modalities. Instead of directly projecting different modalities into a unified space, our multimodal fusion module limits the representations of different modalities independent while leverages bilinear pooling for fusion and incorporates contrastive learning as additional constraints. Furthermore, the decision fusion module delivers the learned weighted average over the predictions of all modalities to better incorporate the complementarity of different modalities. Our approach has been demonstrated to be effective through empirical evaluations on several real-world datasets. The implementation code is available online at https://github.com/HestiaSky/IMF-Pytorch.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Information systems** → **Data mining**.

KEYWORDS

link prediction, knowledge graph, multimodal fusion, contrastive learning

*Xiangyu Zhao and Yong Zhang are corresponding authors.

WWW '23, April 30–May 5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583554

ACM Reference Format:
Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: Interactive Multimodal Fusion Model for Link Prediction. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583554

1 INTRODUCTION

Knowledge Graph (KG) stores rich knowledge and is essential for many real-world applications, such as question answering [14, 41, 52], urban computing [46, 49] and recommendation systems [6, 35, 36]. Typically, a KG consists of relational triples, which are represented as *<head entity, relation, tail entity>* [24]. Nevertheless, KGs are inevitably incomplete due to the complexity, diversity and mutability of knowledge. To fix this gap, the problem of link prediction is studied so as to predict potential missing triples [4].

Traditional link prediction models, including translation-based [4, 38] and neural network methods [21, 23], suffered from the structural bias problem among triples. Recently, some studies [26, 28, 39] addressed this problem by enriching the dataset and proposing new models to capture multimodal information for link prediction. However, the performances of such studies were limited as they projected all modalities into a unified space with the same relation to capture the commonality, which might fail to preserve specific information in each modality. As a result, they could not effectively model the complicated interactions between modalities to capture the complementarity.

To address the above issue, we incline to learn the knowledge comprehensively rather than separately, which is similar to how humans think. Take the scenario in Figure 1 as an example, such a model might also get the wrong prediction that *LeBorn James* `playsFor` *Golden States Warriors* based on the similarity with *Stephen Curry* of the common `bornIn` relation to *Akron, Ohio* in graph structure. Meanwhile, it is difficult for visual features to express fine-grained semantics and the only conclusion is that *LeBorn James* is a basketball player. Also, it might also make the outdated prediction of *Cleveland Cavaliers* due to 'played' in the second sentence (more consistent with `playsFor` than 'joined' in the third sentence) in the textual description. Nevertheless, by integrating the knowledge, it is easy to get the correct answer *Log Angeles Lakers* with the interaction between complementary information of structural, visual and

Figure 1: An example of link prediction which may be hard to predict without interaction of multimodal information.

textual highlighted in Figure 1. Since the knowledge learned from different modalities is diverse and complex, it is very challenging to effectively integrate multimodal information.

In this paper, we propose a novel **I**nteractive **M**ultimodal **F**usion Model (IMF) for multimodal link prediction over knowledge graphs. IMF can learn the knowledge separately in each modality and jointly model the complicated interactions between different modalities with a two-stage fusion which is similar to the natural recognition process of human beings introduced above. In the multimodal fusion stage, we employ a bilinear fusion mechanism to fully capture the complicated interactions between the multimodal features with contrastive learning. For the basic link prediction model, we utilize the relation information as the context to rank the triples as predictions in each modality. In the final decision fusion stage, we integrate predictions from different modalities and make use of the complementary information to make the final prediction. The contributions of this paper are summarized as follows:

- We propose a novel two-stage fusion model, IMF, that is effective in integrating complementary information of different modalities for link prediction.
- We design an effective multimodal fusion module to capture bilinear interactions with contrastive learning for jointly modeling the commonality and complementarity.
- We demonstrate the effectiveness and generalization of IMF with extensive experiments on four widely used datasets for multimodal link prediction.

2 METHODOLOGY

Formally, a knowledge graph is defined as $\mathcal{G} = \langle \mathcal{E}, \mathcal{R}, \mathcal{T} \rangle$, where $\mathcal{E}$ and $\mathcal{R}$ indicate sets of entities and relations, respectively. $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ represents relational triples of the KG. In multimodal KGs, each entity in KGs is represented by multiple features from different modalities. Here, we define the set of modalities $\mathcal{K} = \{s, v, t, m\}$ where s, v, t, m denote structural, visual, textual and multimodal modality, respectively. Due to the complexity of real-world knowledge, it is almost impossible to take all the triples into account. Therefore, given a well-formulated KG, the *Link Prediction* task aims at predicting missing links between entities. Specifically, link prediction models expect to learn a score function of relational triples to estimate the likelihood of a triple, which is always formulated as $\psi : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$.

2.1 Overall Architecture

In order to fully exploit the complicated interaction between different modalities, we propose a two-stage fusion model instead of simply considering the multimodal information separately in a unified vector space. As shown in Figure 2, IMF consists of four key components:

1 The Modality-Specific Encoders are used for extracting structural, visual and textual features as the input of multimodal fusion stage.
2 The Multimodal Fusion Module, which is the first fusion stage, effectively models bilinear interactions between different modalities based on *Tucker* decomposition and contrastive learning.
3 The Contextual Relational Model calculates the similarity of contextual entity representations to formulate triple scores as modality-specific predictions for decision fusion stage.
4 The Decision Fusion Module, which is the second fusion stage, takes all the similarity scores from structural, visual, textual and multimodal models into account to make the final prediction.

2.2 Modality-Specific Encoders

In this subsection, we first introduce the pre-trained encoders used for different modalities. These encoders are not fine-tuned during training and we treat them as fixed feature extractors to obtain the modality-specific entity representations. Note that IMF is a general framework and it is straightforward to replace them with other up-to-date encoders or add ones for new modalities into IMF.

2.2.1 Structural Encoder. From the most basic view, the structural information of KG, we employ a Graph Attention Network (GAT)[1] [33] with TransE loss.

Specifically, our GAT encoder takes L1 distance of neighbor aggregated representations as energy function of triples, which is $E(h, r, t) = ||\mathrm{h} + \mathrm{r} - \mathrm{t}||$. In the training process, we minimize the following Hinge loss (1):

$$\mathcal{L}_{GAT} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{T}'} \max\{0, \gamma + E(h, r, t) - E(h', r, t')\} \quad (1)$$

where γ is margin hyper-parameter and $\mathcal{T}'$ denotes set of negative triples derived from $\mathcal{T}$. $\mathcal{T}'$ is created by randomly replacing head

[1] https://github.com/Diego999/pyGAT

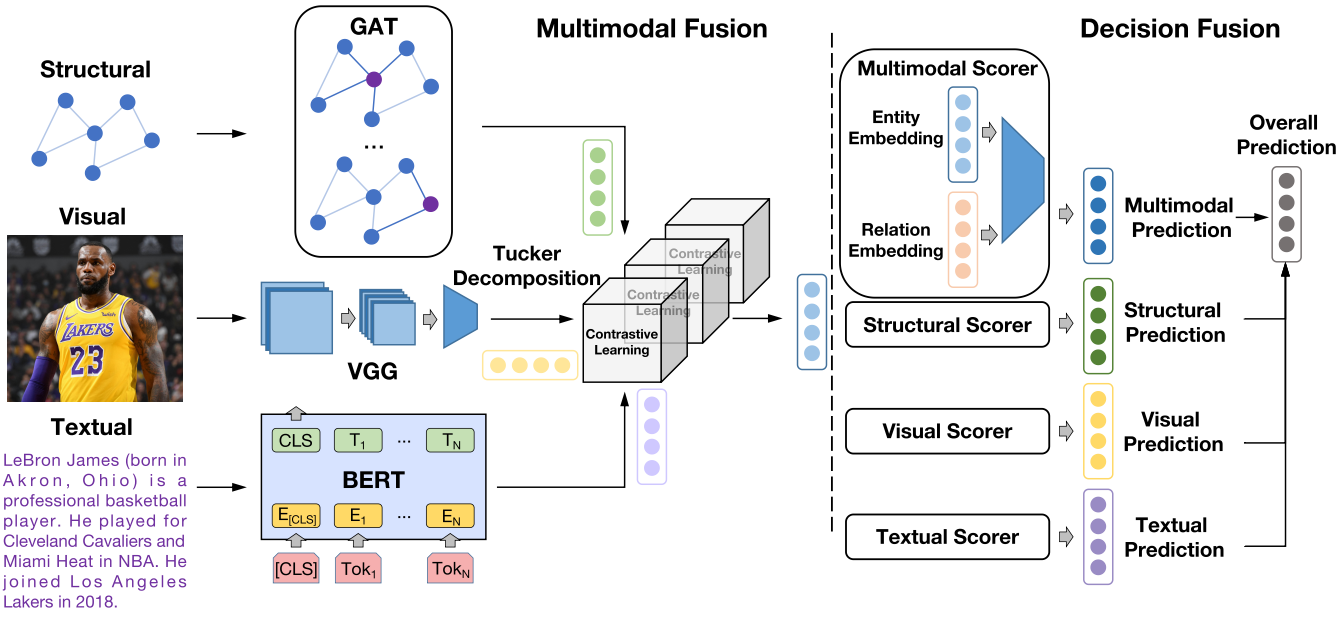

Figure 2: Overall architecture of IMF. The left part represents different modality-specific encoders to extract latent features and the multimodal fusion module to integrate multimodal representations. The right part represents the contextual relational model decoders to get the similarity score and the decision fusion module to make the final prediction on all modalities.

or tail entities of triples in $\mathcal{T}$, which is (2):

$$\mathcal{T}' = \{(h', r, t)|h' \in \mathcal{E}\backslash h\} \cup \{(h, r, t')|t' \in \mathcal{E}\backslash t\} \tag{2}$$

2.2.2 Visual Encoder. Visual features are greatly expressive while providing different views of knowledge from traditional KGs. To effectively extract visual features, we utilize VGG16[2] pre-trained on *ImageNet*[3] to get image embeddings of corresponding entities following [20]. Specifically, we take outputs of the last hidden layer before softmax operation as visual features, which are 4096-dimensional vectors.

2.2.3 Textual Encoder. Entity descriptions contain much richer but more complex knowledge than pure KGs. To fully extract the complex knowledge, we employ BERT [11] as the textual encoder, which is very expressive to get description embeddings of corresponding entities. The textual features are 768-dimensional vectors, i.e., pooled outputs of pre-trained BERT-Base model[4].

2.3 Multimodal Fusion

The multimodal fusion stage aims to effectively get multimodal representations, which fully capture the complex interactions between different modalities. Many existing multimodal fusion methods have achieved promising results in many tasks like VQA (Visual Question Answering). However, most of them aim at finding the commonality to get more precise representations by modality projecting [9, 12] or cross-modal attention [25]. These types of methods will suffer from the loss of unique information in different modalities and can not achieve sufficient interaction between modalities.

[2]https://github.com/machrisaa/tensorflow-vgg
[3]https://image-net.org/
[4]https://github.com/huggingface/transformers

To this end, we propose to employ the bilinear models, which have a strong ability to realize full parameters interaction as the cornerstone to perform the fusion of multimodal information. Specifically, we extend the *Tucker* decomposition, which decomposes the tensor into a core tensor transformed by a matrix along with each mode to 4-mode factors as expressed in Equation (3):

$$\mathcal{P} = (((\mathcal{P}_c \times \mathrm{M}_s) \times \mathrm{M}_v) \times \mathrm{M}_t) \times \mathrm{M}_d \tag{3}$$

where $\mathrm{M}_s \in \mathbb{R}^{d_s \times t_s}$, $\mathrm{M}_v \in \mathbb{R}^{d_v \times t_v}$, $\mathrm{M}_t \in \mathbb{R}^{d_t \times t_t}$, $\mathrm{M}_d \in \mathbb{R}^{\mathcal{D} \times t_d}$ denotes transformation matrix and $\mathcal{P}_c \in \mathbb{R}^{t_s \times t_v \times t_t \times t_d}$ denotes a smaller core tensor.

In such a situation, entity embeddings are first projected into a low-dimensional space and then fused with the core tensor $\mathcal{P}_c$. Following [3], we further reduce the computation complexity by decomposing the core tensor $\mathcal{P}_c$ to merge representations of all modalities into a unified space with element-wise product. The detailed calculation process is expressed as Equation (4):

$$\mathrm{e}_m = \tilde{\mathrm{e}}_s^{\mathrm{T}} \mathrm{M}_d^s * \tilde{\mathrm{e}}_v^{\mathrm{T}} \mathrm{M}_d^v * \tilde{\mathrm{e}}_t^{\mathrm{T}} \mathrm{M}_d^t \tag{4}$$

where $\tilde{\mathrm{e}}_k = \mathrm{ReLU}(\mathrm{e}_k \mathrm{M}_k) \in \mathbb{R}^{t_k}$ denotes latent representations and $\mathrm{e}_k \in \mathbb{R}^{d_k}$ is the original embedding representations and $\mathrm{M}_d^k \in \mathbb{R}^{t_k \times t_d}$ is decomposed transformation matrix for each modality $k \in \{s, v, t\}$.

However, the multimodal bilinear fusion has no bound limitation while the gradient produced by the final prediction result can only implicitly guide parameter learning. To alleviate this problem, we add constraints to limit the correlation between different modality representations of the same entity to be stronger. Therefore, we further leverage contrastive learning [7, 16, 42] between different

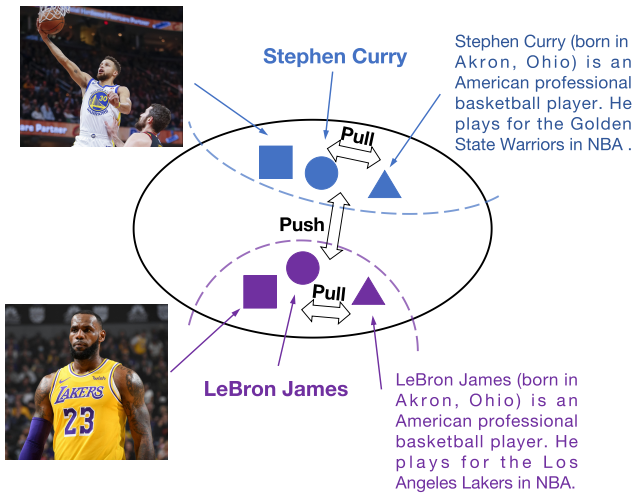

Figure 3: Example of multimodal contrastive learning. The distance between the representations of the same entity in different modalities is minimized, while the distance between the representations of different entities is maximized.

entities and modalities as an additional learning objective for regularization. In the settings of contrastive learning, we take the pairs of representations of the same entity of different modalities as positive samples and the pairs of representations of different entities as negative samples. As shown in Figure 3, we aim at limiting the distance of negative samples to be larger than positive samples to enhance multimodal fusion, which is:

$$d(f(x), f(x^+)) << d(f(x), f(x^-)) \tag{5}$$

where $d(\cdot,\cdot)$ denotes the distance measure and $f(\cdot)$ denotes the embedding function. The superscript $+,-$ represent the positive and negative samples, respectively.

Specifically, we randomly sample N entities from the entity set as a minibatch and define contrastive learning loss upon it. The positive pairs are naturally obtained with the same entities while the negative pairs are constructed by negative sharing [8] of all other entities. We take the latent representations $\tilde{\mathrm{e}}_k = \mathrm{ReLU}(\mathrm{e}_k \mathrm{M}_k) \in \mathbb{R}^{t_k}$ and leverage cosine similarity $d(u,v) = -\mathrm{u}^\top \mathrm{v}/||\mathrm{u}||\mathrm{v}||$ as distance measure. Then we have the following contrastive loss function for each entity i:

$$\mathcal{L}_{CLi} = \frac{1}{3N} \sum_{p,q \in \mathcal{M}} \sum_{j=1}^{N} d(e_i^p, e_i^q) - d(e_i^p, e_j^q) + 2 \tag{6}$$

where $\mathcal{M} = \{(s,v),(s,t),(v,t)\}$ is set of modality pairs.

2.4 Contextual Relational Model

After obtaining representations of each modality and multimodal, we then design a contextual relational model, which takes relations in triples as contextual information for scoring, to get the predictions. Note that this relational model can be easily replaced by any scoring function like TransE.

Due to the variety and complexity of relations in KGs, we argue that improving the degree of parameter interaction [32] is crucial for better modeling the relational triples. The degree of parameter interaction means the calculation ratio of each parameter to all other parameters. For example, dot product could achieve $1/d$ degree while cross product could achieve $(d-1)/d$ degree. Based on this assumption, we propose to use bilinear outer product between entity and relation embeddings to incorporate contextual information into entity representations. Instead of taking relations as input as in previous studies, our contextual relational model utilizes relations to provide context in the transformation matrix of entity embeddings. Then, entity embeddings are projected using the contextual transformation matrix to get *contextual embeddings*, which are used for calculating similarity with all candidate entities. The learning objective is to minimize the binary cross-entropy loss. For each modality $k \in \mathcal{K}$, the computation details are shown as Equation (7) to Equation (9):

$$\hat{\mathrm{e}}_k = \mathrm{e}_k^\top \mathrm{W}_k^r + \mathrm{b} = \mathrm{e}_k^\top \mathrm{W}_k \mathrm{r} + \mathrm{b}_k \tag{7}$$

$$\mathrm{y}_k = \sigma(\mathrm{cosine}(\mathrm{e}_k, \hat{\mathrm{e}}_k)) = \sigma(\frac{\mathrm{e}_k \cdot \hat{\mathrm{e}}_k}{|\mathrm{e}_k||\hat{\mathrm{e}}_k|}) \tag{8}$$

$$\mathcal{L}_k = -\frac{1}{N} \sum_{i=1}^{N} (t_i \cdot \log(y_{i,k}) + (1-t_i) \cdot \log(1-y_{i,k})) \tag{9}$$

where e_k and $\hat{\mathrm{e}}_k$ are original and contextual entity embeddings respectively; $\mathrm{W}_k^r = \mathrm{W}_k \mathrm{r}$ denotes contextual transformation matrix which is obtained by matrix multiplication of weight matrix W_k and relation vectors r while b_k is a bias vector; σ is sigmoid function and $\mathrm{y}_k = [y_{1,k}, y_{2,k}, ..., y_{N,k}]$ is final prediction of modality k.

2.5 Decision Fusion

Existing multimodal approaches mainly focus on projecting different modality representations into a unified space and predicting with commonality between modalities, which will fail to preserve the modality-specific knowledge. We alleviate this problem in the decision fusion stage by joint learning and combining predictions of different modalities to further leverage the complementarity.

Under the multimodal settings, we assign different contextual relational models for each modality and utilize their own results for training in different views. Recall the contrastive learning loss in Equation (6), the overall training objective is to minimize the joint loss shown in Equation (10):

$$\mathcal{L}_{Joint} = \gamma_s \mathcal{L}_s + \gamma_v \mathcal{L}_v + \gamma_t \mathcal{L}_t + \gamma_m \mathcal{L}_m + \mathcal{L}_{CL} \tag{10}$$

where $\mathcal{L}_k$ denotes binary cross entropy loss for modality k as Equation (9) and γ_k is a learned weight parameter.

To better illustrate the whole training process of IMF, we describe it via the pseudo-code of the optimization algorithm. As shown in Algorithm 1, we first obtain the pre-trained encoders of structural, visual and textual and utilize them for entity embeddings (line 3-5). Since the pre-trained models are much larger and more complex than IMF, they are not fine-tuned and their outputs are directly used as inputs of IMF. The multimodal embeddings are obtained by multimodal fusion while contrastive learning is applied to further enhance the fusion stage (line 9-11). During training, each modality delivers its own prediction and loss via the modality-specific scorers (line 12), and then the joint prediction and loss are computed based on all modalities including multimodal ones (line 14).

Algorithm 1 Optimization Algorithm.

Input: Multimodal Knowledge Graph $\mathcal{G}$
Output: Trained Model $\mathcal{M}$
Pre-train structural encoder GAT on $\mathcal{G}$ with the loss in Equation(1)
Obtain pre-trained visual encoder VGG16 and textual encoder BERT-base
Initialize the entity embeddings $\mathrm{E}_s, \mathrm{E}_v, \mathrm{E}_t$ in $\mathcal{M}$ with the outputs of pre-trained encoders
while not converge **do**
  Sample a batch of entities from $\mathcal{G}$
  for Entity e in batch **do**
    Obtain the structural, visual, textual embeddings $\mathrm{e}_s, \mathrm{e}_v, \mathrm{e}_t$ of entity e
    Compute the multimodal fused embeddings e_m of entity e with Equation (4)
    Compute the contrastive learning loss $\mathcal{L}_{CL}$ with Equation (6)
    Compute the loss $\mathcal{L}_s, \mathcal{L}_v, \mathcal{L}_t, \mathcal{L}_m$ with modality-specific scorers via Equation (7) - Equation (9)
    Compute the joint loss $\mathcal{L}_{Joint}$ with the above losses $\mathcal{L}_s, \mathcal{L}_v, \mathcal{L}_t, \mathcal{L}_m, \mathcal{L}_{CL}$ via Equation (10)
    Update model parameters of $\mathcal{M}$ by minimizing $\mathcal{L}_{Joint}$
  end for
end while
return $\mathcal{M}$

For inference, we propose to jointly consider the predictions of each modality as well as multimodal ones. Specifically, the overall predictions are shown in Equation (11):

$$\mathrm{y}_{Joint} = \frac{\gamma_s \mathrm{y}_s + \gamma_v \mathrm{y}_v + \gamma_t \mathrm{y}_t + \gamma_m \mathrm{y}_m}{\gamma_s + \gamma_v + \gamma_t + \gamma_m} \tag{11}$$

where γ_k denotes weight for modality k as same as Equation (10) while the values in y are in [0, 1].

3 EXPERIMENTAL SETUP

3.1 Datasets

In this paper, we use four public datasets to evaluate our model. All the datasets consist of three modalities: structural triples, entity images and entity descriptions. DB15K, FB15K and YAGO15K datasets are obtained from MMKG[5] [20], which is a collection of multimodal knowledge graph. Specifically, we utilize the relational triples as structural features, entity images as visual features and we extract the entity descriptions from Wikidata [34] as textual features. FB15K-237[6] [31] is a subset of FB15K, the visual and textual features in FB15K can be directly reused. Each dataset is split with 70%, 10% and 20% for training, validation and test. The detailed statistics are shown in Table 1.

In the process of evaluation, we consider four metrics of valid entities to measure the model performance following previous studies: (1) mean rank (MR); (2) mean reciprocal rank (MRR); (3) hits ratio (Hits@1 and Hits@10).

[5]https://github.com/nle-ml
[6]https://www.microsoft.com/en-us/download/details.aspx?id=52312

Datasets	#Ent.	#Rel.	#Train	#Valid	#Test
DB15K	14,777	279	69,319	9,903	19,806
FB15K	14,951	1,345	414,549	59,221	118,443
YAGO15K	15,283	32	86,020	12,289	24,577
FB15K-237	14,541	237	272,115	17,535	20,466

Table 1: Statistics of datasets.

3.2 Baselines

To demonstrate the effectiveness of our model, we choose two types of methods for comparison, which are monomodal methods and multimodal methods.

For monomodal models, we take the baselines including:

- **TransE** [4] defines relations as transformations between entities and designs an *energy function* of relational triples as scoring function.
- **ConvE** [10] converts 1D entity and relation embeddings into 2D embeddings and utilizes Convolutional Neural Network (CNN) to model the interactions between entities and relations.
- **ConvKB** [23] employs CNN on the concatenated embeddings of relational triples to compute the triple scores.
- **CapsE** [22] utilizes Capsule Network [27] to capture the complex interactions between entities and relations for prediction.
- **RotatE** [29] introduces rotation operations between entities to represent relations in the complex space to infer symmetry, antisymmetry, inversion and composition relation patterns.
- **QuatE** [43] extends rotation of the knowledge graph embeddings in the complex space into the quaternion space to obtain more degree of freedom.
- **KBAT** [21] leverages Graph Attention Network (GAT) [33] as encoder to aggregate neighbors and employs ConvKB as decoder to compute triple scores.
- **TuckER** [1] applies *Tucker* decomposition to capture the high-level interactions between entity and relation embeddings.
- **HAKE** [45] projects entities into polar coordinate system to model hierachical structures for incorporating semantics.

For multimodal models, we take the baselines including:

- **IKRL** [39] utilizes the TransE energy function as scoring function on each pair of modalities for joint prediction.
- **MKGC** [28] extends IKRL with combination of different modalities to explicitly deliver alignment between modalities.
- **MKBE** [26] employs DistMult [40] as scoring function and designs Generative Adversarial Network (GAN) [13] to predict missing modalities.

For the ablation study, we design three variants of IMF: IMF (w/o MF) utilizes only structural information; IMF (w/o DF) simply takes multimodal representations for training and inference without decision fusion; IMF (w/o CL) removes the contrastive learning loss.

3.3 Implementation Details

The experiments are implemented on the server with an Intel Xeon E5-2640 CPU, a 188GB RAM and four NVIDIA GeForce RTX 2080Ti GPUs using PyTorch 1.6.0. The model parameters are initialized

	DB15K				FB15K				YAGO15K			
	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10
TransE	1128	0.256	13.7	46.9	108	0.495	43.7	77.4	971	0.161	5.1	38.4
ConvE	729	0.312	21.9	50.7	64	0.745	67.0	87.3	714	0.267	16.8	42.6
TuckER	693	0.341	24.3	53.8	40	0.795	74.1	89.2	689	0.281	18.3	45.7
IKRL	984	0.222	11.1	42.6	83	0.594	48.4	76.8	854	0.139	4.8	31.7
MKGC	981	0.208	10.8	41.9	79	0.601	49.2	77.1	939	0.129	4.1	29.7
MKBE	747	0.332	23.5	51.3	48	0.783	70.4	87.8	633	0.273	17.5	42.3
IMF (w/o MF)	687	0.319	21.8	51.2	62	0.752	69.2	86.6	764	0.213	11.4	35.3
IMF (w/o DF)	541	0.443	38.1	57.3	51	0.791	73.9	90.1	527	0.297	21.3	46.3
IMF (w/o CL)	483	0.481	42.3	59.9	29	0.833	78.1	90.8	501	0.289	20.5	45.9
IMF	**478***	**0.485***	**42.7***	**60.4***	**27***	**0.837***	**78.5***	**91.4***	**488***	**0.345***	**27.6***	**49.0***

Table 2: Evaluation results on multimodal DB15K, FB15K and YAGO15K datasets from MMKG. "*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

with Xavier initialization and are optimized using Adam [15] optimizer. The evaluation is conducted under the **Rondom** settings [30], where the correct triples are placed randomly in test set and the negative sampling are correctly employed without test leakage.

For DB15K, FB15K and YAGO15K, we obtain the results by running all the baselines with their released codes. For FB15K-237, we directly obtain the results of TransE, ConvE, ConvKB, CapsE, RotatE, KBAT and TuckER from the re-evaluation work [30] and run the models of QuatE, HAKE, IKRL, MKGC and MKBE with their released codes.

Note that the methods with other enhancing techniques, such as data augmentation [2, 17, 50, 51] or AutoML [18, 37, 44, 47, 48] are orthogonal to our approach for comparison.

4 EXPERIMENTAL RESULTS

4.1 Overall Performance

As shown in Table 2 and Table 3, we can observe that:

- IMF significantly outperforms all the baselines. The performance gain is at most 42% for MRR on DB15K while is also more than 20% for all the evaluation metrics on average.
- State-of-the-art monomodal methods employ a variety of complex models to improve the expressiveness and capture latent interactions. However, the results illustrate that the performance is highly limited by the structural bias of the nature of knowledge graph itself. Although these methods have already achieved promising results, IMF can easily outperform them by a significant margin with a much simpler model structure, which amply demonstrates the effectiveness.
- In comparison with multimodal methods that treat the features of different modalities separately, our IMF jointly learning from different modalities with the two-stage fusion, which is beneficial in modeling the commonality and complementarity simultaneously.

Overall, our proposed IMF can model more comprehensive interactions between different modalities with both commonality and complementarity thanks to the effective fusion of multimodal information and thus achieve significant improvement of link prediction on KGs.

	FB15K-237			
	MR	MRR	H@1	H@10
TransE	357	0.294	-	46.5
ConvE	244	0.325	23.7	50.1
ConvKB	309	0.243	-	42.1
CapsE	403	0.150	-	35.6
RotatE	177	0.338	24.1	53.3
QuatE	176	0.311	22.1	49.5
KBAT	223	0.232	13.6	42.8
TuckER	162	0.353	26.1	53.6
HAKE	-	0.346	25.0	54.2
IKRL	193	0.309	23.2	49.3
MKGC	187	0.297	22.9	49.4
MKBE	158	0.347	25.8	53.2
IMF (w/o MF)	188	0.324	23.4	51.8
IMF (w/o DF)	149	0.356	26.5	55.7
IMF (w/o CL)	138	0.371	27.8	57.1
IMF	**134***	**0.389***	**28.7***	**59.3***

Table 3: Evaluation results on FB15K-237. "*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

4.2 Ablation Study

Table 4 shows the evaluation results of leveraging different modality information on FB15K-237, where S denotes structural information; V denotes visual information of images and T denotes textual information of descriptions. We can see that by introducing visual or textual information, the performance is significantly improved. The significant performance gain brought by multimodal fusion module not only demonstrates the effectiveness of our approach, but also indicates the potential of integrating multimodal information in KG.

To verify the effectiveness of decision fusion, we choose a case of <*LeBron James*, `playsFor` > and visualize the prediction scores of each modality as Figure 4 shows. Due to biases in each modality, the

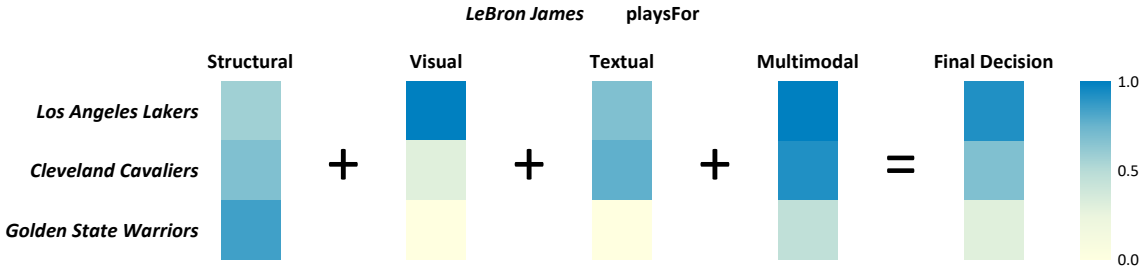

Figure 4: Visualization of prediction scores in decision fusion.

	FB15K-237			
	MR	MRR	H@1	H@10
S	188	0.324	23.4	51.8
S+V	143	0.367	27.4	55.4
S+T	139	0.374	28.1	58.6
S+V+T	**134**	**0.389**	**28.7**	**59.3**

Table 4: Evaluation results with different modality combinations on FB15K-237.

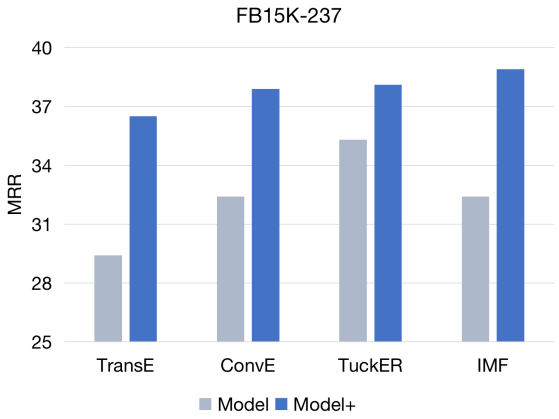

Figure 5: MRR (%) improvement of different basic models on FB15K-237 with IMF.

prediction of monomodal is inevitable error-prone. The results in Table 2 and Table 3 also demonstrate the effectiveness of applying decision fusion to ensemble the specific latent features of each modality.

Besides, the performance comparison between IMF (w/o CL) and IMF in Table 2 and Table 3 illustrates the necessity of contrastive learning for more robust results, especially in the scenario with fewer training samples and relation types.

From the results shown above, we can see that each component in our propose IMF has a significant contribution to the overall performance and it is beneficial to capture the commonality and complementarity between different modalities.

4.3 Generalization

In order to evaluate the generalization of our proposed approach, we simply replace the scoring function (contextual relational model) with existing methods such as TransE, ConvE and TuckER. The results in Figure 5 illustrate that our proposed framework of two-stage fusion is general enough to be applied to any link prediction model for further improvement.

4.4 Parameter Analysis

Figure 6 shows the performance influence of embedding size for IMF. From the picture, we can see that the embedding size plays an important role in the model performance. Meanwhile, it is worthy of note that a larger embedding size not always results in better performance due to the overfitting problem, especially in the datasets with fewer relation types like YAGO15K. Considering the performance and the efficiency, the best choices of embedding size for these three datasets are 256, 256 and 128, respectively.

4.5 Case Study

In order to illustrate the effectiveness of our IMF model in a more intuitive way, we apply *t-SNE* to reduce dimension and visualize the contextual entity representations of basketball players in five different basketball teams. We can see in Figure 7 that the representations of basketball players are messed up with monomodal information due to the biases. However, with the help of interactive multimodal fusion, IMF can effectively capture complicated interactions between different modalities.

5 RELATED WORK

5.1 Knowledge Embedding Methods

Knowledge embedding methods have been widely used in graph representation learning tasks and have achieved great success on knowledge base completion (a.k.a link prediction). Translation-based methods aim at finding the transformation relationships from source to target. TransE [4], the most representative translation-based model, projects entities and relations into a unified vector space and minimizes the *energy function* of triples. Following this route, many translation-based methods have emerged. TransH [38] formulates the translating process on relation-specific hyperplanes. TransR [19] projects entities and relations into separate spaces.

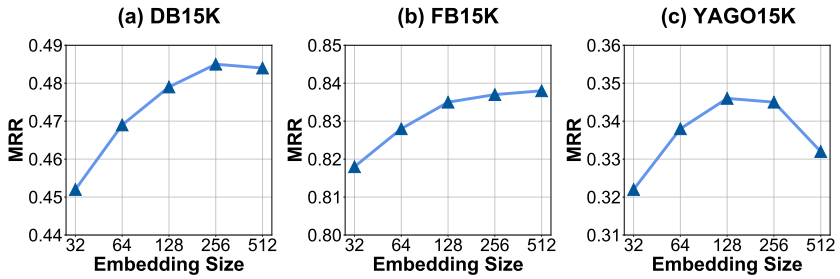

Figure 6: Performance influence of different embedding size.

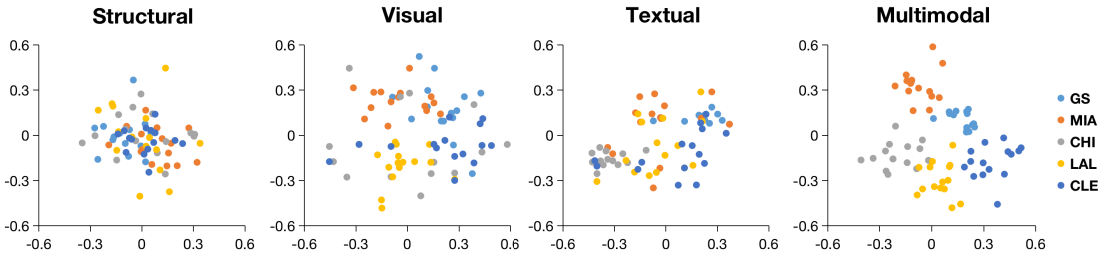

Figure 7: Visualization of low-dimensional representations for basketball players under the context `playsFor`. Each colored node denotes a basketball player and the different colors denote five basketball teams.

Recently, some neural network methods have shown promising results in this task. ConvE [10] and ConvKB [23] utilize Convolutional Neural Network (CNN) to increase parameter interaction between entities and relations. KBAT [21] employ Graph Neural Networks (GNN) as the encoder to aggregate multi-hop neighborhood information.

However, all these methods above utilize only structural information, which is not sufficient for more complicated situations in real world. By incorporating multimodal information in the training process, our approach is able to improve the representations with external knowledge.

5.2 Multimodal Methods

Leveraging multimodal information has yielded extraordinary results in many NLP tasks [3]. DeViSE [12] and Imagined [9] propose to integrate multimodal information with modality projecting which learns a mapping from one modality to another. FiLM [25] extends cross-modal attention mechanism to extract textual-attentive features in visual models. MuRel [5] utilizes pair-wise bilinear interaction between modalities and regions to fully capture the complementarity. IKRL [39] is the first attempt at multimodal knowledge representation learning, which utilizes image data of the entities as extra information based on TransE. MKGC [28] combines textual and visual features extracted by domain-specific models as additional multimodal information compared to IKRL. MKBE [26] creates multimodal knowledge graphs by adding images, descriptions and attributes, and employs DistMult [40] as scoring function.

Although these approaches did incorporate multimodal information to improve performance, they cannot take full advantage of it as they fail to effectively model interactions between modalities.

6 CONCLUSION

In this paper, we study the problem of link prediction over multimodal knowledge graphs. Specifically, we aim at improving the interaction between different modalities. To reach this goal, we propose the IMF with a two-stage framework to enable effective fusion of multimodal information by (i) utilizing bilinear fusion to fully capture the complementarity between different modalities and contrastive learning to enhance the correlation between different modalities of the same entity to be stronger; and (ii) employing an ensembled loss function to jointly consider the predictions of multimodal representations. Experimental results on several benchmarking datasets demonstrate the effectiveness of our proposed model. Besides, we also conduct in-depth exploration to illustrate the generalization of our proposed method and the potential opportunity to apply it in real applications.

However, there are still some limitations of IMF, which are left to future works. For example, IMF requires the integrity of all the modalities and an additional component to predict the missing modalities may be useful to tackle this limitation. Besides, designing appropriate components to support more different kinds of modalities or propose a more lightweight fusion model to replace the bilinear model for better efficiency is also feasible.

ACKNOWLEDGEMENTS

This research was partially supported by the National Key R&D Program of China (No.2020AAA0109603), APRC - CityU New Research Initiatives (No.9610565, Start-up Grant for New Faculty of City University of Hong Kong), SIRG - CityU Strategic Interdisciplinary Research Grant (No.7020046, No.7020074), HKIDS Early Career Research Grant (No.9360163), Huawei Innovation Research Program and Ant Group (CCF-Ant Research Fund).

REFERENCES

[1] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *EMNLP*. 5184–5193.

[2] Robert Bamler, Farnood Salehi, and Stephan Mandt. 2019. Augmenting and Tuning Knowledge Graph Embeddings. In *UAI*, Vol. 115. 508–518.

[3] Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*. 2631–2639.

[4] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NeurIPS*. 2787–2795.

[5] Rémi Cadène, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *CVPR*. 1989–1998.

[6] Jingfan Chen, Wenqi Fan, Guanghui Zhu, Xiangyu Zhao, Chunfeng Yuan, Qing Li, and Yihua Huang. 2022. Knowledge-enhanced Black-box Attacks for Recommendations. In *SIGKDD*. 108–117.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, Vol. 119. 1597–1607.

[8] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On Sampling Strategies for Neural Network-based Collaborative Filtering. In *SIGKDD*. 767–776.

[9] Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined Visual Representations as Multimodal Embeddings. In *AAAI*. 4378–4384.

[10] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2D Knowledge Graph Embeddings. In *AAAI*. 1811–1818.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[12] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NeurIPS*. 2121–2129.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*. 2672–2680.

[14] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge Graph Embedding Based Question Answering. In *WSDM*. 105–113.

[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[16] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*. 9694–9705.

[17] Xinhang Li, Zhaopeng Qiu, Xiangyu Zhao, Zihao Wang, Yong Zhang, Chunxiao Xing, and Xian Wu. 2022. Gromov-Wasserstein Guided Representation Learning for Cross-Domain Recommendation. In *CIKM*. 1199–1208.

[18] Weilin Lin, Xiangyu Zhao, Yejing Wang, Tong Xu, and Xian Wu. 2022. AdaFS: Adaptive Feature Selection in Deep Recommender System. In *SIGKDD*. 3309–3317.

[19] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *AAAI*. 2181–2187.

[20] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *ESWC*. 459–474.

[21] Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. 2019. Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs. In *ACL*. 4710–4723.

[22] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2020. A Capsule Network-based Model for Learning Node Embeddings. In *CIKM*. 3313–3316.

[23] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. 2018. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In *NAACL-HLT*. 327–333.

[24] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* 104, 1 (2016), 11–33.

[25] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*. 3942–3951.

[26] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *EMNLP*. 3208–3218.

[27] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic Routing Between Capsules. In *NeurIPS*. 3856–3866.

[28] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In **SEMEAVL*. 225–234.

[29] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.

[30] Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha P. Talukdar, and Yiming Yang. 2020. A Re-evaluation of Knowledge Graph Completion Methods. In *ACL*. 5516–5522.

[31] Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference.

[32] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, Nilesh Agrawal, and Partha P. Talukdar. 2020. InteractE: Improving Convolution-Based Knowledge Graph Embeddings by Increasing Feature Interactions. In *AAAI*. 3009–3016.

[33] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.

[34] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[35] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. In *CIKM*. 417–426.

[36] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *SIGKDD*. 950–958.

[37] Yejing Wang, Xiangyu Zhao, Tong Xu, and Xian Wu. 2022. Autofield: Automating feature selection in deep recommender systems. In *WWW*. 1977–1986.

[38] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *AAAI*. 1112–1119.

[39] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *IJCAI*. 3140–3146.

[40] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR*.

[41] Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In *ACL*. 1321–1331.

[42] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal Contrastive Training for Visual Representation Learning. In *CVPR*. 6995–7004.

[43] Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion Knowledge Graph Embeddings. In *NeurIPS*. 2731–2741.

[44] Yongqi Zhang, Quanming Yao, Wenyuan Dai, and Lei Chen. 2020. AutoSF: Searching Scoring Functions for Knowledge Graph Embedding. In *ICDE*. 433–444.

[45] Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning Hierarchy-Aware Knowledge Graph Embeddings for Link Prediction. In *AAAI*. 3065–3072.

[46] Xiangyu Zhao, Wenqi Fan, Hui Liu, and Jiliang Tang. 2022. Multi-Type Urban Crime Prediction. In *AAAI*, Vol. 36. 4388–4396.

[47] Xiangyu Zhao, Haochen Liu, Wenqi Fan, Hui Liu, Jiliang Tang, and Chong Wang. 2021. AutoLoss: Automated Loss Function Search in Recommendations. In *SIGKDD*. 3959–3967.

[48] Xiangyu Zhao, Haochen Liu, Wenqi Fan, Hui Liu, Jiliang Tang, Chong Wang, Ming Chen, Xudong Zheng, Xiaobing Liu, and Xiwang Yang. 2021. Autoemb: Automated embedding dimensionality search in streaming recommendations. In *ICDM*. IEEE, 896–905.

[49] Xiangyu Zhao and Jiliang Tang. 2017. Modeling Temporal-Spatial Correlations for Crime Prediction. In *CIKM*. ACM, 497–506.

[50] Zhi Zheng, Zhaopeng Qiu, Hui Xiong, Xian Wu, Tong Xu, Enhong Chen, and Xiangyu Zhao. 2022. DDR: Dialogue Based Doctor Recommendation for Online Medical Service. In *SIGKDD*. 4592–4600.

[51] Zhi Zheng, Zhaopeng Qiu, Tong Xu, Xian Wu, Xiangyu Zhao, Enhong Chen, and Hui Xiong. 2022. CBR: Context Bias aware Recommendation for Debiasing User Modeling and Click Prediction. In *WWW*. 2268–2276.

[52] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*. 4623–4629.