Data Science and Engineering
https://doi.org/10.1007/s41019-023-00208-9

RESEARCH PAPER

Probing the Impacts of Visual Context in Multimodal Entity Alignment

Meng Wang[1] · Yinghui Shi[2] · Han Yang[3] · Ziheng Zhang[4] · Zhenxi Lin[4] · Yefeng Zheng[4]

Received: 7 December 2022 / Revised: 8 March 2023 / Accepted: 19 March 2023
© The Author(s) 2023

Abstract
We study the problem of multimodal embedding-based entity alignment (EA) between different knowledge graphs. Recent works have attempted to incorporate images (visual context) to address EA in a multimodal view. While the benefits of multimodal information have been observed, its negative impacts are non-negligible as injecting images without constraints brings much noise. It also remains unknown under what circumstances or to what extent visual context is truly helpful to the task. In this work, we propose to learn entity representations from graph structures and visual context, and combine feature similarities to find alignments at the output level. On top of this, we explore a mechanism which utilizes classification techniques and entity types to remove potentially un-helpful images (visual noises) during alignment learning and inference. We conduct extensive experiments to examine this mechanism and provide thorough analysis about impacts of the visual modality on EA.

Keywords Entity alignment · Multimodality · Visual context · Knowledge graph

1 Introduction

Entity alignment (EA) is a task aiming to find entities from different knowledge graphs (KGs) that refer to the same real-world object. It plays an important role in KG construction and knowledge fusion as KGs are often independently created and suffer from incompleteness. Most existing models for EA leverage graph structures and/or side information of entities such as name and attributes along with KG embedding techniques to achieve alignment [1, 2]. Several recent methods enrich entity representations by incorporating images, a natural component of entity profiles in many KGs such as DBpedia [3] and Wikidata [4], to address EA in a multimodal view [5–7].

While experimental results have demonstrated that incorporating visual context benefits the EA task [5, 7], it is worth noting that the use of entity images may introduce noises. An error analysis in EVA [7] pointed out that hundreds of source entities were correctly matched to their counterparts before injecting images but were mismatched with images present. Different visual representations of equivalent entities could be potential noises that induce mismatches, and there are various reasons for the visual inconsistency between two equivalent entities. One major reason is that entities naturally have multiple visual representations. As shown in Fig. 1, images (visual context) at left are dissimilar from their counterparts at right, yet they refer to same real-world entities. In addition, the incompleteness of visual data is also a challenging issue for multimodal EA, as reported in [7] that ca. 15–50% entities in the most commonly used benchmark DBP15K [8] are not provided with images.

✉ Yinghui Shi
shiyinghui@seu.edu.cn

Meng Wang
wangmengsd@outlook.com

Han Yang
han.yang6@zeekrlife.com

Ziheng Zhang
zihengzhang@tencent.com

Zhenxi Lin
chalerislin@tencent.com

Yefeng Zheng
yefengzheng@tencent.com

[1] College of Design and Innovation, Tongji University, Shanghai, China

[2] School of Cyber Science and Engineering, Southeast University, Nanjing, China

[3] ZEEKR Intelligent Technology Holding Ltd., Shanghai, China

[4] Tencent Jarvis Lab, Shenzhen, China

Published online: 04 April 2023

Fig. 1 Thumbnail examples of DBpedia entities. **A** and **C** correspond to entities *Oakland_(Californie)* and *Little_Mix* in the French version of DBpedia, respectively. **B** and **D** correspond to entities *Oakland*, *_California* and *Little_Mix* in the English version of DBpedia, respectively

The aforementioned observations raise a doubt: to what extent or under what circumstances is visual context truly helpful to the EA task? Is there a way to filter potential noises and better use entity images? To investigate the above issues, in this work, we propose MMEA-s+v, a simple approach which combines embedding similarities between entities from structural and visual modalities at the output level. In order to fully exploit visual context, we explore a mechanism with classification techniques and entity types to identify potential visual noises and meanwhile generate binary entity mask vectors, which are used with MMEA-s+v to filter images during alignment learning and inference.

In summary, our main contributions are three-fold: (1) To the best of our knowledge, we are the first to investigate the positive and negative aspects of incorporating visual context for EA. We provide insights on actual visual noises that tend to induce misalignment in the multimodal EA. (2) We explore a mechanism with classification techniques and entity types to locate potential visual noises, and conduct extensive experiments to examine this mechanism. (3) We construct the multimodal version of DBP15K which contains a full set of entity images. With the proposed dataset, we hope to facilitate the community in the development of multimodal learning apporaches for KGs.

2 Related Work

Embedding-based approaches for entity alignment (EA) can be generally divided into two categories: that only utilized graph structures and that used additional side information of entities [2]. Among the first category, MTransE [9] adopted TransE [10] to encode language-specific KGs in separate embedding spaces and learned a transformation to align counterpart entities across embeddings. IPTransE [11] and BootEA [12] embedded two KGs in a unified space and bootstrapped the labeled alignments iteratively. Among the second category, GCN-Align [13], JAPE [8] and AttrE [14] used attribute triples in the KGs to refine structural embeddings. MultiKE [15] explored more types of features. It learned entity embeddings from three different views including entity names, relations and attributes. HMAN [16] further exploited literal descriptions of entities to boost performance. UEA [17] utilized useful features from side information in an unsupervised framework to perform EA in the open world.

Recently, a few attempts have been made to incorporate entity images into KGs and build multimodal embeddings for EA. MMEA [5] applied TransE to learn structural embeddings for entities, and utilized image features to learn visual representations. It integrated multiple representations of entities via common space learning. HMEA [6] adopted the hyperbolic graph convolutional networks (HGCNs) to learn structural and visual embeddings of entities separately, then merged them in the hyperbolic space by a weighted Mobius addition. EVA [7] employed GCNs [18] to learn structural representations for entities, and used feed-forward networks to learn embeddings from image, relation and attribute features, respectively. Then it fused embeddings of different modalities by a trainable weighted concatenation. MCLEA [19] considered task-oriented modality and utilized contrastive learning to model the intra-modal and inter-modal interactions for each entity representation. Although existing multimodal entity alignment approaches have shown promising performance, all of them ignored the negative impact of leveraging visual context for EA.

3 Method

We start with the task definition and notations. A KG is denoted as $G = (E, R, T, I)$, where E, R, T, I are the sets of entities, relations, triples and images, respectively. Given a source KG $G_1 = (E_1, R_1, T_1, I_1)$ and a target KG $G_2 = (E_2, R_2, T_2, I_2)$, multimodal entity alignment (MMEA) aims to find every pair (e_1, e_2) where $e_1 \in E_1$, $e_2 \in E_2$ and e_1 and e_2 refer to the same real-world object. To solve this task, we adopt different encoders to encode structural information and visual context of entities, and propose a late fusion mechanism, which combines embedding similarity scores at the output level to find alignment. We name this approach as MMEA-s+v. For comparison, we also present two variants MMEA-avg and MMEA-cat, which adopt different early fusion strategies and learn multimodal joint embeddings to achieve alignment. The main structures of the three variants

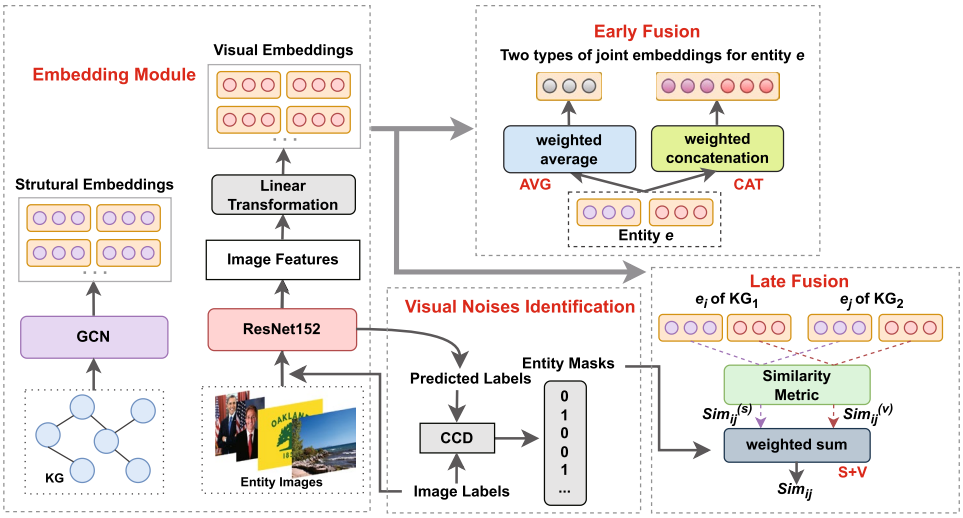

Fig. 2 An illustration of the framework, including the entity embedding module, two multimodal early fusion strategies, the visual noises identification and late fusion mechanisms

are shown in Fig. 2. We further explore a mechanism to filter potential visual noises and generate entity mask vectors, which are used with MMEA-s+v, aiming to exploit visual context for EA. Section 3.1 details about how to identify visual noises. Sections 3.2 and 3.3 focus on entity representation learning, alignment learning and inference.

3.1 Visual Noise Identification

We observe that in most cases visual representations of entities vary largely from a type to another, while they are less different within a type. Based on the findings, we take entity types as classes of images to train a classifier, and use it to identify images whose predicted class is semantically distant from their actual class, i.e., visual noises. To this end we obtain entity types and inter-class conflicts from the ontology of KGs, and design mask vectors to store identification results.

3.1.1 Entity Types

The ontology of KGs usually contains properties and hierarchical classes, and defines subsumption relationships between classes and class disjointness optionally [20]. Types (classes) are often organized in a hierarchical tree structure in the ontology of a KG, and an entity is often associated to a set of types. For example, as shown in Fig. 3, the entity *Barack Obama* in French DBpedia has four types declared (we do not include the root *owl#Thing*): *Agent*, *Person*, *Politician* and *President*, with *Agent* as the most generic type and *President* as the most specific and a leaf node. As we observe that entities of fine-grained types like *President* and *Senator* are more semantically different than visually different, and therefore, we take the type of each entity at most at the fourth level (*Politician* in this example) as the label of its image. We also empirically find the choice of the fourth level, rather than the third or the fifth, yields better classification performance.

3.1.2 Inter-class Conflicts

To measure the semantic discrepancy between the predicted and real classes of an entity image, and inspired by OntoEA [21], we use a class conflict dictionary (CCD) to store the inter-class conflicts. Given two classes a and b, we calculate their conflict degree as $C[a, b]$. For better illustration, we let V denote the hierarchical class tree in which each node refers to a unique class and o the root (typically *owl#Thing*), and define S_x^c as the set of children (subclasses) of node x and S_x^d as the set of all the descendants of x in V, respectively. We assume that all subclasses of the root in V are mutually disjoint, which is in accordance with the design intent for the class hierarchy, and we regard any two descendants of two disjoint classes as disjoint. Let D denote the set of all disjoint class pairs, thus $D = \{(a, b) \mid a, b \in S_o^c, a \neq b \text{ or } \forall c_1, c_2 \in S_o^c, a \in S_{c_1}^d, b \in S_{c_2}^d, c_1 \neq c_2\}$. Given two classes a and b, we firstly determine if $a \equiv b$ or $a \in S_b^d$ or $b \in S_a^d$, and set $C[a, b] = 0$ if they satisfy the condition, which ensures that a class does not conflict with itself or its descendant class, otherwise we look up D and set $C[a, b] = 1$ if $(a, b) \in D$, i.e., two disjoint classes are treated as conflicted. If neither of the above two conditions is met, we follow OntoEA and calculate $C[a, b]$ as:

$$C[a, b] = 1 - \frac{|S(a) \cap S(b)|}{|S(a) \cup S(b)|}, \tag{1}$$

where $S(a)$ and $S(b)$ denote the sets of classes passed by routing from a and b to the root class, respectively, and $|\cdot|$ denotes the set cardinality.

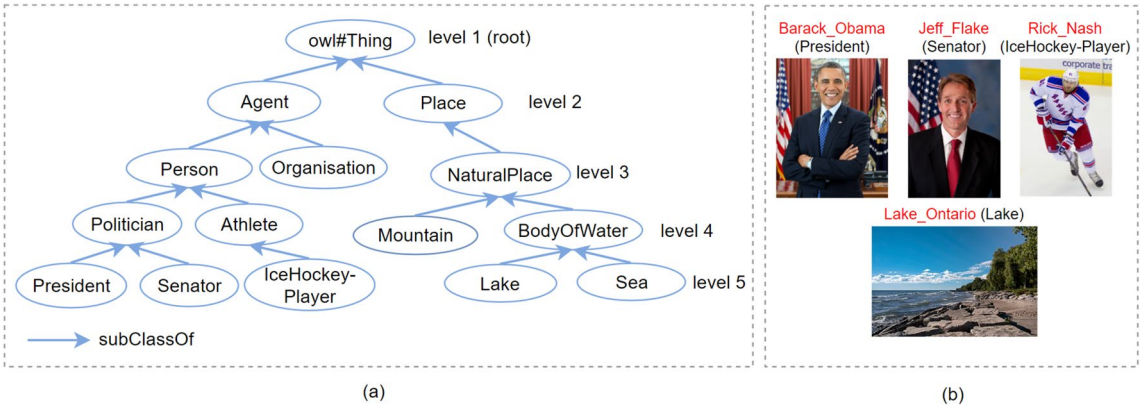

Fig. 3 Subfigure **a** is an example of hierarchical classes. Subfigure **b** presents four entities (denoted by red texts), their finest types in the parentheses and their thumbnails. Because we take entity types at most at the fourth level as image labels, *Barack_Obama* and *Jeff_Flake* share the same label *Politician*, while labels of *Rich_Nash* and *Lake_Ontario* are *Athlete* and *BodyOfWater*, respectively

3.1.3 Entity Mask

We use $\mathbf{M}$ as an entity mask vector and denote $\mathbf{M}_{e_i}$ as the mask value of the ith entity e_i in E. If the image of e_i is determined as potential noise, we set $\mathbf{M}_{e_i} = 0$, which means e_i is masked and its image should be filtered in the training or test phase; otherwise, we set $\mathbf{M}_{e_i} = 1$. Note that the length of $\mathbf{M}$ depends on the sum of the number of source entities and the number of target entities in a dataset. We initialize $\mathbf{M}$ with all zeros and update it iteratively. Specifically, given a conflict degree threshold λ, for $e \in E$, we feed its corresponding image to a classifier to obtain top k predictions (denoted as p_1, ..., p_k), and if the minimum conflict degree between the predictions and the actual class (denoted as g) of e is no greater than λ, i.e., $\min_{1 \le i \le k} \{C[p_i, g]\} \le \lambda$, we reset the mask value of e to be 1.

3.2 Entity Embedding

To better analyze the impacts of visual context on MMEA, we only model two modalities in the entity embeddings, i.e., graph structures and visual context.

3.2.1 Structural Embedding

Graph convolutional networks (GCNs) have proven to be effective in capturing information from graph structures and have been used for embedding-based EA recently [1]. Formally, given as input the adjacency matrix $\mathbf{A}$ of a KG and randomly initialized feature matrix $\mathbf{H}^{(0)}$ of its entities, a multi-layer GCN iteratively updates entity representations from the ith layer to the $(i+1)$th layer with the following propagation rule:

$$\mathbf{H}^{(i+1)} = \phi\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(i)}\mathbf{W}^{(i)}\right), \tag{2}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and $\mathbf{I}$ is an identity matrix, $\hat{\mathbf{D}}$ is the diagonal degree matrix of $\hat{\mathbf{A}}$, $\mathbf{W}^{(i)}$ denotes learnable parameters in the ith layer and ϕ is the activation function ReLU. Following previous works [13, 16], we adopt GCNs to encode the neighborhood information of entities and take the output of the last GCN layer as the structural embeddings.

3.2.2 Visual Embedding

We choose ResNet-152 [22] pre-trained on the ImageNet [23] recognition task as the initial image classifier and fine-tune it with our datasets for EA. The fine-tuning details are given in Sect. 4.1. The fine-tuned model is used to extract image features. We feed each image $i \in I$, through a forward pass and take the output of last layer before logits as its feature vector. Then, we project the feature into a low-dimensional space by a linear transformation to obtain visual embedding $\mathbf{e}_v$:

$$\mathbf{e}_v = \mathbf{W}_v \cdot \mathrm{ResNet}(i) + \mathbf{b}_v, \tag{3}$$

where $\mathbf{W}_v$ is the projection matrix and $\mathbf{b}_v$ is the bias vector.

3.2.3 Multimodal Representation

Given an entity e, its structural embedding $\mathbf{e}_s$ and visual embedding $\mathbf{e}_v$, we present two strategies to combine $\mathbf{e}_s$ and $\mathbf{e}_v$ into a multimodal embedding $\mathbf{e}$ as the joint representation of entity e.

(1) Weighted concatenation. Following the same setting in EVA [7], we calculate $\mathbf{e}$ as:

$$\mathbf{e} = \frac{e^{w_s}}{e^{w_s} + e^{w_v}}\mathbf{e}_s \oplus \frac{e^{w_v}}{e^{w_s} + e^{w_v}}\mathbf{e}_v, \tag{4}$$

where w_s and w_v represent the weight of structural modality and the weight of visual modality, respectively, and both are trainable during learning. The symbol $\oplus$ means concatenation of embeddings. We denote the variant using this kind of fusion as MMEA-cat, which is the same as the setting in EVA where only structural information and visual context are kept.

(2) Weighted averaging. We have $\mathbf{e} = \sum_{i\in\{s,v\}} w_i\mathbf{e}_i$, where w_i is calculated by:

$$w_i = \frac{\cos\left(\mathbf{e}_i, \bar{\mathbf{e}}\right)}{\sum_{j\in\{s,v\}} \cos\left(\mathbf{e}_j, \bar{\mathbf{e}}\right)}, \quad i \in \{s, v\}, \tag{5}$$

and $\bar{\mathbf{e}} = \frac{1}{2}\left(\mathbf{e}_s + \mathbf{e}_v\right)$. By assigning weights to modality-specific entity embeddings, this kind of combination allows the model to emphasize important modalities. We denote the corresponding variant which adopts this fusion strategy as MMEA-avg.

3.3 Alignment Learning and Inference

This section presents details about alignment learning and inference. We integrate G_1 and G_2 as one KG and learn both structural embeddings and visual embeddings of entities in E_1 and E_2 in a unified space. For notations, we let E_s and E_t denote the sets of source entities and the corresponding target entities, respectively, where $E_s \subseteq E_1$, $E_t \subseteq E_2$ and $|E_s| = |E_t|$. We rearrange the elements in both sets in order that the ith entity in E_s corresponds to the ith in E_t. We denote P as the set of all aligned pairs, i.e., $P = \{(e_1, e_2) \mid e_1 \equiv e_2, e_1 \in E_s, e_2 \in E_t\}$, and $\mathbf{M} \in \mathbb{R}^{|E_s|+|E_t|}$ as the entity mask used to filter potential noisy images. The training and test sets are obtained by splitting P with a ratio r.

3.3.1 Alignment Learning

Let $\hat{E}_s$ and $\hat{E}_t$ denote the source entities and target entities, respectively, in the training set. We align each modality separately. For the structural modality s, we compute a similarity matrix $\mathbf{Sim}^{(s)} = \langle \hat{\mathbf{E}}_s^{(s)}, \hat{\mathbf{E}}_t^{(s)} \rangle \in \mathbb{R}^{|\hat{E}_s|\times|\hat{E}_t|}$, where $\hat{\mathbf{E}}_s^{(s)}$ ($\hat{\mathbf{E}}_t^{(s)}$) represents the structural embeddings of entities in $\hat{E}_s$ ($\hat{E}_t$), and each entry $\mathbf{Sim}_{ij}^{(s)}$ corresponds to the cosine similarity between the ith entity in $\hat{E}_s$ and jth in $\hat{E}_t$. To better punish hard negatives and mitigate the hubness problem [24], we choose the HAL loss [25] as the objective function and apply it to obtain the loss of structural modality $\mathcal{L}^{(s)}$ and train the structural embeddings:

$$\mathcal{L}^{(s)} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\alpha}\log\left(1 + \sum_{m\neq i} e^{\alpha\mathbf{Sim}_{mi}^{(s)}}\right) + \frac{1}{\alpha}\log\left(1 + \sum_{n\neq i} e^{\alpha\mathbf{Sim}_{in}^{(s)}}\right) - \log\left(1 + \beta\mathbf{Sim}_{ii}^{(s)}\right)\right), \tag{6}$$

where α, β are temperature scales and N is the batch size. Likewise, we compute $\mathbf{Sim}^{(v)}$ and $\mathcal{L}^{(v)}$ for the visual modality v. Thus the final loss $\mathcal{L}$ of MMEA-s+v is formulated as:

$$\mathcal{L} = \mathcal{L}^{(s)} + \mathcal{L}^{(v)}. \tag{7}$$

To apply the entity mask vectors $\mathbf{M}$ to MMEA-s+v in alignment learning, we obtain a new set of alignment pairs $P' = \{(e_1, e_2) \mid e_1 \equiv e_2, e_1 \in \hat{E}_s, e_2 \in \hat{E}_t, \mathbf{M}_{e_1} = 1, \mathbf{M}_{e_2} = 1\}$ with P and $\mathbf{M}$, determine from P' new sets of source entities and target entities, denoted by $\tilde{E}_s$ and $\tilde{E}_t$, respectively, and compute the visual similarity matrix as $\mathbf{Sim}^{(v)} = \langle \tilde{\mathbf{E}}_s^{(v)}, \tilde{\mathbf{E}}_t^{(v)} \rangle \in \mathbb{R}^{|\tilde{E}_s|\times|\tilde{E}_t|}$.

For MMEA-avg and MMEA-cat, because we use the multimodal embeddings of entities to find alignment, we also optimize the joint representations and calculate a multimodal loss $\mathcal{L}^{(mm)}$ similar to Eq. (6). Then the final loss $\mathcal{L}$ of MMEA-avg/MMEA-cat is:

$$\mathcal{L} = \mathcal{L}^{(s)} + \mathcal{L}^{(v)} + \mathcal{L}^{(mm)}. \tag{8}$$

3.3.2 Inference

Given source entity set $\bar{E}_s$ and target entity set $\bar{E}_t$ used for inference, we compute $\mathbf{Sim}^{(s)} = \langle \bar{\mathbf{E}}_s^{(s)}, \bar{\mathbf{E}}_t^{(s)} \rangle$ and $\mathbf{Sim}^{(v)} = \langle \bar{\mathbf{E}}_s^{(v)}, \bar{\mathbf{E}}_t^{(v)} \rangle$, where $\mathbf{Sim}^{(r)}, \mathbf{Sim}^{(v)} \in \mathbb{R}^{|\bar{E}_s|\times|\bar{E}_t|}$ are cosine similarity matrices for the structural and visual modalities, respectively. For MMEA-s+v, we simply combine them by a weighted addition to obtain the final similarity matrix $\mathbf{Sim} = w \cdot \mathbf{Sim}^{(s)} + (1 - w) \cdot \mathbf{Sim}^{(v)}$, where $w \in (0, 1)$ is a hyper-parameter to balance the two modalities.

To hinder the potential negative impact of visual information when measuring the similarity between two entities, i.e., pulling closer two nonequivalent entities or pushing farther two identical entities, we use the entity mask vector on top of MMEA-s+v. Specifically, we define the similarity score between the ith entity e_i in $\bar{E}_s$ and the jth e_j in $\bar{E}_t$, i.e., the (i, j) entry of $\mathbf{Sim}$ as:

$$\mathbf{Sim}_{ij} = \begin{cases} w \cdot \mathbf{Sim}_{ij}^{(s)} + (1 - w) \cdot \mathbf{Sim}_{ij}^{(v)} & \text{if } \mathbf{M}_{e_i} = 1 \text{ and } \mathbf{M}_{e_j} = 1 \\ \mathbf{Sim}_{ij}^{(s)} & \text{otherwise} \end{cases}. \tag{9}$$

Equation (9) illustrates the principal idea of fusing the two modalities: for source entity e_i and candidate target entity e_j, the in-between similarity is predicted from both aspects of

Table 1 Statistics of image coverage

Source	FR-EN		JA-EN		ZH-EN	
	FR	EN	JA	EN	ZH	EN
Image covered (by DBpedia)	13,858	14,174	12,739	13,741	15,910	14,125
Image covered (by web source)	5794	5816	7011	6035	3421	5441
No. of entities with images	19,652	19,990	19,750	19,776	19,322	19,566
All entities	19,661	19,993	19,814	19,780	19,388	19,572

knowledge only when their images are regarded as potentially useful; otherwise it is solely based on the structural similarity.

For MMEA-avg and MMEA-cat, the cosine similarity matrix **Sim** is simply computed based on multimodal embeddings of entities in $\bar{E}_s$ and $\bar{E}_t$, i.e., $\mathbf{Sim} = \langle \bar{\mathbf{E}}_s, \bar{\mathbf{E}}_t \rangle$. After obtaining **Sim**, we further use cross-domain similarity local scaling (CSLS) [24] to post-process it. Then for $e_i \in \bar{E}_s$, we retrieve the similarity scores of the ith row in **Sim**, rank them in a descending order, and take the top ranked entity as the match.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

We construct the multimodal version of DBP15K and evaluate the methods on this benckmark. DBP15K is a widely used cross-lingual dataset extracted from DBpedia (2016-04) and contains three bilingual subsets: Chinese-English (ZH-EN), Japanese-English (JA-EN), and French-English (FR-EN). Each subset has 15K aligned entity pairs. DBpedia has provided links of thumbnails for many entities; however, it does not cover all of them. Statistics show that ca. 50–85% entities in DBP15K have images [7]. To solve the problem of data incompleteness, for (almost) every entity without an image in DBP15K, We construct a request URL with its surface name, obtain top 10 image URLs ranked by the keyword "selectedIndex" from Bing Images search and download the images. In our experiments, we take the most relevant image (with "selectedIndex = 0" in the URL) as the visual representation of an entity. The statistics of image coverage are presented in Table 1. To retrieve entity types, we query the classes of each entity with *rdf* : *type* via a public SPARQL endpoint.[1] We also obtain the subsumption, disjoint relationships between classes, which are explicitly defined by *rdfs* : *subClassOf* and *owl#disjointWith* properties in the DBpedia ontology, respectively.

4.1.2 Implementation Details

Alignment We employ a three-layer GCN (including the input layer) to encode structural information of entities. The dimensions of the input and hidden layers are both set to 400. For MMEA-cat and MMEA-s+v, we set both the dimension of the output layer of GCN and the dimension of visual embeddings to 200. Whereas for MMEA-avg, we set both the output dimension of GCN and the dimension of visual embeddings to 400, so that its final embedding dimension is the same as MMEA-cat. For all there variants, we adopt AdamW to update parameters, and set the learning rate to 5×10^{-4}. When calculating losses, we set $\alpha = 5$, $\beta = 10$ for $\mathcal{L}^{(s)}$, and $\alpha = 15$, $\beta = 10$ for $\mathcal{L}^{(v)}$ and $\mathcal{L}^{(mm)}$. We train MMEA-avg and MMEA-s+v for 1000 epochs, while we only train MMEA-cat for 600 epochs because we observe an evident overfitting after epoch 600. For MMEA-s+v, we set $w = 0.5$ as the weight of structural similarities between entities during inference.

Following conventions, we use 30% of the aligned pairs for training and the remaining for evaluation, and choose H@1 (Hits@1), H@10 (Hits@10) and mean reciprocal rank (MRR) as the evaluation metrics. For the proposed variants, MMEA-avg, MMEA-cat and MMEA-s+v, we conduct five experiments with different random seeds and present the averaged results along with their standard deviations.

Classification We collect unique entities from all three subsets of DBP15K, filter those either without a type or an image, and use the remaining entities E' as indices to retrieve their images and labels. For each split of DBP15K, we fine-tune a classifier based on the pre-trained ResNet152 [22], and build the test and training data from $E_s \cup E_t$ and $E' \setminus (E_s \cup E_t)$, respectively. We adopt stochastic gradient descent (SGD) to update parameters of classifiers with a learning rate of 0.001 and a momentum of 0.9. We set the batch size to 32 and the number of epochs for training to 25. During test, we obtain top 5 predictions for each entity image, and calculate the mask value of an entity based on its groundtruth class and predictions.

[1] http://dbpedia.org/sparql.

Table 2 Entity alignment results on DBP15K. Bold denotes the best results, and underline denotes the averaged results under five experiments with different random seeds. The gray shading and blue shading represents our proposed variants, including RelEA,VisEA, MMEA-avg, MMEA-cat and MMEA-s+v, and $\frac{Means}{\pm Stds.}$ are shown

Methods	FR-EN			JA-EN			ZH-EN		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
Only use graph structures									
MTransE [9]	0.224	0.556	0.335	0.279	0.575	0.349	0.308	0.614	0.364
IPTransE [11]	0.333	0.685	0.451	0.367	0.693	0.474	0.406	0.735	0.516
SEA [27]	0.400	0.797	0.533	0.385	0.783	0.518	0.424	0.796	0.548
MuGNN [26]	0.495	0.870	0.621	0.501	0.857	0.621	0.494	0.844	0.611
RelEA	0.504 ±.005	0.826 ±.004	0.616 ±.005	0.505 ±.005	0.797 ±.006	0.608 ±.005	0.479 ±.005	0.772 ±.007	0.582 ±.006
AliNet [29]	0.552	0.852	0.657	0.549	0.831	0.645	0.539	0.826	0.628
Leverage graph structures and side information (except visual context)									
JAPE [8]	0.324	0.667	0.430	0.363	0.685	0.476	0.412	0.745	0.490
GCN-Align [13]	0.373	0.745	0.532	0.399	0.745	0.546	0.413	0.744	0.549
HMAN [16]	0.543	0.867	-	0.565	0.866	-	0.537	0.834	-
MultiKE [15]	0.639	0.712	0.665	0.393	0.489	0.426	0.509	0.576	0.532
Leverage visual context and/or graph structures									
VisEA	0.495 ±.003	0.531 ±.003	0.507 ±.003	0.371 ±.001	0.411 ±.001	0.387 ±.001	0.375 ±.001	0.425 ±.002	0.396 ±.001
MMEA-s+v	0.712 ±.005	0.901 ±.003	0.779 ±.004	0.627 ±.005	0.858 ±.005	0.711 ±.004	0.612 ±.006	0.837 ±.006	0.693 ±.005
MMEA-avg	0.716 ±.005	0.901 ±.002	0.782 ±.004	**0.643** ±.003	0.863 ±.006	**0.723** ±.003	**0.624** ±.013	0.830 ±.013	0.699 ±.011
MMEA-cat	**0.725** ±.004	**0.914** ±.005	**0.793** ±.004	0.641 ±.003	**0.869** ±.007	**0.723** ±.003	**0.624** ±.003	**0.845** ±.006	**0.702** ±.005

4.1.3 Comparative Methods

To investigate the effectiveness of unimodal data, we develop two variants named RelEA using only structural information (relational triples), and VisEA using only visual context (entity images) to achieve alignment, respectively. To generally verify the effectiveness of incorporating visual context, we compare MMEA-s+v with RelEA and other public structure-based EA approaches including: MTransE, IPTransE, MuGNN [26], SEA [27] and AliNet. To compare the effects of leveraging visual context with that of using other types of side information, such as entity names and/or attributes, we include other methods, which are JAPE, GCN-Align, HMAN and MultiKE. Note that for fair comparison, the results of HMAN are from its variant that only uses training data in DBP15K as alignment signals.

Recent multimodal approaches for entity alignment, such as MCLEA, EVA, MSNEA [28], etc. use three or more types of information including structural data, numerical/attribute triples, visual knowledge and surface names of entities to improve alignment performance. Because our work focuses on probing the impact of visual context, and, therefore, only utilized graph structure and visual context to conduct experiments, for fair comparison we have not included these methods (Table 2).

4.2 Alignment Results and Analyses

4.2.1 Performance Comparison

We analyze the alignment results from the following perspectives: (1) comparison between our unimodal variants (i.e., RelEA and VisEA) and baselines; (2) comparison between our multimodal variants and the rest of methods to verify the effects of visual context; (3) comparison of different fusion strategies.

(1) our variant RelEA using only structural information is comparable to other structure-based approaches, and even surpasses two models using additional side information, JAPE and GCN-Align. We think that many factors, such as the choice of models to learn structural embeddings for entities, the choice of loss functions, and the settings of hyperparameters for training, have an impact on the model performance. In detail, both MTransE and RelEA only used relational triples to embed entities, however, the former utilized TransE and the latter adopted GCNs. We deem that model capacity limits the quality of entity embeddings that

Table 3 Alignment results and absolute improvements (Improv.) against RelEA under different settings on DBP15K, obtained with random seed as 2021 in the experiments

Settings	FR-EN			JA-EN			ZH-EN		
	No	Hits@1	Improv.	No	Hits@1	Improv.	No	Hits@1	Improv.
RelEA	–	0.506	–	–	0.497	–	–	0.477	–
$\lambda = 0$	26,090	0.663	+ 15.7%	25,532	0.593	+ 9.6%	25,092	0.587	+ 11.0%
$\lambda = 0.4$	27,298	0.681	+ 17.5%	26,805	0.605	+ 10.8%	26,561	0.597	+ 12.0%
$\lambda = 0.67$	27,740	0.687	+ 18.1%	27,234	0.606	+ 10.9%	27,435	0.606	+ 12.9%
$\lambda = 1$	29,479	0.715	+ 20.9%	29,498	0.620	+ 12.3%	28,979	0.618	+ 14.1%
Spec	28,685	0.769	+ 26.3%	28,387	0.701	+ 20.4%	28,079	0.683	+ 20.6%

"No." denotes the number of entities with images

directly determine alignment accuracy. Traning (or inference) settings and design of losses also affect performance. Both the uses of HAL loss (rather than the simple margin-based ranking loss) and CSLS in RelEA greatly improved performance, which may explain why it outperforms GCN-Align that additionaly used attribute triples in learning entity embeddings. VisEA performs worse than RelEA on all there datasets, indicating that leveraging visual context alone is insufficient to achieve satisfactory results.

(2) The proposed three multimodal variants, i.e., MMEA-avg, MMEA-cat and MMEA-s+v, all outperform other baselines using structural and/or side information, indicating that visual context is equally useful as other side information like entity attributes and names. They gain over RelEA 20.8–22.1% (absolute) improvement of Hits@1 on FR-EN, 12.2–13.8% improvement of Hits@1 on JA-EN and 13.3–14.5% improvement of Hits@1 on ZH-EN, respectively, which demonstrates that the incorporation of the visual context can substantially improve the EA system.

(3) As for the modality combination strategies, MMEA-cat achieves slightly better results than MMEA-avg and MMEA-s+v. We think this is due to the attention mechanism during modality fusion, which allows automatic learning of modality weights. Overall there is no prominent difference in the effectiveness among the three strategies.

4.2.2 Impacts of Filtering Entity Images

To maximize the benefits that the incorporation of visual context brings to EA, we selectively combine the feature similarities based on MMEA-s+v during inference with precomputed entity masks to filter potential visual noises. The range of values of the class conflict ratio λ, calculated according to the rules and Eq. (1) presented in Sect. 3.1.2, is a finite set $\{0, 0.4, 0.5, 0.6, 0.67, 1\}$. We choose $\lambda \in \{0, 0.4, 0.67, 1\}$ to calculate mask values of entities with classification results. $\lambda = 0$ corresponds to the strictest setting and $\lambda = 1$ is the no-masking setting, where no entity images are filtered. Bigger λ indicates that more image pairs are involved and the visual context has more influence on alignment prediction during inference. Additionally, we design a special mask based on the alignment result obtained when $\lambda = 1$. Specifically, we reset the mask value of an entity to be 0 if it is correctly matched by only structural similarity but is missed by a joint decision of the two modalities.

We conduct experiments under the above different settings and present the results in Table 3. As shown in Table 3, Hits@1 increases as λ is set larger and the no-masking setting ($\lambda = 1$) outperforms the strictest setting by 2.7–5.2%. We consider that it is mainly attributed to the relatively low quality of visual context. Nevertheless, filtering visual noises is non-trival, as we observe an average performance gain of 6.7% in Hits@1 with the special masks over the no-masking setting. It is also clear that filtering visual noises with the special masks achieves obvious improvements comparing with MMEA-cat and MMEA-avg, which implicitly weaken the impacts of visual noises with their weighted concatenation and weighted average mechanisms to generate better multimodal embeddings. We further analyze the change of errors after visual context is injected under three settings, i.e., the strictest (mask), the no-masking and the special (Spec.). As shown in Fig. 4, on all three datasets the use of special masks greatly reduces errors while retaining as much benefits as no-masking settings bring. The observations suggest such complexity of the problem that the model will not necessarily output better results with visual context considered. They also prove that the visual noise filtering is beneficial to the multimodal entity alignment. The key challenge lies in locating real visual noises.

4.3 Classification Performance and Analysis

The classification accuracies along with numbers of images in training and test sets for each split of DBP15K are reported in Table 4. We collect classification results altogether and merge them for general analysis. For better understanding, we take nodes at the second level of the hierarchical class tree as base classes, and then use them to group fine-grained types, i.e., image labels used in the classification experiments. Note that we additionally treat *Person* and *Organization*, which are subclasses of *Agent*, as two base

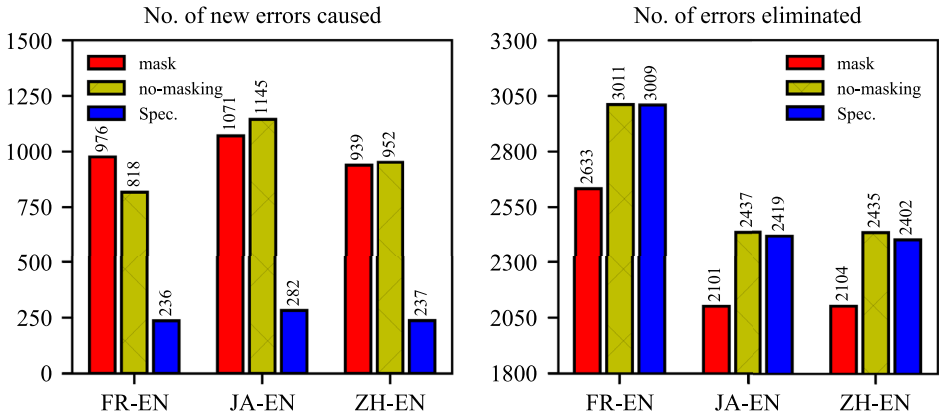

Fig. 4 Number of new errors caused (left) and number of errors eliminated (right) with the use of images on DBP15K. Different colors indicate the results from different settings (cf. Sect. 4.2.2)

Table 4 Entity image classification results on the DBP15K dataset

	Training images	Test images	Classes	Hits@1	Hits@5
FR-EN	54,117	29,479	76	0.513	0.828
JA-EN	54,799	29,498	82	0.509	0.821
ZH-EN	55,146	28,979	82	0.480	0.805

classes, as they are drastically different in both semantics and visual representations. A total of 17 base classes are identified and including their descendants, the total number of classes is 76 for FR-EN and 82 for JA-EN and ZH-EN (cf. Table 4). Among them, Top 4 base classes (together with their descendants) *Person*, *Place*, *Work* and *Organisation* cover 92% of all test entities over three datasets. Figure 5 illustrates the distribution of accuracy and number of test (entity) images with respect to all classes.

We summarize the classification errors into two kinds: (1) the predicted class of an (entity) image and its true class are relatively close and in the same group, i.e., one is the super class of the other or they are siblings or cousins, and (2) the predicted class and the true class are disjoint. We find that without the first kind of errors, the accuracies of four base classes *Person*, *Place*, *Organization*, and *Work* rise from 0.53, 0.65, 0.36 and 0.31 to 0.91, 0.83, 0.51 and 0.52, respectively, which indicates that entities of *Person* or *Place* are more visually distinguishable, while entities of *Organization* and *Work* have less stable visual characteristics. By investigating the mispredictions, we identify several reasons that may explain the poor classification performance on many classes, which also provides insights into the quality of

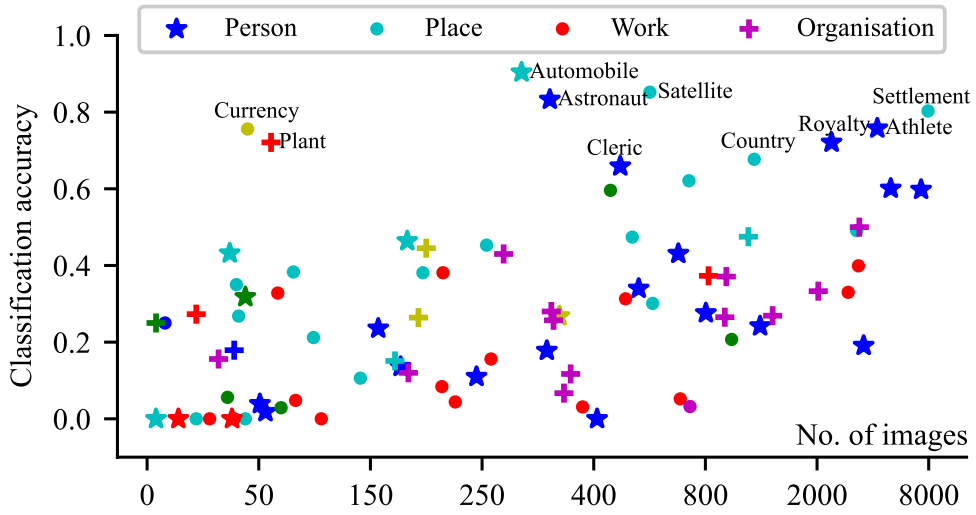

Fig. 5 The distribution of classification accuracy and the number of test images w.r.t. all classes. Each base class is denoted with a unique marker. The same markers scattered at different positions denote fine-grained types that share a common base class, such as blue stars denoting *Royalty*, *Athlete* and *Cleric* sharing the base class *Person*, and cyan dots denoting *Country* and *Settlement* sharing the base class *Place*. Because of limited space, we only present top 4 base classes and explicitly annotate top 10 classes (ranked by classification accuracy) beside their markers

visual data used for MMEA. First, an image provided for an entity can be irrelevant to the entity itself. Second, the visual representations of entities of some classes are unstable. For example, entities of type *Single* or *Album* often have covers as their thumbnails, and these covers often vary widely from one to another depending on the design styles which are also easily misclassified into other classes like *Artist* and *Settlement*. Third, it is difficult to find accurate visual representations for conceptual entities, namely the entities referring to cognitive objects instead of physical objects. A typical type is *MusicGenre*, and its accuracy is as low as 0.03.

4.4 Study of Alignment Errors

In this subsection, we analyze alignment errors and investigate how visual context impacts entity alignment. Generally, The incorporation of entity images can reduce thousands of errors; but on the other hand, it also brings in much noise leading to many new mismatches, as illustrated in Fig. 4. Overall, it improves the alignment performance.

For the positive impact, we find that visual context is particularly helpful when structural information is insufficient to make correct alignment predictions. This finding is supported by the observation that among 3011 newly aligned entity pairs under the no-masking setting on FR-EN, 78% of them have a summed degree below the mean value of the summed degrees of all aligned entity pairs (i.e., long-tailed entities), and a lower degree of an entity indicates less structural information available to learn reliable structural embeddings.

To gain some insights into the negative impact of injecting visual context, we take results of FR-EN as an example and collect new errors occurred under the no-masking setting. These new errors shed light on true visual noises that should be filtered. Among the 818 errors on FR-EN, 139 source entities have mask values of 0 s, meaning that the top 1 predicted class of their image by the classifier is disjoint with their actual (entity) type, and that 139 errors could be reduced if these images are filtered. The remaining 679 errors are mostly about source entities with mask values of 1 s, which we divide into three categories for detailed analysis: (1) The first category contains 436/679 source entities where both the mask values of their aligned counterparts and their predicted matches are 1 s, and 80% of the mismatches are between entities of same or very close types, such as siblings, with *Person* and *Place* as two largest base classes. These mismatches are quite difficult to address because these entity types show relatively stable visual characteristics and the corresponding entity images are less visually distinguishable from those of the same types. (2) The second category includes 154/679 source entities where one of the mask values of their aligned counterparts and their predicted matches is 0, indicating that inappropriate or inconsistent images induced mismatches and these errors could be avoided when the noises are excluded. (3) Errors of the last category, making up about 9% of the total errors, are about source entities mismatched to entities without images, which means these images are not as useful as structural information in multimodal entity alignment.

5 Conclusion

This paper investigated impacts of incorporating visual context (entity images) for multimodal entity alignment. We proposed to learn entity embeddings from structural information and visual context, and integrate feature similarities at the output level. On top of this fusion strategy, we further explored a mechanism which uses image classification techniques and entity types to filter potential noises, and conducted extensive experiments to examine this mechanism. We found that visual context overall is beneficial and that while challenging, filtering noises can further boost performance. We experimentally proved that selectively using visual context brings the most benefits to EA, though the results largely depend on the quality of visual data. Our work also examined the quality of entity images in some multimodal KGs, which has not been inspected by existing studies.

Acknowledgements We thank Chun Zhang for visual data crawling and collection during our study.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by MW, YS, ZZ and ZL. The first draft of the manuscript was written by YS and revised by all authors. All authors contributed to the critical revision of the final manuscript.

Funding This work was supported by the National Key Research and Development Program of China (No. 2022YFF0712400), the National Natural Science Foundation of China (No. 62276063), and the Natural Science Foundation of Jiangsu Province under Grants No. BK20221457.

Availability of Data and Materials The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest All authors disclosed no relevant relationships.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Sun Z, Zhang Q, Hu W, Wang C, Chen M, Akrami F, Li C (2020) A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc VLDB Endow 13:2326–2340
2. Zhang Z, Chen J, Chen X, Liu H, Xiang Y, Liu B, Zheng Y (2020) An industry evaluation of embedding-based entity alignment. In: Proceedings of the 28th international conference on computational linguistics, Barcelona, Spain, pp 179–189
3. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, Van Kleef P, Auer S (2015) DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. Semant Web 6(2):167–195
4. Vrandečić D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. Commun ACM 57(10):78–85
5. Chen L, Li Z, Wang Y, Xu T, Wang Z, Chen E (2020) MMEA: entity alignment for multi-modal knowledge graph. In: International conference on knowledge science, engineering and management, pp 134–147
6. Guo H, Tang J, Zeng W, Zhao X, Liu L (2021) Multi-modal entity alignment in hyperbolic space. Neurocomputing 461:598–607
7. Liu F, Chen M, Roth D, Collier N (2021) Visual pivoting for (unsupervised) entity alignment. In: Proceedings of the AAAI conference on artificial intelligence, pp 4257–4266
8. Sun Z, Hu W, Li C (2017) Cross-lingual entity alignment via joint attribute-preserving embedding. In: International semantic web conference, pp 628–644
9. Chen M, Tian Y, Yang M, Zaniolo C (2017) Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, pp 1511–1517
10. Bordes A, Usunier N, García-Durán A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems, pp 2787–2795
11. Zhu H, Xie R, Liu Z, Sun M (2017) Iterative entity alignment via joint knowledge embeddings. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, pp 4258–4264
12. Sun Z, Hu W, Zhang Q, Qu Y (2018) Bootstrapping entity alignment with knowledge graph embedding. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, pp 4396–4402
13. Wang Z, Lv Q, Lan X, Zhang Y (2018) Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 349–357
14. Trisedya BD, Qi J, Zhang R (2019) Entity alignment between knowledge graphs using attribute embeddings. In: Proceedings of the AAAI conference on artificial intelligence, pp 297–304
15. Zhang Q, Sun Z, Hu W, Chen M, Guo L, Qu Y (2019) Multi-view knowledge graph embedding for entity alignment. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, pp 5429–5435
16. Yang H-W, Zou Y, Shi P, Lu W, Lin JJ, Sun X (2019) Aligning cross-lingual entities with multi-aspect information. In: Proceedings of the 2019 conference on empirical methods in natural language processing, pp. 4430–4440
17. Zhao X, Zeng W, Tang J, Li X, Luo M, Zheng Q (2022) Toward entity alignment in the open world: an unsupervised approach with confidence modeling. Data Scie Eng 7:1–14
18. Kipf T, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: International conference on learning representations
19. Lin Z, Zhang Z, Wang M, Shi Y, Wu X, Zheng Y (2022) Multi-modal contrastive representation learning for entity alignment. In: Proceedings of the 29th international conference on computational linguistics, pp 2572–2584. https://aclanthology.org/2022.coling-1.227
20. Hao J, Chen M, Yu W, Sun Y, Wang W (2019) Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1709–1719
21. Xiang Y, Zhang Z, Chen J, Chen X, Lin Z, Zheng Y (2021) OntoEA: ontology-guided entity alignment via joint knowledge graph embedding. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021, pp 1117–1128
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
23. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 248–255
24. Conneau A, Lample G, Ranzato M, Denoyer L, J'egou H (2018) Word translation without parallel data. In: International conference on learning representation
25. Liu F, Ye R, Wang X, Li S (2020) HAL: improved text-image matching by mitigating visual semantic hubs. In: Proceedings of the AAAI conference on artificial intelligence, pp 11563–11571
26. Cao Y, Liu Z, Li C, Liu Z, Li J-Z, Chua T-S (2019) Multi-channel graph neural network for entity alignment. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1452–1461
27. Pei S, Yu L, Hoehndorf R, Zhang X (2019) Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In: The world wide web conference, pp 3130–3136
28. Chen L, Li Z, Xu T, Wu H, Wang Z, Yuan NJ, Chen E (2022) Multi-modal siamese network for entity alignment. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pp 118–126
29. Sun Z, Wang C, Hu W, Chen M, Dai J, Zhang W, Qu Y (2020) Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In: Proceedings of the AAAI conference on artificial intelligence, pp 222–229