MMpedia: A Large-scale Multi-modal Knowledge Graph

Yinan Wu[1], Xiaowei Wu[1], Junwen Li[1], Yue Zhang[1], Haofen Wang[2], Wen Du[3], Zhidong He[3], Jingping Liu[1], and Tong Ruan[1](✉)

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
[2] College of Design and Innovation, Tongji University, Shanghai, China
[3] DS Information Technology, Shanghai, China
{y21220035}@mail.ecust.edu.cn,{jingpingliu,ruantong}@ecust.edu.cn

Abstract. Knowledge graphs serve as crucial resources for various applications. However, most existing knowledge graphs present symbolic knowledge in the form of natural language, lacking other modal information, e.g., images. Previous multi-modal knowledge graphs have encountered challenges with scaling and image quality. Therefore, this paper proposes a highly-scalable and high-quality multi-modal knowledge graph using a novel pipeline method. Summarily, we first retrieve images from a search engine and build a new Recurrent Gate Multimodal model to filter out the non-visual entities. Then, we utilize entities' textual and type information to remove noisy images of the remaining entities. Through this method, we construct a large-scale multi-modal knowledge graph named MMpedia, containing 2,661,941 entity nodes and 19,489,074 images. As we know, MMpedia has the largest collection of images among existing multi-modal knowledge graphs. Furthermore, we employ human evaluation and downstream tasks to verify the usefulness of images in MMpedia. The experimental result shows that both the state-of-the-art method and multi-modal large language model (e.g., VisualChatGPT) achieve about a 4% improvement on Hit@1 in the entity prediction task by incorporating our collected images. We also find that the multi-modal large language model is hard to ground entities to images. The dataset[4] and source code of this paper are available at https://github.com/Delicate2000/MMpedia.

Keywords: Multi-modal · Knowledge graph · Entity grounding.

1 Introduction

Knowledge Graph (KG) is an important resource and has been applied to various applications such as text classification [6], recommendation [52] and question answering [1]. KGs (e.g., DBpedia [21] and Wikidata [44]) contain a large volume of symbol knowledge. The symbol knowledge is usually represented in the form

[4] https://zenodo.org/record/7816711

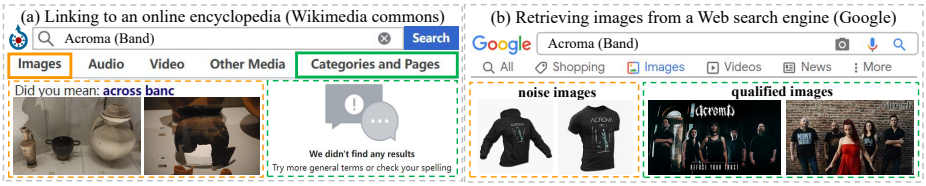


Fig. 1. Traditional MMKG construction methods

of RDF triples $< h, r, t >$, where h and t are the head and tail entity respectively, and r is the relation between h and t.

Problem Statement. However, most existing KGs illustrate the entity in the form of natural language without other modal information such as visual or audio [57]. This results in two problems. **(1)** In the cognitive domain, this situation limits machines' ability to know the physical world. For example, for human beings, we form the concept of *cat* based on the experience of living with a cat. However, for machines, it is challenging to understand what *cat* is as humans do, since symbols or text alone can not bridge the entity *cat* with the experience of cats. Hence, it is necessary to ground entities in KGs to corresponding images, which provides visual experiences for machines. **(2)** In the application domain, grounding entities in KGs to images can enhance machines' performance on various NLP tasks, including relation extraction (RE) [18], named entity recognition [5] and recommendation [38]. In most cases, the integration of visual features has the potential to resolve issues that are difficult to be comprehended from symbolic and textual representations. For example, in RE, given the sentence *JustinBieber (JB) and HaileyBaldwin (HB) arriving at LIV club* from the MNRE dataset [56], it is challenging to determine the social relation between "JB" and "HB" because the text does not provide any semantics of their relation. Fortunately, with the additional information (e.g., age and gender) from images of "JB" and "HB", the relation (Couple) is easier to be inferred.

Hence, in this paper, we aim to help machines understand *what the entity is* by providing high-quality images for KGs.

Limits of Previous MMKGs. Several multi-modal KGs (MMKGs) with entities grounded to images have been proposed. These MMKGs are constructed by collecting images from *online encyclopedias* (OEs) or *web search engines* (WSEs) while either of them still has limitations in providing sufficient and high-quality images for entities.

The first category considers *OEs* (e.g, Wikipedia) as the visual source since they provide images (e.g, Wikimedia commons[5]) as auxiliary information to depict entities. MMKGs built through this category include IMGpedia [13] and Visualsem [2] with data-linking and image-text matching methods. The images in them are relatively reliable and come with textual annotations. However, these MMKGs are hard to scale **due to the limited number of entities in OEs**.

[5] `http://commons.wikimedia.org`

For example, given an entity *Acroma_ (band)* in DBpedia, we can not find its images because it is absent from Wikimedia commons as shown in Figure 1(a).

To improve scalability, the second category considers *WSEs* (e.g., Google) as the visual source and ground the entity to its retrieved Top-K images. MMKGs alone this line include Imagegraph [30], MMKG [26] and Richpedia [46], which are constructed mainly through two methods: (1) generating unambiguous queries with the entity type information from triples [30,26] and (2) employing clustering and ranking information to select images retrieved from WSEs [46]. Nevertheless, these MMKGs suffer from relatively low image quality due to two reasons. Firstly, **both (1) and (2) overlook the removal of non-visual entities, which leads to mismatched images.** Non-visual entities lack a clear visual representation and can not be described in images. For example, given the entity *Idealism*, it is difficult to find an image that accurately reflects it. In contrast, entities with specific visual representations are known as visual entities (e.g., Cat). Secondly, **both (1) and (2) are limited to filtering noisy images retrieved from WSEs.** For example, even with the unambiguous query Acroma (band) generated by (1), some high-ranked images that do not match the corresponding entity still remain as shown in Figure 1(b). Furthermore, there are many noisy images and they may belong to the same class (e.g., shirt in Figure 1(b)), making it challenging to remove them via (2).

Our idea and contribution. In this paper, we construct a large-scale MMKG named MMpedia, which is both highly-scalable and high-quality. This MMKG is built by a novel pipeline method that retrieves images from WSEs (the second category) to ensure the scalability. To ensure image quality, we address the above two issues: (1) non-visual entities in KG and (2) noisy images in WSEs. Specifically, to solve (1), we model the non-visual entity filtering task as a binary classification problem to judge whether the entity is visualizable. In this task, we build a new Recurrent Gate Multi-modal model (RGMM) where the classifier receives the multi-modal features extracted from multiple images and text. To solve (2), we implement a double-filtering process. Firstly, we filter images not depicting the given entity with the text information. To this end, we employ a pre-trained image-text model (e.g, CLIP [32]) to compute the matching score between the textual description and retrieved images. Secondly, we introduce CV models to compare the types of objects in images with the pre-defined entity type. Note that the type information is not leveraged in the query to match with the context of images for two reasons. First, many noisy images with the context containing type-based query are retrieved from WSEs. For example, the context of *shirt.img* in Figure 1(b) is *Acroma Band T-Shirt*, which includes the entity *Acroma* and type *Band*. Second, even with the type-based query, WSEs would return images whose context does not include the type, making the type information useless. For example, for the query *Johnny G (Cyclist)*, the contexts of most retrieved images do not have the type *Cyclist*. In contrast, our approach removes noisy images directly using visual information, rather than relying on the context. Our contributions are summarized as follows:

- We propose a novel pipeline method to construct MMKGs, which consists of the following steps: entity information collection, non-visual entity filtering, entity-image matching and entity type detection.
- We construct a MMKG named MMpedia containing 2,661,941 entities and 19,489,074 images. As we know, MMpedia has the biggest image dataset among existing MMKGs. The accuracy of our images reaches 84.91% after human evaluation.
- Experimental results verify the effectiveness of our proposed method and collected images. In the entity prediction task, both the state-of-the-art method and multi-modal large language model (e.g., VisualChatGPT [50]) achieve about a 4% improvement on Hit@1 by incorporating our collected images.

2 Related Work

We first introduce existing two opposite MMKG construction methods. One is to label images with symbols and another is to grounding entities to images. Then we introduce a closely related task cross-modal retrieval.

Labeling images with symbols can be mainly classified into two categories. The first way is to directly extract visual entities and relations from an image. Chen et al. [8] propose NEIL to automatically extract generic relations from online images. Krishna et al. [20] construct Visual Genome with the images from YFCC100M [40] and MS-COCO [24]. However, they can only obtain limited relation categories. To address this problem, the second way is to extract knowledge from multi-modal information [22,49]. GAIA [22] and Resin [49] first extract event knowledge from multimedia news and then link them to KGs. Although they enrich relation categories, this way requires multi-modal data and a pre-defined schema for different event types, which restricts the scale of MMKGs.

Grounding entities to images mainly includes two groups. One way is to collect images from OEs. Ferrada et al. construct IMGpedia [13] by linking the entity to Wikimedia Commons. Alberts et al. build VisualSem [2] that regards Babelnet [29] as the visual source and addresses the known issue of noisy images [10,4] via image-text matching. Images in OEs are commonly more qualified than those retrieved from WSEs. However, this way is hard to provide images for all entities due to entity differences between OEs and KGs. Another way is to collect images from WSEs. Onoro et al. [30] collect images for FB15K [3] and construct ImageGraph for answering visual-relational queries. Based on DBpedia, Yago [37] and FB15K, Liu et al. [26] retrieve Top-20 images from WSEs and build MMKG. Wang et al. [46] construct Richpedia via employing K-means on images and remaining Top-20 images of each cluster. Although these works provide rich visual resources for KGs, they have limitations on image quality.

Cross-modal retrieval (CMR) is mainly classified into two groups according to the textual query: (1) object-centric and (2) scene-centric [17]. The former compares the objects in the given text with the object in images for CMR. For example, Corbiere et al. [11] retrieve images for fashion-related objects by training two independent uni-modal models with weakly annotated data. Wang et

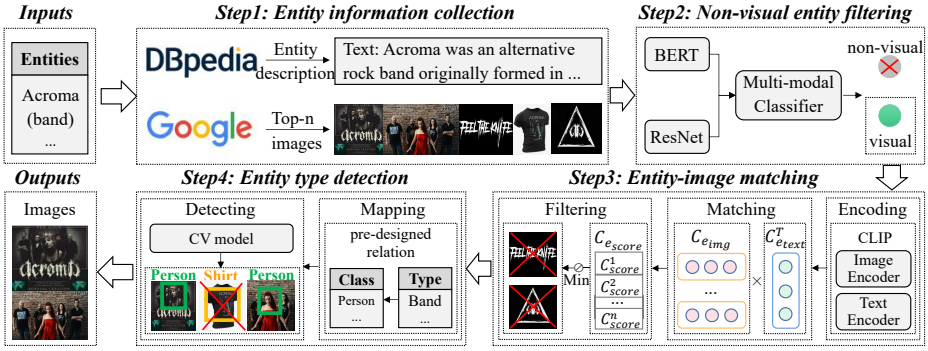


Fig. 2. The frame of our proposed pipeline method. We first collect entity information and remove non-visual entities with a multi-modal classifier. Then, we take entities' textual and type information to remove noisy images.

al. [45] propose SCAN to retrieve images based on the given food objects. The latter considers the relation between multiple objects to retrieve the images. Liu et al. [25] explicitly model objects and relations with GSMN. Mafla et al. [28] propose StacMR, which utilizes GCN to obtain context representation of images and scene text. Cheng et al. [9] present ViSTA to encode image patches and scene text with mid-level fusion. However, both of them focus on abstract concepts (e.g., man) and are limited to grounding a specific entity to images.

3 MMpedia construction

In this paper, we aim to construct a MMKG via providing high-quality images for entities in KGs. For example, given the entity *Acroma_(band)*, we expect to collect images about its members or live performances. To this end, we propose a novel four-step pipeline method, as shown in Figure 2.

3.1 Entity information collection

In this step, we aim to collect entities' textual and visual information for the subsequent non-visual entity filtering and removal of noisy images. To acquire textual information, we retrieve it from KGs as they provide high-quality abstracts for entities. To obtain sufficient candidate images, we build a crawler and retrieve images from a WSE. Specifically, given an entity, we first replace its special characters with space as the query. Then we input the query into a WSE and collect Top-n returned images. For example, the query for the entity "Juan_Pablo_Plada" is "Juan Pablo Plada" because WSEs (e.g., Google) are confused by the character "_".

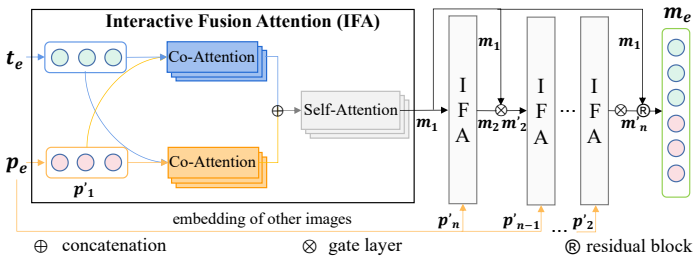


Fig. 3. The multi-modal fusion process of our Recurrent Gate Multi-modal model.

3.2 Non-visual entity filtering

Based on the collected entity information, we expect to remove the non-visual entities which can not be characterized visually. To this end, we regard the task of non-visual entity filtering as a binary classification problem $f(m_e) = 0/1$. Given an entity e, the input is its retrieved Top-n images and textual description and the output is 0 (*non-visual*) or 1 (*visual*). f is denoted as a multi-modal classifier and $\boldsymbol{m_e}$ represents the embedding of multi-modal information.

Since WSEs easily introduce noisy images for entities and existing multi-modal fusion methods have limitations in processing multiple images mixed with noise data, we propose a Recurrent Gate based Multi-modal Model (RGMM) as shown in Figure 3. The core idea of the model is a recurrent structure which employs the Interactive Fusion Attention (IFA) module and gate mechanism to select useful information for multi-modal fusion at each iteration.

Uni-modal feature extraction. Given the Top-n images and text of an entity e, we utilize pre-trained uni-modal models to extract n image features $\boldsymbol{p_e}$ and text feature $\boldsymbol{t_e}$. Specifically, to achieve $\boldsymbol{p_e}$, we first obtain the embedding $\boldsymbol{p'_i}$ of each image P_i by a visual feature extractor (e.g., ResNet [16]). Then we feed $\boldsymbol{p'_i}$ into a fully connected layer and return the transformed image representation $\boldsymbol{p_i}$. Finally, we treat the list $[\boldsymbol{p_1}, ..., \boldsymbol{p_n}]$ as $\boldsymbol{p_e}$. This process can be formalized as

$$\boldsymbol{p_i} = ResNet(P_i) \in \mathbb{R}^{d_p}, \boldsymbol{p'_i} = \boldsymbol{W_p}\boldsymbol{p_i} + \boldsymbol{b_p} \in \mathbb{R}^{d_t}, \boldsymbol{p_e} = [\boldsymbol{p_1}, ..., \boldsymbol{p_n}], \quad (1)$$

where $\boldsymbol{W_p} \in \mathbb{R}^{d_p * d_t}$, $\boldsymbol{b_p} \in \mathbb{R}^{d_t}$ are learnable parameters .

To achieve $\boldsymbol{t_e}$, we first concatenate the text of e with the special tokens $\langle CLS \rangle$, $\langle SEP \rangle$ and feed it into a pre-trained language model (e.g., BERT [12]) to obtain the text representation $\boldsymbol{T'}$. Then, we employ average pooling on $\boldsymbol{T'}$ to obtain $\boldsymbol{t_e}$ [34]. The process is computed as

$$\boldsymbol{T'} = BERT([\langle CLS \rangle, w_1, ..., w_k, \langle SEP \rangle]), \boldsymbol{t_e} = \frac{\sum_{i=0}^{k+1} \boldsymbol{t_i}}{k+2} \in \mathbb{R}^{d_t}, \quad (2)$$

where $[w_1, ..., w_k]$ is a sequence of tokens from e's text and $\boldsymbol{t_i} \in \boldsymbol{T'}$ is the embedding of the corresponding token.

Interactive Fusion Attention (IFA). After achieving two kinds of features $\boldsymbol{p_e}$ and $\boldsymbol{t_e}$, we obtain the initial multi-modal representation $\boldsymbol{m_1}$ with $\boldsymbol{p_1} \in \boldsymbol{p_e}$

and $\boldsymbol{t_e}$. To this end, we build a IFA module to merge multi-modal information. Specifically, we first employ two independent co-attention [27] layers for $\boldsymbol{p_1}$ and $\boldsymbol{t_e}$. One refines $\boldsymbol{p_1}$ with the textual information in $\boldsymbol{t_e}$ and another refines $\boldsymbol{t_e}$ with the visual information in $\boldsymbol{p_1}$. The process is defined as

$$\boldsymbol{p'_1} = MHAtt(\boldsymbol{Q} = \boldsymbol{W_{Q_p}}\boldsymbol{p_1}, \boldsymbol{K} = \boldsymbol{W_{K_p}}\boldsymbol{t_e}, \boldsymbol{V} = \boldsymbol{W_{V_p}}\boldsymbol{t_e})_h \in \mathbb{R}^{d_t}, \tag{3}$$

$$\boldsymbol{t'}_{\boldsymbol{e}} = MHAtt(\boldsymbol{Q} = \boldsymbol{W_{Q_t}}\boldsymbol{t_e}, \boldsymbol{K} = \boldsymbol{W_{K_t}}\boldsymbol{p'}_{\boldsymbol{1}}, \boldsymbol{V} = \boldsymbol{W_{V_t}}\boldsymbol{p'}_{\boldsymbol{1}})_h \in \mathbb{R}^{d_t}, \tag{4}$$

where $MHAtt(\cdot)_h$ is h heads' attention mechanism and $\boldsymbol{W_Q}, \boldsymbol{W_K}, \boldsymbol{W_V}$ are learnable parameters. Then we concatenate the co-attention outputs $\boldsymbol{p'_1}$ and $\boldsymbol{t'_e}$ and fuse them with a self-attention layer, which is formalized as

$$\boldsymbol{m_1} = SAtt(\boldsymbol{p'_1} \oplus \boldsymbol{t'_e}) = MHAtt(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \boldsymbol{p'_1} \oplus \boldsymbol{t'_e})_h \in \mathbb{R}^{d_t}, \tag{5}$$

where $\oplus$ represents concatenation. We denote Eq. (3) to (5) as IFA.

Recurrent structure. After achieving $\boldsymbol{m_1}$, we obtain the final multi-modal representation $\boldsymbol{m_e}$ by iteratively fusing $\boldsymbol{p_i}, 2 \leq i \leq n$ into $\boldsymbol{m_1}$ with IFA and a gate mechanism. To begin with, we reverse the list of image features $[\boldsymbol{p_2}, ..., \boldsymbol{p_n}]$ to $\boldsymbol{S} = [\boldsymbol{p_n}, \boldsymbol{p_{n-1}}, ..., \boldsymbol{p_2}]$ as the input of IFA. The reason is that the recurrent structure tends to forget previously input information [54] and we expect RGMM to lay emphasis on the features of high-ranked images sorted by WSEs. Next, at the i-th step, we first feed the i-th image feature $\boldsymbol{S}[\boldsymbol{i}]$ and the multi-modal fusion result at $(i-1)$-th step $\boldsymbol{m'_{i-1}}$ into IFA to obtain the multi-modal representation $\boldsymbol{m_i}$, which is formalized as

$$\boldsymbol{m_i} = IFA(\boldsymbol{m'_{i-1}}, \boldsymbol{S}[\boldsymbol{i}]). \tag{6}$$

We then input $\boldsymbol{m'_{i-1}}$ and $\boldsymbol{m_i}$ into the gate layer and outputs $\boldsymbol{m'_i}$, which is also the input of $(i+1)$-th step. The gate layer is defined as

$$\boldsymbol{Z} = Sigmoid(\boldsymbol{W_m}\boldsymbol{m_i} + \boldsymbol{b_m}) \in \mathbb{R}^{d_t}, \boldsymbol{m'_i} = \boldsymbol{Z} \odot \boldsymbol{m_i} + (\mathbb{1} - \boldsymbol{Z}) \odot \boldsymbol{m'_{i-1}} \in \mathbb{R}^{d_t}, \tag{7}$$

where $\boldsymbol{W_m} \in \mathbb{R}^{d_t * d_t}$, $\boldsymbol{b_m} \in \mathbb{R}^{d_t}$ are learnable parameters, $\mathbb{1} \in \mathbb{R}^{d_t}$ donates as an all-ones vector and $\odot$ represents element-wise production. Finally, we feed $\boldsymbol{m_1}$ and the final multi-modal fusion result $\boldsymbol{m'_n}$ into a residual block to obtain $\boldsymbol{m_e}$, which reinforces the visual information in the Top-1 image.

After obtaining $\boldsymbol{m_e}$, we feed it into a binary classifier. The classifier consists of two fully connected layers and a softmax function. If the classifier outputs 0, we judge the entity as non-visualizable.

3.3 Entity-image matching

In most cases, some images retrieved from WSEs (e.g., Google) do not depict the corresponding entity. For example, given the query "Acroma (band)", WSE returns some images of Acroma's previous Facebook logo. Hence, we introduce entity-image matching and employ a pre-trained image-text model named CLIP

[32] to remove these images. For each entity, we treat its textual description and retrieved images as input and CLIP outputs their matching score.

Specifically, given an entity e, we first feed its textual description T_e into the text encoding part of CLIP and return the embedding $\boldsymbol{c_{e_{text}}} \in \mathbb{R}^{d_c}$. Then we encode e's retrieved images $[P_1, ..., P_n]$ with the visual encoding part of CLIP. After obtaining the embedding of text $\boldsymbol{c_{e_{text}}} \in \mathbb{R}^{d_c}$ and images $\boldsymbol{c_{e_{img}}} \in \mathbb{R}^{n*d_c}$, we employ outer product on them to compute the image-text matching degree. The process can be formulated as

$$\boldsymbol{c_{e_{text}}} = Enc_{text}(T_e), \boldsymbol{c_{e_{img}}} = [\boldsymbol{c^1_{img}}, ..., \boldsymbol{c^n_{img}}] = Enc_{image}([P_1, ..., P_n]), \quad (8)$$

$$\boldsymbol{c_{e_{score}}} = [c^1_{score}, ..., c^n_{score}] = \boldsymbol{c_{e_{text}}} \boldsymbol{c^T_{e_{img}}} \in \mathbb{R}^{d_n}, \quad (9)$$

where $c^i_{score} \in \mathbb{R}$ represents the matching score between the text T_e and image P_i. If c^i_{score} is lower than the pre-defined threshold, we remove P_i.

3.4 Entity type detection

Although we have removed noisy images not depicting the corresponding entity with model CLIP and the text information, some remaining images may still not be the appropriate visual representation. For example, for the entity *Acroma*, images such as a shirt with "*Acroma*" and a WordArt "*Acroma*" are considered valid by CLIP. These images illustrate the given entity but do not allow us to associate *Acroma* with "band". Hence, in this paper, we take the type information of entities to conduct further filtering. The core idea is to employ CV models to detect the entity class from a candidate image and assess whether the result aligns with the type information.

Specifically, given an entity e and one of its candidate images P_i, we first retrieve e's type information A_e from KGs (e.g., DBpedia). Then we map A_e to the expected entity classes $C'_e = [C'_1, C'_2, ..., C'_n]$ using a manually constructed type-to-class list $L_{A \to C}$ (e.g., Band $\to$ [Person]), where *class* is from COCO [24] and imagenet dataset [35]. After obtaining C'_e, we employ pre-trained CV models YOLO [33] and VGG [36] to identify entity classes $C_e = [C_1, C_2, ..., C_m]$ from P_i. Finally, we calculate the intersection of C'_e and C_e to determine whether P_i should be removed. This process is formalized as

$$Y = \Omega(L_{A \to C}(A_e), CV(P_i)) \quad (10)$$

where $\Omega(\cdot)$ denotes the Boolean function judging whether the intersection is an empty set. If the output is true, we remove the P_i.

4 MMpedia Analysis

In this section, we first report the dataset statistics of MMpedia and typical MMKGs. Then we give a detailed analysis of the image quality and diversity.

MMpedia statistics. We perform our proposed method on the KG DBpedia, which has a well-defined ontology and contains 7,195,709 entity nodes,

Table 1. Comparison between MMpedia and typical MMKGs.[6]

KG	Nodes	Images	Triples(KG)
IMGpedia [13]	14,765,300	14,765,300	-
Imagegraph [30]	14,870	829,931	564,010
MMKG[26]	45,011	37,479	814,127
Richpedia[46]	29,985	2,915,770	-
VisualSem[2]	89,896	938,100	1,481,007
MMpedia (Ours)	2,661,941	**19,489,074**	**5,960,965**

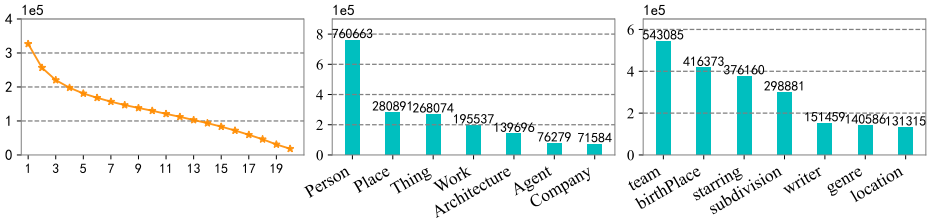


Fig. 4. The distribution of images per node (left) and most common numerical categories of entities (middle) and relations (right) in MMpedia.

633 relation categories, and 21,687,345 triples. Based on this KG, we construct a MMKG named MMpedia, including 598 relation categories, 5,960,965 triples and 19,489,074 images for 2,661,941 entity nodes. Table 1 reports the statistic of our MMpedia and other typical MMKGs. MMpedia has the biggest image dataset among existing MMKGs. Note that IMGpedia has the most entities while it is built by data linking without powerful means to supervise the image quality. To better understand MMpedia, we report the distribution of images per node and high-frequency entity and relation categories in Figure 4. Around 45% of entities have one to five images and each entity has 7.3 images on average. For entities, we note that *Person* is the most numerous entity type of all 362 categories, accounting for 28.57%. The number of *Place*, *Thing*, *Work*, and *ArchitecturalStructure* also exceeds 10^5. For relations, we observe that *team*, *birthPlace*, *starring*, *subdivision*, *writer*, *genre* and *location* take a high proportion in total of 598 categories, all exceeding 10^5.

Image quality. Since there is no ground truth, we employ manual and automatic evaluation to verify the image quality in MMpedia. For manual evaluation, we invite three CV research students. The criteria is that if an image reflects what the corresponding entity is, it is labeled as 1. Otherwise, it is labeled as 0. Before manual evaluation begins, we conduct a test for all participants. To this end, we crawl 1,000 image-text pairs from Wikipedia and randomly select 100 correct and 100 incorrect pairs for each participant to evaluate. We start the manual evaluation when every participant achieves a test accuracy of 95%. Dur-

[6] We report triples of relations between entities in KG. The triples in IMGpedia and Richpedia are relations between entities and images.

ing the manual evaluation process, we randomly select 500 entities with 3,415 images. We also add 200 noisy images to assess the quality of the evaluation, which provides a basis for final accuracy calculations. The three participants recall 0.98, 0.96 and 0.99 of these noisy images, respectively. The current MMpedia achieves 84.91% accuracy on the weighted average and 81.14% on T@3, where T@k means an image is labeled as 1 by k people. The Fleiss' kappa [14] is 0.836, showing the consistency of human evaluation. Additionally, to evaluate the quality of images associated with "nodes pairs", we randomly select 500 pairs of Top-1 images corresponding to the head-tail entities, which are sorted by the proposed pipeline method. The average accuracy is 88.20% and the Fleiss' kappa is 0.859. For automatic evaluation, we introduce two downstream tasks to verify the image quality in section 5.2.

Image diversity. Similarly, we employ human evaluation on 3,415 images of 500 entities to verify the image diversity of MMpedia. We first evaluate each entity's diversity by calculating the percentage of similar image pairs. For example, given an entity e with n_e images, we will build $n_p = 0.5 * n_e * (n_e - 1)$ image pairs. If there are s_e similar image pairs, the diversity score d of e will be $d = \frac{n_p - s_e}{n_p}$. Then, we compute the average diversity of each entity as the diversity score of the whole dataset. Finally, our current MMpedia reaches the average diversity score of 90.07% and the Fleiss' kappa is 0.807.

5 Experiment

Through the experiment, we expect to demonstrate the effectiveness of our proposed pipeline method and collected images. We first report implementation details of the pipeline method. Then we introduce *entity prediction* and *relation prediction* to verify that our collected images are helpful for downstream tasks. Finally, we give a detailed analysis on MMpedia construction.

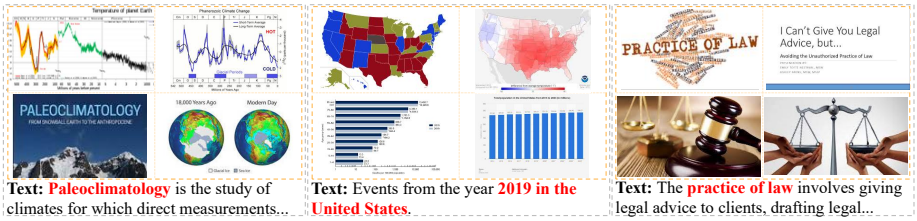
5.1 Implementation details

We give detailed information about each step in the proposed method, including the input-output, data analysis and hyperparameter settings.

Entity information collection collects 3,494,367 entities with the information of type, textual description and candidate images. First, for 7,195,709 entity nodes in DBpedia, we remove 2,600,603 entity nodes that are similar to others (e.g., *Herbowo* and *Herbowo__Tenure__1*). Second, for the remaining 4,595,106 entities, we take SPARQL API[7] to retrieve the corresponding 3,668,041 textual descriptions and 4,264,106 type information from DBpedia and Wikidata. We remove entities missing the abstract or type. Finally, we crawl Top-20 images from Google for each entity. Since some entities have less than 20 images in Google, there are 66,399,460 images for remained 3,494,367 entities.

Non-visual entity filtering judges 3,136,997 entities into visualizable and 357,370 entities into non-visualizable. We employ ResNet50 and BERT$_{BASE}$ to

[7] `https://dbpedia.org/sparql`



Text: **Paleoclimatology** is the study of climates for which direct measurements...

Text: Events from the year **2019 in the United States**.

Text: The **practice of law** involves giving legal advice to clients, drafting legal...

(a) Images and entities eliminated in the step of non-visual entity filtering

Text: **Motor Launch** (ML) is a small military vessel in Royal Navy service. It was...
Filterd: An image of 1910 Mathis launch

Text: **Carl Swenson** is an American former cyclist and cross-country skier. He...
Filterd: An image of Karl Swenson

Text: **Rover Meteor** was a short-lived 2½-litre or 2-litre medium-sized car made...
Filterd: An image of Rolls-Royce Meteor

(b) Images eliminated (left) and remained (right) in the step of entity-image matching

Text: **Weedflower** is a 2006 American children's historical novel by Cynthia...
Type: Book

Text: **Novembre** is an Italian heavy metal band, formed in Rome in 1990.
Type: Band

Text: **Pratibhasthali** is a girls school started by Acharya Vidyasagar. Schools are...
Type: School

(c) Images eliminated (left) and remained (right) in the step of entity type detection

Fig. 5. Case studies of each step in proposed pipeline methods. The entities in KG are marked in red, while entities depicted by noisy images are marked in blue.

embed the Top-5 images and the text respectively, where $d_p = 2048$ and $d_t = 768$. During the training process, we run for 50 epochs with a batch size of 32. We choose AdamW as the optimizer and the learning rate is 1e-4.

Since the model is supervised and there is no public labeled data for non-visual entity filtering, we construct a dataset based on Wordnet. We first sample 200 entities from Wordnet and research the path between the root node r and them in the 'hyponymy' hierarchy. Given an entity e and its path, we observe two regularities: (1) If $pathLength|(e,r)| \leq 5$ and the node "Abstraction" appears in the path, e is commonly to be 0 (non-visualizable) and (2) If e is a leaf node and the node "Abstraction" not appears in the path, e is commonly to be 1 (visualizable). Based on (1) and (2), we crawled 2,142 entities and give them unsupervised labels. Then we collect their textual information and images from DBpedia and Google, respectively. Finally, we invite three volunteers to revise the unsupervised label based on the criteria that if the Top-5 images of an entity reflect it, its label should be 1, and vice versa. The Fleiss' kappa is 0.798 and we revise the unsupervised label of an entity if it is corrected by three volunteers at the same time. Note that these volunteers are different with those in section 4.

Finally, we collect 1,182 visual entities and 960 non-visual entities and randomly split them as 1328/406/408 for training, validation and testing respectively. Our classifier reaches the F1 score of 92.88% on the test dataset.

To intuitively understand non-visualizable and visualizable, we give some cases. As shown in Figure 5(a), it is hard to find an image reflecting *Paleoclimatology*, which is a scientific discipline, not a data table or a globe. By contrast, images (right) reflect the corresponding entity, as shown in Figure 5(b) and 5(c).

Entity-image matching remains 2,785,592 entity nodes and 22,274,190 images. We introduce the pre-trained CLIP to perform *Entity-image matching*. We sample the images of 500 entities and conduct a statistical analysis on the CLIP results. Finally, we define $Min_{CLIP} = 29$ as the threshold since we observe that most noisy images have a CLIP score of lower than 29. Figure 5(b) gives the cases to intuitively demonstrate the effectiveness of this step.

Entity type detection remains 2,661,941 entity nodes and 19,489,074 images. We first manually construct a type-to-class list containing 1,179 mappings, where *type* is from DBpedia containing 141 entity type information and *class* is from COCO and ImageNet containing 1080 image recognition classes (e.g., Ship $\rightarrow$ [Boat, Fireboat, Ocean liner]). Then we introduce YOLOv5 and VGG19 to perform image recognition. The input is 10,293,162 candidate images and 1,306,957 entity type information. For each image, the recognized entity class consists of all results from YOLOv5 and Top-3 ones from VGG19. Figure 5(c) gives the cases to intuitively demonstrate the effectiveness of this step.

5.2 Downstream tasks

To verify the usefulness of MMpedia, we employ its images in two real-world tasks: (1) *entity prediction* and (2) *relation prediction* [53]. We conduct the experiment on DB15K [26] which is a sub-graph of DBpedia. Since there are non-visual entities in DB15K, we need to filter the triples. Specifically, we first remove the triples if the head-tail entity can not find corresponding images in MMpedia. Then we further filter the triples containing one-shot relation or head-tail entity. Finally, we remain 23,055 triples and split them as 16,384/ 3,389/ 3,282 for training, validation and testing. The splitting principle is that entities and relations in validation and test sets need to appear in the training set. The vocabulary size is 5,239 and the number of relation categories is 158.

Entity prediction. Given a triple fact $< h, r, t >$, entity prediction requires models to complete the missing head or tail information. Taking the tail entity prediction as an example, the input is the image of h and the textual information of $< h, r >$. For each test example, we first replace t with all candidate entities and then record the ranking of them in descending order based on the predicted scores. We report four metrics: MRR, MR, Hit@1, and Hit@10, where MR and MRR are the mean rank and reciprocal rank of all correct entities, respectively and Hit@k represents the proportion of correct entities existing in Top-k.

First, to verify whether our collected images reflect the corresponding entity, we design an A/B testing. The input of experiment A_1 is h and r

Table 2. The performance of BERT-based models on the entity prediction task. We highlight the data using our collected images in gray. ↑ means that higher values provide better performance while ↓ means that lower values provide better performance.

Method	Head Entity prediction				Tail Entity prediction			
	MRR↑	MR↓	Hit@1↑	Hit@10↑	MRR↑	MR↓	Hit@1↑	Hit@10↑
BERT	10.94	439.43	5.00	22.73	23.67	157.37	14.78	40.77
+ResNet50+Noise	10.91	441.36	5.09	22.09	23.61	152.84	14.20	42.23
+ResNet50+Our	**12.27**	**423.43**	**5.94**	24.95	**25.44**	**147.27**	**16.33**	**43.93**
ViLT+Noise	10.70	639.71	4.88	22.58	22.90	249.83	14.47	40.74
ViLT+Our	12.08	596.34	5.45	**26.02**	24.60	226.54	16.30	42.47

Table 3. The result of SOTA MKGC and KGC models on the entity prediction task.

Method	Head Entity prediction				Tail Entity prediction			
	MRR↑	MR↓	Hit@1↑	Hit@10↑	MRR↑	MR↓	Hit@1↑	Hit@10↑
Translational Distance Models								
ComplEX	22.98	1476.31	17.40	33.18	20.23	2125.34	15.39	29.77
RotatE	26.14	784.91	20.69	36.53	39.16	579.77	31.23	53.84
LineaRE	26.34	418.47	21.43	36.02	34.38	309.99	27.57	47.54
RSME+Google	25.90	622.46	20.78	35.01	40.78	308.95	33.00	55.48
RSME+Our	26.93	547.08	21.57	36.41	42.10	274.74	34.34	57.10
MoSE+Google	29.04	338.24	21.25	42.60	43.84	123.16	33.79	62.87
MoSE+Our	29.99	**329.28**	22.73	**43.27**	45.62	**122.13**	35.89	63.92
Pre-trained Language Models								
KG-BERT	3.68	543.54	0.91	7.46	7.53	493.58	2.43	17.06
MKGformer+Google	29.01	379.07	22.82	40.13	44.62	135.50	35.83	61.61
MKGformer+Our	**30.05**	371.30	**23.85**	42.02	**48.17**	128.20	**39.49**	**65.14**

while experiment B_1 has two kinds of input: (1) h, r and h's image in MMpedia (+Our) and (2) h, r and an image of another entity (+Noise). For the input "The r of h is [MASK]" of experiment A_1, we employ $BERT_{base}$ as the backbone, where a classifier is connected to the [MASK] representation. For the experiment B_1, we introduce BERT+ResNet50 [43] and ViLT [19] to predict t. As shown in table 2, BERT+ResNet50 and ViLT with (+Our) outperform BERT, indicating that image features are helpful for *entity prediction*. Moreover, both methods with (+Noise) achieve no significant improvement than BERT, demonstrating that the improvement is mainly due to the input image rather than the added visual encoder. Hence, our collected images provide effective visual information.

Second, to evaluate whether our collected images improve the performance of state-of-the-art (SOTA) multi-modal knowledge graph completion (MKGC) models, we design an A/B testing. For each MKGC model, the input of experiment A_2 is h, r and h's image crawled from Google (+Google) while B_2 is h, r and h's image from MMpedia (+Our). We introduce two SOTA MKGC models MoSE [55] and MKGformer [7]. Following them, we

Table 4. The results of relation prediction.

Methods	MRR↑	MR↓	Hit@1↑	Hit@3↑	Hit@10↑
ComplEX	41.48	26.11	24.38	55.30	69.35
RotatE	65.51	5.29	50.79	76.93	91.99
KG-BERT	73.36	2.95	57.86	88.48	96.92
RSME+Google	68.34	4.03	51.86	83.21	93.60
RSME+Our	69.51	3.76	53.05	84.89	94.64
MoSE+Google	72.24	6.20	59.08	83.58	93.48
MoSE+Our	74.54	6.07	63.01	85.10	93.69
MKGformer+Google	78.96	2.20	65.90	91.62	98.35
MKGformer+Our	**80.34**	**2.12**	**68.31**	**91.74**	**98.57**

also introduce four uni-modal KGC models ComplEX [42], RotatE [39], LineaRE [31], KG-BERT [53] and one MKGC model RSME [47]. As shown in Table 3, MKGC models +Our outperform other methods. Compared with MKGC models +Google, MKGC models +Our achieve at most 3.5% improvement on Hit@1, indicating that our collected images enhance MKGC models' performance.

Relation prediction. Given a triple $< h, ?, t >$, models are required to complete the missing r. The input is h, t and two images of h and t respectively. The evaluation metrics are the same as those in *entity prediction*.

To evaluate whether our collected images are useful to improve MKGC models' performance on relation prediction, we design an A/B testing. For each MKGC model, the input of experiment A_3 is h, t and images (+ Google) while B_3 is h, t and images (+Our). As shown in Table 4, MKGC models outperform uni-modal KGC models, indicating that the visual information is beneficial for *relation prediction*. Compared with MKGC models +Google, MKGC models +Our achieve at most 4.0% improvement on Hit@1. Hence, our collected images enhance the model's performance on *relation prediction*.

5.3 Detailed analysis

In this section, we make a detailed analysis on *non-visual entity filtering*, *entity type detection* and the multi-modal large language model (M-LLM).

To verify whether images reflect non-visual entities, we still design an A/B testing. For experiment A_4, the input is h and r. For experiment B_4, the input is h, r and two kinds of h's image: (1) +Google and (2) filtered by steps 3 and 4 (+Our w/o 2). We first select 6,657 triples from DB15K where h is non-visualizable. Then we split the triples into 4,644/970/1043 for training, validation and testing. We denote this dataset as D_{nv}. Finally, we also employ BERT-based models to perform tail entity prediction. As shown in Table 5, either (+Google) or (+Our w/o 2) do not enhance BERT's performance, showing the necessity of filtering non-visual entities. To evaluate our proposed RGMM, we compare it with some typical SOTA multi-modal interaction methods. The baselines contain BERT, MAS [48], MCAN [51], MIFN [23] and DRMM [41]. Among them, MIFN

Table 5. Tail entity prediction on D_{nv}. We denote w/o k as removing the step k.

Methods	Input	MRR↑	MR↓	Hit@1↑	Hit@3↑	Hit@10↑
BERT	(h, r)	**42.21**	53.77	**30.29**	46.40	**68.55**
BERT+ResNet50	+Google	41.49	**52.04**	29.62	45.83	67.31
	+Our w/o 2	41.89	59.47	29.91	**46.60**	68.36
ViLT	+Google	39.99	104.76	29.15	46.02	60.88
	+Our w/o 2	40.26	94.64	29.82	45.35	60.79

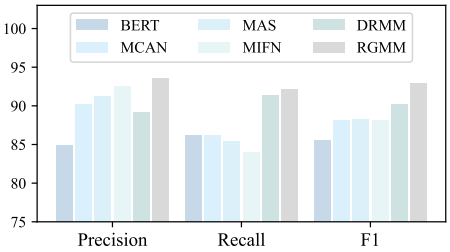


Fig. 6. Non-visual entity filtering.

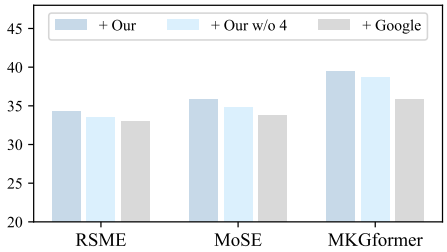


Fig. 7. Entity type detection.

and DRMM can process multiple images. We train all models on the dataset depicted in the section 5.1 with the same hyperparameters. As shown in Figure 6, RGMM can filter non-visual entities more effectively.

To evaluate whether entity type detection improves the image quality, we compare MKGC models' performance on three kinds of images: (1) +Our, (2) filtered without *entity type detection* (+Our w/o 4) and (3) +Google. For the dataset depicted in section 5.2, we first replace the images of 844 entities with those filtered via *entity type detection*. Then we also employ MKGC models to conduct tail entity prediction. As shown in Figure 7, the performance of MKGC models decreases on Hit@1, showing the effectiveness of *entity type detection*.

To evaluate whether M-LLMs generate high-quality images for the given entity, we introduce VisualChatGPT (VCG). The input is the prompt *"Please generate an image of [entity]. [entity's abstract]"* and the output is a generated image. We sample 200 entities from DBpedia and invite the participants in section 4 to evaluate the images generated by VCG. VCG achieves an average accuracy of 0.29 and the Flessi's Kappa is 0.870. The reasons for error cases are mainly classified into two groups. The first group is an image depicting another entity of the same type as the given entity, accounting for 59%. For example, given the entity *Masuisuimatamaalii Tauaua-Pauaraisa*, VCG generates a *another_person.jpg* as shown in Figure 8. The second group is an image of another entity appeared in the given abstract, accounting for 28%. For example, as shown in Figure 8, given the company *Dean Markley*, VCG generates a *guitar.jpg*, where *guitar* appears in the given abstract. Hence, grounding entities to images remains a challenge for M-LLMs. **To evaluate whether our collected images are helpful for M-LLMs**, we randomly select 200 triples and compare

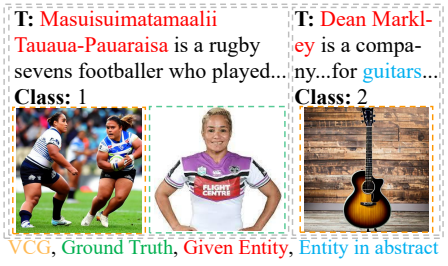


Fig. 8. Case study of VCG.

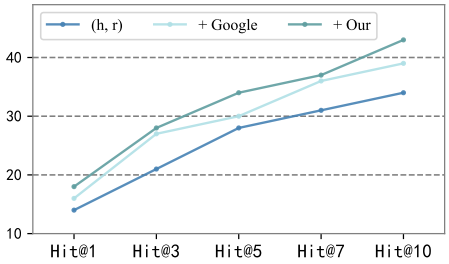


Fig. 9. The result of VCG.

the performance of VCG on three kinds of input: (1) h and r, (2) h, r +Our and (3) h, r +Google. VCG is asked to reorder the list of candidate t based on the given h, r. The prompt consists of task definition, one positive example and two negative examples [15]. As shown in Figure 9, our collected images improve VCG's performance on *tail entity prediction*.

6 Conclusion

In this paper, we present a large-scale MMKG named MMpedia. To this end, we propose a novel pipeline method, which first collects images from a WSE and filters non-visual entities with a multi-modal classifier, and then leverage entities' textual and type information to remove noisy images. Through the pipeline method, MMpedia is constructed, containing 2,661,941 entities and 19,489,074 images. As we know, MMpedia boasts the largest number of images among existing MMKGs. Extensive experiments are conducted to demonstrate the effectiveness of our proposed method. Furthermore, the images in MMpedia are helpful for different downstream tasks.

Acknowledgements This work was supported by the Shanghai Municipal Special Fund for Promoting High-quality Development of Industries (2021-GZL-RGZN-01018) and the Shanghai Sailing Program (23YF1409400).

References

1. Aghaei, S., Raad, E., Fensel, A.: Question answering over knowledge graphs: A case study in tourism. IEEE Access (2022)
2. Alberts, H., Huang, T., Deshpande, Y., Liu, Y., Cho, K., Vania, C., Calixto, I.: Visualsem: a high-quality knowledge graph for vision and language. arXiv preprint arXiv:2008.09150 (2020)
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems **26** (2013)

4. Calabrese, A., Bevilacqua, M., Navigli, R.: Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4680–4686 (2020)
5. Chen, D., Li, Z., Gu, B., Chen, Z.: Multimodal named entity recognition with image attributes and image knowledge. In: Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26. pp. 186–201. Springer (2021)
6. Chen, Q., Wang, W., Huang, K., Coenen, F.: Zero-shot text classification via knowledge graph embedding for social media data. IEEE Internet of Things Journal (2021)
7. Chen, X., Zhang, N., Li, L., Deng, S., Tan, C., Xu, C., Huang, F., Si, L., Chen, H.: Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. arXiv preprint arXiv:2205.02357 (2022)
8. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: Proceedings of the IEEE international conference on computer vision. pp. 1409–1416 (2013)
9. Cheng, M., Sun, Y., Wang, L., Zhu, X., Yao, K., Chen, J., Song, G., Han, J., Liu, J., Ding, E., et al.: Vista: vision and scene text aggregation for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5184–5193 (2022)
10. Colla, D., Mensa, E., Radicioni, D.P., Lieto, A.: Tell me why: Computational explanation of conceptual similarity judgments. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 74–85. Springer (2018)
11. Corbiere, C., Ben-Younes, H., Ramé, A., Ollion, C.: Leveraging weakly annotated data for fashion image retrieval and label prediction. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2268–2274 (2017)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Ferrada, S., Bustos, B., Hogan, A.: Imgpedia: a linked dataset with content-based analysis of wikimedia images. In: International Semantic Web Conference. pp. 84–93. Springer (2017)
14. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
15. Gao, J., Zhao, H., Yu, C., Xu, R.: Exploring the feasibility of chatgpt for event extraction. arXiv preprint arXiv:2303.03836 (2023)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hendriksen, M., Vakulenko, S., Kuiper, E., de Rijke, M.: Scene-centric vs. object-centric image-text cross-modal retrieval: A reproducibility study. arXiv preprint arXiv:2301.05174 (2023)
18. Kang, H., Li, X., Jin, L., Liu, C., Zhang, Z., Li, S., Zhang, Y.: Tspnet: Translation supervised prototype network via residual learning for multimodal social relation extraction. Neurocomputing **507**, 166–179 (2022)
19. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)

20. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
21. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web **6**(2), 167–195 (2015)
22. Li, M., Zareian, A., Lin, Y., Pan, X., Whitehead, S., Chen, B., Wu, B., Ji, H., Chang, S.F., Voss, C., et al.: Gaia: A fine-grained multimedia knowledge extraction system. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 77–86 (2020)
23. Li, Y., Li, J., Jin, H., Peng, L.: Focusing attention across multiple images for multimodal event detection. In: ACM Multimedia Asia, pp. 1–6. Association for Computing Machinery (2021)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10921–10930 (2020)
26. Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., Rosenblum, D.S.: Mmkg: multi-modal knowledge graphs. In: European Semantic Web Conference. pp. 459–474. Springer (2019)
27. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
28. Mafla, A., Rezende, R.S., Gomez, L., Larlus, D., Karatzas, D.: Stacmr: Scene-text aware cross-modal retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2220–2230 (2021)
29. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence **193**, 217–250 (2012)
30. Oñoro-Rubio, D., Niepert, M., García-Durán, A., González, R., López-Sastre, R.J.: Answering visual-relational queries in web-extracted knowledge graphs. arXiv preprint arXiv:1709.02314 (2017)
31. Peng, Y., Zhang, J.: Lineare: Simple but powerful knowledge graph embedding for link prediction. In: 2020 IEEE International Conference on Data Mining (ICDM). pp. 422–431. IEEE (2020)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
33. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
34. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)

36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706 (2007)
38. Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., Wang, Z., Zheng, K.: Multi-modal knowledge graphs for recommender systems. In: Proceedings of the 29th ACM international conference on information & knowledge management. pp. 1405–1414 (2020)
39. Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197 (2019)
40. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
41. Tong, M., Wang, S., Cao, Y., Xu, B., Li, J., Hou, L., Chua, T.S.: Image enhanced event detection in news articles. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9040–9047 (2020)
42. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: International conference on machine learning. pp. 2071–2080. PMLR (2016)
43. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
44. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
45. Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E.p., Hoi, S.C.: Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. IEEE Transactions on Multimedia **24**, 2515–2525 (2021)
46. Wang, M., Wang, H., Qi, G., Zheng, Q.: Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Research **22**, 100159 (2020)
47. Wang, M., Wang, S., Yang, H., Zhang, Z., Chen, X., Qi, G.: Is visual context really helpful for knowledge graph? a representation learning perspective. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2735–2743 (2021)
48. Wang, X., Tian, J., Gui, M., Li, Z., Ye, J., Yan, M., Xiao, Y.: : Prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In: International Conference on Database Systems for Advanced Applications. pp. 297–305. Springer (2022)
49. Wen, H., Lin, Y., Lai, T., Pan, X., Li, S., Lin, X., Zhou, B., Li, M., Wang, H., Zhang, H., et al.: Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. pp. 133–143 (2021)
50. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
51. Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2560–2569 (2021)

52. Yang, Y., Zhu, Y., Li, Y.: Personalized recommendation with knowledge graph via dual-autoencoder. Applied Intelligence **52**(6), 6196–6207 (2022)
53. Yao, L., Mao, C., Luo, Y.: Kg-bert: Bert for knowledge graph completion. arXiv preprint arXiv:1909.03193 (2019)
54. Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., Tian, G.: Do rnn and lstm have long memory? In: International Conference on Machine Learning. pp. 11365–11375. PMLR (2020)
55. Zhao, Y., Cai, X., Wu, Y., Zhang, H., Zhang, Y., Zhao, G., Jiang, N.: Mose: Modality split and ensemble for multimodal knowledge graph completion. arXiv preprint arXiv:2210.08821 (2022)
56. Zheng, C., Wu, Z., Feng, J., Fu, Z., Cai, Y.: Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
57. Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., Xiao, Y., Yuan, N.J.: Multimodal knowledge graph construction and application: A survey. arXiv preprint arXiv:2202.05786 (2022)