applied sciences

MDPI

Article

Multi-Modal Entity Alignment Method Based on Feature Enhancement

Huansha Wang, Qinrang Liu *, Ruiyang Huang and Jianpeng Zhang

Institute of Information Technology, PLA Information Engineering University, Zhengzhou 450001, China; whs123@mail.ustc.edu.cn (H.W.); gisexpert@163.com (R.H.); zjp@ndsc.com.cn (J.Z.)
* Correspondence: qinrangliu@sina.com

Abstract: Multi-modal entity alignment refers to identifying equivalent entities between two different multi-modal knowledge graphs that consist of multi-modal information such as structural triples and descriptive images. Most previous multi-modal entity alignment methods have mainly used corresponding encoders of each modality to encode entity information and then perform feature fusion to obtain the multi-modal joint representation. However, this approach does not fully utilize the multi-modal information of aligned entities. To address this issue, we propose MEAFE, a multi-modal entity alignment method based on feature enhancement. The MEAFE adopts the multi-modal pre-trained model, OCR model, and GATv2 network to enhance the model's ability to extract useful features in entity structure triplet information and image description, respectively, thereby generating more effective multi-modal representations. Secondly, it further adds modal distribution information of the entity to enhance the model's understanding and modeling ability of the multi-modal information. Experiments on bilingual and cross-graph multi-modal datasets demonstrate that the proposed method outperforms models that use traditional feature extraction methods.

Keywords: knowledge graph; multi-modal technology; entity alignment; representation learning

check for updates

Citation: Wang, H.; Liu, Q.; Huang, R.; Zhang, J. Multi-Modal Entity Alignment Method Based on Feature Enhancement. *Appl. Sci.* **2023**, *13*, 6747. https://doi.org/10.3390/app13116747

Academic Editor: Tobias Meisen

Received: 25 April 2023
Revised: 30 May 2023
Accepted: 31 May 2023
Published: 1 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A knowledge graph is a large-scale knowledge database that stores knowledge in a structured manner. It represents entity-related knowledge in the form of triplets, such as <headEntity,Attribute,attributeValue> or <headEntity,Relation,tailEntity> to display the properties and relationship information of entities clearly and intuitively. Large public knowledge graphs such as DBPedia [1] and YAGO [2] have demonstrated their crucial role in tasks such as information retrieval and relationship mining. With the emergence of multi-modal data such as images and videos, researchers have realized that they have more information richness and intuitiveness compared to text data. The excellent performance of large-scale multi-modal pre-training models such as UNITER [3] and ImageBERT [4] also proves that using multi-modal data to train models can enable them to achieve excellent representation capability. Therefore, multi-modal knowledge graphs such as MMKG [5] and RichPedia [6] have emerged in large numbers. Multi-modal knowledge graphs integrate visual data such as images and videos into traditional knowledge graphs and then treat them as entities or descriptive attributes to further enhance the completeness and richness of knowledge graphs. They can also be applied to multi-modal downstream tasks such as visual question answering and image-text generation, thus enhancing the universality of the knowledge graph. Figure 1 is a simple illustration of a multi-modal knowledge graph.

However, due to the constant growth of data and knowledge, multi-modal knowledge graphs are always incomplete. The incompleteness is mainly manifested in two aspects: entity missing and triple missing. Entity missing refers to the absence of entities that should have been included in the knowledge graph, while triple missing refers to the lack of important attributes or relationships of a given entity; both can lead to unsatisfactory downstream

task performance based on the knowledge graph. Entity alignment aims to use models and algorithms to determine whether two entities with different sources and representations refer to the same object in reality, thus reducing the number of redundant entities in the knowledge graph and integrating the triple information of aligned entities. Therefore, entity alignment is an important supporting technology for improving multi-modal knowledge graphs. We denote two pre-aligned multi-modal knowledge graphs as $G_1 = (E_1, R_1, A_1, V_1, T_1)$ and $G_2 = (E_2, R_2, A_2, V_2, T_2)$, where $E_i, R_i, A_i, V_i, T_i, i \in \{1, 2\}$ are the sets of entities, relations, attributes, visual information, and triples, respectively, of two graphs. Multi-modal entity alignment aims to find aligned entity pairs $S = \{(e_1, e_2) \in E_1 \times E_2 | e_1 \Leftrightarrow e_2\}$. The symbol $\Leftrightarrow$ means that the entity e_1 from E_1 and the entity e_2 from E_2 have the same semantic; e_1 and e_2 refer to the same object in reality.

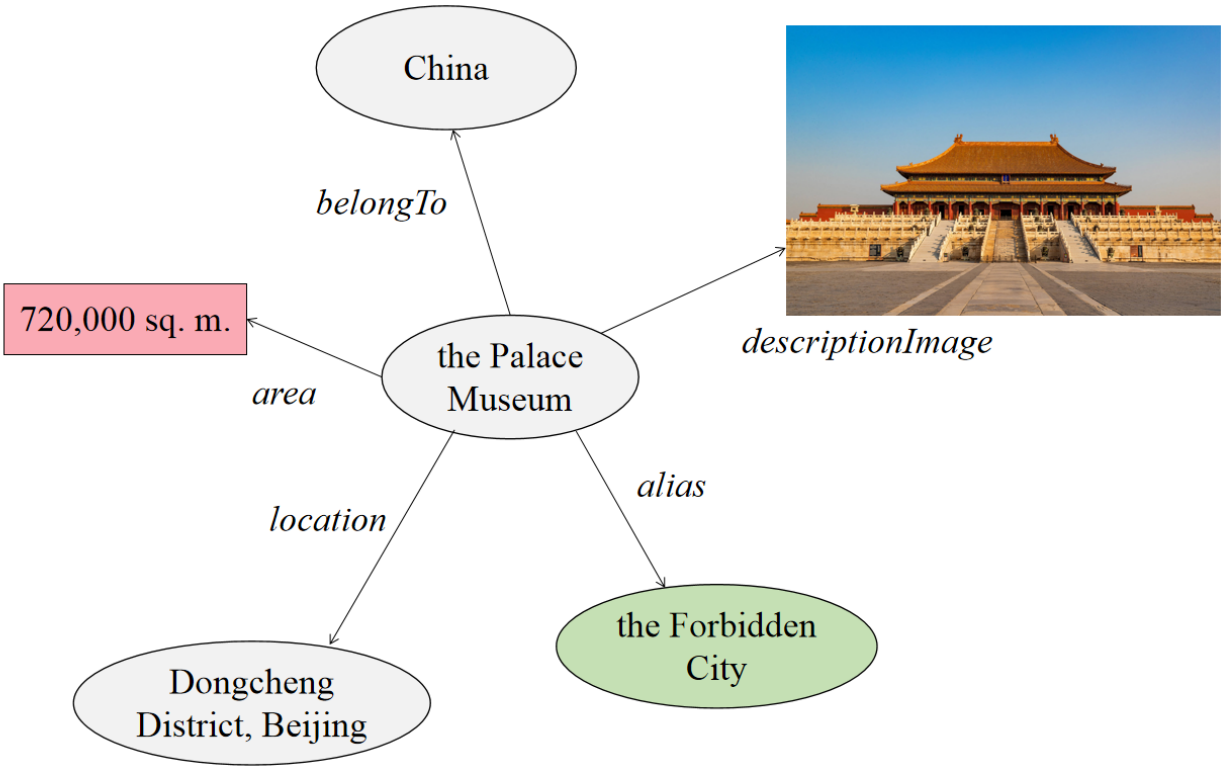

Figure 1. An example of a multi-modal knowledge graph.

Some researchers have proposed effective models in the field of multi-modal entity alignment. For example, MMEA [7] models different modal information of entity attributes and conducts knowledge fusion to achieve the effect of modeling and alignment of multi-modal entities. EVA [8] uses the visual similarity of entities to create an initial seed dictionary, providing a completely unsupervised solution. HMEA [9] models multi-modal representations in the hyperbolic space to improve the entity representation capability. MSNEA [10] also extracts visual, relational, and attribute features of entities separately, integrates visual features based on modal enhancement mechanisms to guide multi-modal feature learning, and adaptively allocates attention weights to capture valuable attributes for alignment. MCLEA [11] jointly models the intra-modality and inter-modality interactions based on contrastive learning after acquiring the attribute features of each modality to improve the model's representation capability. Table 1 shows the comparison of main stream multi-modal entity alignment models. However, the previously mentioned models did not fully utilize the entity information of each modality but only relied on the pre-trained encoders for encoding. This limits their ability to extract deep features from the data. For instance, the equivalent entities that refer to the same movie in the real world may

have different description images (such as different movie posters) in various knowledge graphs, which may result in low similarity if only a pre-trained visual model is used for encoding. However, there may be similar text in the posters, such as the movie name and slogan. By extracting the text information to supplement the entity image information, the alignment accuracy can be further improved.

Table 1. Comparison of multi-modal entity alignment models. Adversarial column refers to whether the model introduces adversarial learning.

Model	Proposed Year	Innovation	Adversarial
PoE [5]	2019	Combine multi-modal Features and matching the underlying semantics	No
MMEA [7]	2020	Obtain the joint representations	No
EVA [8]	2021	Achieve in both supervised and unsupervised Manner	No
HMEA [9]	2021	Model representations in the hyperbolic space	No
MSNEA [10]	2022	Adaptively allocates attention weights to capture valuable attributes	No
MCLEA [11]	2022	introduce contrastive learning	Yes

To further enhance the utilization of entity multi-modal information, this paper proposes **MEAFE**: a **m**ulti-modal **e**ntity **a**lignment method based on **f**eature **e**nhancement. The core idea is to maximize the use of entity visual, textual, and relational modals to enhance the corresponding feature embedding and improve the knowledge representation ability of the model. Firstly, MEAFE enhances the image modality information by cleaning up the low semantic relevant images using a pre-trained multi-modal model, extracting text information possibly present in the entity image using an OCR (Optical Character Recognition) model, and then encodes it with a pre-trained language model as a supplementary representation of the image embedding. Secondly, the GATv2 [12] network is used instead of traditional graph convolutional networks [13] or graph attention networks to extract the neighborhood structure features of the entity. Finally, the distribution of the entity's modal information, including whether the entity has some modal information and its corresponding amount of data, is used as an auxiliary attribute of the entity to enhance the model's understanding and modeling ability of various modal information. In addition, we introduce the intra-modal contrastive loss and multi-modal alignment loss used in the MCLEA to better align entities in different knowledge graphs. We design and conduct experiments to verify the effectiveness of the model and validate it on DBP15K [8], which includes three bilingual datasets, and on two cross kg datasets FB15K-DB15K/YAGO15K [5]. The proposed model achieves state-of-the-art performance, demonstrating the positive effect of multi-modal feature enhancement on entity alignment.

The contributions of this paper are as follows: (1) To address the issue of insufficient utilization of visual information in traditional multi-modal entity alignment models, we propose using pre-trained multi-modal models and OCR models to enhance the visual representation ability, enabling models to extract more entity-related knowledge from visual images. (2) To address the issue of the weak adaptability of vanilla graph attention networks in entity alignment tasks, which affects alignment accuracy, we propose using the dynamic GATv2 instead to improve the ability to extract structural features. (3) We propose the use of the modal distribution information of entities as a supplementary part of entity features in order to improve the model's understanding and modeling ability for multi-modal information.

2. Materials and Methods

This paper proposes MEAFE, a multi-modal entity alignment method based on feature enhancement, which enhances image and structural features with multi-modal pre-training models, OCR models, and GATv2 networks based on traditional models that only use each modal encoder to initially obtain entity representations. Additionally, MEAFE utilizes the entity modal distribution information to improve the understanding and joint modeling ability of the model for various entity attributes. The overall architecture of the model is shown in Figure 2. MEAFE first uses the encoders of each modality, and the feature enhancement method proposed in this paper to obtain embedding of each modality and then performs weighted aggregation for multi-modal information fusion to generate the multi-modal joint representation of entities.

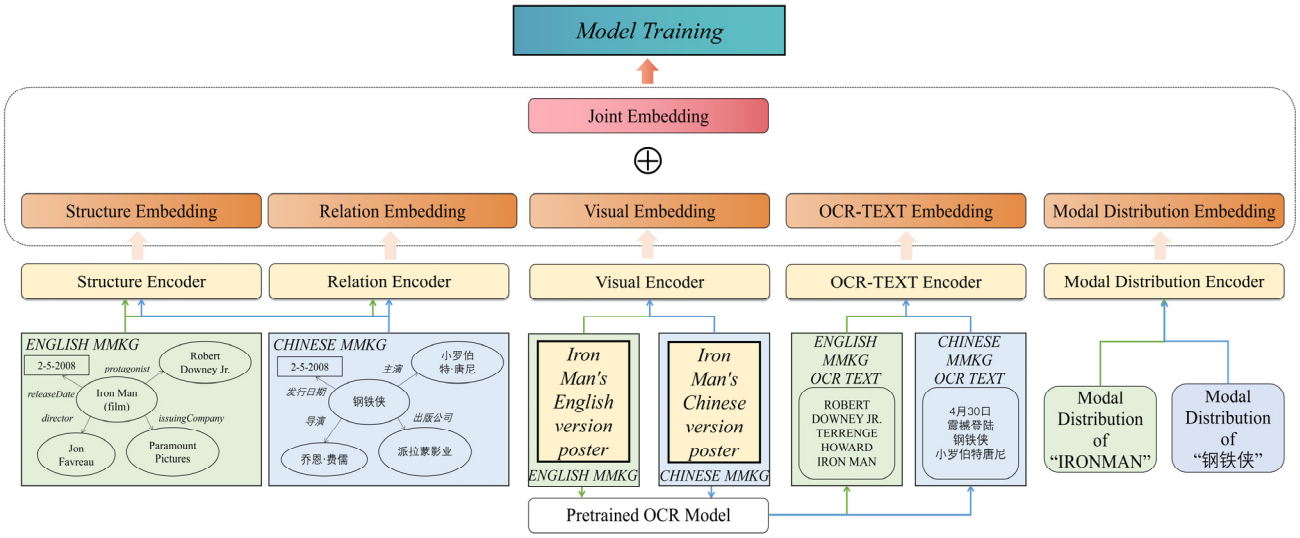

Figure 2. The model architecture of MEAFE. Taking use of MEAFE to align Chinese and English multi-modal knowledge graphs as an example.

2.1. Neighborhood Structure Embedding

Most entity alignment models use graph convolutional networks (GCNs) or graph attention networks (GATs) to structurally model the relationship triples of entities. The formula for using a vanilla GAT [14] to obtain entity structure embedding is shown as follows:

$$e_{ij} = LeakyReLU\left(\vec{a}^{T}\left[\boldsymbol{W}\vec{h}_i \middle|\middle| \boldsymbol{W}\vec{h}_j\right]\right) \tag{1}$$

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k\in N_i} exp(e_{ik})} \tag{2}$$

$$h_i^g = \sigma\left(\sum_{j\in N_i} \alpha_{ij} h_j^g\right) \tag{3}$$

where $\vec{h}_i$ is the original feature of $entity_i$; e_{ij} means similarity coefficient between entities; $\boldsymbol{W}$ is the shared weight matrix; $\vec{a}^{T}$ is the single-layer feed forward neural network; σ is the activation function; $||$ means matrix splicing; α_{ij} means the importance of $entity_j$ to $entity_i$; h_i^g is the hidden state of $entity_i$ by aggregating all its one-hop neighbors N_i; and *softmax* and *LeakyReLU* are corresponding nonlinear functions.

However, when applied to the entity alignment task, the vanilla graph attention network has two problems: the shared weight matrix and the static property.

Firstly, in the entity alignment task, there are often significant differences between the entity nodes in the knowledge graph and their adjacent nodes. For example, the target entity node and its attribute nodes usually have significant structural differences. Applying shared weight matrices to different types of nodes can make it difficult for the model to correctly distinguish between entity nodes and adjacent nodes, resulting in a reduction in the model's representational capacity. To address this problem, MEAFE uses two different weight matrices, $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$, to calculate different entity nodes, which enhances the discrimination ability of the GAT for nodes, and then the network can better extract features. The modified attention coefficient calculation formula is as follows:

$$e_{ij} = LeakyReLU\left(\vec{\boldsymbol{a}}^T\left[\boldsymbol{W_1}\vec{h}_i \middle|\middle| \boldsymbol{W_2}\vec{h}_j\right]\right) \tag{4}$$

$$\alpha_{ij} = softmax(LeakyReLU\left(\vec{\mathbf{a}}^T\left[\mathbf{W_1}\vec{h}_i \middle|\middle| \mathbf{W_2}\vec{h}_j\right]\right)) \tag{5}$$

Secondly, from the finiteness of the relationship between adjacent nodes and entity nodes and the monotonicity of *softmax* and *LeakyReLU*, it can be seen that there exists node j for any node i to maximize $\vec{\boldsymbol{a}}^T\left[\boldsymbol{W}\vec{h}_j\right]$, so that the vanilla graph attention network always tends to give node j the largest attention coefficient while ignoring the different relationships between different input nodes; that is, given a group of nodes and a pre-trained GAT layer, the attention function α has the same maximum tendency node j, so the vanilla GAT is more suitable for mapping all inputs to the constant mapping of the same output. However, it is difficult to build a better model when different query inputs have different correlations with different nodes in the entity alignment task. To put it further, as shown in Formula (3), vanilla graph attention network continuously uses $\boldsymbol{W}$ and $\vec{\boldsymbol{a}}^T$ to perform a linear transformation on vectors, then continuously uses *softmax* and *LeakyReLU* to perform the nonlinear transformation; both can only use once transform override, which, in turn, contributes to the static of vanilla graph attention network. To address this problem, MEAFE attempts to apply dynamic graph attention networks [12] in the multi-modal entity alignment task. Figure 3 shows the attention tendencies of static and dynamic graph attention networks. The main reason for the limited attention of the static graph attention network is that it simply uses the learnable matrices $\vec{\boldsymbol{a}}^T$ and $\boldsymbol{W}$ continuously, so that it can degenerate into a single linear layer. Therefore, the dynamic graph attention network attempts to apply the nonlinear function *LeakyReLU* first and then input the feedforward neural network $\vec{\boldsymbol{a}}^T$ in the calculation of the attention mechanism. The expression is modified as follows:

$$e_{ij} = \vec{\boldsymbol{a}}^T LeakyReLU\left(\left[\boldsymbol{W_1}\vec{h}_i \middle|\middle| \boldsymbol{W_2}\vec{h}_j\right]\right) \tag{6}$$

In the dynamic graph attention network, each query has a different order for the attention coefficients of the keys, so it has a stronger representation ability. Moreover, the dynamic graph attention network avoids the simple continuous use of the learnable matrix $\vec{\boldsymbol{a}}^T$ and $\boldsymbol{W}$, thus preventing the feedforward neural network and the weight matrix from degenerating and decomposing into a single linear layer, realizing a general approximate attention function, and improving the robustness of the model while obtaining stronger representation ability.

Therefore, the dynamic graph attention network can significantly optimize the acquisition of the weight between nodes for the entity alignment tasks in the case of more complex relationships between nodes and different requirements for the ranking of nodes in different neighborhoods, thus improving the entity alignment effect.

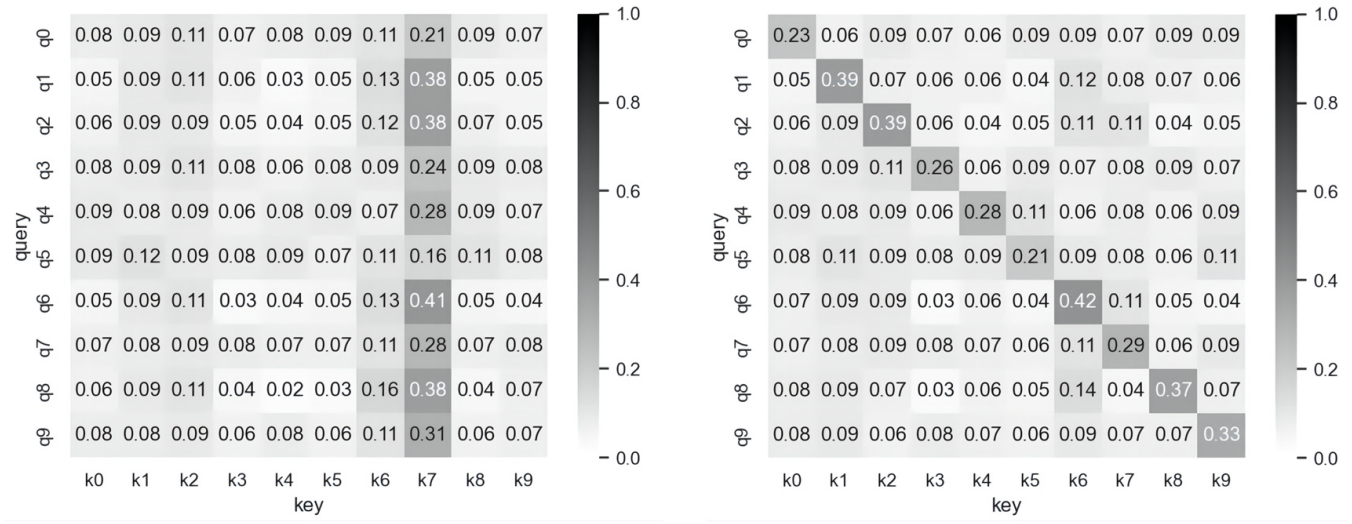

Figure 3. Comparison of attention tendency of static and dynamic graph attention network. (**a**) Schematic of static graph attention network. (**b**) Schematic of dynamic graph attention network.

2.2. Relation, Attribute, and Name Embeddings

To better utilize the triples of the given aligned entity pairs for model training, we extract the information of entity relationships, attributes, and entity names as external knowledge to assist in model training.

We follow the modeling approach of the MCLEA model for the relation, attribute, and name information of entities, treating the three types of entity information as bag-of-words features. Additionally, inputting them into three separate feedforward neural networks for training results in relationships, attributes, and name embedding. These embeddings are calculated as [15]:

$$h_i^l = \boldsymbol{W}^l {u_i}^l + \boldsymbol{b}^l, l \in \{r, a, n\} \tag{7}$$

where $h_i^l, l \in \{r, a, n\}$ is the relation, attribute, and name embeddings of $entity_i$; $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$ are the learnable weights of the bias matrix; ${u_i}^r$ is the bag-of-words relation feature; ${u_i}^a$ is the bag-of-words attribute feature; and ${u_i}^n$ is the name feature obtained by averaging the pre-trained GloVe [16] vectors of name strings.

2.3. Visual Embeddings

MEAFE uses a pre-trained visual model (PVM), e.g., ResNet-152 [17], as a visual encoder to encode the described image of the entities. Additionally, it then uses the final layer output of it as the image feature. After that, the image feature is inputted into a feedforward layer to achieve the original visual embedding:

$$h_i^{v_e} = \boldsymbol{W}^v PVM(Img_i) + \boldsymbol{b}^v \tag{8}$$

where $h_i^{v_e}$ is the original visual embedding of $entity_i$; $\boldsymbol{W}^v$ and $\boldsymbol{b}^v$ are the learnable weights of the bias matrix of corresponding feedforward neural network; PVM means pre-trained visual model; Img_i means the visual image of $entity_i$.

To further explore useful information in entity images, MEAFE additionally uses a multi-modal pre-training model to perform semantic matching on the description images of entities and remove the description images with poor semantic connections. Then, an OCR model is used to extract possible text information from the images as auxiliary knowledge for the visual encoding obtained only by using pre-trained visual models.

We use the multi-modal pre-training model CLIP [18] to encode entity names and images separately and calculate their similarities. For entity images, we directly use CLIP

for image encoding, while for entity names, we modify them to "A photo of entity name" for text encoding, and then delete images with a similarity below the set threshold.

For the cleaned entity image set, we use the pre-trained PaddleOCR (https://github.com/PaddlePaddle/PaddleOCR, accessed on 28 December 2022) model to extract possible text information and retain the detected text with a confidence level higher than γ_1. It can output the text that may exist in the detected image as a list. We then use the pre-trained multi-language BERT [19] to encode the text, taking the pooler output of its last layer as the OCR feature. After obtaining the OCR encoding, we input it into a feedforward neural network for learning to obtain the final OCR embedding:

$$h_i^{v_o} = \boldsymbol{W}^o BERT\left(OCR\left(Img_i\right)\right) + \boldsymbol{b}^o \tag{9}$$

where $h_i^{v_o}$ is the OCR embedding of $entity_i$; $\boldsymbol{W}^o$ and $\boldsymbol{b}^o$ are the learnable weights of the bias matrix of corresponding feedforward neural network; *OCR* means pre-trained PaddleOCR model; *BERT* is text encoder we used; Img_i means the visual image of $entity_i$.

2.4. Modal Distribution Embeddings

To better understand the multi-modal information of entities and conduct more reasonable multi-modal joint modeling, we additionally embed the types of modal information and some countable attribute information of each entity by one-hot coding as f_m, such as the number of triples and associated relationships, and then feed it into feedforward network. The specific modeling method is as follows:

$$h_i^m = \boldsymbol{W}^m f_m + \boldsymbol{b}^m \tag{10}$$

where h_i^m is the modal distribution embedding of $entity_i$; $\boldsymbol{W}^m$ and $\boldsymbol{b}^m$ are the learnable weights of the bias matrix of corresponding feedforward neural network.

2.5. Joint Embeddings

MEAFE then implements a weighted aggregation to integrate the multi-modal features into a multi-modal joint embedding $\boldsymbol{h_i}$:

$$\boldsymbol{h}_i = \oplus_{m \in M} \text{softmax}(w_M) {h_i}^M \tag{11}$$

where $M = \{g, r, a, n, v_e, v_o, m\}$ is the set of modal type of entity; w_M is a trainable attention weight for the modality of M; $\oplus$ means matrix concatenate.

3. Experiment and Discussion

To verify the entity alignment effect of MEAFE and the effectiveness of the improved method proposed herein, we designed and performed experiments based on five different multi-modal entity alignment datasets, and then analyzed and discussed the experimental results.

3.1. Datasets

We adopted five multi-modal entity alignment datasets for training and evaluation, including three bilingual datasets DBP15K (ZH/JA/FR-EN) (https://github.com/nju-websoft/JAPE, accessed on 5 May 2022) [8] and two cross kg datasets FB15K-DB15K/YAGO15K (https://github.com/mniepert/mmkb, accessed on 4 March 2023) [5]. The specific statistics of the datasets are shown in Tables 2 and 3. As for DBP15K, 30% aligned entity pairs are given as the training set, while for cross kg datasets, 20%, 50%, and 80% aligned entity pairs are given [5].

Table 2. The statistics of DBP15K.

Datasets	Language	Entities	Relations	Triplets	Image	Ref.
$DBP15K_{ZH-EN}$	Chinese English	19,388 19,572	9812 8496	318,449 438,360	15,912 14,125	15,000
$DBP15K_{JA-EN}$	Japanese English	19,814 19,780	7181 7219	326,205 414,100	12,739 13,741	15,000
$DBP15K_{FR-EN}$	French English	19,661 19,993	5450 7630	379,823 466,816	14,174 13,858	15,000

Table 3. The statistics of FB15K-DB15K/YAGO15K.

Datasets	KG	Entities	Relations	Triplets	Image	Ref.
FB15K-DB15K	FB15K DB15K	14,951 12,842	1461 504	621,608 137,277	13,444 12,837	12,846
FB15K-YAGO15K	FB15K YAGO15K	14,951 15,404	1461 39	621,608 146,418	13,444 11,194	11,199

3.2. Implementation Details

The hidden size of each layer of GATv2 (the embedding size of h_i^g) was 300, while the embedding size of the other modalities was 100. We used the AdamW [20] optimizer, and set the learning rate to 5×10^{-4}. The number of training epochs was 1000 with early stopping. The batch size was 512. The hyperparameters of γ_1 were set to 80%.

For visual embedding, we used pre-processed visual features as the initial entity image features [7,8]. The visual features of DBP15K were pre-trained by a model that uses ResNet-152 as the backbone, and the visual features of FB15K-DB15K/YAGO15K were pre-trained by a model that uses VGG-16 [21] as the backbone. We used ViT-B/32 as the pre-trained model of CLIP for semantic matching and image set cleaning; we used ch_PP-OCRv3_rec_infer (https://paddleocr.bj.bcebos.com/PP-OCRv3/chinese/ch_PP-OCRv3_rec_infer.tar, accessed on 9 December 2022) for OCR; and we used bert-base-multilingual-cased (https://huggingface.co/bert-base-multilingual-cased, accessed on 20 November 2022) for OCR text encoding. Not all entities of datasets have descriptive images or can detect text information from the image, and for entities without visual information or OCR information, we treated random vectors as $h_i^{v_e}$ and $h_i^{v_o}$.

When performing model training and validation with DBP15K, we only modified the graph attention network of the MEAFE due to the absence of an original entity image. However, when using FB15K-DB15K/YAGO15K, due to the large number of entities, the attention coefficient calculation of the dynamic graph puts forward high requirements on the hardware, so we only increased $h_i^{v_o}$ and h_i^m. For pre-aligned entities, we treated the FB15K entity images from different sources as the original images from FB15K versus DB15K/YAGO15K, respectively.

3.3. Optimization Objective

We adopted the Intra-modal Contrastive Loss (ICL) and Inter-modal Alignment Loss (IAL) methods proposed by the MCLEA to train [10] and enable the model to fully capture the dynamics within and between modalities while maintaining semantic proximity and minimizing modal differences [22]. The ICL and IAL are formulated as follows:

$$\delta_m(u, v) = \exp\left(f_m(u)^T f_m(v)/\tau\right) \tag{12}$$

$$q_m\left(e_1^i, e_2^i\right) = \frac{\delta_m\left(e_1^i, e_2^i\right)}{\delta_m\left(e_1^i, e_2^i\right) + \sum_{e_1^j \in N_1^i} \delta_m\left(e_1^i, e_1^j\right) + \sum_{e_2^j \in N_2^i} \delta_m\left(e_1^i, e_2^j\right)} \tag{13}$$

$$\mathcal{L}_m^{\mathrm{ICL}} = -\mathbb{E}_{i\in\mathcal{B}} \log\left[\frac{1}{2}\left(q_m\left(e_1^i, e_2^i\right) + q_m\left(e_2^i, e_1^i\right)\right)\right] \tag{14}$$

$$\mathcal{L}_m^{\mathrm{IAL}} = \mathbb{E}_{i\in\mathcal{B}} \quad \frac{1}{2}\left[\mathrm{KL}\left(q_o^{'}(e_1^i, e_2^i) \parallel q_m^{'}(e_1^i, e_2^i)\right) + \mathrm{KL}\left(q_o^{'}(e_2^i, e_1^i) \parallel q_m^{'}(e_2^i, e_1^i)\right)\right] \tag{15}$$

where $\mathcal{L}_m^{\mathrm{IAL}}$ and $\mathcal{L}_m^{\mathrm{IAL}}$ are the mean Intra-modal Contrastive Loss (ICL) and Inter-modal Alignment Loss (IAL); $q_m\left(e_1^i, e_2^i\right)$ is the probability distribution of the modality of m for positive pair $\left(e_1^i, e_2^i\right)$.

$\delta_m\left(e_1^i, e_2^i\right)$ is the correlation probability between entities; it is calculated as shown in Equation (12), where $f_m(.)$ is the encoder of the modality m, and τ is the hyperparameter. Additionally, $q_o^{'}\left(e_1^i, e_2^i\right)$ and $q_o^{'}\left(e_2^i, e_1^i\right)$ and $q_m^{'}\left(e_1^i, e_2^i\right)$ and $q_m^{'}\left(e_2^i, e_1^i\right)$ represent the output predictions with two directions of joint embedding and the uni-modal embedding of modality m. KL refers to the KL divergence. When calculating the ICL and IAL, τ is set to 0.1 and 4.0, respectively.

3.4. Evaluation Index

We used *Hit@n*, the Mean Reciprocal Rank (*MRR*), and the Mean Rank (*MR*) to objectively evaluate the entity alignment accuracy of the model. A larger *Hit@n* and *MRR* and a smaller *MR* indicate the better performance of the model.

Hit@n represents the probability that the top n items of the candidate entity alignment possibility rank have correct results; *MRR* represents the average of the reciprocal of correct ranking in the candidate alignment; and *MR* represents the average correct ranking in the candidate alignment. The calculation formulas are as follows:

$$Hits@n = \frac{1}{|S|}\sum_{i}^{|S|}||(rank_i \leq n) \tag{16}$$

$$MRR = \frac{1}{|S|}\sum_{i}^{|S|}\frac{1}{rank_i} \tag{17}$$

$$MR = \frac{1}{|S|}\sum_{i}^{|S|} rank_i \tag{18}$$

where S is the total triplet set; $rank_i$ is the entity alignment prediction ranking of the ith triplet; and $|*|$ is the indicator function.

3.5. Configuration

This paper conducts experimental research based on the TensorFlow 2.0 deep learning framework; the compilation environment is Python 3.7.11; and the operating system is Ubuntu 18.04. The experimental hardware is configured with an Intel (R) (Santa Clara, CA, USA) Xeon (R) gold 6132 2.60 GHz CPU, 256 GB of memory, and an NVIDIA (Santa Clara, CA, USA) Geforce 3090 24 GB GPU.

3.6. Results and Analysis

Tables 4 and 5 report the performance of MEAFE on bilingual datasets DBP15K (ZH/JA/FR-EN) and cross kg datasets FB15K-DB15K/YAGO15K. MEAFE performs the best across all the datasets against other baselines.

Table 4. Comparative results of MEAFE against other baseline methods on three bilingual datasets. The best results are marked in bold and the second best results are underlined.

Model	$DBP15K_{ZH-EN}$			$DBP15K_{JA-EN}$			$DBP15K_{FR-EN}$		
	H@1	*H@10*	*MRR*	*H@1*	*H@10*	*MRR*	*H@1*	*H@10*	*MRR*
MultiKE [23]	0.437	0.516	0.466	0.570	0.643	0.596	0.714	0.761	0.733
HMAN [17]	0.562	0.851	-	0.567	0.969	-	0.540	0.871	-
RDGCN [24]	0.708	0.846	-	0.767	0.895	-	0.886	0.957	-
AttrGNN [25]	0.777	0.920	0.829	0.763	0.909	0.816	0.942	0.987	0.959
BERT-INT [26]	0.968	0.990	0.977	0.964	0.991	0.975	0.992	0.998	0.995
ERMC [27]	0.903	0.946	0.899	0.942	0.944	0.925	0.962	0.982	0.973
MCLEA [11]	0.972	0.996	0.981	0.986	**0.999**	0.991	**0.997**	**1.00**	**0.998**
MEAFE (Ours)	**0.973**	**0.997**	**0.982**	**0.987**	**0.999**	**0.992**	**0.997**	**1.00**	**0.998**

Table 5. Comparative results of MEAFE against other baseline methods on two cross kg datasets. The best results are marked in bold and the second best results are underlined.

Training Set Ratio	Model	FB15K-DB15K			FB15K-YAGO15K		
		H@1	*H@10*	*MRR*	*H@1*	*H@10*	*MRR*
20%	PoE [5]	0.126	0.251	0.170	0.113	0.229	0.154
	HMEA [9]	0.127	0.369	-	0.105	0.313	-
	MMEA [7]	0.265	0.541	0.357	0.234	0.480	0.317
	EVA [8]	0.134	0.338	0.201	0.098	0.276	0.158
	MCLEA [11]	0.445	0.705	0.534	0.388	0.641	0.474
	MEAFE (Ours)	**0.617**	**0.817**	**0.686**	**0.567**	**0.745**	**0.628**
	Improv. best	0.172	0.112	0.152	0.179	0.104	0.154
50%	PoE	0.464	0.658	0.533	0.347	0.536	0.414
	HMEA	0.262	0.581	-	0.265	0.581	-
	MMEA	0.417	0.703	0.512	0.403	0.645	0.486
	EVA	0.223	0.471	0.307	0.240	0.477	0.321
	MCLEA	0.573	0.800	0.652	0.543	0.759	0.616
	MEAFE (Ours)	**0.712**	**0.880**	**0.771**	**0.678**	**0.837**	**0.734**
	Improv. best	0.139	0.080	0.119	0.135	0.078	0.118
80%	PoE	0.666	0.820	0.721	0.573	0.746	0.635
	HMEA	0.417	0.786	-	0.433	0.801	-
	MMEA	0.590	0.869	0.685	0.598	0.839	0.682
	EVA	0.370	0.585	0.444	0.394	0.613	0.471
	MCLEA	0.730	0.883	0.784	0.653	0.835	0.715
	MEAFE (Ours)	**0.804**	**0.934**	**0.851**	**0.740**	**0.884**	**0.791**
	Improv. best	0.074	0.051	0.067	0.087	0.049	0.076

Table 4 reports the performance of MEAFE against the supervised baselines on DBP15K. Compared to MCLEA, MEAFE achieved parity or improvement in all indicators on the DBP15K. Although the improvement is not significant due to the high completion of the model for this dataset, it still proves the adaptability of dynamic graph attention networks for entity alignment tasks, achieving improvements even when only modifying the single-layer network without adding external information.

Table 5 reports the performance of MEAFE against the other baselines on the cross kg datasets FB15K-DB15K/YAGO15K. When the training set ratio was set to 20%, MEAFE compared to the optimal baseline (MCLEA), H@1 increased by 0.172, H@10 improved by 0.112, and MRR by 0.152 on FB15K-DB15K, while H@1 increased by 0.179, H@10 improved by 0.104, and MRR by 0.154 on FB15K-YAGO15K. When the training set ratio was set to 50% and 80%, all indicators also achieved a certain improvement. Additionally, as the proportion of the training set decreased, the extent of improvement in the entity alignment

effect increased. The experimental results demonstrate the effectiveness of extracting textual information that may be included in the visual modal information and adding additional embedding of entity modal distribution for multi-modal entity alignment.

We analyzed the fusion weights of various model information during multi-modal fusion, and the results are shown in Figure 4. On two kg datasets, MEAFE tends to give higher weights to image embedding and structural embedding, which is consistent with our preliminary assumption. Our newly added OCR embedding has similar attention weights to attribute, relation, and name embedding, proving its effectiveness. The attention weight of the modal distribution embedding is lower, possibly due to its lower original embedding dimension.

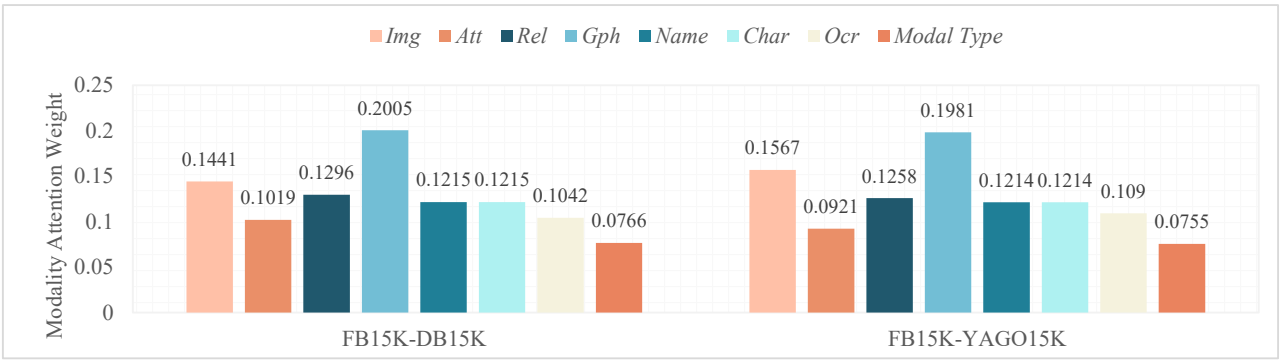

Figure 4. Modality attention weight on two cross kg datasets.

3.7. Ablation Study

To verify the effectiveness of our proposed correction schemes, we designed additional ablation experiments on FB15K-DB15K/YAGO15K, and the experimental results are shown in Table 6. Where MEAFE (none), MEAFE (only MD), and MEAFE (only OCR), respectively, refer to MEAFE models that do not add any information, only add modal distribution information, and only add OCR information.

Table 6. Ablation experiment result. The best results are marked in bold and the second best results are underlined.

Training Set Ratio	Model	FB15K-DB15K			FB15K-YAGO15K		
		H@1	H@10	MRR	H@1	H@10	MRR
20%	MEAFE (none)	0.456	0.720	0.547	0.430	0.667	0.511
	MEAFE (only MD)	0.464	0.731	0.543	0.397	0.640	0.480
	MEAFE (only OCR)	0.592	0.802	0.665	0.566	**0.752**	**0.630**
	MEAFE	**0.617**	**0.817**	**0.686**	**0.567**	0.745	0.628
50%	MEAFE (none)	0.603	0.825	0.680	0.578	0.778	0.650
	MEAFE (only MD)	0.612	0.827	0.689	0.572	0.772	0.642
	MEAFE (only OCR)	0.695	0.869	0.713	0.648	0.812	0.705
	MEAFE	**0.712**	**0.880**	**0.771**	**0.678**	**0.837**	**0.734**
80%	MEAFE (none)	0.730	0.893	0.788	0.659	0.845	0.724
	MEAFE (only MD)	0.724	0.902	0.787	0.669	0.844	0.730
	MEAFE (only OCR)	0.788	0.919	0.835	0.734	**0.888**	0.788
	MEAFE	**0.804**	**0.934**	**0.851**	**0.740**	0.884	**0.791**

Compared with models without any additional optimization, adding only OCR information or only MD information encouraged a certain improvement in the entity alignment effect of the model, and both were weaker than models that added both information simultaneously. Due to the richness of OCR information, the improvement is significant, while MEAFE with only MD information was lower in some indicators than MEAFE without

any modifications. The possible reason for this situation is that the original features of the entity modal distribution information have a lower dimension and contain less information, making it difficult to significantly affect the entity alignment effect of the local model alone. However, when used as an additional supplement to OCR information, it still improved the model's understanding and modeling ability for multi-modal information to a certain extent.

4. Conclusions

Aiming at the problem that most existing multi-modal entity alignment models do not effectively utilize the multi-modal information of aligned entity pairs, resulting in poor alignment performance, this paper proposes a multi-modal entity alignment method based on feature enhancement, called MEAFE. Its core idea is to maximize the use of entity visual, textual, and relational modalities to enhance the corresponding feature embedding and improve the knowledge representation ability of the model. MEAFE adopts multi-modal pre-trained models, OCR models, and GATv2 networks to enhance the information extraction ability of entity structural triplets and image descriptions, respectively, to obtain more effective multi-modal representations, and analyzing the modal distribution of entities to enhance the modeling ability and understanding of entity information.

MEAFE can more accurately align entities that refer to the same real object within the multi-modal knowledge graph from different sources, thereby removing redundant entities within the graph and integrating non overlapping attributes of aligned entity pairs. So as to further improve the knowledge richness of the graph, then improving the effectiveness of downstream tasks based on knowledge graphs.

At present, when MEAFE processes the numerical attributes and relationship information of entities (<headEntity,attributeKey,attributeValue>), it only uses bag-of-words to encode the key of numerical triplets and relationship triplets without utilizing their specific value. In our future work, we will strive to improve the model's ability to analyze and process numerical attribute information of entities to better utilize the values of a large number of numerical relationship triplets as supervisory information for model training.

Author Contributions: Conceptualization, H.W. and R.H.; methodology, H.W.; software, H.W.; validation, H.W., Q.L., R.H. and J.Z.; formal analysis, H.W. and Q.L.; investigation, H.W.; resources, H.W.; data curation, H.W.; writing—original draft preparation, H.W.; writing—review and editing, H.W., Q.L. and R.H.; visualization, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant 62002384.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/Ccccitrus/MEAFE (accessed on 30 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.S.; et al. DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
2. Mahdisoltani, F.; Biega, J.; Suchanek, F. YAGO3: A Knowledge Base from Multilingual Wikipedias. In Proceedings of the CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 4–7 January 2015.
3. Chen, Y.C.; Li, L.; Yu, L.; Kholy, A.E.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: Learning UNiversal Image-TExt Representations. *arXiv* **2019**. [CrossRef]
4. Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; Sacheti, A. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv* **2020**. [CrossRef]
5. Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; Rosenblum, D.S. MMKG: Multi-modal Knowledge Graphs. In *The Semantic Web*; Springer International Publishing: Cham, Switzerland, 2019; pp. 459–474.

6. Wang, M.; Wang, H.; Qi, G.; Zheng, Q. Richpedia: A Large-Scale, Comprehensive Multi-Modal Knowledge Graph. *Big Data Res.* **2020**, *22*, 100159. [CrossRef]
7. Chen, L.; Li, Z.; Wang, Y.; Xu, T.; Wang, Z.; Chen, E. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In Proceedings of the KSEM 2020, LNAI 12274, Hangzhou, China, 28–30 August 2020; pp. 134–147.
8. Liu, F.; Chen, M.; Roth, D.; Collier, N. Visual Pivoting for (Unsupervised) Entity Alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 4257–4266.
9. Guo, H.; Tang, J.; Zeng, W.; Zhao, X.; Liu, L. Multi-modal entity alignment in hyperbolic space. *Neurocomputing* **2021**, *461*, 598–607. [CrossRef]
10. Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N.J.; Chen, E. Multi-modal Siamese Network for Entity Alignment. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 118–126.
11. Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; Zheng, Y. Multi-modal Contrastive Representation Learning for Entity Alignment. *arXiv* **2022**, arXiv:2209.00891.
12. Brody, S.; Alon, U.; Yahav, E. How Attentive are Graph Attention Networks? In Proceedings of the Tenth International Conference on Learning Representations, {ICLR} 2022, Virtual Event, 25–29 April 2022.
13. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3837–3845.
14. Veličković, P.; Cucurull, G.; Casanova, A.; Romeo, A.; Liò, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
15. Yang, H.-W.; Zou, Y.; Shi, P.; Lu, W.; Lin, J.; Sun, X. Aligning Cross-Lingual Entities with Multi-aspect Information. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4431–4441.
16. Pennington, J.; Socher, R.; Manning, C. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 770–778.
18. 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 Proceedings of the ICML 2021, Virtual, 18–24 July 2021; pp. 8748–8763.
19. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
20. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, {ICLR} 2019, New Orleans, LA, USA, 6–9 May 2019.
21. Simonyan, K.; Zisserman, A. Very deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
22. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G.E. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual, 13–18 July 2020; Volume 119, pp. 1597–1607.
23. Zhang, Q.; Sun, Z.; Hu, W.; Chen, M.; Guo, L.; Qu, Y. Multi-view Knowledge Graph Embedding for Entity Alignment. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 5429–5435. Available online: https://www.ijcai19.org/ (accessed on 30 January 2023).
24. Wu, Y.; Liu, X.; Feng, Y.; Wang, Z.; Yan, R.; Zhao, D. Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 5278–5284. Available online: https://www.ijcai.org/proceedings/2019/733 (accessed on 5 May 2022).
25. Liu, Z.; Cao, Y.; Pan, L.; Li, J.; Liu, Z.; Chua, T.-S. Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Virtual, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 6355–6364.
26. Tang, X.; Zhang, J.; Chen, B.; Yang, Y.; Chen, H.; Li, C. BERT-INT: A BERT-based interaction model for knowledge graph alignment. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 3174–3180.
27. Yang, J.; Wang, D.; Zhou, W.; Qian, W.; Wang, X.; Han, J.; Hu, S. Entity and Relation Matching Consensus for Entity Alignment. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Online, 1–5 November 2021; pp. 2331–2341.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.