applied sciences

MDPI

Article

MDSEA: Knowledge Graph Entity Alignment Based on Multimodal Data Supervision

Jianyong Fang [1,2] and Xuefeng Yan [1,*]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China; jy_fang@189.cn
[2] Jiangsu Automation Research Institute, Lianyungang 222000, China
* Correspondence: yxf@nuaa.edu.cn

Abstract: With the development of social media, the internet, and sensing technologies, multimodal data are becoming increasingly common. Integrating these data into knowledge graphs can help models to better understand and utilize these rich sources of information. The basic idea of the existing methods for entity alignment in knowledge graphs is to extract different data features, such as structure, text, attributes, images, etc., and then fuse these different modal features. The entity similarity in different knowledge graphs is calculated based on the fused features. However, the structures, attribute information, image information, text descriptions, etc., of different knowledge graphs often have significant differences. Directly integrating different modal information can easily introduce noise, thus affecting the effectiveness of the entity alignment. To address the above issues, this paper proposes a knowledge graph entity alignment method based on multimodal data supervision. First, Transformer is used to obtain encoded representations of knowledge graph entities. Then, a multimodal supervised method is used for learning the entity representations in the knowledge graph so that the vector representations of the entities contain rich multimodal semantic information, thereby enhancing the generalization ability of the learned entity representations. Finally, the information from different modalities is mapped to a shared low-dimensional subspace, making similar entities closer in the subspace, thus optimizing the entity alignment effect. The experiments on the DBP15K dataset compared with methods such as MTransE, JAPE, EVA, DNCN, etc., all achieve optimal results.

Keywords: multimodal data supervision; entity alignment; knowledge graph; transformer

Citation: Fang, J.; Yan, X. MDSEA: Knowledge Graph Entity Alignment Based on Multimodal Data Supervision. *Appl. Sci.* **2024**, *14*, 3648. https://doi.org/10.3390/app14093648

Academic Editor: Tobias Meisen

Received: 28 March 2024
Revised: 19 April 2024
Accepted: 22 April 2024
Published: 25 April 2024

1. Introduction

With the development of cross-disciplinary research between knowledge engineering and multimodal learning, multimodal knowledge graphs (KG) [1] have become increasingly crucial as a means to assist computers in understanding the entity background knowledge in many artificial intelligence applications, such as question answering systems [2], recommendation systems [3], natural language understanding [4], and scene graph generation [5]. In recent years, many researchers have constructed numerous multimodal knowledge graphs targeting different domains and languages. Some of the widely used ones include DBpedia, YAGO, and Freebase, which store vast amounts of knowledge and can support various downstream applications. However, most real-world KGs are highly incomplete, primarily because they are often constructed from single data sources. To facilitate knowledge fusion, the task of knowledge graph entity alignment (EA) has received increasing attention from researchers [6]. EA aims to identify equivalent entities across KGs while addressing challenges such as multiple languages, heterogeneous graph structures, and different naming conventions.

Early EA was mostly heuristic, and entity mapping was constructed using techniques such as logical reasoning and lexical matching. The recent EA methods are often based

on embeddings, learning an embedding space to represent the KG to be aligned so that similar entities are located closer while dissimilar entities are far apart, thereby mitigating heterogeneity issues [7]. Specifically, the existing methods can be classified into two categories: (1) translation-based EA methods, employing methods like TransE [8] based on the translation of KG embeddings to capture entity structural information from relation triplets; and (2) graph neural network (GNN)-based EA methods, primarily utilizing methods like graph convolution network (GCN) [9] and GAT [10] for aggregating the neighborhood entity features. In addition to the above-mentioned methods, the effectiveness of EA can be enhanced through various strategies, such as parameter sharing [11] (sharing entity embeddings across KGs, explicitly linking the seed sequences across multiple heterogeneous KGs), iterative learning (IL) [12] (iteratively proposing more alignment seeds from unaligned entities), attribute value encoding [13], collectively stable matching of interdependent alignment decisions [14], or guiding EA through ontology patterns [15].

Translation-based methods primarily learn embeddings based on the translation assumption within each KG. For instance, TransE regards a relation r as a translation from the head entity h to the tail entity t and confirms that a correct knowledge triplet should satisfy the $|h + r - t|$ function, specifically extracting a vector from both the entity matrix and the relationship matrix, performing L_1 or L_2 operations, and obtaining a result that approximates the vector of another entity in the entity matrix, thereby achieving the representation of the relationship between the existing triplets in the knowledge graph through word vectors. Methods like MTransE [16] and ITransE [17] introduce linear transformations to improve the EA performance of KG with multiple mapping relationships at the cost of increasing the model complexity. JAPE [18] and RSNs [19] use parameter sharing to maintain the same embedding between pre-aligned entities. Additionally, Transedge [20] and BootEA [21] integrate entity embeddings into relational embeddings, which together serve as relational representations to solve "one-to-many" and "many-to-one" problems. However, due to the fact that embedding based on triples is constrained on a single triplet, it is difficult to capture global graph structure information, which makes it difficult to achieve overall consistency during the alignment process.

To address the aforementioned issues, embeddings based on GNN are utilized to achieve local subgraph-level consistency. The first endeavor in this direction is GCN-align [22], which utilizes the entity relationships in each KG to construct the network structure of GCN. This method embeds multiple languages into a unified vector space and discovers entity alignment based on the distance between the entities in the embedding space. However, GCN-align is mainly aimed at aligning isomorphic graphs, but its processing ability for heterogeneous graphs is weak, resulting in the loss of heterogeneous edge information. Therefore, in recent years, many studies have attempted to integrate edge information into GCN to enhance the relationship perception ability of the model. MuGNN [23] and NAEA [24] introduce attention mechanisms to learn different weights for different types of relationships. HMAN [25] uses GCNs to combine multiple aspects of entity information, including topological connections, relationships, and attributes, to learn entity embeddings. RDGCN [26] merges relational information via the attentional interaction between the original graph and the dual relationship graph, and further captures adjacent structures to learn better entity representations. MRAEA [27] models cross-linguistic entity embeddings directly by focusing on the metasemantics of the incoming and outgoing neighbors and their connection relationships regarding the nodes. PSR [28] proposes a simplified graph encoder with relation graph sampling, which achieves high performance, scalability, and robustness through symmetric non-negative alignment loss and incremental semi-supervised learning. All these efforts demonstrate the importance of relation information in entity alignment. However, these methods do not consider the role of edge alignment in EA and only consider the integrated semantic information of the relationships regarding entity embeddings. Additionally, some recent works incorporate extra external information as weak supervision signals. For instance, EVA [29] proposes a structure-aware uncertainty sampling strategy that can measure the

uncertainty of each entity in KG and its impact on the adjacent entities. JEANS [30] jointly represents multilingual KG and text corpora in a shared embedding scheme and seeks to improve the alignment of entities and text with accompanying supervisory signals. Furthermore, CG-MuAlign [31] employs a designed attention mechanism to facilitate the collaborative alignment of positive information from the entity neighborhood and comparably effective negative messages, achieving a joint alignment of multiple types for the entity. ActiveEA [32] designs an active learning framework to create seed comparisons with large amounts of information in order to obtain more effective EA models at lower annotation costs.

With more and more research beginning to explore how to combine visual content in the internet with EA, a new trend is to associate images with entity names to enrich the information of entity pairs. At present, the research mainly focuses on designing fusion methods suitable for cross-modal data to achieve cross-modal EA. Chen et al. [33] generated entity representations of relational knowledge, visual knowledge, and digital knowledge and integrated them through a multimodal knowledge fusion module. Chen et al. [34] employed a modal enhancement mechanism to integrate visual features to guide relational feature learning, and adaptively assigned attention weights to capture valuable attributes for alignment. Lin et al. [35] learned multiple individual representations from multiple modalities and then performed contrastive learning to jointly model the interactions within and between modalities. However, these methods learn multimodal fusion weights at the knowledge graph level, ignoring the intra-modal differences for each entity (such as node degree or relationship quantity) and inter-modal preferences (such as modality absence or ambiguity). This is crucial in real-world EA scenarios since knowledge graphs (especially MMKG) discovered from the internet or professional domains inevitably contain errors and noise, such as those with unrecognized images. Additionally, intra-modal feature differences and inter-modal phenomena such as modality absence, imbalance, or ambiguity are common in KGs. These shortcomings affect their robustness to some extent.

Through research, it was found that the current entity alignment methods have the following three issues:

(1) The existing entity alignment methods focus more on the entity alignment of traditional textual knowledge graphs. Some research embeds knowledge graphs from different sources into a low-dimensional space and achieves entity alignment by calculating the similarity between entities, yielding good results. However, these methods only utilize single-modal data (text) and ignore other modal data (images), thus failing to fully exploit the entity feature information in other modal data.

(2) Traditional cross-modal entity alignment methods often require extensive manual data annotation or carefully designed alignment features. For example, Zhang [36] proposed an adaptive co-attention network that selected Twitter as the data source, crawled and annotated a dataset containing images, and controlled the preference level of each word for the images and text using gate and filter mechanisms. While these traditional entity alignment methods can achieve high alignment effectiveness, they require a considerable amount of manual annotation, resulting in time wastage and increased labor costs. Moreover, the entity features designed by such methods often lack scalability and universality.

(3) Multimodal pre-trained language models achieve cross-modal entity alignment by pre-training on a large amount of unlabeled data. However, this method mostly focuses on global image and text features and is designed only for English text–image pairs. Models like CLIP pre-trained language models do not model the fine-grained relationships between text and images, which are valuable in domain-specific multimodal knowledge graph cross-modal EA tasks. Additionally, image–text pairs often contain noise in practice.

Based on these issues, this paper proposes a knowledge graph entity alignment method based on multimodal data supervised (MDSEA). It first uses Transformer to obtain knowledge graph entity encoding representations. Then, it employs a multimodal

supervised method for knowledge graph entity representation learning, ensuring that the vector representation of the entities contains rich multimodal semantic information, enhancing the generalization ability of the learned entity representation. Finally, it maps information from different modalities to a shared low-dimensional subspace, making similar entities closer in the subspace, thus optimizing the effect of the entity alignment. The main contributions are as follows:

(1) An embedding-based cross-lingual entity alignment method was proposed that uses Transformer to obtain knowledge graph entity encoding representations. Under multimodal information supervision, different models of information are mapped to a shared low-dimensional subspace to achieve entity alignment.
(2) We proposed a multimodal supervised strategy for knowledge graph entity representation learning, ensuring that the vector representation of the entities contains rich multimodal semantic information, enhancing the generalization ability of the learned entity representation.
(3) We evaluated the proposed method on a real cross-lingual dataset from DBpedia. The experimental results showed that the proposed method outperforms several cross-lingual entity alignment methods on Hits@1, Hits@10, and MRR. The framework is simple, fast, and has strong interpretability.

2. Method

We define a knowledge graph as $G = \{E, R, A, I, T\}$, where E, R, A, and I represent the set of entities, relationships, attributes, and images, respectively, and $T = \{E, R, E\}$ is the set of relationship triplets. Given two knowledge graphs $G_s = \{E_s, R_s, A_s, I_s, T_s\}$ and $G_t = \{E_s, R_t, A_t, I_t, T_t\}$, EA aims to identify entity pairs (e_s, e_t), where $e_s \in E_s$, $e_t \in E_t$. The model framework is illustrated in Figure 1. Given two multimodal knowledge bases (KBs), the model learns vector embeddings representing different KBs and expects closely embedded entities with potential alignment. The specific algorithmic steps are as follows:

(1) Firstly, the Transformer is utilized to obtain encoding representations of knowledge graph entities.
(2) Then, multimodal data supervision is employed for learning knowledge graph entity representations, ensuring that the vector representations of entities contain rich multimodal semantic information, thus enhancing the generalization capability of the learned entity representations.
(3) Entity embeddings are obtained for all entities, followed by the computation of similarities between all pairs of entities, which are then constrained using neighborhood component analysis (NCA) loss. Iterative learning helps to expand the training set.

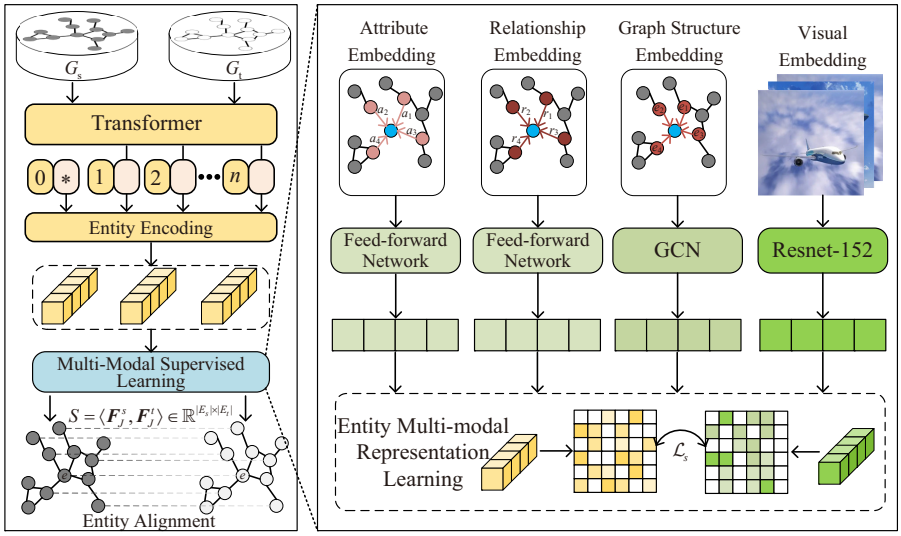


Figure 1. The framework of the proposed MDSEA.

2.1. Transformer-Based Knowledge Graph Entity Encoding

This section elaborates on how entities from two knowledge graphs, denoted as G_s and G_t, are embedded into low-dimensional vectors. The L_T layer of the Transformer is employed as the entity encoder to extract entity features. This layer is composed of multi-head attention (MHA) and feed-forward network (FFN) blocks. When a token sequence $\{w_1, \ldots, w_n\}$ is embedded into a word embedding matrix $\boldsymbol{F}_T \in \mathbb{R}^{n \times d_T}$, the entity encoding is computed as follows:

$$\begin{aligned} \boldsymbol{F}_{T,l} &= \boldsymbol{F}_T + \boldsymbol{T}_p \\ \tilde{\boldsymbol{F}}_{T,l} &= \mathrm{LN}(\mathrm{FFN}(\boldsymbol{F}_{T,l-1})) + \boldsymbol{F}_{T,l-1}, l = 1, \ldots, L_T \\ \boldsymbol{F}_{T,l} &= \mathrm{LN}(\mathrm{FFN}(\tilde{\boldsymbol{F}}_{T,l})) + \tilde{\boldsymbol{F}}_{T,l}, l = 1, \ldots, L_T \end{aligned} \tag{1}$$

where $\boldsymbol{T}_p$ represents positional embeddings, $\mathrm{LN}(\cdot)$ denotes layer normalization, $\boldsymbol{F}_{T,l}$ is the hidden feature of the entity at the l-th layer.

MHA is utilized to compute the weighted hidden states for each head, which are then concatenated as

$$\begin{aligned} &\mathrm{MHA}(\boldsymbol{x}) = [\mathrm{head}_1, \cdots, \mathrm{head}_\mathrm{h}]\boldsymbol{W}_o \\ &\mathrm{head}_i = \mathrm{Attn}(\boldsymbol{x}\boldsymbol{W}_{q,i}, \boldsymbol{x}\boldsymbol{W}_{k,i}, \boldsymbol{x}\boldsymbol{W}_{v,i}) = \mathrm{Attn}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) \end{aligned} \tag{2}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{d \times d}$ and d represents the dimensionality of the hidden embeddings. $d_h = d/N_h$ is typically set in MHA.

The FFN consists of two layers of linear transformations with a ReLU activation function:

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{x}\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2 \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_m}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_m \times d}$.

2.2. Multimodal Supervised Learning Network

After obtaining entity encodings using an entity encoder, the entities are fine-tuned to incorporate multimodal information representations through multimodal supervised learning. Specifically, for the relationship and attribute information of entities, features are extracted using a feed-forward network, while graph structure information is acquired using GCN, and image information is extracted using ResNet.

Relationship and Attribute Embedding: Since modeling relationships and attributes using GCNs might result in contaminated entity representations due to noise interference from neighbors [25], a simple feed-forward network is employed to map relationship and attribute features in a low-dimensional space:

$$\begin{aligned} \boldsymbol{F}_R &= \boldsymbol{W}_R \cdot \mathbf{R} + \boldsymbol{b}_R \\ \boldsymbol{F}_A &= \boldsymbol{W}_A \cdot \mathbf{A} + \boldsymbol{b}_A \end{aligned} \tag{4}$$

where $\boldsymbol{W}_R$ and $\boldsymbol{W}_A$ are parameter matrices for the relationship features $\boldsymbol{F}_R$ and attribute features $\boldsymbol{F}_A$, respectively.

Graph Structure Embedding: To mimic the structural similarity between G_s and G_t, capturing the proximity of entities and relationships, a GCN is employed to extract graph structural information. Specifically, a graph can be defined as $G = (V, b)$, where V is a series of nodes $\{v_1, v_2 \ldots, v_G\}$, and b represents the edge set. The entire feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times G}$ comprises N feature vectors, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^T$. The sparse symmetric adjacency matrix, denoted by $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, reflects the connection between each pair of nodes. $\boldsymbol{A}_{ij}$ can be computed using the following radial basis function (RBF):

$$\boldsymbol{A}_{ij} = \exp\left(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{\gamma_1}\right) \tag{5}$$

where the parameter γ_1 is empirically set to control the width of the RBF. The diagonal matrix is defined as $\boldsymbol{D} = diag(d_1, d_2, \ldots, d_N)$, where $d_i = \sum_{j=1}^{M} \boldsymbol{A}_{ij}$ represents the sum of the i-th row of the adjacency matrix.

The multi-layer GCN at the l-th layer is represented as

$$\boldsymbol{H}^{(l+1)} = [\widetilde{\boldsymbol{D}}^{-\frac{1}{2}} \widetilde{\boldsymbol{M}} \widetilde{\boldsymbol{D}}^{-\frac{1}{2}} \boldsymbol{H}^{(l)} \boldsymbol{W}^{(l)}] \tag{6}$$

where $[\cdot]_+$ represents the ReLU activation function, $\widetilde{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{I}_N$ is the adjacency matrix of $G_s \bigcup G_t$ plus the identity matrix (self-connections), $\widetilde{\boldsymbol{D}}$ is the trainable layer-specific weight matrix, $\boldsymbol{H}^{(l)} \in \mathbb{R}^{N \times D}$ is the output of the previous layer of GCN, where N is the number of entities and D is the feature dimensionality. $\boldsymbol{H}^{(0)}$ is randomly initialized, and the output of the last layer of GCN is used as the embedded graph structure $\boldsymbol{F}_G$.

Visual Embedding: Resnet-152 has been pre-trained on the ImageNet recognition task and serves as the feature extractor for all images. For each image, we use trainable Resnet-152 to extract image features and use the output of the last layer as a feature representation to obtain visual embedding:

$$\boldsymbol{F}_I = \boldsymbol{W}_I \cdot \mathrm{ResNet}(I) + \boldsymbol{b}_I \tag{7}$$

The visual representations extracted by Resnet are expected to capture both low-level similarity and high-level semantic correlation between images.

In multimodal supervised learning, the feature similarity matrix of entity embeddings from different modalities is utilized as supervision information, and the following objective function is minimized:

$$\mathcal{L}_s^i = ||\boldsymbol{F}_T^s - S^{(1,2)} \boldsymbol{F}_T^t||_F^2 + \gamma_2 (||\boldsymbol{F}_T^s - S^{(1)} \boldsymbol{F}_T^s|| + ||\boldsymbol{F}_T^t - S^{(2)} \boldsymbol{F}_T^t||_F^2) \tag{8}$$

where $i \in \{R, A, G, I\}$ represents four different embeddings, and γ_2 is a hyperparameter used to balance the similarity between KBs and their internal similarity. $S^{(1,2)} = \langle \boldsymbol{F}_T^s, \boldsymbol{F}_T^t \rangle \in \mathbb{R}^{|E_s| \times |E_t|}$, $S^{(1)} = \langle \boldsymbol{F}_T^s, \boldsymbol{F}_i^s \rangle \in \mathbb{R}^{|E_s| \times |E_s|}$, and $S^{(2)} = \langle \boldsymbol{F}_T^t, \boldsymbol{F}_i^t \rangle \in \mathbb{R}^{|E_t| \times |E_t|}$. The objective is to minimize $\mathcal{L}_s$ to tightly embed semantically similar entities across KGs.

2.3. Knowledge Graph Entity Alignment

First, obtain all entity embeddings $\boldsymbol{F}_J$ obtained through multimodal data supervised learning, then compute the similarity of all entity pairs, and constrain them using the NCA loss. Simultaneously, use IL to expand the training set.

Embedding Alignment: Let $\boldsymbol{F}_j^s$ and $\boldsymbol{F}_j^t$, respectively, represent the embeddings of the source entity E_s and the target entity E_t. Compute their cosine similarity matrix:

$$S = \langle \boldsymbol{F}_J^s, \boldsymbol{F}_J^t \rangle \in \mathbb{R}^{|E_s| \times |E_t|} \tag{9}$$

where each entry S_{ij} corresponds to the cosine similarity between the i-th entity in E_s and the j-th entity in E_t.

NCA Loss: Inspired by the NCA-based text–image matching method proposed in [37], a similar form of NCA loss is adopted. It measures the importance of samples using local and global statistics and penalizes hard negatives with a soft weighting scheme. The formula for the NCA loss is as follows:

$$\mathcal{L}_{\mathrm{NCA}} = \sum_{i=1}^{N} \left(\log \sum_{y_i = y_j} e^{S_{ij}} - \log \sum_{k=1}^{N} e^{S_{ik}} \right) \tag{10}$$

where N is the number of samples, S_{ij} is the cosine similarity between entity sample pairs.

Applying NCA loss for classification in the context of producing matches between two sets of entities:

$$\mathcal{L}_s = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{\alpha}\log\left(1+\sum_{m\neq i}e^{\alpha S_{mi}}\right)+\frac{1}{\alpha}\log\left(1+\sum_{n\neq i}e^{\alpha S_{in}}\right)-\log(1+\beta S_{ii})\right) \quad (11)$$

where α, β are hyperparameters; M is the number of pivots in a mini-batch. This loss is applied separately to each modality and also to the merged multimodal representation as shown in Equation (10). The joint loss is written as

$$\mathcal{L}_{\text{Joint}} = \sum_{i}^{n}\mathcal{L}_s^i + \mathcal{L}_s^T \quad (12)$$

where $\mathcal{L}_s^i$ represents the loss term supervised learning under different embeddings, with losses $\mathcal{L}_s^R, \mathcal{L}_s^A, \mathcal{L}_s^I, \mathcal{L}_s^G; \mathcal{L}_s^T; \mathcal{L}_s^T$ applied to the multimodal representation $\boldsymbol{F}_J$.

Iterative Learning: In order to improve learning with few training points, this paper adopts an IL strategy to propose more alignment seeds from unaligned entities. Specifically, for each iteration, a new round of proposals is created. Each pair of cross-graph entities, which are nearest neighbors to each other, is proposed and added to the candidate list. If a proposed entity pair remains each other's nearest neighbors in consecutive k rounds (i.e., trial stage), they are permanently added to the training set. Thus, the candidate list is refreshed every $K_e \cdot K_s$ times.

3. Experiment

In this section, we conducted experiments on three subsets of the DBP15K dataset (Section 3.1), compared the entity alignment effects (Section 3.3.1) of different methods under the same experimental settings (Section 3.2), and provided an efficiency analysis of the model (Section 3.3.2). At the same time, we also conducted a detailed study on the ablation experiments of different modules of MDSEA (Section 3.3.3).

3.1. Experiment Dataset

The DBP15K dataset is a multilingual dataset containing English, Chinese, Japanese, and French, shown in Table 1. It is constructed from the multilingual versions of DBpedia, a large-scale multilingual knowledge base that includes language interlinks from English entities to entities in other languages. During the construction of the DBP15K dataset, 15,000 popular entities were extracted separately from English to Chinese, Japanese, and French, and these were used as reference alignments. The extraction strategy involved randomly selecting a language interlink pair, where the involved entities had at least four relation triples, and then extracting relation and attribute information triples for the selected entities. The number of entities involved in each language far exceeds 15,000, with attribute triples contributing significantly to the dataset.

Table 1. DBP15K dataset distribution.

Dataset	KG	Entity	Relationship	Attribute	Relationship Triplet	Attribute Triplet	Figure	Entity Pairs
DBP15K_{ZH-EN}	ZH	19,388	1701	8111	70,414	248,035	15,912	15,000
	EN	19,572	1323	7173	95,142	343,218	14,125	
DBP15K_{JA-EN}	JA	19,814	1299	5882	77,214	248,991	12,739	1500
	EN	19,780	1,153	6066	93,484	320,616	13,741	
DBP15K_{FR-EN}	FR	19,661	903	4547	105,998	273,825	14,174	15,000
	EN	19,993	1208	6422	115,722	351,094	13,858	

In this experiment, three datasets from DBP15K were utilized: DBP15K_{ZH-EN} (Chinese to English), DBP15K_{JA-EN} (Japanese to English), and DBP15K_{FR-EN} (French to En-

glish). Each of these datasets contains approximately 400,000 triples and 15,000 pre-aligned entity pairs, with 30% used as seed alignments (Rs = 0.3). The English, French, and Japanese versions of the entities contain images provided by DBpedia, while Chinese images are extracted from the original Chinese Wikipedia dumps. Additionally, not all entities have images; only around 50–85% of entities have images. For entities without images, a random vector sampled from a normal distribution is assigned, parameterized by the mean and standard deviation of other images.

3.2. Experimental Parameter Settings

The experimental platform utilized a server equipped with an Intel i9-12900k CPU (Intel Corporation, Santa Clara, CA, USA) and Nvidia RTX 3080Ti GPU (Nvidia Corporation, Santa Clara, CA, USA). The proposed algorithm was implemented using the Adam optimizer in PyTorch. For training on the DBP15K dataset, the number of epochs was set to 500, with a learning rate of 0.001. To ensure fair comparison, the experimental setup employed the same training/testing split as common methods: 30% of pre-aligned entities were used for training, while the remaining 70% of anchor links were used for testing, with 20% of training entity pairs reserved for validation. To demonstrate the model's stability, the visual encoder was set to ResNet-152 with visual feature dimensions of 2048. Each experiment was conducted 10 times, and the results were averaged to reduce randomness.

In the experiments, the effectiveness of multimodal entity alignment was evaluated as an indicator of the proposed model's performance. Specifically, three common metrics were employed: the average percentage of triplets ranked 1 in the test samples (Hits@1), the average percentage of triplets ranked below 10 in the test samples (Hits@10), and the mean reciprocal rank (MRR).

The experimental parameters, including regularization parameters and network model parameters, were adjusted within given ranges to maximize classification accuracy. To achieve this, a 10-fold cross-validation was performed on the training set to determine parameter combinations for different methods. Additionally, feature dimensionality was identified as a key parameter affecting the quality of the final learned feature representations. Therefore, the optimal feature dimensionality was determined by testing values ranging from 100 to 500 at intervals of 5, based on the best classification performance on the training set.

3.3. Experimental Analysis

The proposed method was compared with common entity alignment methods such as MTransE [16], JAPE [18], EVA [29], and DNCN [1], using 30% of the EA labels for training. The experimental results are presented below along with the corresponding analyses.

3.3.1. Experimental Results Analysis

Table 2 reports the results regarding EA. The results indicate that the proposed method outperforms the other models, achieving the best performance. Specifically, the proposed method improved Hits@1 by over 20% compared to the baseline methods. When incorporating visual information, the proposed method achieved a 4–15% improvement in Hits@1 over the other methods. This suggests that combining visual representations can effectively enhance cross-lingual entity representations to infer their correspondences.

Table 2. Cross-language EA results on DBP15K.

Methods	DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MTransE	30.83	61.41	0.364	27.86	57.45	0.349	24.41	55.55	0.335
JAPE	41.18	74.46	0.490	36.25	68.50	0.476	32.39	66.68	0.430
EVA	59.44	83.44	0.680	63.12	85.85	0.712	66.52	88.40	0.747
DNCN	72.10	87.90	0.775	72.13	88.58	0.781	74.84	88.53	0.790
MDSEA	76.81	90.35	0.814	76.92	94.63	0.832	76.51	94.67	0.834

Compared to the multimodal methods, the proposed method, by introducing multimodal data supervision, enables the model to learn more generalized and robust representations, aiding in addressing noise and variations, thus achieving good entity alignment performance even in complex data scenarios.

3.3.2. Experimental Efficiency Analysis

To further understand the model, this study investigated the efficiency behavior of several algorithms on a dataset with the same 500 epochs and an early stopping strategy. As shown in Table 3, the proposed method consistently outperforms the other algorithms throughout the entire training process and excels in balancing convergence time and performance.

Table 3. Comparison of efficiency results of different methods on the DBP15K$_{ZH-EN}$.

Methods		MTransE	JAPE	EVA	DNCN	MDSEA
MRR	50 epoch	0.173	0.191	0.232	0.345	0.612
	150 epoch	0.241	0.276	0.347	0.574	0.784
	250 epoch	0.335	0.424	0.669	0.716	0.803
	500 epoch	0.364	0.490	0.680	0.775	0.814

From Figure 2, it can be observed that the proposed method achieves optimal performance the fastest among the compared baselines and surpasses the other algorithms throughout the entire process.

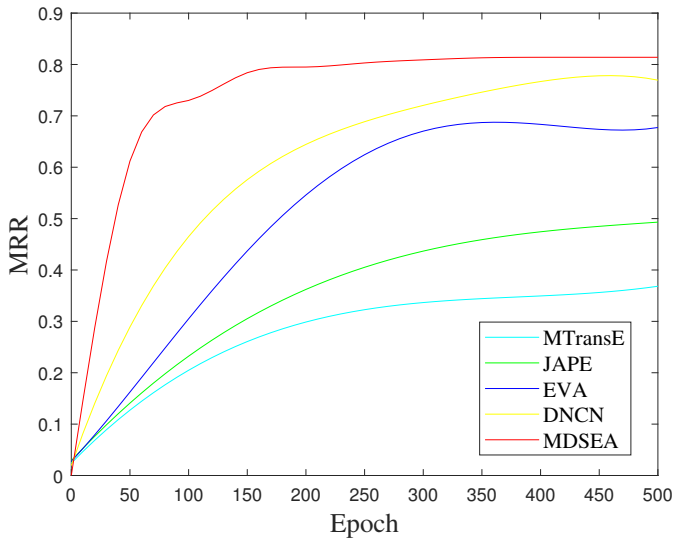


Figure 2. Comparison of efficiency results of different methods on the DBP15K$_{ZH-EN}$.

3.3.3. Ablation Study

To demonstrate the effectiveness of each module in the proposed method, two variants of the knowledge graph entity alignment method based on multimodal data, MDSEA-A and MDSEA-B, were proposed. The final classification performance under different modules was compared. The specific module selections are shown in Table 4, and the alignment results are presented in Table 5.

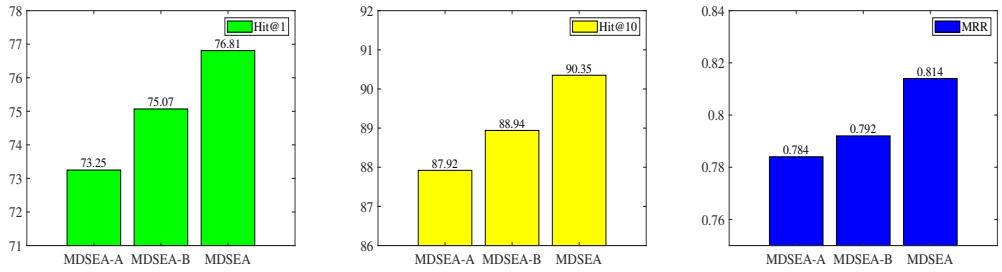
Table 4. Selection of different modules in MDSEA.

Module	MDSEA-A	MDSEA-B	MDSEA
MDS		✓	✓
MWF			✓

Where MDS represents the multimodal data supervision module, and MWF represents the multimodal weighted fusion module. The ablation comparisons are shown in Table 5 and Figure 3.

Table 5. The alignment effect of different modules in the DBP15K$_{ZH-EN}$.

Module	MDSEA-A	MDSEA-B	MDSEA
Hits@1	73.25	75.07	76.81
Hits@10	87.92	88.94	90.35
MRR	0.784	0.792	0.814



(**a**) (**b**) (**c**)

Figure 3. The alignment effect of different modules in the DBP15K$_{ZH-EN}$. (**a**) Hits@1; (**b**) Hits@10; (**c**) MRR.

Compared to MDSEA-A, MDSEA-B improves the Hit@1 on DBP15K$_{ZH-EN}$ by 1.82%, indicating that the multimodal supervision strategy enriches the vector representations of entities with rich multimodal semantic information, enhancing the generalization ability of the learned entity representations. Compared to MDSEA-B, MDSEA improves the Hit@1 on DBP15K$_{ZH-EN}$ by 1.74%, attributed to the multimodal weighted fusion strategy reducing noise from different modal information.

4. Conclusions

This article proposes a knowledge graph entity alignment method based on multimodal supervised learning. Firstly, it utilizes Transformer to obtain the encoded representations of knowledge graph entities. Then, it employs a multimodal supervised learning approach for knowledge graph entity representation learning. This ensures that the vector representations of the entities contain rich multimodal semantic information, thereby enhancing the generalization capability of the learned entity representations. Finally, it maps information from different modalities into a shared low-dimensional subspace, making similar entities closer in the subspace, thus optimizing the entity alignment effect. The proposed method is compared with common entity alignment methods, and the results demonstrate its superiority over the state-of-the-art baseline methods. In addition, due to the lack of some visual modalities in the dataset, the multimodal supervised learning of the model is limited to some extent. Therefore, we will conduct in-depth research on this issue in subsequent work.

Author Contributions: Conceptualization, J.F. and X.Y.; methodology, J.F.; validation, J.F. and X.Y.; formal analysis, J.F. and X.Y.; writing—original draft preparation, J.F. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Joint Fund of National Natural Science Foundation of China and Civil Aviation Administration of China (No. U2033202).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Zhang, Y.; Li, Y.; Wei, X.; Yang, Y.; Liu, L.; Murphey, Y. L. Graph matching for knowledge graph alignment using edge-coloring propagation. *Pattern Recogn.* **2023**, *144*, 109851. [CrossRef]
2. Yu, J.; Zhu, Z.; Wang, Y.; Zhang, W.; Hu, Y.; Tan, J. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recogn.* **2020**, *108*, 107563. [CrossRef]
3. Xie, C.; Zhang, L.; Zhong, Z. Entity alignment method based on joint learning of entity and attribute representations. *Appl. Sci.* **2023**, *13*, 5748. [CrossRef]
4. Wang, H.; Liu, Q.; Huang, R.; Zhang, J. Multi-modal entity alignment method based on feature enhancement. *Appl. Sci.* **2023**, *13*, 6747. [CrossRef]
5. Lin, B.; Zhu, Y.; Liang, X. Atom correlation based graph propagation for scene graph generation. *Pattern Recogn.* **2022**, *122*, 108300. [CrossRef]
6. Zhang, X.; Zhang, R.; Chen, J.; Kim, J.; Mao, Y. Semi-supervised entity alignment with global alignment and local information aggregation. *IEEE Trans. Knowl. Data. Eng.* **2023**, *35*, 10464–10477. [CrossRef]
7. Sun, Z.; Hu, W.; Wang, C.; Wang, Y.; Qu, Y. Revisiting embedding-based entity alignment: a robust and adaptive method. *IEEE Trans. Knowl. Data. Eng.* **2022**, *35*, 8461–8475. [CrossRef]
8. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Proceedings of the Conference and Workshop on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2787–2795.
9. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the International Conference on Learning Representations, Lugano, Switzerland, 9 September 2016.
10. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
11. Huang, W.; Liu, J.; Li, T.; Ji, S.; Wang, D.; Huang, T. FedCKE: Cross-domain knowledge graph embedding in federated learning. *IEEE Trans. Big Data* **2022**, *9*, 792–804. [CrossRef]
12. Li, H.; Han, Z.; Zhu, H.; Qian, Y. A novel embedding model for knowledge graph entity alignment based on graph neural networks. *Appl. Sci.* **2023**, *13*, 5876. [CrossRef]
13. Trisedya, B.D.; Qi, J.; Zhang, R. Entity alignment between knowledge graphs using attribute embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 297–304.
14. Zeng, W.; Zhao, X.; Tang, J.; Lin, X. Collective entity alignment via adaptive features. In Proceedings of the IEEE 36th International Conference on Data Engineering, Dallas, TX, USA, 20–24 April 2020; pp. 1870–1873.
15. Xu, Y.; Zhong, J.; Zhang, S.; Li, C.; Li, P.; Guo, Y.; Zhang, Y. A domain-oriented entity alignment approach based on filtering multi-type graph neural networks. *Appl. Sci.* **2023**, *13*, 9237. [CrossRef]
16. Chen, M.; Tian, Y.; Yang, M.; Zaniolo, C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia, 19–25 August 2017; pp. 1511–1517.
17. Zhu, H.; Xie, R.; Liu, Z.; Sun, M. Iterative entity alignment via joint knowledge embeddings. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia, 19–25 August 2017; pp. 4258–4264.
18. Sun, Z.; Hu, W.; Li, C. Cross-lingual entity alignment via joint attribute-preserving embedding. In Proceedings of the International Semantic Web Conference, Vienna, Austria, 21–25 October 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 628–644.
19. Guo, L.; Sun, Z.; Hu, W. Learning to exploit long-term relational dependencies in knowledge graphs. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2505–2514.
20. Sun, Z.; Huang, J.; Hu, W.; Chen, M.; Guo, L.; Qu, Y. Transedge: Translating relation contextualized embeddings for knowledge graphs. In Proceedings of the International Semantic Web Conference, Auckland, New Zealand, 26–30 October, 2019; pp. 612–629.
21. Sun, Z.; Hu, W.; Zhang, Q.; Qu, Y. Bootstrapping entity alignment with knowledge graph embedding. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 4396–4402.
22. Wang, Z.; Lv, Q.; Lan, X.; Zhang, Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 349–357.
23. Cao, Y.; Liu, Z.; Li, C.; Li, J.; Chua, T. S. Multi-channel graph neural network for entity alignment. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1452–1461.
24. Zhu, Q.; Zhou, X.; Wu, J.; Tan, J.; Guo, L. Neighborhood-aware attentional representation for multilingual knowledge graphs. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 1943–1949.

25. 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, Hong Kong, China, 3–7 November 2019; pp. 4422–4432.
26. Wu, Y.; Liu, X.; Feng, Y.; Wang, Z.; Yan, R.; Zhao, D. Relation-aware entity alignment for heterogeneous knowledge graphs. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
27. Mao, X.; Wang, W.; Xu, H.; Lan, M.; Wu, Y. MRAEA: An efficient and robust entity alignment approach for cross-lingual knowledge graph. In Proceedings of the 13th International Conference on Web Search and Data Mining, Houston, TX, USA, 3–7 February 2020; pp. 420–428.
28. Mao, X.; Wang, W.; Wu, Y.; Lan, M. Are negative samples necessary in entity alignment? An approach with high performance, scalability and robustness. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual, 1–5 November 2021; pp. 1263–1273.
29. 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; pp. 4257–4266.
30. Chen, M.; Shi, W.; Zhou, B.; Roth, D. Cross-lingual entity alignment with incidental supervision. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Virtual, 19–23 April 2021; pp. 645–658.
31. Zhu, Q.; Wei, H.; Sisman, B.; Zheng, D.; Faloutsos, C.; Dong, X.L.; Han, J. Collective multi-type entity alignment between knowledge graphs. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 2241–2252.
32. Liu, B.; Scells, H.; Zuccon, G.; Hua, W.; Zhao, G. ActiveEA: Active learning for neural entity alignment. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3364–3374.
33. Chen, L.; Li, Z.; Wang, T.; Xu, T.; Wang, Z.; Chen, E. MMEA: Entity alignment for multi-modal knowledge graph. In Proceedings of the Knowledge Science, Engineering and Management: 13th International Conference, Hangzhou, China, 28–30 August 2020; pp. 134-147.
34. 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.
35. Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; Zheng, Y. Multi-modal contrastive representation learning for entity alignment. In Proceedings of the International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 2572–2584.
36. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; p. 32.
37. Liu, F.; Ye, R.; Wang, X.; Li, S. Hal: Improved text-image matching by mitigating visual semantic hubs. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11563–11571.