

Multi-modal Contrastive Representation Learning for Entity Alignment

Zhenxi Lin¹, Ziheng Zhang¹, Meng Wang², Yinghui Shi³, Xian Wu¹, Yefeng Zheng¹

¹Tencent Jarvis Lab, Shenzhen, China

²School of Computer Science and Engineering, Southeast University, Nanjing, China

³School of Cyber Science and Engineering, Southeast University, Nanjing, China

{chalerislin, zihengzhang}@tencent.com

{meng.wang, shiyinghui}@seu.edu.cn

{kevinxwu, yefengzheng}@tencent.com

Abstract

Multi-modal entity alignment aims to identify equivalent entities between two different multi-modal knowledge graphs, which consist of structural triples and images associated with entities. Most previous works focus on how to utilize and encode information from different modalities, while it is not trivial to leverage multi-modal knowledge in entity alignment because of the modality heterogeneity. In this paper, we propose MCLEA, a **M**ulti-modal **C**ontrastive **L**earning based **E**ntity **A**lignment model, to obtain effective joint representations for multi-modal entity alignment. Different from previous works, MCLEA considers task-oriented modality and models the inter-modal relationships for each entity representation. In particular, MCLEA firstly learns multiple individual representations from multiple modalities, and then performs contrastive learning to jointly model intra-modal and inter-modal interactions. Extensive experimental results show that MCLEA outperforms state-of-the-art baselines on public datasets under both supervised and unsupervised settings.¹

1 Introduction

Knowledge graphs (KGs) such as DBpedia (Lehmann et al., 2015) and YAGO (Mahdisoltani et al., 2015) employ the graph structure to represent real-world knowledge, where the concepts are represented as nodes and the relationships among concepts are represented as edges. KGs have been widely applied to knowledge-driven applications to boost their performance, like recommendation system (Cao et al., 2019b), information extraction (Han et al., 2018) and question answering (Lan et al., 2021). In recent years, an increasing amount of knowledge has been represented in multi-modal formats, such as MMKG (Liu et al., 2019) and Richpedia (Wang

et al., 2020). These multi-modal KGs usually contain images as the visual modality, like profile photos, thumbnails, or posters. The augmented visual modality has shown the significant capability to improve KG-based applications (Chen et al., 2020a). It was also proven that the incorporation of visual modality can enhance the contextual semantics of entities and also achieve improved KG embeddings (Wang et al., 2021).

Due to the large scope of real-world knowledge, most KGs are often incomplete, and multiple different KGs are usually complementary to one another. As a result, integrating multiple KGs into a unified one can enlarge the knowledge coverage and can also assist in refining KG by discovering the potential flaws (Chen et al., 2020a). To integrate heterogeneous multi-modal knowledge graphs, the task of multi-modal entity alignment (MMEA) is therefore proposed, which aims to discover equivalent entities referring to the same real-world object. Several previous MMEA works have shown that the inclusion of visual modality in modeling helps to improve the performance of entity alignment. For instance, MMEA (Chen et al., 2020a) and EVA (Liu et al., 2021) proposed distinct multi-modal fusion modules to integrate entity representations from multiple modalities into joint embeddings and they achieved state-of-the-art performance.² However, these methods mainly utilize existing representations from different modalities, and customized representation learning for EA is not fully explored. In addition, existing methods only explore the use of diverse multi-modal representations to enhance the contextual embedding of entities, the inter-modal interactions are often neglected in modeling.

To address aforementioned problems, we propose MCLEA, a **M**ulti-modal **C**ontrastive **L**earning based **E**ntity **A**lignment model, which effectively

¹The source code is available at <https://github.com/lzxlin/MCLEA>.

²To distinguish the model MMEA from the task MMEA, we use EA to denote multi-modal entity alignment for the rest of the paper.

integrates multi-modal information into joint representations for EA. MCLEA firstly utilizes multiple individual encoders to obtain modality-specific representations for each entity. The individually encoded information includes neighborhood structures, relations, attributes, surface forms (i.e., entity names), and images. Furthermore, we introduce contrastive learning into EA with intra-modal contrastive loss and inter-modal alignment loss. Specifically, intra-modal contrastive loss aims at distinguishing the embeddings of equivalent entities from the ones of other entities for each modality, thus generating more appropriate representations for EA. Inter-modal alignment loss, on the other hand, aims at modelling inter-modal interactions and reducing the gaps between modalities for each entity. With these two losses, MCLEA can learn discriminative cross-modal latent embeddings and ensure potentially equivalent entities close in the joint embedding space, regardless of the modality. MCLEA is also generic as it can support a wide variety of modalities. Moreover, it combines multiple losses and simultaneously learns multiple objectives using task-dependent uncertainty.

The contributions of this paper are three-fold: (i) We propose a novel method, called MCLEA, to embed information from different modalities into a unified vector space and then obtain discriminative entity representations based on contrastive learning for entity alignment. (ii) We propose two novel losses to explore intra-modal relationships and inter-modal interactions, ensuring that to-be-aligned entities between different KGs are semantically close with minimum gaps between modalities. (iii) We experimentally validate the effectiveness and superiority of MCLEA as it achieves state-of-the-art performance on several public datasets in both supervised and unsupervised settings. The overall results also suggest that our MCLEA is capable of learning more discriminative embedding space for multi-modal entity alignment.

2 Related Work

2.1 Multi-modal Knowledge Graph

While many efforts (Mahdisoltani et al., 2015; Lehmann et al., 2015) have been made to achieve large-scale KGs, there are just a few attempts to enrich KGs with multiple modalities. For example, MMKG (Liu et al., 2019) and Richpedia (Wang et al., 2020) utilized the rich visual resources (mainly images) to construct multi-modal knowl-

edge graphs. Their target was to enrich the KG information via appending sufficient and diverse images to textual entities but they also brought challenges to the KG embedding methods. Unlike traditional KG embedding methods, multi-modal KG embedding methods model the textual and visual modalities at the same time (Zhang et al., 2019; Wang et al., 2021). For example, Zhang et al. (2019) proposed MKHAN to exploit multi-modal KGs with hierarchical attention networks on question answering, and Wang et al. (2021) proposed RSME to selectively incorporate visual information during the KG embedding learning process.

2.2 Multi-modal Entity Alignment

Recent studies for entity alignment mostly focused on exploring the symbolic similarities based on various features, including entity names (Wu et al., 2019; Zhang et al., 2019), attributes (Liu et al., 2020), descriptions (Zhang et al., 2019; Tang et al., 2020) and ontologies (Xiang et al., 2021). Most of them started with transforming entities from different KGs into a unified low-dimensional vector space by translation-based models or graph neural networks and then discovered their counterparts based on the similarity metrics of entity embeddings. Some surveys summarized that additional KG information, if appropriately encoded, could further improve the performance of EA methods (Sun et al., 2020b; Zhang et al., 2020). Some previous attempts even proposed to guide these embedding-based EA models with probabilistic reasoning (Qi et al., 2021). With such findings and the increasing popularity of multi-modal knowledge graphs, how to incorporate visual modality in EA, namely multi-modal entity alignment, has begun to draw research attention but the attempts are limited. The pioneer method PoE (Liu et al., 2019) combined the multi-modal features and measured the credibility of facts by matching the underlying semantics of entities. Afterward, MMEA (Chen et al., 2020a) was proposed to integrate knowledge from different modalities (relational, visual, and numerical) and obtain the joint entity representations. Another method termed EVA (Liu et al., 2021) leveraged visual knowledge and other auxiliary information to achieve EA in both supervised and unsupervised manner. Alternatively, the method HMEA (Guo et al., 2021) modeled structural and visual representations in the hyperbolic space, while Masked-MMEA (Shi et al., 2022)

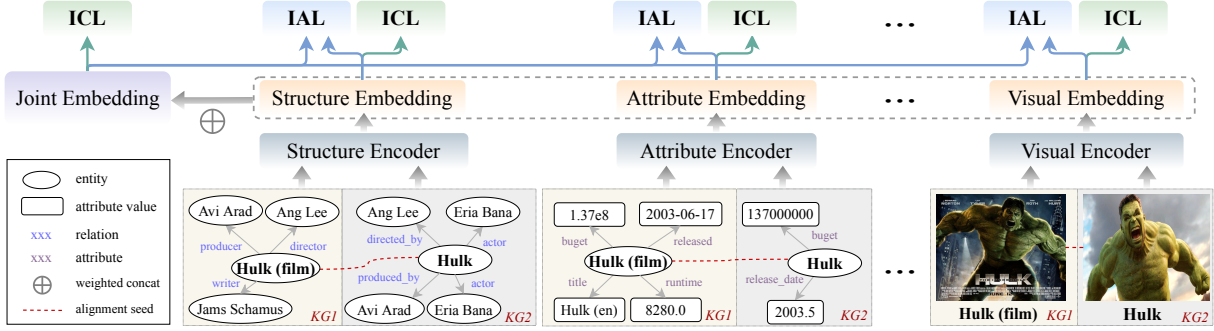


Figure 1: The overall architecture of MCLEA, which combines multiple modalities (§ 3.1) and learns through two proposed losses (§ 3.2), intra-modal contrastive loss (ICL) and inter-modal alignment loss (IAL).

discussed the impacts of visual modality and proposed to incorporate a selectively masking technique to filter potential visual noises. These methods mainly utilize multi-modal representations to enhance the contextual embedding of entities, nevertheless, customized entity representations for EA and inter-modal interactions are often neglected in modeling. Different from previous methods, our proposed MCLEA can learn both intra-modal and inter-modal dynamics simultaneously by the proposed contrastive objectives, expecting to learn more discriminative and abundant entity representation for EA.

3 Proposed Method

We start with the problem formulation and the notations. A multi-modal KG is denoted as $G = (E, R, A, V, T)$, where E, R, A, V, T are the sets of entities, relations, attributes, images, and triples, respectively. Given $G_1 = (E_1, R_1, A_1, V_1, T_1)$ and $G_2 = (E_2, R_2, A_2, V_2, T_2)$ as two KGs to be aligned, the aim of EA is to find aligned entity pairs $\mathcal{A} = \{(e_1, e_2) | e_1 \equiv e_2, e_1 \in E_1, e_2 \in E_2\}$, where we assume a small set of entity pairs \mathcal{S} (seeds) are given as training data. The overall architecture of the proposed MCLEA is shown in Figure 1, and its primary components, multi-modal embeddings, and contrastive representation learning will be detailed in the following sections.

3.1 Multi-Modal Embeddings

Multi-modal KGs often depict various features with multiple modalities (or views), which are complementary to each other. We investigate different embeddings from different modalities for MCLEA, including neighborhood structures, relations, attributes, names (often termed as surface forms in previous work (Liu et al., 2021)), and images. Each

modality is processed with an individual encoder network adapted to the nature of the signal. Furthermore, these uni-modal embeddings are aggregated with a simple weighted mechanism to form a joint embedding. Theoretically, MCLEA can support more modalities, e.g., numerical values (Chen et al., 2020a), which will be left in our future work.

3.1.1 Neighborhood Structure Embedding

The graph attention network (GAT) is a typical neural network that directly deals with structured data (Velickovic et al., 2018). Hence, we leverage GAT to model the structural information of G_1 and G_2 , shown as “**Structure Encoder**” in Figure 1. Specifically, the hidden state $\mathbf{h}_i \in \mathbb{R}^d$ (d is the hidden size) of entity e_i by aggregating its one-hop neighbors \mathcal{N}_i with self-loop is formulated as:

$$\mathbf{h}_i = \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{h}_j \right), \quad (1)$$

where \mathbf{h}_j is the hidden state of entity e_j , $\sigma(\cdot)$ denotes the ReLU nonlinearity, and α_{ij} denotes the importance of entity e_j to entity e_i , which is calculated with the self-attention:

$$\alpha_{ij} = \frac{\exp(\eta(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \oplus \mathbf{W}\mathbf{h}_j]))}{\sum_{u \in \mathcal{N}_i} \exp(\eta(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \oplus \mathbf{W}\mathbf{h}_u]))}, \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$ denotes the weight matrix, $\mathbf{a} \in \mathbb{R}^{2d}$ is a learnable parameter, \oplus is the concatenation operation and η is the LeakyReLU nonlinearity. Motivated by (Li et al., 2019), we restrict \mathbf{W} to be a diagonal matrix to reduce computations, thus increasing the scalability of the model. To stabilize the learning process of self-attention (Velickovic et al., 2018), we perform K ($K = 2$) heads of independent attention of Eq. (1) in parallel, and concatenate these features to obtain the structure

embedding of entity e_i :

$$\mathbf{h}_i^g = \bigoplus_{k=1}^K \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{h}_j \right), \quad (3)$$

where α_{ij}^k is the normalized attention coefficient computed by the k -th attention. In practice, we apply a two-layer GAT model to aggregate the neighborhood information within multiple hops, and use the output of the last GAT layer as the neighborhood structure embedding.

3.1.2 Relation, Attribute, and Name Embeddings

Because the vanilla GAT operates on unlabeled graphs, it is unable to properly model relational information in multi-relational KGs. To alleviate this issue, we follow (Yang et al., 2019) and regard the relations of entity e_i as bag-of-words feature and feed it into a simple feed-forward layer to obtain the relation embedding \mathbf{h}_i^r . For the simplicity and consistency of MCLEA, we adopt the same approach for the attribute embedding \mathbf{h}_i^a and the name embedding \mathbf{h}_i^n for entity e_i . Therefore, these embeddings are calculated as:

$$\mathbf{h}_i^l = \mathbf{W}_l \mathbf{u}_i^l + \mathbf{b}_l, \quad l \in \{r, a, n\}, \quad (4)$$

where \mathbf{W}_l and \mathbf{b}_l are learnable parameters, \mathbf{u}_i^r is the bag-of-words relation feature, \mathbf{u}_i^a is the bag-of-words attribute feature, and \mathbf{u}_i^n is the name feature obtained by averaging the pre-trained GloVe (Pennington et al., 2014) vectors of name strings. To avoid the out-of-vocabulary challenges brought by the extensive proper nouns (e.g., person names) and the limited vocabulary of word vectors, we further incorporate the character bigrams (Mao et al., 2021) of entity names as auxiliary features for name embedding.

3.1.3 Visual Embedding

We adopt the pre-trained visual model (PVM), e.g., ResNet-152 (He et al., 2016), to learn visual embedding, shown as “**Visual Encoder**” in Figure 1. We feed the image v_i of entity e_i into the pre-trained visual model and use the final layer output before logits as the image feature. The feature is sent through a feed-forward layer to get the visual embedding:

$$\mathbf{h}_i^v = \mathbf{W}_v \cdot \text{PVM}(v_i) + \mathbf{b}_v. \quad (5)$$

3.1.4 Joint Embedding

Next, we implement a simple weighted concatenation by integrating the multi-modal features into a single compact representation $\hat{\mathbf{h}}_i$ for entity e_i :

$$\hat{\mathbf{h}}_i = \bigoplus_{m \in \mathcal{M}} \left[\frac{\exp(w_m)}{\sum_{j \in \mathcal{M}} \exp(w_j)} \mathbf{h}_i^m \right], \quad (6)$$

where $\mathcal{M} = \{g, r, a, n, v\}$ and w_m is a trainable attention weight for the modality of m . L2-normalization is performed on the input embeddings before the weighted concatenation.

The current joint embeddings are coarse-grained and there are no interactions between modalities. Therefore, two training strategies are designed for learning the dynamics within (intra-) and between (inter-) modalities.

3.2 Contrastive Representation Learning

As the core of MCLEA, we propose two novel losses on the uni-modal and joint representations to sufficiently capture the dynamics within and between modalities while preserving semantic proximity and minimizing the modality gap.

3.2.1 Intra-modal Contrastive Loss (ICL)

Inspired by recent work on contrastive learning (CL) (Chen et al., 2020b; Khosla et al., 2020), we devise an intra-modal contrastive loss (ICL) that enforces the input embedding to respect the similarity of entities in the original embedding space. Meanwhile, ICL allows MCLEA to distinguish the embeddings of the same entities in different KGs from those of other entities for each modality.

Given that \mathcal{S} can be naturally regarded as positive samples, whereas any non-aligned pairs can be regarded as negative samples due to the convention of 1-to-1 alignment constraint (Sun et al., 2018). Formally, for the i -th entity $e_1^i \in E_1$ of minibatch \mathcal{B} , the positive set is defined as $P_i = \{e_2^i | e_2^i \in E_2\}$, where (e_1^i, e_2^i) is an aligned pair. The negative set includes two types, inner-graph unaligned pairs from the source KG G_1 and cross-graph unaligned pairs from the target KG G_2 , defined as $N_1^i = \{e_1^j | \forall e_1^j \in E_1, j \neq i\}$ and $N_2^i = \{e_2^j | \forall e_2^j \in E_2, j \neq i\}$, respectively. Both N_1^i and N_2^i come from minibatch \mathcal{B} . These two types of negative samples are designed to constrain the joint embedding space, in which the semantically similar entities from the same KG stay close-by and the aligned entities from two KGs map to proximate points. Overall, we define the alignment

probability distribution $q_m(e_1^i, e_2^i)$ of the modality of m for each positive pair (e_1^i, e_2^i) as:

$$q_m(e_1^i, e_2^i) = \frac{\delta_m(e_1^i, e_2^i)}{\delta_m(e_1^i, e_2^i) + \sum_{e_1^j \in N_1^i} \delta_m(e_1^i, e_1^j) + \sum_{e_2^j \in N_2^i} \delta_m(e_1^i, e_2^j)}, \quad (7)$$

where $\delta_m(u, v) = \exp(f_m(u)^T f_m(v) / \tau_1)$, $f_m(\cdot)$ is the encoder of the modality m , and τ_1 is a temperature parameter. Especially, L2-normalization is performed on the input feature embedding before computing the inner product. Notably, the distribution of Eq. (7) is directional and asymmetric for each input; the distribution for another direction is thus defined similarly for $q_m(e_2^i, e_1^i)$. The ICL can be formulated as:

$$\mathcal{L}_m^{\text{ICL}} = -\mathbb{E}_{i \in \mathcal{B}} \log \left[\frac{1}{2} (q_m(e_1^i, e_2^i) + q_m(e_2^i, e_1^i)) \right]. \quad (8)$$

We apply ICL on each modality separately and also on the combined multi-modal representation as specified in Eq. (6). ICL is performed in contrastive supervised learning to learn intra-modal dynamics for more discriminative boundaries for each modality in the embedding space.

3.2.2 Inter-modal Alignment Loss (IAL)

Since the embeddings of different modalities are separately trained with ICL, their representations are not aligned and it is difficult to model the complex interaction between modalities solely with the fusion module. To alleviate this, we further propose inter-modal alignment loss (IAL), which targets at reducing the gap between the output distribution over different modalities, so that the MCLEA can model inter-modal interactions and obtain more meaningful representations.

We consider the joint embedding as the comprehensive representation due to its fusion of multi-modal features; therefore, we attempt to transfer the knowledge from the joint embedding back to uni-modal embedding so that the uni-modal embedding could better utilize the complementary information from others. Concretely, we minimize the bidirectional KL divergence over the output distribution between joint embedding and uni-modal embedding:

$$\mathcal{L}_m^{\text{IAL}} = \mathbb{E}_{i \in \mathcal{B}} \left[\frac{1}{2} [\text{KL}(q'_o(e_1^i, e_2^i) \parallel q'_m(e_1^i, e_2^i)) + \text{KL}(q'_o(e_2^i, e_1^i) \parallel q'_m(e_2^i, e_1^i))] \right], \quad (9)$$

where $q'_o(e_1^i, e_2^i)$, $q'_o(e_2^i, e_1^i)$ and $q'_m(e_1^i, e_2^i)$, $q'_m(e_2^i, e_1^i)$ represent the output predictions with

two directions of joint embedding and the uni-modal embedding of modality m , respectively. Their calculation are similar to Eq. (7) but with a temperature parameter τ_2 . We only back-propagate through $q'_m(e_1^i, e_2^i)$, $q'_m(e_2^i, e_1^i)$ in Eq. (9) as knowledge distillation (Hinton et al., 2015).

IAL aims at learning interactions between different modalities within each entity, which concentrates on aggregating the distribution of different modalities and thus reduces the modality gap. Some approaches (Zhang et al., 2019; Chen et al., 2020a) attempt to learn a common space by imposing alignment constraints on the features between different modalities, but they introduce noises due to semantic heterogeneity. Different from these approaches, we distill the useful knowledge from the output prediction of multi-modal representation to the uni-modal representation, while maintaining relatively modality-specific features of each modality.

3.3 Optimization Objective

The overall loss of the MCLEA is as follows,

$$\mathcal{L} = \mathcal{L}_o^{\text{ICL}} + \sum_{m \in \mathcal{M}} \alpha_m \mathcal{L}_m^{\text{ICL}} + \sum_{m \in \mathcal{M}} \beta_m \mathcal{L}_m^{\text{IAL}}, \quad (10)$$

where $\mathcal{M} = \{g, r, a, n, v\}$, $\mathcal{L}_o^{\text{ICL}}$ denotes the ICL operated on joint embedding, α_m and β_m are hyper-parameters that balance the importance of losses. However, manually tuning these hyper-parameters is expensive and intractable. Instead, we treat MCLEA as a multi-task learning paradigm and then use homoscedastic uncertainty (Kendall et al., 2018) to weigh each loss automatically during model training. We adjust the relative weight of each task in the loss function by deriving a multi-task loss function based on maximizing the Gaussian likelihood with task-dependant uncertainty. Due to space limits, we only show the derived result and leave the detailed derivation process in the Appendix. The loss in Eq. (10) can be rewritten as:

$$\mathcal{L} = \mathcal{L}_o^{\text{ICL}} + \sum_{m \in \mathcal{M}} \left(\frac{1}{\alpha_m^2} \mathcal{L}_m^{\text{ICL}} + \frac{1}{\beta_m^2} \mathcal{L}_m^{\text{IAL}} + \log \alpha_m + \log \beta_m \right), \quad (11)$$

where α_m and β_m are automatically learned during training.

To overcome the lack of training data, we incorporate a bi-directional iterative strategy used in (Liu et al., 2021) to iteratively add new aligned seeds during training. In the inference, we use the cosine similarity metric between joint embeddings of entities to determine the counterparts of entities.

MCLEA can be extended to the unsupervised setting, in which the pseudo-alignment seeds are discovered based on feature similarities of entity names (Mao et al., 2020a; Ge et al., 2021) or entity images (Liu et al., 2021), accordingly resulting in different unsupervised versions of MCLEA.

4 Experiments

4.1 Experimental Setup

Datasets. Five EA datasets are adopted for evaluation, including three bilingual datasets ZH/JA/FR-EN versions of DBP15K (Liu et al., 2021) and two cross-KG datasets FB15K-DB15K/YAGO15K (Liu et al., 2019). The detailed dataset statistics are listed in Table 5 in the Appendix. Note that not all entities have corresponding images and for those without images, MCLEA would assign random vectors for the visual modality, as the setting of EVA (Liu et al., 2021). As for DBP15K, 30% reference entity alignments are given as \mathcal{S} while as for cross-KG datasets, 20%, 50%, and 80% reference entity alignments are given (Liu et al., 2019).

Baselines. We compare the proposed MCLEA against 19 state-of-the-art EA methods, which can be classified into four categories: 1) *structure-based methods* that solely rely on structural information for aligning entities, including BootEA (Sun et al., 2018), MUGNN (Cao et al., 2019a), KECG (Li et al., 2019), NAEA (Zhu et al., 2019), and AliNet (Sun et al., 2020a); 2) *auxiliary-enhanced methods* that utilize auxiliary information to improve the performance, including MultiKE (Zhang et al., 2019), HMAN (Yang et al., 2019), RDGCN (Wu et al., 2019), AttrGNN (Liu et al., 2020), BERT-INT (Tang et al., 2020) and ERMC (Yang et al., 2021); 3) *multi-modal methods* that combine the multi-modal features to generate entity representations, including PoE (Liu et al., 2019), MMEA (Chen et al., 2020a), HMEA (Guo et al., 2021), and EVA (Liu et al., 2021); 4) *unsupervised methods*, including RREA (Mao et al., 2020b), MRAEA (Mao et al., 2020a), EASY (Ge et al., 2021), and SEU (Mao et al., 2021).

Implementation Details. The hidden size of each layer of GAT is 300, while the embedding size of the other modalities is 100. We use the AdamW optimizer with a learning rate of 5×10^{-4} to update the parameters. The number of training epochs is 1000 with early-stopping and the batch size is 512. The hyper-parameters τ_1, τ_2 are set to 0.1

and 4.0, respectively. To keep in line with previous works, we use the same entity name translations and word vectors provided by Xu et al. (2019) for bilingual datasets. As for cross-KG datasets, we do not consider surface forms for fair comparison. For visual embedding, we adopt the pre-processed image features provided by Liu et al. (2021) and Chen et al. (2020a) for bilingual datasets and cross-KG datasets, where the former uses ResNet-152 as the pre-trained backbone, while the latter uses VGG-16. Previous work has revealed that surface forms are quite helpful for entity alignment (Liu et al., 2020). For fair comparison, we divide the supervised methods on bilingual datasets into two groups based on whether surface forms are used, and we implement an MCLEA variant (MCLEA \dagger) where the name embedding is removed. For the unsupervised setting, we implement two variants, MCLEA-V and MCLEA-N, which generate pseudo-alignment seeds based on the similarities of images and names, respectively.

Evaluation. We rank matching candidates of each to-be-aligned entity and use the metrics of Hits@1 (H@1), Hits@10 (H@10), and mean reciprocal rank (MRR). In the following tables, the best results are **in bold** with the second best results underlined, and “Improv. best %” denotes the relative improvement of MCLEA over the best baseline.

4.2 Overall Results

Table 1, Table 2, and Table 3 report the performance of MCLEA against different baselines on different datasets with different settings. Overall, MCLEA and its variants mostly perform the best across all the datasets on all the metrics.

Table 1 reports the performance of MCLEA against the supervised baselines on bilingual datasets in the settings of w/ and w/o surface forms. Compared with the first group without using surface forms, MCLEA \dagger brings about 5.2% to 7.2% relative improvement in H@1 over the best baseline EVA. The superiority of MCLEA confirms that the proposed contrastive representation learning substantially promotes the performance. Specifically, compared with the second group with surface forms involvement, there are two notable observations. On one hand, MCLEA shows a clear improvement when combined with name embedding, suggesting that entity names provide useful clues for entity alignment, which has been revealed in

	Models	DBP15K _{ZH-EN}			DBP15K _{JA-EN}			DBP15K _{FR-EN}		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
w/o SF	BootEA (Sun et al., 2018)	.629	.847	.703	.622	.854	.701	.653	.874	.731
	KECG (Li et al., 2019)	.478	.835	.598	.490	.844	.610	.486	.851	.610
	MUGNN (Cao et al., 2019a)	.494	.844	.611	.501	.857	.621	.495	.870	.621
	NAEA (Zhu et al., 2019)	.650	.867	.720	.641	.873	.718	.673	.894	.752
	AliNet (Sun et al., 2020a)	.539	.826	.628	.549	.831	.645	.552	.852	.657
	EVA (Liu et al., 2021)	.761	.907	.814	.762	.913	.817	.793	.942	.847
	MCLEA† (Ours)	.816	.948	.865	.812	.952	.865	.834	.975	.885
	Improv. best %	7.2	4.5	6.3	6.6	4.3	5.9	5.2	3.5	4.5
w/ SF	MultiKE (Zhang et al., 2019)	.437	.516	.466	.570	.643	.596	.714	.761	.733
	HMAN (Yang et al., 2019)	.562	.851	–	.567	.969	–	.540	.871	–
	RDGCN (Wu et al., 2019)	.708	.846	–	.767	.895	–	.886	.957	–
	AttrGNN (Liu et al., 2020)	.777	.920	.829	.763	.909	.816	.942	.987	.959
	BERT-INT (Tang et al., 2020)	<u>.968</u>	<u>.990</u>	<u>.977</u>	<u>.964</u>	<u>.991</u>	<u>.975</u>	<u>.992</u>	<u>.998</u>	<u>.995</u>
	ERMC (Yang et al., 2021)	.903	.946	.899	.942	.944	.925	.962	.982	.973
	MCLEA (Ours)	.972	.996	.981	.986	.999	.991	.997	1.00	.998
	Improv. best %	0.4	0.6	0.4	2.3	0.8	1.6	0.5	0.2	0.3

Table 1: Comparative results of MCLEA without (w/o) and with (w/) surface forms (SF) against strong supervised methods on three bilingual datasets, and † denotes an MCLEA variant without name embedding.

previous work (Zhang et al., 2019; Liu et al., 2020; Ge et al., 2021). On the other hand, MCLEA still shows slightly better performance than the best baseline BERT-INT with 0.4% to 2.3% relative improvement in H@1 nevertheless with far fewer parameters (13M *vs.* 110M). This also reveals that MCLEA can effectively model robust entity representations instead of attaching over-parameterized encoders. Noteworthily, BERT-INT relies heavily on entity descriptions to fine-tune BERT, but entity descriptions may not be available for every entity, and collecting them is labor-intensive, limiting the scope of its application.

Table 2 shows the comparison of multi-modal methods on two cross-KG datasets, which provides direct evidence of the effectiveness of MCLEA. When 20% training seeds are given, MCLEA outperforms the best baseline MMEA with 67.9% higher in H@1, 30.3% higher in H@10, and 49.6% higher in MRR. The performance gains are still significant when 50% and 80% alignment seeds are given. It is worth noting that the performance gains reach the highest in the 20% setting and MCLEA (20%) obtains comparable results to EVA (80%), indicating that MCLEA could better utilize the minimum number of alignment seeds to obtain effective representations. We also find that MMEA greatly outperforms EVA, we speculate that the cross-KG datasets are quite heterogeneous (w.r.t. the number of relations) compared to bilingual datasets, as shown in Table 5 in the Appendix, and the structure encoder of EVA struggles to model heterogeneous

information and EVA cannot utilize the numerical knowledge in cross-KG datasets, which is well exploited in MMEA.

	Models	FB15K-DB15K			FB15K-YAGO15K		
		H@1	H@10	MRR	H@1	H@10	MRR
20%	PoE	.126	.251	.170	.113	.229	.154
	HMEA	.127	.369	–	.105	.313	–
	MMEA	.265	.541	.357	.234	.480	.317
	EVA*	.134	.338	.201	.098	.276	.158
	MCLEA (Ours)	.445	.705	.534	.388	.641	.474
	Improv. best %	67.9	30.3	49.6	65.8	33.5	49.5
50%	PoE	.464	.658	.533	.347	.536	.414
	HMEA	.262	.581	–	.265	.581	–
	MMEA	.417	.703	.512	.403	.645	.486
	EVA*	.223	.471	.307	.240	.477	.321
	MCLEA (Ours)	.573	.800	.652	.543	.759	.616
	Improv. best %	23.5	13.8	22.3	34.7	17.7	26.7
80%	PoE	.666	.820	.721	.573	.746	.635
	HMEA	.417	.786	–	.433	.801	–
	MMEA	.590	.869	.685	.598	.839	.682
	EVA*	.370	.585	.444	.394	.613	.471
	MCLEA (Ours)	.730	.883	.784	.653	.835	.715
	Improv. best %	9.6	1.6	8.7	9.2	-0.4	4.8

Table 2: Experimental results on two cross-KG datasets where $X\%$ represents the percentage of reference entity alignments used for training. The symbol * denotes the reproduced results.

When compared to the unsupervised methods in Table 3, both MCLEA variants perform slightly better than the best baseline with performance gain in H@1 varying from 0.7% to 6.7%. Note that using image (-V) or name (-N) similarities to produce seeds leads to almost identical results, demonstrating the effectiveness of such simple rules to enable MCLEA in the unsupervised setting.

Models	DBP15K _{ZH-EN}			DBP15K _{JA-EN}			DBP15K _{FR-EN}		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
MRAEA (Mao et al., 2020a)	.778	.832	–	.889	.927	–	.950	.970	–
RREA (Mao et al., 2020b)	.822	.964	–	.918	.978	–	.963	.992	–
EASY (Ge et al., 2021)	.898	.979	.930	.943	.990	.960	.980	.998	.990
SEU (Mao et al., 2021)	.900	.965	.924	.956	.991	.969	.988	.999	.992
MCLEA-V (Ours)	.959	.995	.974	.977	.999	.987	.990	1.00	.994
MCLEA-N (Ours)	.960	.994	.974	.983	.999	.990	.995	1.00	.997
Improv. best %	6.7	1.6	4.7	2.8	0.8	2.2	0.7	0.1	0.5

Table 3: Unsupervised experimental results on three bilingual datasets, where -V and -N denote different methods to generate pseudo-alignment seeds.

4.3 Model Analysis

Ablation study. The ablation experiments are performed on two bilingual datasets and the results are presented in Table 4. We first examine the individual contribution of different modalities. The removal of different modalities has varying degrees of performance drop, and entity names have shown the primary importance with the most significant drop, which is in line with the previous findings (Mao et al., 2020a; Ge et al., 2021). The structural information also shows its stable effectiveness across different datasets and other modalities make a slight contribution to MCLEA. Especially, visual information can play a more pronounced role in the absence of surface forms (Chen et al., 2020a; Liu et al., 2021). Furthermore, we inspect various training strategies in MCLEA. It dramatically degrades the performance when removing the ICL from MCLEA, which indicates the importance of ICL in learning the intra-modal proximity. The IAL learns the interdependence between different modalities and is also beneficial to our model. Training MCLEA without the iterative strategy and replacing the uncertainty mechanism with uniform weights (i.e., w/o uncertainty) also cause decreases in performance. Overall, the ablation experiments validate the involvement of these modalities and training strategies with empirical evidence.

Impact of hyper-parameters τ_1, τ_2 . We investigate the effects of hyper-parameters τ_1, τ_2 on DBP15K_{ZH-EN}. As shown in Figure 2, different values of τ_1 have drastic effects on MCLEA, especially in terms of H@1 and MRR, which is because τ_1 controls the strength of penalties on hard negative samples and an appropriate τ_1 is conducive to learning discriminative entity embeddings. On the other hand, we observe lower variance in the performance w.r.t. τ_2 and the performance saturates when

	Models	DBP15K _{ZH-EN}			DBP15K _{JA-EN}		
		H@1	H@10	MRR	H@1	H@10	MRR
	MCLEA	.972	.996	.981	.986	.999	.991
Modalities	w/o structure	.883	.956	.909	.947	.980	.959
	w/o relation	.967	.995	.978	.985	.999	.991
	w/o attribute	.961	.994	.974	.983	.999	.991
	w/o name	.816	.948	.865	.812	.952	.865
	w/o visual	.968	.994	.978	.985	.999	.991
Training	w/o ICL	.782	.892	.818	.813	.909	.844
	w/o IAL	.966	.995	.977	.980	.998	.987
	w/o iter. strategy	.942	.991	.960	.964	.995	.976
	w/o uncertainty	.969	.996	.980	.984	.999	.990

Table 4: Ablation study on two bilingual datasets.

$\tau_2 = 4.0$. The KL divergence establishes the associations between different modalities, within which τ_2 regulates the softness of the alignment distribution produced by input embedding and transfers the generalization capability of the joint embedding to uni-modal embedding.

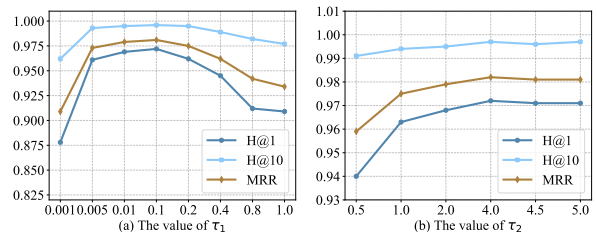


Figure 2: Performance comparison with different values of τ_1, τ_2 .

Similarity Distribution of Representations. To investigate the effectiveness of entity representations, we experiment MCLEA with and without ICL/IAL on DBP15K_{ZH-EN} and produce the visualization in Figure 3 by averaging the similarity distribution of the test entities and their predicted counterparts for different modalities. It can be observed that in every modality, especially in structure and name, it holds a high top-1 similarity and a large similarity variance. More importantly, it meets our expectation that contrastive learning

(ICL and IAL) enables more discriminative entity learning in the joint representations.

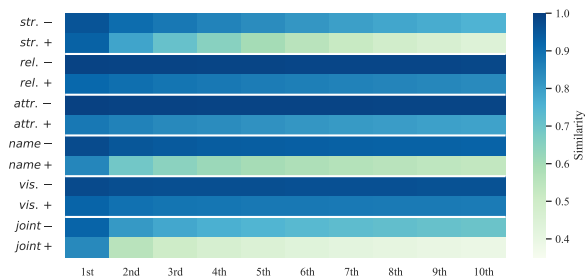


Figure 3: Similarity visualization of representations of test entities and their top-10 predicted counterparts. The vertical axis represents different modalities with (+) and without (-) ICL/IAL and the horizontal axis represents the index of ranked predictions.

5 Conclusion

This paper presented a novel method termed MCLEA to address the multi-modal entity alignment. MCLEA utilizes multi-modal information to obtain the joint entity representations and it is composed of two losses, intra-modal contrastive loss, and inter-modal alignment loss, to explore the intra-modal relationships and cross-modal interactions, respectively. We experimentally validated the state-of-the-art performance of MCLEA in several public datasets and its capability of learning more discriminative embedding space for entity alignment. For future work, we plan to explore more side information such as entity descriptions to boost alignment performance.

References

- Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, and Tat-Seng Chua. 2019a. **Multi-Channel Graph Neural Network for Entity Alignment**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1452–1461, Florence, Italy. Association for Computational Linguistics.
- Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019b. **Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences**. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 151–161. ACM.
- Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020a. **MMEA: Entity Alignment for Multi-modal Knowledge Graph**. In *KSEM*.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. **A Simple Framework for Contrastive Learning of Visual Representations**. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Congcong Ge, Xiaoze Liu, Lu Chen, Baihua Zheng, and Yunjun Gao. 2021. **Make It Easy: An Effective End-to-End Entity Alignment Framework**. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 777–786, New York, NY, USA. Association for Computing Machinery.
- Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. **Multi-modal Entity Alignment in Hyperbolic Space**. *Neurocomputing*, 461:598–607.
- Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. **Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text**. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), USA, February 2-7, 2018*, pages 4832–4839. AAAI Press.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. **Deep Residual Learning for Image Recognition**. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. **Distilling the Knowledge in a Neural Network**. *ArXiv preprint*, abs/1503.02531.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. **Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics**. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, USA, June 18-22, 2018*, pages 7482–7491. IEEE Computer Society.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. **Supervised Contrastive Learning**. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. **A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions**. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4483–4491. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia—A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 6(2).
- Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. **Semi-supervised Entity Alignment via Joint Knowledge Embedding Model and Cross-graph Model**. 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)*, pages 2723–2732, Hong Kong, China. Association for Computational Linguistics.
- Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual Pivoting for (Unsupervised) Entity Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4257–4266.
- Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *The Semantic Web*, pages 459–474, Cham. Springer International Publishing.
- Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, Zhiyuan Liu, and Tat-Seng Chua. 2020. **Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6355–6364, Online. Association for Computational Linguistics.
- Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. **YAGO3: A Knowledge Base from Multilingual Wikipedias**. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org.
- Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. **From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2853, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020a. **MRAEA: An Efficient and Robust Entity Alignment Approach for Cross-lingual Knowledge Graph**. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 420–428. ACM.
- Xin Mao, Wenting Wang, Huimin Xu, Yuanbin Wu, and Man Lan. 2020b. **Relational Reflection Entity Alignment**. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1095–1104. ACM.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **GloVe: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Zhiyuan Qi, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Yuejia Xiang, Ningyu Zhang, and Yefeng Zheng. 2021. **Unsupervised Knowledge Graph Alignment by Probabilistic Reasoning and Semantic Embedding**. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2019–2025. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Yinghui Shi, Meng Wang, Zhang Ziheng, Zhenxi Lin, and Yefeng Zheng. 2022. Probing the Impacts of Visual Context in Multimodal Entity Alignment. In *Web and Big Data - 6th International Joint Conference, APWeb-WAIM 2022*.
- Zequan Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. **Bootstrapping Entity Alignment with Knowledge Graph Embedding**. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4396–4402. ijcai.org.
- Zequan Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020a. Knowledge Graph Alignment Network with Gated Multi-hop Neighborhood Aggregation. In *AAAI*, volume 34.
- Zequan Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020b. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. In *Proceedings of the 46th International Conference on Very Large Databases*, pages 2326–2340, Tokyo, Japan.
- Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. 2020. **BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment**. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3174–3180. ijcai.org.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. **Graph Attention Networks**. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: A Large-scale, Comprehensive Multi-modal Knowledge Graph. *Big Data Research*, 22.

- Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. [Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective](#). *MM '21*, page 2735–2743, New York, NY, USA. Association for Computing Machinery.
- Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. [Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5278–5284. ijcai.org.
- Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. [OntoEA: Ontology-guided Entity Alignment via Joint Knowledge Graph Embedding](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1117–1128, Online. Association for Computational Linguistics.
- Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019. [Cross-lingual Knowledge Graph Alignment via Graph Matching Neural Network](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3156–3161, Florence, Italy. Association for Computational Linguistics.
- Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. [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)*, pages 4431–4441, Hong Kong, China. Association for Computational Linguistics.
- Jinzhu Yang, Ding Wang, Wei Zhou, Wanhui Qian, Xin Wang, Jizhong Han, and Songlin Hu. 2021. [Entity and Relation Matching Consensus for Entity Alignment](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2331–2341.
- Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. [Multi-view Knowledge Graph Embedding for Entity Alignment](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5429–5435. ijcai.org.
- Ziheng Zhang, Hualuo Liu, Jiaoyan Chen, Xi Chen, Bo Liu, Yuejia Xiang, and Yefeng Zheng. 2020. [An Industry Evaluation of Embedding-based Entity Alignment](#). In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 179–189, Online. International Committee on Computational Linguistics.
- Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. 2019. [Neighborhood-Aware Attentional Representation for Multilingual Knowledge Graphs](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1943–1949. ijcai.org.

Appendix

A Derivation for Adaptively Weighted Multi-task Loss

In this section, we treat Eq. (10) as a multi-task loss function and combine multiple objectives using homoscedastic uncertainty (Kendall et al., 2018), allowing us to automatically learn the relative weights of each loss.

Firstly, the ICL can actually be regarded as a classification loss with negative log-likelihood, i.e., predicting whether two entities are equivalent. Here, we rewrite the loss function of ICL as follows (for simplicity, here we omit the modality index and the inner-graph negative samples, and only consider the unidirectional version):

$$\begin{aligned}\mathcal{L}^{\text{ICL}} &= -\mathbb{E}_{i \in \mathcal{B}} \log q(e_1^i, e_2^i) \\ &= -\mathbb{E}_{i \in \mathcal{B}} \log P(c = 1 | f^{\mathbf{W}}(e_1^i, e_2^i)),\end{aligned}\quad (12)$$

where $c = 1$ means that the two input entities are equivalent, otherwise $c = 0$; $f^{\mathbf{W}}(\cdot, \cdot)$ is the model output with parameter \mathbf{W} . Following (Kendall et al., 2018), we adapt the negative log-likelihood to squash a scaled version of the model output with an uncertainty scalar σ through a softmax function:

$$\begin{aligned}& -\log P(c = 1 | f^{\mathbf{W}}(e_1^i, e_2^i), \sigma) \\ &= -\log \text{Softmax}\left(\frac{1}{\sigma^2} f^{\mathbf{W}}(e_1^i, e_2^i)\right) \\ &= -\frac{1}{\sigma^2} f^{\mathbf{W}}(e_1^i, e_2^i) + \log \sum_{j \neq i} \exp\left(\frac{1}{\sigma^2} f^{\mathbf{W}}(e_1^i, e_2^j)\right),\end{aligned}\quad (13)$$

where e_2^j with $j \neq i$ is the cross-graph negative samples defined in the main paper.

Applying the same assumption in (Kendall et al.,

2018):

$$\begin{aligned}& \frac{1}{\sigma} \sum_{j \neq i} \exp\left(\frac{1}{\sigma^2} f^{\mathbf{W}}(e_1^i, e_2^j)\right) \\ & \approx \left(\sum_{j \neq i} \exp\left(f^{\mathbf{W}}(e_1^i, e_2^j)\right)\right)^{\frac{1}{\sigma^2}},\end{aligned}\quad (14)$$

we can simplify Eq. (13) to:

$$\begin{aligned}& -\log P(c = 1 | f^{\mathbf{W}}(e_1^i, e_2^i), \sigma) \\ & \approx -\frac{1}{\sigma^2} f^{\mathbf{W}}(e_1^i, e_2^i) + \frac{1}{\sigma^2} \log \sum_{j \neq i} \exp(f^{\mathbf{W}}(e_1^i, e_2^j)) + \log(\sigma) \\ & = -\frac{1}{\sigma^2} \log P(c = 1 | f^{\mathbf{W}}(e_1^i, e_2^i)) + \log(\sigma),\end{aligned}\quad (15)$$

where σ can be interpreted as the relative weight of the loss and automatically learned with stochastic gradient descent.

On the other hand, the IAL defines the KL divergence over the output distribution between joint embedding and uni-modal embedding (we omit the modality index and only consider the unidirectional version for simplicity):

$$\begin{aligned}\mathcal{L}^{\text{IAL}} &= \mathbb{E}_{i \in \mathcal{B}} \text{KL}(q'_o(e_1^i, e_2^i) \parallel q'(e_1^i, e_2^i)) \\ &= \mathbb{E}_{i \in \mathcal{B}} q'_o(e_1^i, e_2^i) \log \frac{q'_o(e_1^i, e_2^i)}{q'(e_1^i, e_2^i)} \\ &= \mathbb{E}_{i \in \mathcal{B}} [q'_o(e_1^i, e_2^i) \log q'_o(e_1^i, e_2^i) \\ & \quad - q'_o(e_1^i, e_2^i) \log q'(e_1^i, e_2^i)],\end{aligned}\quad (16)$$

where $q'_o(e_1^i, e_2^i)$ and $q'(e_1^i, e_2^i)$ represent the output predictions of joint embedding and the uni-modal embedding, respectively. Since we only back-propagate through $q'(e_1^i, e_2^i)$ in Eq. (9), \mathcal{L}^{IAL} is equivalent to calculating the cross-entropy loss over the two distributions:

$$\mathcal{L}^{\text{IAL}} = -q'_o(e_1^i, e_2^i) \log q'(e_1^i, e_2^i).\quad (17)$$

Dataset	KG	#Ent.	#Rel.	#Attr.	#Rel tr.	#Attr tr.	#Image	#Ref.
DBP15K _{ZH-EN} (Liu et al., 2021)	ZH	19,388	1,701	8,111	70,414	248,035	15,912	15,000
	EN	19,572	1,323	7,173	95,142	343,218	14,125	
DBP15K _{JA-EN} (Liu et al., 2021)	JA	19,814	1,299	5,882	77,214	248,991	12,739	15,000
	EN	19,780	1,153	6,066	93,484	320,616	13,741	
DBP15K _{FR-EN} (Liu et al., 2021)	FR	19,661	903	4,547	105,998	273,825	14,174	15,000
	EN	19,993	1,208	6,422	115,722	351,094	13,858	
FB15K-DB15K (Liu et al., 2019)	FB15K	14,951	1,345	116	592,213	29,395	13,444	12,846
	DB15K	12,842	279	225	89,197	48,080	12,837	
FB15K-YAGO15K (Liu et al., 2019)	FB15K	14,951	1,345	116	592,213	29,395	13,444	11,199
	YAGO15K	15,404	32	7	122,886	23,532	11,194	

2583
Table 5: Dataset Statistics.

Therefore, similar to ICL, we can automatically learn the relative weight of IAL for each modality through task-dependent uncertainty. As mentioned above, the total loss in Eq. (10) can be rewritten as:

$$\mathcal{L} = \mathcal{L}_o^{\text{ICL}} + \sum_{m \in \mathcal{M}} \left(\frac{1}{\alpha_m^2} \mathcal{L}_m^{\text{ICL}} + \log \alpha_m \right) + \sum_{m \in \mathcal{M}} \left(\frac{1}{\beta_m^2} \mathcal{L}_m^{\text{IAL}} + \log \beta_m \right), \quad (18)$$

where α_m, β_m are learnable parameters. Large α_m (β_m) will decrease the contribution of $\mathcal{L}_m^{\text{ICL}}$ ($\mathcal{L}_m^{\text{IAL}}$) for the m -th modality, whereas small α_m (β_m) will

increase its contribution.

B Dataset Statistics

The detailed dataset statistics are listed in Table 5, including the number of entities (#Ent.), relations (#Rel.), attributes (#Attr.), number of relation triples (#Rel tr.) and attribute triples (#Attr tr.), number of images (#Image), and number of reference entity alignments (#Ref.). It is worth noting that not all entities have the associated images or the equivalent counterparts in the other KG.