RESEARCH PAPER

Faster Zero-shot Multi-modal Entity Linking via Visual-Linguistic Representation

Qiushuo Zheng[1], Hao Wen[2], Meng Wang[2,3], Guilin Qi[2,3,†] & Chaoyu Bai[1]

[1]School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China

[2]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China

[3]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 211189, China

Keywords: Knowledge Graph; Multi-modal Learning; Poly Encoders

Citation: Zheng Q.S., Wen H., Wang M. et al. Faster Zero-shot Multi-modal Entity Linking via Visual-Linguistic Representation. Data Intelligence 4(3), 493-508 (2022). DOI: 10.1162/dint_a_00146

Received:Nov.11, 2022 Revised: Jan. 10, 2022 Accepted: Feb.12, 2022

ABSTRACT

Multi-modal entity linking plays a crucial role in a wide range of knowledge-based modal-fusion tasks, i.e., multi-modal retrieval and multi-modal event extraction. We introduce the new *ZEro-shot Multi-modal Entity Linking* (ZEMEL) task, the format is similar to multi-modal entity linking, but multi-modal mentions are linked to unseen entities in the knowledge graph, and the purpose of zero-shot setting is to realize robust linking in highly specialized domains. Simultaneously, the inference efficiency of existing models is low when there are many candidate entities. On this account, we propose a novel model that leverages visual-linguistic representation through the co-attentional mechanism to deal with the ZEMEL task, considering the trade-off between performance and efficiency of the model. We also build a dataset named ZEMELD for the new task, which contains multi-modal data resources collected from Wikipedia, and we annotate the entities as ground truth. Extensive experimental results on the dataset show that our proposed model is effective as it significantly improves the precision from 68.93% to 82.62% comparing with baselines in the ZEMEL task.

1. INTRODUCTION

Traditional entity linking tasks usually focus on a single modal, such as text [1], image [2] or video [3]. However, contemporary society spreads news through multimedia, in this situation, the multi-modal entity

† Corresponding author: Guilin Qi (Email: gqi@seu.edu.cn; ORCID: 0000-0002-1957-6961).

linking task has emerged in our sight. The existing works [4] utilize the object detection and relation classification as the main method to achieve visual scene understanding, but these works still detect visual objects in coarse-grained concept level, i.e., categories. In many practical scenarios, such as news-reading and e-shopping, we require entity level detection for the fine-grained scenes understanding, named as multi-modal entity linking.

Taking Figure 1 as an example, the input of multi-modal entity linking contains contexts of different modal, we can use the consistency of semantic representations among the same entity mention in different modals, to jointly learn the entity feature representation and link it to the corresponding knowledge graph entity. Recently, substantial improvements to state-of-the-art benchmarks for multi-modal entity linking have been achieved by utilizing different modal features.

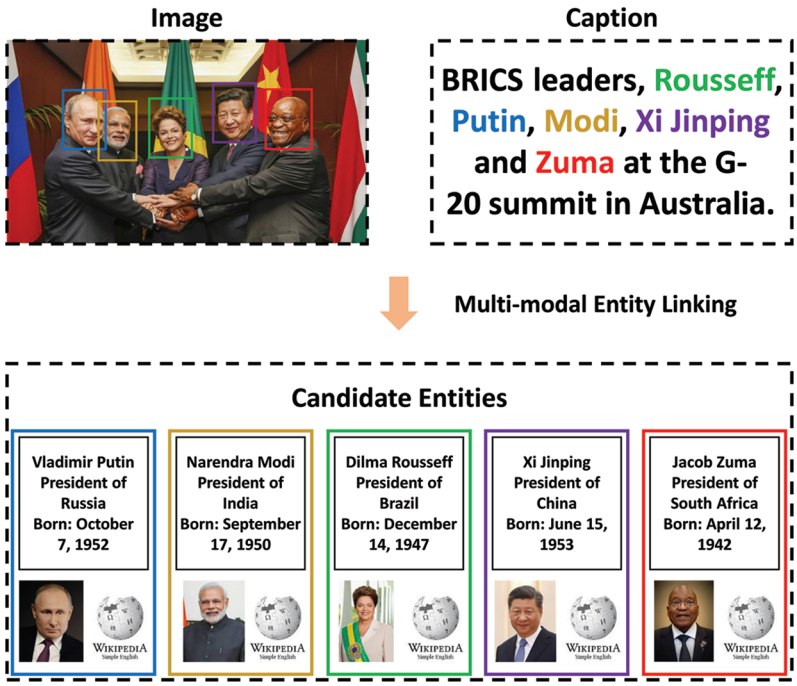


Figure 1. Example of multi-modal entity linking.

Although most of the previous work focused on linking to the general knowledge graph, it is usually desirable to link to specialized entity sets such as medical fields, industrial specific scenarios, and e-shopping platforms. Unfortunately, labeled data is not readily available and usually expensive to obtain for these specialized entity domains. Therefore, we need to develop a zero-shot multi-modal entity linking system, which can be extended to unseen specialized entities. Without the alias tables and frequency statistics, the task becomes substantially more challenging.

Simultaneously, existing entity linking systems have an urgent problem to be solved: with the introduction of deep neural networks and pre-trained language models, entity linking models become more and more complex, real-time response speed becomes relatively slow, and it is difficult to adapt to the requirement of the real scene. Therefore, we envisage using comprehensive metrics to measure multi-modal entity linking models, not only considering the performance of models but also measuring its response speed when applied in real-time systems. Measured metrics in our task via two axes: linking quality and linking speed, as scoring many candidates can be prohibitively slow.

In our paper, we propose a novel *zero-shot multi-modal entity linking* (ZEMEL) task, and construct a new dataset ZEMELD for it. The ZEMEL task mainly contains the representation of multi-modal context and the selection of candidate entities, which is similar to visual commonsense reasoning [5] and visual question answering [6] tasks. Each sample of data contains a linguistic caption and a visual image, and the corresponding entity linkings in visual-linguistic-knowledge graph modals are given as ground truth. We introduce a two-stage approach for the ZEMEL task, based on fine-tuned visual-linguistic representation architecture. In the first stage, we encoder the multi-modal contexts and candidate entities respectively via the visual-linguistic co-attentional mechanism, and get the aggregation vector for each encoder module. In the second stage, we integrate the features of each candidate entity into the context representation vector through an additional learned attention mechanism, to represent more global features. We evaluate the performance on ZEMELD dataset, our model achieves a nearly 12 point absolute gain on the test data, driven largely by the co-attentional mechanism.

The main contribution of our paper can be summarized as follows:

- We are the first to consider the zero-shot multi-modal entity linking task and have constructed a new dataset ZEMELD for evaluation.
- We propose a novel model to encoder the multi-modal context and candidates utilizing visual-linguistic co-attentional mechanism and simultaneously design a poly-architecture decoder module that achieves faster inference speed.
- We conduct extensive experiments to evaluate our model. Results on the constructed dataset show that our proposed method is effective as it significantly improves the precision from 68.93% to 82.62% comparing with the baselines.

The rest of this paper is organized as follows: Section 2 contains an analysis of the related work, Section 3 describes the process of dataset construction, we introduce formally the problem and present our model in Section 4, Section 5 shows our experimental results, and our conclusion is in Section 6.

2. RELATED WORK

This section discusses the existing related research in the following aspects: multi-modal entity linking, visual-linguistic representation, and efficiency of transformer architectures.

2.1 Multi-modal Entity Linking

The multi-modal entity linking task is mapping the objects in visual scenes to the corresponding entities in the knowledge graph (KG) by leveraging the different modality features, i.e., visual features, linguistic features, and KG features. [2] demonstrated the first comprehensive and open-source multimedia knowledge extraction system, which realized multi-modal tasks including multi-modal entity linking. [7] proposed an unsupervised algorithm for object detection in images, entity recognition in texts, entity linking to ontology, and entity mention in aligned visual-texts. [8] built a deep multi-modal network for social media posts disambiguation with the feature extracts from both the text and image contexts. [9] proposed a novel method to solve the named entity recognition problem for tweets that contain multi-modal information.

However, current models can not link the entities which have never been seen before, considering the ZEMEL task is necessary.

2.2 Visual-Linguistic Representation

BERT [10] has demonstrated effective representation learning using self-supervised tasks, the pre-trained model can then be fine-tuning for a variety of supervised tasks. The existing models [11, 12, 13, 14, 15] employ BERT-like objectives to learn multi-modal representations from a concatenated-sequence of visual region features and language token embeddings. A single-stream approach takes visual input and text into a BERT-like transformer-based encoder, i.e., VisualBERT [16], VL-BERT [14] and Unicoder-VL [15]. Two-stream approaches need an additional fusion step, i.e., ViLBERT [17] and LXMERT [18] employ two modality-specific streams for images.

With the help of the visual-linguistic pre-trained model, we can achieve better results than current models in multi-modal tasks using joint representation.

2.3 Efficiency of Transformer Architectures

The current models aimed to pursue excellent accuracy performance, making the structure more and more complex, resulting in the inability to meet the requirements of real-time systems. [19] proposed a Bi-architecture transformer as a general sentence encoder. The architecture is learned with multiple tasks including the unsupervised Skip-Thought task [20], the supervised conversation input-response task [21], and the supervised sentence classification SNLI task [22]. [23] study the Poly-architecture model to give an improved trade-off between efficiency and accuracy based on Bi-encoder models and Cross-encoder models, which learns global rather than token level self-attention features.

Efficiency is an important indicator to measure the models, introducing the poly-encoders into the ZEMEL task can greatly reduce the response time of the model and obtain higher real-time efficiency.

3. DATASET CONSTRUCTION

Due to the absence of data, it is an urgent mission to construct a comprehensive and standard zero-shot multi-modal entity linking dataset. We construct a novel dataset using multi-modal data from Wikipedia, the dataset contains 22K images about 16K entities. The general statistics of our ZEMELD dataset are given in Table 1.

Table 1. Brief statistics of ZEMELD dataset.

ZEMELD dataset statistics:	
Number of images	22,156
Number of entities	28,930
Number of unique entities	17,428
Number of training data	15,500
Number of validation data	3,000
Number of testing data	3,656
Average caption character length	48.98
Average caption word length	8.01
Max unique entities in one image	5

3.1 Entity and Image Collection

We generate a list of 80K entities from Wikipedia and collect relevant multi-modal description information. In total, we collect 30K relevant images with captions from Wikipedia.

3.2 Image Preprocessing

First of all, we screen the image quality and remove those images with a low pixel. Second, we delete images that do not contain an entity and too many entities. The former does not meet our experimental requirements, and the latter often has problems with occlusion and deformation, which is not meaningful for the proposed task.

3.3 Human Annotation

We use the object detection to mark all the entities in the images, for each detected entity, the bounding box consists of 4 predictions: x_1, y_1, x_2, and y_2. Then we provide the preprocessed images to human annotators, we ask the human annotators to identify the entity of the object bounding box and give the entry linking of the entities in Wikipedia. Some examples of the annotated dataset are shown in Figure 2.

Figure 2. A selection of images from ZEMELD dataset. The bounding box position of each entity is marked in the image, and entity names and bounding box coordinates are given below.

4. METHODOLOGY

In this section, we first define the ZEMEL task, then we detail how to leverage multi-modal resources to solve the ZEMEL task, finally, we introduce how to make our model faster through the settings of poly-architecture. Our ZEMEL model is shown in Figure 3.

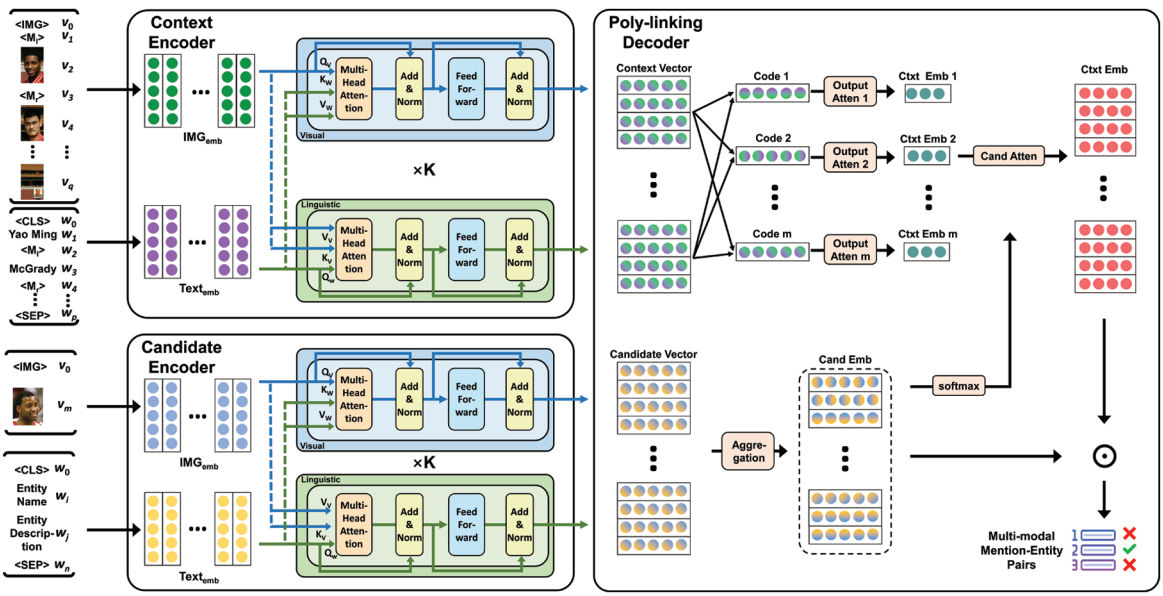


Figure 3. Overview of ZEMEL model, which consists of three parts independently, namely the context encoder module, the candidate encoder module and the poly-linking module.

4.1 Definition and Task Formulation

We assume that there is a multi-modal entity knowledge graph $\mathcal{E} = \{(u_n, t_n, e_n)\}_{n\in(1,K)}$, where u_n is a visual description of entity e_n, t_n is a linguistic description of entity e_n and the count of dictionary entity is K. Given an input of multi-modal data, including linguistic document $\mathcal{D} = \{w_1, \ldots, w_p\}$ of words and visual image $\mathcal{I} = \{v_1, \ldots, v_q\}$ of object regions, we can get a list of linguistic entity mentions $M^T = \{m_1{}^T, \ldots, m_a{}^T\}$ from $\mathcal{D}$ and visual entity mentions $M^V = \{m_1{}^V, , m_b{}^V\}$ from $\mathcal{I}$, the output of a multi-modal entity linking model is a list of mention-entity pairs $\mathcal{P} = \{(m_i{}^T, m_j{}^V, e_k)\}_{i,j\in[1,min(a,b)]}$, where each entity is an entry in a multi-modal knowledge graph.

We define a set as $\mathcal{S} = (D_\mathcal{S}, I_\mathcal{S}, E_\mathcal{S})$, that $D_\mathcal{S}$, $I_\mathcal{S}$ are respectively the linguistic documents and the visual descriptions in the set, and $E_\mathcal{S}$ is a multi-modal entity dictionary related to $\mathcal{S}$. ZEMEL task is similar to multi-modal entity linking task definition, except that knowledge graphs and entities are separated in training and testing time. Formally, we denote S_{train} and S_{test} to be the set in training and testing, it is required that $S_{train} \cap S_{test} = \varnothing$.

4.2 Context Encoder Module

Joint visual-linguistic embedding is the basic for the ZEMEL task, where multi-modal inputs are simultaneously processed for joint visual-linguistic understanding.

Our context encoder aims to learn the joint representations of visual and linguistic content from paired labeled static images and corresponding descriptive captions. Our context encoder is mainly used to jointly pay attention to different modal features, and form a multi-modal context encoding representation. We use an architecture based on a deep multi-modal transformer ViLBERT [17] which has achieved state-of-the-art performance on some multi-modal tasks.

The input of our context encoder module consists of two separate streams for vision and language, given an image $\mathcal{I}$ represented as a list of visual regions $v_1, \ldots, v_q$, and a linguistic description $\mathcal{D}$ represented as a list of words $w_1, \ldots, w_p$. We generate visual feature embedding of v_i by pre-trained object detection network, the spatial feature is represented by bounding box position and the fraction of bounding box covered, we project the spatial feature to the visual embedding space and summed them. The linguistic tokens w_j are encoded by the pre-trained language model. For each modal, the representation of mentions is surrounding by the context. Specifically, we reconstruct the input of each mention example as:

$$[CLS]\ \text{ctxt}_l\ [M_l]\ \text{mention}\ [M_r]\ \text{ctxt}_r\ [SEP]$$

where mention, ctxt_l, ctxt_r are the token representations of the mention, the left, and the right context respectively. In particular, we use the special tokens $[M_l]$ and $[M_r]$ to tag the mention.

Our context encoder consists of two parallel BERT-based models that operate on image regions and linguistic segments. Each steam is a series of co-attentional transformer layers, we introduce it to make the

information exchange between modals possible. We learn from the co-attention mechanism proposed in ViLBERT to interconnect between vision and language. Given the intermediate visual and linguistic representations H_V^l and H_W^l, our context encoder block computes query, key, and value matrices normally as a standard transformer block. Especially, the key and value matrices from each modal are sent to the other modal's transformer block as the input matrices, each modal transformer block uses interactive query, key, and value for multi-head self-attention. As a result, the multi-head self-attention block produces attention vectors for each modal that depends on another modal image-conditional language attention is performed in the visual stream, and language-conditional image attention is performed in the language stream.

These input tokens are encoded to produce final representations h_{v1}, …, h_{vq} for vision and h_{w1}, …, v_{wp} for language, we simplify them to h_0, …, h_T for layers. Let H^l be a matrix with rows $h_0^l,...,h_T^l$ corresponding to the intermediate representations after the l-th layer.

Finally, the visual and linguistic output can be obtained through the multi-head self-attention context encoder, and we contact the output to get a final fusion vector for multi-modal context representation.

4.3 Candidate Encoder Module

In particular, because of the zero-shot setting of the ZEMEL task, we do not have alias tables like standard entity linking task, we design a novel approach for candidate generation, it is the fusion of indicator-based and rule-based algorithms.

For indicator, we use BM25[24], a variant of TF-IDF, to measure the similarity between mentions and candidate entity names for training and evaluation. For rules, we generate the candidate entity using a partial matching strategy. The rules are following:

1. The entity name has several words with the context entity mention in common.
2. The entity name is wholly contained in or contains the context entity mention.
3. The entity name exactly pairs the first letters of all words in the context entity mention.

Finally, we merge the results obtained from both algorithms as candidate entities. The input of our candidate encoder also includes vision modal and language modal as:

[*CLS*] entity name [*ENT*] entity description [*SEP*]

where [*ENT*] is a special token to separate entity name and entity description. The visual input is the image information of the candidate entity, and the linguistic input includes the entity name and entity description. We use the same structure as our context encoder to generate the vector representation of the candidate entities. For each entity mention, we have K candidate entities depending on our experimental setting.

4.4 Poly-linking Decoder Module

After we get context vector and candidate vector, we obtain the right entity through a ranking model. Since real-time systems need to consider responding speed, we propose a poly-architecture linking model, which increases the speed of our model significantly without losing too much precision. The candidate is represented by a vector, as in a Bi-encoder, which allows caching candidates to infer faster, while the input context is encoded with the candidates jointly, as in a Cross-encoder, allowing global information to be extracted.

Our model aims to obtain the best results of both worlds from Bi-architecture and Cross-architecture, it uses two separate multi-modal transformers for context and candidate like a Bi-architecture. Each candidate entity is encoded into a separate vector representation y_{cand_i}, so our model can be implemented using a pre-computed cache for the candidate. After that, the candidate vectors are aggregated into candidate embedding for attentional representation. However, the context vector, which is much longer and more complex than the candidate vector, is represented by m codes ($y^1_{cand},...,y^m_{cand}$) instead of one long vector in Bi-encoder, where m is a hyper-parameter affecting the inference speed of entity linking. Our model trains m learned parameter codes (p_1, …, p_m), which p_i reflects the weight of each position in previous layer, to obtain the m global codes. Formulaically, we get y^i_{ctxt} as follow:

$$y^i_{ctxt} = \sum_{j=1}^{N} w_j^{p_i} \cdot h_j \tag{1}$$

$$(w_1^{p_i},...,w_N^{p_i}) = \text{softmax}(p_i \cdot h_1,...,p_i \cdot h_N) \tag{2}$$

In the training process, we randomly initialize the m context codes, they are updated in the fine-tuning step iteratively. We use the vector representation of each candidate entity y_{cand_i} as a query to generate a context representation based on the candidate attention mechanism as follow:

$$y_{ctxt} = \sum_{i=1}^{m} v_i \cdot y^i_{ctxt} \tag{3}$$

$$(v_1,...,v_m) = \text{softmax}(y_{cand_i} \cdot y^1_{ctxt},...,y_{cand_i} \cdot y^m_{ctxt}) \tag{4}$$

The final score of a candidate $cand_i$ is calculated by dot-product between the vector representation of context and candidate entity as follow:

$$s(ctxt, cand_i) = y_{ctxt} \cdot y_{cand_i} \tag{5}$$

Our poly-architecture model is trained to minimize the cross-entropy loss, in which logits are the scores of other candidates $y_{ctxt} \cdot y_{cand_i},\ldots,y_{ctxt} \cdot y_{cand_n}$, where $cand_1$ is the correct label and the others are false candidate entity.

5. EXPERIMENTS

In this section, we perform an empirical study of our model with state-of-the-art methods on the ZEMELD dataset to test our model architectures.

5.1 Experiments Setting

Task: Given a visual description and an accompanying caption, our goal is to link both the image bounding box and the textual mentions to the corresponding KG entities in Wikipedia. For our task is zero-shot learning, the entities in the test set have never been seen before in the training set.

Task Settings: In this work, our goal is to link the multi-modal mentions to the corresponding KG entities, while the object detection and named entity recognition is not our main objective. Therefore, we conduct two sub-tasks of ZEMEL to evaluate our proposed method. **(1) V-T-KG linking:** Without given the correspondence between visual modal and linguistic modal, predict the linking among three modals, and the linked entities in the three modals are required to be correct at the same time. **(2) V&T-KG linking:** Given the multi-modal entity mentions and their correspondences between them, only predict the entities of knowledge graph modal in the links.

Evaluation Metrics: The primary metric of our evaluation is the precision, recall, and micro-F_1 score of the ZEMEL task. We measure metrics where each test example has N possible candidates to select from, abbreviated to -/N.

5.2 Baselines

We report the performance of the following state-of-the-art multi-modal entity linking models and named entity disambiguation models as baselines and configurations of our proposed model. We re-implemented the baselines for ZEMEL models.

- **GAIA-VEL:** [2] constructs a fine-grained multi-modal knowledge extraction system, which realizes the multi-modal entity linking function.
- **VTKEL:** [7] presents a purely unsupervised algorithm for the solution of the Visual-linguistic-Knowledge Entity Linking tasks.
- **DZMNED:** [8] uses an attentional LSTM model for multi-modal named entity disambiguation task in social media posts.
- **CBCFuFiC:** [9] propose a model for entity linking in tweets that contain multi-modal information.
- **(proposed) ZEMEL:** the model proposed in our paper as described in Figure 3.

5.3 Results and Discussion

Main Results: Table 2 shows the precision, recall, and micro-F_1 score under the situations of 5 and 100 candidate entities on the ZEMELD dataset. For the V-T-KG linking task, we measure precision, recall, and

micro-F_1 score for experiments. For the V&T-KG linking task, because the correspondences between vision bounding boxes and linguistic mentions are given, we only measure the precision. Our model achieves the precision of 85.46% in the V&T-KG task and 82.62% in the V-T-KG task. From the holistic perspective, we find that given the correspondence between vision and language can improve the precision of the results.

Table 2. Test performance of our proposed model and models from prior work on our dataset. The evaluation task contain two sub-tasks, V-T-KG linking and V&T-KG linking.

Sub-Task	Model	Metric					
		Pre/5	R@1/5	Micro-F_1/5	Pre/100	R@1/100	Micro-F_1/100
V-T-KG	GAIA-VEL	64.95%	70.13%	67.44%	65.94%	74.16%	69.81%
V-T-KG	VTKEL	60.16%	69.29%	64.40%	62.98%	70.94%	66.72%
V-T-KG	DZMNED	64.84%	72.84%	68.61%	66.68%	74.49%	70.37%
V-T-KG	CBCFuFiC	68.93%	72.94%	70.88%	72.99%	73.81%	73.40%
V-T-KG	**ZEMEL**	**82.62%**	**80.47%**	**81.53%**	**83.67%**	**81.94%**	**82.80%**
V&T-KG	GAIA-VEL	70.19%	–	–	71.42%	–	–
V&T-KG	VTKEL	64.08%	–	–	66.84%	–	–
V&T-KG	DZMNED	71.97%	–	–	73.84%	–	–
V&T-KG	CBCFuFiC	72.51%	–	–	72.57%	–	–
V&T-KG	**ZEMEL**	**85.46%**	–	–	**86.93%**	–	–

Compared with the state-of-the-art models in our full task (V-T-KG linking), our model achieves 81.53% on the micro-F_1 metric of 5 candidate entities and 82.80% on the micro-F_1 metric of 100 candidate entities, which achieves a nearly 12 point absolute gain on a recently introduce multi-modal entity linking benchmark. In general, the result of 100 candidate entities is better than the result of 5 candidate entities. From the experimental results, we can conclude that the setting of more candidate entities can make the correct entity appear in the candidate set as much as possible.

Inference Time Efficiency: An important motivation for our model is to achieve better results than Bi-architecture while performing at more reasonable speeds than Cross-architecture. We design speed experiments to determine the precision and inference time for different architecture models in the situation of 1k candidate entities. We perform these experiments on both CPU and GPU setups. CPU computations are run on a 32 core Intel processor CPU E5-2620V4. GPU computations are run on a single NVIDIA 2080Ti.

We show the trade-off of precision and efficiency of our model in Table 3, greatly improving the efficiency of our model by losing a little precision, to achieve real-time faster ZEMEL. In Table 3, we show the average time per example for each architecture, the difference in timing between the Bi-architecture and our model is rather minimal, but our model will spend more time training. Nevertheless, both models are still tractable. The Cross-architecture, however, is 2 orders of magnitude slower than the Bi-architecture and our model, rendering it intractable for real-time inference. Thus, our model, given their desirable performance and speed trade-off, are the preferred model.

Table 3. The balance between precision and efficiency of our model.

	Pre	T_{train}	T_{cpu}	T_{gpu}
Bi-	80.76%	1.4h	130ms	20ms
Cross-	84.65%	9.8h	20.6s	3.9s
ours	**83.92%**	**2.0h**	**150ms**	**23ms**

Parameter Sensitivity Experiment: In this section, we evaluate our model on different settings of the parameters.

First, we compared the results achieved by our model with different counts of candidate images. It can be seen from the Figure 4 that the results of ten candidate images are better than one candidate image. The increase of candidate image counts can improve the precision of results, but the effect is relatively limited.

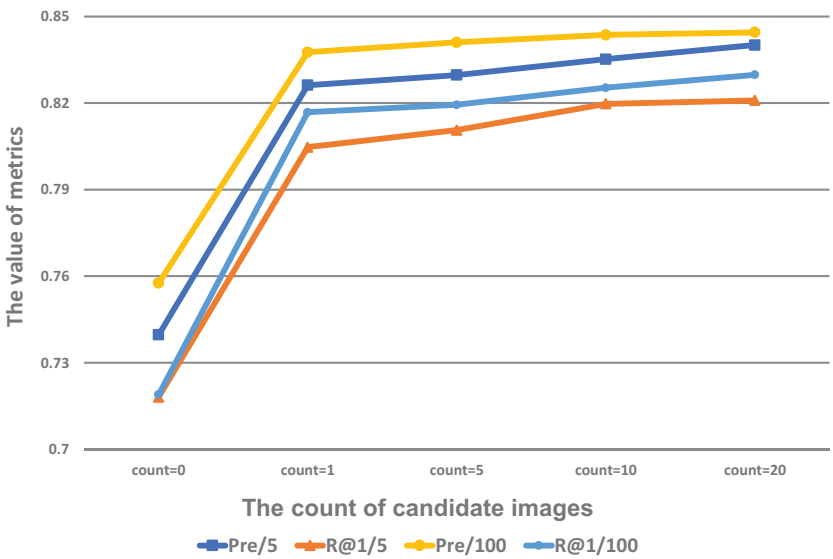


Figure 4. Example of multi-modal entity linking.

Second, we evaluate our model on different hyper-parameter m, which influences the representation of context fusion vector. From the experimental results, we can see that the larger m, the higher the precision obtained from the experiment. The larger m will make the context representation more comprehensive and sufficient.

Table 4. The influence of hyper-parameter m.

m	4	16	64	360
Pre	81.19%	81.49%	81.97%	82.62%

6. CONCLUSION

In this paper, we introduce a new task for ZEMEL and construct a multi-modal dataset named ZEMELD for it. This dataset can be used as a shared benchmark for multi-modal entity linking research focused on specialized domains where entities in the test set have not been seen in the training process. A strong baseline is proposed by combining visual-linguistic representation with poly-encoder architecture for faster ZEMEL in inference time. The experimental results show that our model achieves state-of-the-art results and has real-time speeds. Moreover, through extensive additional experiments, we demonstrate the efficacy of our model and prove the influence of hyper-parameters on experimental results.

In the future, a possible improvement direction is to incorporate NIL recognition and mention detection. We also expect the model can solve more entity types for the generalization of our model. The candidate generation phase leaves notable room for improvement.

AUTHOR CONTRIBUTIONS

Qiushuo Zheng (qiushuo_zheng@seu.edu.cn): responsible for task definition, model training, model tuning and paper writing. Wen Hao (wenhao7841@seu.edu.cn): responsible for data collection, data processing and model training. Meng Wang (meng.wang@seu.edu.cn): responsible for motivation proposal, task definition and paper modification. Guilin Qi (gqi@seu.edu.cn): responsible for idea formation, motivation proposal, model tuning and paper modification. Chaoyu Bai (baichaoyu@seu.edu.cn): responsible for data collection, data processing and model training.

REFERENCES

[1] Broscheit, S.: Investigating entity knowledge in Bert with simple neural end-to-end entity linking, arXiv preprint arXiv:2003.05473 (2020).
[2] Li, M. et al.: Gaia: A fine-grained multimedia knowledge extraction system, In: Annual Meeting of the Association for Computational Linguistics, 2020, pp. 77–86.
[3] Li, Y., Yang X., Luo J.: Semantic video entity linking based on visual content and metadata. In: IEEE International Conference on Computer Vision, pp. 4615–4623 (2015).
[4] Johnson, J. et al.: Image retrieval using scene graphs. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3668–3678 (2015).
[5] Zellers, R. et al.: From recognition to cognition: Visual commonsense reasoning. In: IEEE International Conference on Computer Vision and Pattern Recognition. pp. 6720–6731 (2019).
[6] Saqur, R., Narasimhan K.: Multimodal graph networks for compositional generalization in visual question answering, Conference and Workshop on Neural Information Processing Systems 33 (2020).
[7] Dost, S. et al.: Sperduti On visual-textual-knowledge entity linking. In: IEEE International Conference on Semantic Computing, IEEE, pp. 190–193 (2020).
[8] Moon, S., Neves L., Carvalho V.: Multi-modal named entity disambiguation for noisy social media posts. In: Annual Meeting of the Association for Computational Linguistics, pp. 2000–2008 (2018).

[9] Zhang, Q. et al.: Adaptive co-attention network for named entity recognition in tweets. In: The AAAI Conference on Artificial Intelligence, pp. 5674–5681 (2018).
[10] Devlin, J. et al.: Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[11] Sun, C. et al.: A joint model for video and language representation learning. In: IEEE International Conference on Computer Vision, pp. 7464–7473 (2019).
[12] Qi, D. et al.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, arXiv preprint arXiv:2001.07966 (2020).
[13] Chen, Y.-C et al.: Uniter: Learning universal image-text representations, arXiv preprint arXiv:1909.11740 (2019).
[14] Su, W. et al: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019).
[15] Li, G. et al.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Annual Meeting of the Association for Computational Linguistics, pp. 11336–11344 (2020).
[16] Li, L. H. et al.: Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).
[17] Lu, J. et al.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Conference and Workshop on Neural Information Processing Systems, pp.13–23 (2019).
[18] Tan, H., Bansal M. Lxmert: Learning cross-modality encoder representations from transformers, arXiv preprint arXiv:1908.07490 (2019).
[19] Cer, D. et al.: Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).
[20] Kiros, R. et al.: Fidler, Skip-thought vectors. In: Conference and Workshop on Neural Information Processing Systems, pp. 3294–3302 (2015).
[21] Henderson, M. et al.: Efficient natural language response suggestion for smart reply, arXiv preprint arXiv:1705.00652 (2017).
[22] Bowman, S. R., Angeli G., Potts C.: Manning, A large annotated corpus for learning natural language inference, arXiv preprint arXiv:1508.05326 (2015).
[23] Humeau, S. et al.: Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multisentence scoring, arXiv preprint arXiv:1905.01969 (2019).
[24] Pérez-Iglesias, J. et al.: Integrating the probabilistic models bm25/bm25f into lucene, arXiv preprint arXiv:0911.5046 (2009).

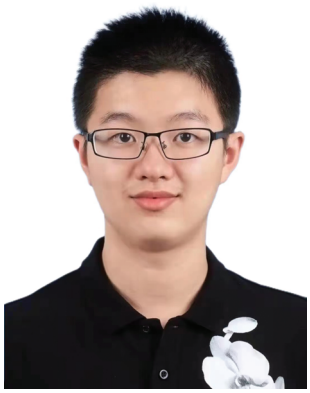
AUTHOR BIOGRAPHY

Qiushuo Zheng is a graduate student at Southeast University. He received a bachelor's degree from Southeast University. His main research interests are multi-modal learning and downstream applications of knowledge graph.

Hao Wen is an undergraduate student at the School of Computer Science and Engineering, Southeast University. Currently, my research interests mainly include Information retrieval, entity linking and multi-media research.

Wang Meng is an assistant professor in the Knowledge Graph & AI Research Group, School of Computer Science and Engineering, Southeast University, China. He is also a SEU Zhishan Young Scholar. He obtained the doctoral degree from the Department of Computer Science and Technology, Xi'an Jiaotong University in 2018, under the supervision of Prof. Jun Liu. He was a visiting scholar, working with Prof. Xue Li and Prof. Xiaofang Zhou, in the DKE lab at University of Queensland, Australia in 2016. His research area is in the knowledge graph, semantic search, NLP, and cross-modal data.

Qi Guilin, professor and doctoral supervisor of Southeast University, director of the Institute of cognitive intelligence of Southeast University, was supported by the six talent peak programs of Jiangsu Province. At present, he is the deputy director of the language and Knowledge Computing Professional Committee of Chinese information society and the deputy director of the knowledge organization professional committee of China Science and technology information society. He is a visiting professor at Griffith University in Australia (November 2011 February 2012 and June 2013 July 2013) and a visiting professor at the first university of Toulouse in France (January 2013 February 2013). He graduated from Yichun University in Mathematics in 1998, obtained a master's degree in mathematics and Information Department of Jiangxi Normal University in 2002 and a doctor's degree in computer science from Queen's University of Belfast in 2006.
ORCID: 0000-0002-1957-6961



Chaoyu Bai is a graduate student at Southeast University. He received a bachelor's degree from Nanjing University of Posts and Telecommunications. His main research interests are multi-modal learning and information extrction.