Multimodal Knowledge Learning for Named Entity Disambiguation

Anonymous ACL submission

Abstract

With the popularity of online social medias in recent years, massive-scale multimodal information has brought new challenges to traditional Named Entity Disambiguation (NED) tasks. Recently, Multimodal Named Entity Disambiguation (MNED) is proposed to link ambiguous mentions with the textual and visual contexts to a predefined knowledge graph. Recent attempts handle these issues mainly by annotating multimodal mentions and adding multimodal features to traditional NED models. These methods still suffer from 1) lack of multimodal annotation data against the huge scale of unlabeled corpus and 2) failing to model multimodal information at knowledge level. In this paper, we explore a pioneer study on leveraging multimodal knowledge learning to address the MNED task. Specifically, we propose a knowledge-guided transfer learning strategy to extract unified representation from different modalities and enrich multimodal lnowledge in a Meta Learning way which is much easier than collecting ambiguous mention corpus. Then we propose an Interactive Multimodal Learning Network (IMN), which is capable of fully utilizing the multimodal information in both mention and knowledge side. To verify the validity of the proposed method, we implemented comparisons on a public large-scale MNED dataset based on Twitter KB. Experimental results show that our method is superior to the state-of-the-art multimodal methods.

1 Introduction

Nowadays, online social medias have become more and more important in our daily life. And valuable information to understand users and their preferences is hidden in the massive-scale user-generated content. However, how to extract such information from these social media posts is extremely challenging because the posts are always in unstructured texts and images. Named Entity Disambiguation is such a critical task for extracting structured information, which maps ambiguous mentions from free-form texts to specific entities in a predefined knowledge graph. NED can benefit many downstream applications such as recommender systems, personal assistance, question answering,etc (Dredze et al., 2010).

Existing researches on NED mainly focus on texts only and have been proved to be successful for well-formed text. However, as the popularity of incorporating a mix of text and images in social media platforms (e.g. Twitter[1], Instargram[2], Snapchat[3], etc.), more ambiguous mentions appear in short and noisy text. Thus the cross-modal ambiguity makes traditional text-only NED methods more difficult to link them correctly due to enormous number of mentions arising from incomplete and inconsistent expressions. In many of such cases, it is impossible to disambiguate entities from text alone. For example, The mention *Swift* is completely ambiguous only from the textual context in Fig 1. It is difficult to distinguish whether *Swift* refers to **Taylor Swift** or **Ben Swift** for lacking of critical information in the text. Furthermore, the target person **Ben Swift** cannot be directly recognized from the image alone through face recognition techniques due to the obstruction of eyes, hats and other objects. However, by considering both mutimodal contexts in the post and historical data of the entity, the correct entity **Ben Swift** can be disambiguated from the candidates. That is, the textual features and visual features can complement each other.

Although some recent works has been proposed for the MNED task (Moon et al., 2018; Adjali et al., 2020a,b), there also exist some shortcomings. First, sufficient annotated corpus with both texts and images is required to train a multimodal model. How-

[1] https://twitter.com/
[2] https://www.instagram.com/
[3] https://www.snapchat.com/

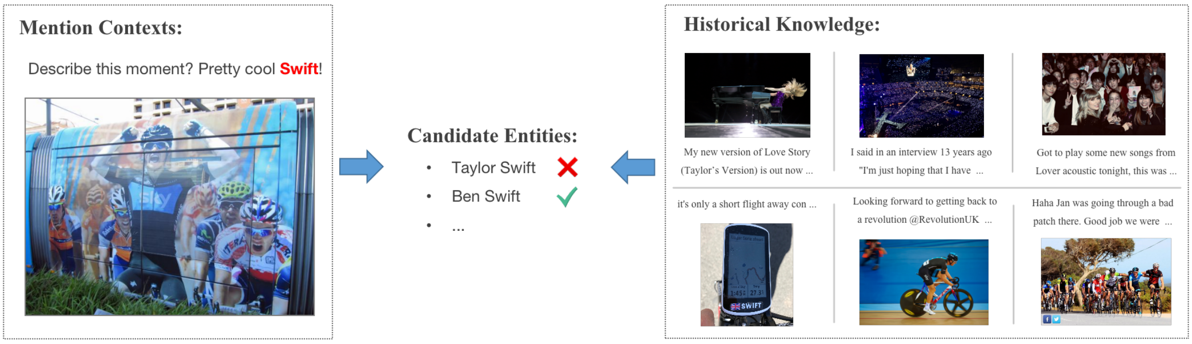

Figure 1: The example of MNED task with historical knowledge. Because of the insufficiency of information, the mention *Swift* is completely ambiguous only from the textual context. And the correct entity *Ben Swift* can be disambiguated by considering mutimodal contexts in the post and historical knowledge.

ever, the multimodal training data requires the annotation of all ambiguous mentions with the context of both texts and images in a post, which is costly to collect and annotate in practice (Abuczki and Ghazaleh, 2013). As such, The lack of sufficient training data would limit the performance of neural models. Second, previous works mainly learn from the mutimodal mention contexts, and do not exploit available information at the knowledge level which contains useful description and historical data with visual features.

In this paper, we focus on solving MNED tasks at the knowledge level and the training process consists of three steps: knowledge-guided pre-training, knowledge prototype construction and interactive learning. To reduce the dependence on annotated data, we firstly train a mutimodal feature extractor by implementing a knowledge-guided transfer learning strategy to make full use of unsupervised mutimodal corpus. After that we enrich multimodal information at the knowledge level using a Meta Learning aggregation method. This keeps both entities and mentions are multimodal which only requires a small number of knowledge annotation. Finally, we unifiedly integrate different modalities using an **I**nteractive **M**ultimodal learning **N**etwork (IMN), which is able to flexibly utilize the multimodal information from both mention contexts and knowledge graph. Our contributions are summarized as follows:

- We propose a knowledge-guided pre-train model to reduce the dependence on multimodal annotated data by transfer learning. To the best of our knowledge, this is the first time to introduce mutimodal pre-train model in MNED task.
- We propose a Meta Learning method to utilize multimodal information at the knowledge level. With the Meta Learning method and pre-train model, only a small number of annotation knowledge is required to distinguish candidate entities.
- We conducted comparative experiments on a public large-scale MNED dataset. Experimental results show the advantages of our pre-training method and the Meta Learning network outperforms state-of-the-art MNED methods.

2 Related Work

Multimodal Learning As an efficient mechanism of leveraging contextual information from multiple modalities in parallel, multimodal learning has been applied in a wide range of tasks in recent years (Elliott et al., 2015; Specia et al., 2016). In previous works, representation of different modalities was mostly obtained separately. For visual representation, CNN-based models such as VGG (Simonyan and Zisserman, 2014) , Google Inception (Szegedy et al., 2016), ResNet (He et al., 2016) are widely adopted in many multimodal tasks. Textual features are mostly represented by language models such as GloVe (Pennington et al., 2014), GPT (Radford et al., 2018), XLNet (Yang et al., 2019) etc. Recently, with the success of pre-train and self-supervised learning (Misra et al., 2016; Xie et al., 2017b), several mutimodal transfer learning methods and architectures (Yu et al., 2021; Gao et al., 2020; Lu et al., 2019b; Qi et al., 2020) have been proposed, and have achieved state-of-the-art results on various vision language tasks, including Visual Question Answering, Visual Commonsense Reasoning, Region-to-Phrase Grounding, Image-text Retrieval, etc. VideoBERT (Sun et al., 2019) learns joint distributions over sequences of visual and linguistic tokens as multimodal features. Vision-and-Language BERTs (Lu et al., 2020, 2019a; Gao et al., 2020) extend BERT architecture to adapt multimodal input by extracting RoIs from images and regards as image tokens. Although these pre-train models can learn unsupervised features in unsupervised corpus, they still need further improvement in tasks that require additional knowledge. And we argue that the self-supervised models still requires guidance of knowledge.

Named Entity Disambiguation Traditional NED methods mainly focus on text-only corpus which can be divided into two categories, local methods and global methods (Barrena et al., 2018; Ganea and Hofmann, 2017). For local methods, each mention is disambiguated separately via hand-crafted features (Bunescu and Paşca, 2006; Mihalcea and Csomai, 2007) and contextual representations learned by neural networks (He et al., 2013; Eshel et al., 2017). Global methods(Nguyen et al., 2016; Le and Titov, 2018) jointly disambiguate mentions by taking into account the topical coherence among the referred entities in the same document(Fang et al., 2019). For the MNED task, the work from (Moon et al., 2018) is the first to utilize multimodal mention contexts via weighting the embeddings of images and words based on attention mechanism. The previous multimodal works primarily depend on sufficient training data with fully annotations on all mention modalities which is costly in practice(Abuczki and Ghazaleh, 2013). Although Moon et al. (2018) involve a zero-shot layer in their model to allow for disambiguation of unseen entities during training, the performance is limited if the multimodal information is incomplete in the training data. Inspired by recent success on multimodal knowledge graph (Xie et al., 2017a; Mousselly-Sergieh et al., 2018; Pezeshkpour et al., 2018),we aim at handle MNED tasks at the knowledge level, which is much easier than collecting and annotating multimodal corpus.

3 Proposed Method

3.1 Task Definition

Formally, the inputs of the MNED task are a set of multimodal posts $P = \{p^{(1)}, p^{(2)}, ..., p^{(n)}\}$ and a predefined knowledge graph $G = (E, R, H)$ that is composed of the entity set E, the relation set R and relative historical data of entities. Each input post $p \in P$ is denoted as $p = \{p_m, p_t, p_v\}$, where p_m is a mention that needs to be disambiguated, p_t is a sequence of words surrounding the mention in the post, and p_v is an image associated in the post. Note that the mention p_m can be obtained by other tasks such as Named Entity Recognition (Lample et al., 2016), which is beyond the scope of this paper. Then the target of MNED is to find the ground truth entity $\hat{e} \in E$ that p_m corresponds to.

3.2 Knowledge-Guided Pre-train Model

Before dealing with the input multimodal posts, we firstly build a pre-trained model to capture the inherent relationship between images and texts which is guided by the knowledge graph. In this transfer learning way, the model can better understand the content of different modalities and is helpful to overcome insufficient of annotated mutimodal corpus.

End-to-end architecture The pretrain model is composed of four parts, textual representation, visual representation, transformer encoder and training with adaptive loss. The multimodal inputs consist of textual and visual representation which is tokenized into a token and patch sequence according to WordPieces and Object Detection methods. We use the standard BERT(Devlin et al., 2018) pre-process method to get the textual sequence. Unlike traditional pipeline image representation techniques, We use an end-to-end method to obtain the visual representation. DEtection TRansformer(DETR)(Carion et al., 2020) approaches object detection as a direct set prediction problem which directly output the final set of objects in parallel. Given an input image, we take the fixed-length vector sequence of the output layer of DETR decoder as the visual representation. Each of the vectors corresponds to one image patch, we regard each patch as an "patch token".

The concatenation of the text token sequence and image patch sequence consists of the pre-train model inputs. A pre-trained standard Transformer (Vaswani et al., 2017) is adopted as the matching backbone network of the pre-train model. The information of text tokens and image patches thus interact freely in multiple self attention layers. In order to ensure the mutimodal comprehension ability as well as sensitiveness at the knowledge of the pre-train model, we exploit three tasks in the train process.

Mention Masked Language Modeling(MMLM) Different from previous random word masking, our mention masking is directed by the knowledge graph. For mention tokens, we mask it with a probability of 85%. For other tokens are masked out with the probability of 15%. We apply the Whole Word Masking (WWM) strategy to mask out all the text tokens corresponding to a word at once. Finally, the MLM task is to minimize the cross-entropy loss, written as

$$L_{mm} = -\sum_{t_i \in p_t} \log P(t_i | t_{\setminus i}, \theta) \tag{1}$$

Where θ is trainable parameters, $Pre(t_i | t_{\setminus i}, \theta)$ is denotes the probability of the masked-out token t_i predicted by the model, given surrounding tokens $t_{\setminus i}$ in the post p.

Patch Masked Image Modeling(PMIM) Similar to MMLM, we mask out certain patches in a patch sequence (Gao et al., 2020). Given an image patch sequence $v = \{v_1, v_2, ..., v_n\}$ generate by DETR, we randomly mask out patches with the probability of 15%. The masked patch features are set to zero vectors. PMIM is to predict the distribution over the masked-out patch features. The MPM training is supervised by minimizing the KL-divergence between the distributions of patch features.

$$L_{pm} = -\sum_{v_i \in p_v} KL(v_i, Pre(v_i | v_{\setminus i}, \theta))) \tag{2}$$

Image and Text Alignment Modeling(ITAM) In the ITAM task, the hidden output of the token [CLS] is fed into a scoring function to indicate whether the text and image data are in the same post. Given a knowledge graph, the negative sample are randomly selected from similar posts such as tweets posted by candidate entities and tweets with the same mention. The hinge-based bi-directional ranking loss (Lee et al., 2018; Faghri et al., 2018; Karpathy and Fei-Fei, 2015) is the most popular objective function for image and text alignment, which can be formulated as follows:

$$\begin{aligned} L_{am} = -\sum_{p_{v^-}, p_{t^-}} \{\max[0, m - S(p_v, p_t) + S(p_v, \\ p_{t^-})] + \max[0, m - S(p_v, p_t) + S(p_{v^-}, p_t)]\} \end{aligned} \tag{3}$$

where m is a margin constraint, (v^-, u^-) are negative pairs. $S(\cdot)$ is a scoring function. The objective function is specifically trained attempts to pull positive image-text pairs close and push negative ones away which contribute to distinguish between mention contexts and candidate entities. The pre-training model is trained to recover the different modal information with three objectives and the three objectives are jointly optimized. Thus, the overall pre-training objective L is:

$$L = L_{mm} + L_{pm} + L_{am} \tag{4}$$

For more implementation details, see related description in appendix.

3.3 Knowledge Prototype Construction

In spite of the multimodal mention contexts, We believe that multi-modal information at the knowledge level is potentially important for MNED tasks. Different from the previous textual representation methods, we prefer to establish multimodal representation at the knowledge level. Given an entity, we construct a small-scale support set which is composed of related annotation knowledge for each modality respectively. Then a scoring model (see section 4) to measure the correlation between query set and support set is adopted for meta learning. As an entity is associated with many related historical posts containing images and texts, We simply select a part of the representative timeline tweets as the support set. Specifically, we adopt three modalities representations to depict an entity based on timeline posts. The visual prototype of each entity e_v is acquired by aggregating the features of the k representative corresponding images. And features of an image can generated by many image identification such as ResNet-101 (He et al., 2016). Similarly, the textual prototype of each entity e_t is acquired by pre-trained language models such as Bert (Devlin et al., 2018). Meanwhile, the joint prototype of each entity e_o can be acquired by the hidden state of the pre-training model described in previous subsections.

To select most representative support set from a large number of historical data, we build a similarity graph for each modality. The vertexes of the similarity graph are feature vectors obtained in previous steps. And the edges are the cosine similarity between the vertexes. Then top-k representative results are acquired by calculating the

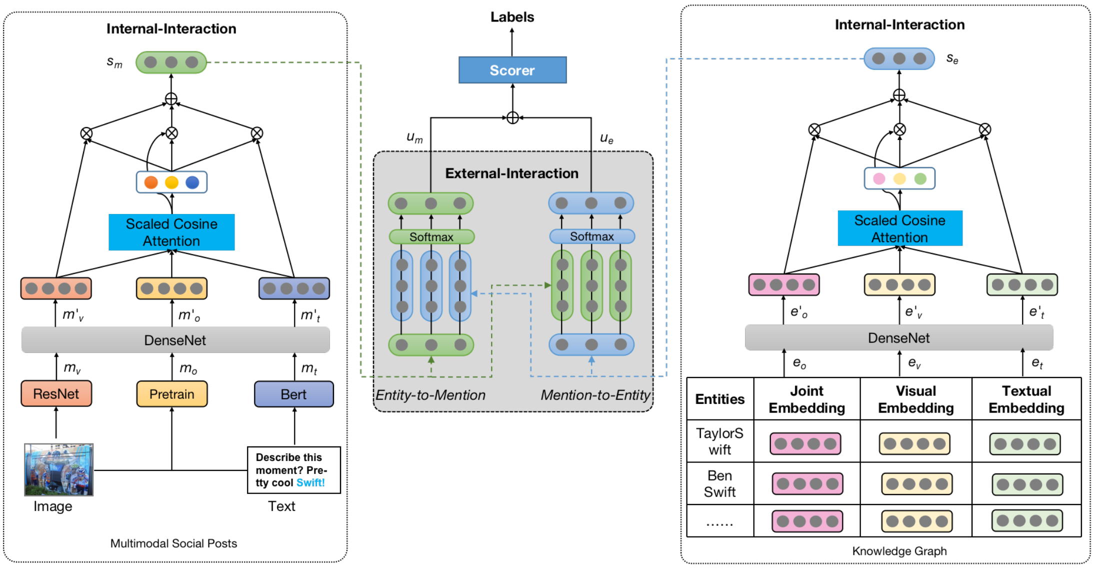

Figure 2: The overview of the IMN with internal and external interactive component. The internal-interaction component extracts the attentions of different modalities within mention contexts and the candidate entity respectively. The external-interaction component conduct a bidirectional interaction across mention contexts and the candidate entity.

PageRank score (Page et al., 1999) of each vertex in the similarity graph. The multimodal prototypes of an entity can be acquired by averaging the feature vectors of the top-k PageRank vertexes, and we perform L2 regularization on each prototype. Finally, each entity is represented with three different modalities $e = \{e_v, e_t, e_o\}$.

For the multimodal posts, three different feature extractors is applied to obtain query set embeddings. For each post $p = \{p_m, p_t, p_v\}$, the visual embedding m_v and textual embedding m_t is generated by the same method used in entity representation process. The joint embedding m_j of the mention p_m is acquired by pre-trained model in section 3.2. Thus, each mention is embedded with three modalities $m = \{m_v, m_t, m_o\}$.

3.4 Interactive Multimodal Learning Network

The architecture of IMN is shown in Figure 3. IMN adopts the idea of decoupling for modular design which has strong flexibility and applicability for different forms of input. In general, IMN consists of three components: Internal-Interaction, External-Enteraction and a score component which conduct a bidirectional interaction of the different modalities across mention contexts and knowledge graph.

3.4.1 Internal-Interaction

The inputs of IMN include two parts: multimodal mention contexts and the candidate entity prototypes. The internal-interaction component is utilized to explore the effect of different modalities within each part of inputs respectively.

Firstly, We adopt a Dense layer (Huang et al., 2017) to map multimodal embeddings to a unified representation space. The outputs of the Dense layer are denoted as $m' = \{m'_v, m'_t, m'_j\}$. To evaluate the effect of different modalities, a scaled cosine attention mechanism is performed on the feature representations m' as follows:

$$q = [q_v; q_t; q_l] = W_q \cdot [m'_v; m'_t; m'_o] \tag{5}$$

$$k = [k_v; k_t; k_l] = W_k \cdot [m'_v; m'_t; m'_o] \tag{6}$$

$$\alpha_{i,j} = \frac{exp(cos(q_i, k_j))}{\sum_j exp(cos(q_i, k_j))} \quad \forall i, j \in \{v, t, o\} \tag{7}$$

where q and k are queries and keys for calculating the scaled cosine attention, W_q and W_k are the weight matrices, $\alpha_{i,j}$ denotes the attention weights on multimodal embeddings.

Then the final embeddings of the input multimodal mention contexts s_m can be achieved by stacking weighted multimodal embeddings.

$$s_m = [\sum_i \alpha_{i,j} m'_j] \quad \forall i, j \in \{v, t, l\} \tag{8}$$

Similarly, the internal-interaction for the extended knowledge graph is performed with the multimodal representations of the entities obtained in

Section 3.4 and the output embedding of each entity is denoted as s_e.

3.4.2 External-Interaction

The external-interaction component implements a bidirectional interaction which can deal with the effect of different modalities from mention contexts to the knowledge graph and vice versa. We denote the two directions of effect as *entity-to-mention* and *mention-to-entity*, respectively.

To evaluate the effect of *entity-to-mention*, we take s_m as queries and s_e as keys respectively. Then we utilize the scaled cosine attention mechanism to obtain interactive results.

$$q = [q_v; q_t; q_o] = W_q \cdot s_m \tag{9}$$

$$k = [k_v; k_t; k_o] = W_k \cdot s_e \tag{10}$$

$$\alpha_{i,j} = \frac{exp(cos(q_i, k_j))}{\sum_j exp(cos(q_i, k_j))} \quad \forall i, j \in \{v, t, o\} \tag{11}$$

Then the final representations of mention contexts with the effect of different modalities from the knowledge graph u_m can be obtained as follows.

$$u_m = [\sum_{j \in \{s,t,v\}} \alpha_{i,j} k_j] \quad \forall i \in \{v, t, o\} \tag{12}$$

By switching the queries and keys, we can get the final representations of the entities u_e with the *mention-to-entity* effect. Then u_m and u_e are concatenated to predict the matching score of the corresponding mention m and the entity e. The scorer function is as follows.

$$f(m, e) = tanh(W_y[u_m; u_e] + b_y) \tag{13}$$

where W_y and b_y are the weight matrix and bias term, respectively. The scorer function evaluates the probability distribution of the ground-truth labels for matching pairs (m, e), where the labels belong to $[-1, 1]$.

3.4.3 Training

Given a set of multimodal posts which contain mentions and their corresponding entities, the training process is to minimize the ranking loss between the positive and negative pairs. Intuitively, the model is trained to produce a higher score between the representations of multimodal mention contexts and the ground-truth entity. Then the loss function is defined as:

$$\tau = \sum_{e^- \in E} max(\gamma + f(m, e^+) - f(m, e^-), 0) \tag{14}$$

where e^+ is the ground-truth corresponding entity of mention contexts m and e^- is the incorrect entity. γ is a margin parameter that controls the amount of difference between $f(m, e^+)$ and $f(m, e^-)$.

4 Experiments

4.1 Datasets

Measurement	Value
# multimodal input posts	85K
# distinct mentions in posts	1678
# entities in the knowledge graph	68K
# timeline tweets in the knowledge graph	2M
avg. length of posts	20.59
avg.# mentions in a post	1.15
avg.# candidate entities for each mention	17.24
avg.# timeline tweets of an entity	121

Table 1: Key statistics of the MNED dataset.

We conduct comparative experiments on a public multimodal entity disambiguation dataset (Adjali et al., 2020a) which collects text and images to jointly build a corpus of tweets with ambiguous mentions along with a Twitter KB defining the entities. The entities in the corpus are composed of popular twitter users including people, companies, and organizations. The overall statistics can be seen in table 1 and more details of the dataset construction can be found in appendix section.

4.2 Experimental Settings

Hyperparameters For the pre-train model, We use the default parameters of DETR and Bert(base) in which the number of negative examples is set to 5, the margin of ITAM is 0.3 and the training steps is 1M. For knowledge prototype construction, we keep 10 PageRank results as the support set of each modality, other parameters adopt the default configuration of original feature extraction model. For IMN, the mapped size is 300, the margin of the loss function is 0.2 and the epoch is 100 with a validation set for early stopping. We update the parameters using Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001, the dropout rate is 0.2, the score function is tanh.

Evaluation Metrics For evaluation, we use standard micro P@1 accuracy(Adjali et al., 2020b; Moon et al., 2018) and R@3 (Moon et al., 2018) recall as metrics in our experiments. P@1 can intuitively reflect the precision of results. R@3 evaluates the matching quality by measuring whether the ground-truth entity is highly ranked.

4.3 Results and Analysis

4.3.1 Baselines

We compare our IMN model with both machine learning methods and multimodal deep learning methods. These benchmark methods are introduced as follows:

- **DZMNED** (Moon et al., 2018): The first proposed method for MNED by considering multimodal contexts, which adopts a CNN-LSTM hybrid network with modality attention.
- **ET** (Adjali et al., 2020b): A feature-based machine learning model use the combination of multimodal features to build an Extra-Trees classifier for MNED task.
- **JMEL** (Adjali et al., 2020b): The state-of-the-art method which extract the features of different modalities and learn a joint representation of tweets with a fully connected neural network.

4.3.2 Main Results

Table 2 shows the results of our model compared with baselines. In general, our IMN model achieves significant improvements over all the baselines on both P@1 and R@3 with the mutimodal dataset[4]. It can be observed that the pretrain methods are at an absolute advantage in both P@1 adn R@3, which shows advantage of transfer learning and the necessity of jointly representing multimodal features for MNED task. Comparing to the multimodal method such as JMEL with traditional textual and visual representation methods, our model achieves 1.9% absolute improvement on P@1. The improvements indicate that the interaction between multiple modalities also adds performance gain by capturing the effect of different modalities from both the posts and the knowledge graph. In addition, adding more multimodal features can still supplement MNED tasks, even that the pre-trained representation already contain multimodal information. This proves that the information of different modes can complement each other.

Model	modals			result	
	text	image	joint	P@1(%)	R@3(%)
ET	✓	✓		67.1	-
JMEL	✓	✓		80.3	-
DEMNED	✓	✓		80.14	94.18
IMN(base)	✓	✓		82.23	94.54
IMN(joint)			✓	81.19	93.84
IMN(img)		✓	✓	82.40	94.61
IMN(txt)	✓		✓	82.44	94.83
IMN	✓	✓	✓	**83.99**	**95.04**

Table 2: Comparison results with baselines on the multimodal dataset. The best performance is denoted with bold text and "✓" indicates that features of the corresponding modal are included in the input.

To investigate the effect of each component in our model, we conduct a set of ablation experiments as shown in Table 3. *IMN* is the complete proposed model. The notation '-' means removing some part of the model. From the experimental results we can observe that the performance drops significantly when both interactions are removed, which demonstrates the effectiveness of our interactive model. The performance drops considerably by removing one of the interactions (i.e. *Internal-Interaction* or *External-Interaction*). This proves the multimodal information from both the posts and the entities is helpful for the MNED task.

4.3.3 Ablation Study

Model	Results	
	P@1(%)	R@3
IMN	**83.99**	**95.04**
- External-Interaction	83.16	93.07
- Internal-Interaction	82.26	94.89
- Both Interactions	82.10	94.50
- Knowledge Guided	83.07	94.95

Table 3: Ablation tests for MNED. "-" means removing corresponding component of the model.

We also investigated the necessity of knowledge guidance in the pre-training process. Firstly, We implement the same mask strategy of Bert by treating mentions as normal words. Then, negative examples of each case are randomly selected from all tweets. We can observe that the overall accuracy will be reduced to a certain extent in Table 3. The

[4]We select the same feature extractors used in baselines respectively to ensure the fairness of comparison. Since we have reached a consistent conclusion, the difference of extractors is not reflect

Modal Side	Mention Modals			Entity Modals			Results	
	text	image	joint	text	image	joint	P@1(%)	R@3(%)
Single Modal	✓			✓			79.84	94.03
		✓			✓		77.56	91.93
			✓			✓	**81.19**	94.16
Mention Side	✓	✓		✓			80.38	94.16
	✓		✓	✓			80.80	94.14
	✓	✓	✓	✓			**81.11**	**94.26**
Entity Side	✓			✓	✓		82.38	94.59
	✓			✓		✓	82.19	94.81
	✓			✓	✓	✓	**83.21**	**95.00**

Table 4: Results of the Multimodalitiy Analysis. Single Modal indicates the effect of different modals when used alone. Mention Side and Entity Side refer to the enrichment means of multimodal information on the mention and the knowledge side respectively.

result shows that the structure and historical information in the knowledge graph can be learned by a pre-train manner and is helpful to improve the effect of the MNED task.

4.3.4 Multimodalitiy Analysis

In this part, we perform a series of experiments to evaluate the performance of our model on dealing with the multimodal features on different input sides. As shown in Table 4, the pre-trained features are significantly outperform other single-modal features. Besides, we enrich multimodal features on the mention side and the entity side respectively. Results show that adding multimodal features from both sides can improve the model effect, and the multimodal features on the entity side has a more obvious contribution to the improvement of results. This points out a new direction for data annotating of MNED tasks: we can put the focus of data annotation on the production of multimodal knowledge, even if the input mention does not have multimodal contexts. In this way, the multimodal annotation dependence on the mention side can be greatly reduced.

4.3.5 Aggregating Statistics

In order to further study the effect of different methods for entity support set construction, we conduct comparative experiments using different K values and two aggregation strategies and the results are shown in Figure 3. We can observe that the effect of PageRank method is significantly outperform random method especially for a small number of K values. It indicates that the features selected by the PageRank method are more representative and the influence of noise on the result is reduced to some extent. The point can be inferred from the experimental results that it is significant to improve the quality of multimodal knowledge rather than rely on accumulating features.

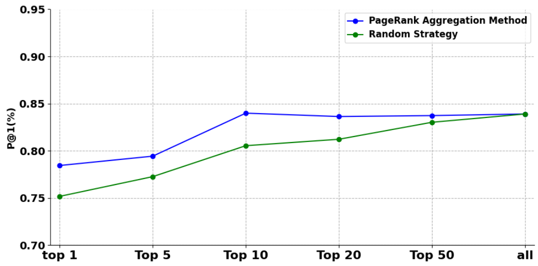

Figure 3: Results corresponding to different aggregation strategies. The abscissa represents the final aggregated number of entity historical data and the ordinate represents the corresponding precision.

5 Conclusion

We propose to solve MNED task at the knowledge level through Mutimodal Transfer Learning and Meta Learning. With large-scale unsupervised data and a small amount of annotated knowledge, our model significantly outperforms the state-of-the-art MNED methods. Experimental results show that enrich multimodal features at the knowledge level is more conducive to improving the effect of MNED models compared with mention contexts annotation.

There are still many points worth continuing to explore. In particular, the structural information in the knowledge graph which can be learned by knowledge representation models such as transE may also be useful. Besides, the prototype aggregation method still needs further exploration with graph learning models such as GCN etc.

References

Ágnes Abuczki and Esfandiari Baiat Ghazaleh. 2013. An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, 9:86–98.

Omar Adjali, romaric Besancon, olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020a. Building a multimodal entity linking dataset from tweets. In *International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Omar Adjali, romaric Besancon, olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020b. Multimodal entity linking for tweets. In *European Conference on Information Retrieval (ECIR)*, Lisbon, Portugal.

Ander Barrena, Aitor Soroa, and Eneko Agirre. 2018. Learning text representations for 500k classification tasks on named entity disambiguation. In *CoNLL*, pages 171–180.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *COLING*, pages 277–285.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In *CoNLL*, pages 58–68.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The World Wide Web Conference*, pages 438–447.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *EMNLP*, pages 2619–2629.

Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2251–2260. Association for Computing Machinery.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *ACL*, pages 30–34.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*, pages 4700–4708.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL-HLT*, pages 260–270.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *ACL*, pages 1595–1604.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visio linguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *CIKM*, pages 233–242.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *ACL*, pages 2000–2008.

Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Joint Conference on Lexical and Computational Semantics,SEM@NAACL-HLT*, pages 225–234.

Thien Huu Nguyen, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo, and Mohammad Sadoghi. 2016. Joint learning of local and global features for entity linking via neural networks. In *COLING*, pages 2310–2320.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding multimodal relational data for knowledge base completion. In *EMNLP*, pages 3208–3218.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pages 543–553.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017a. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017b. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

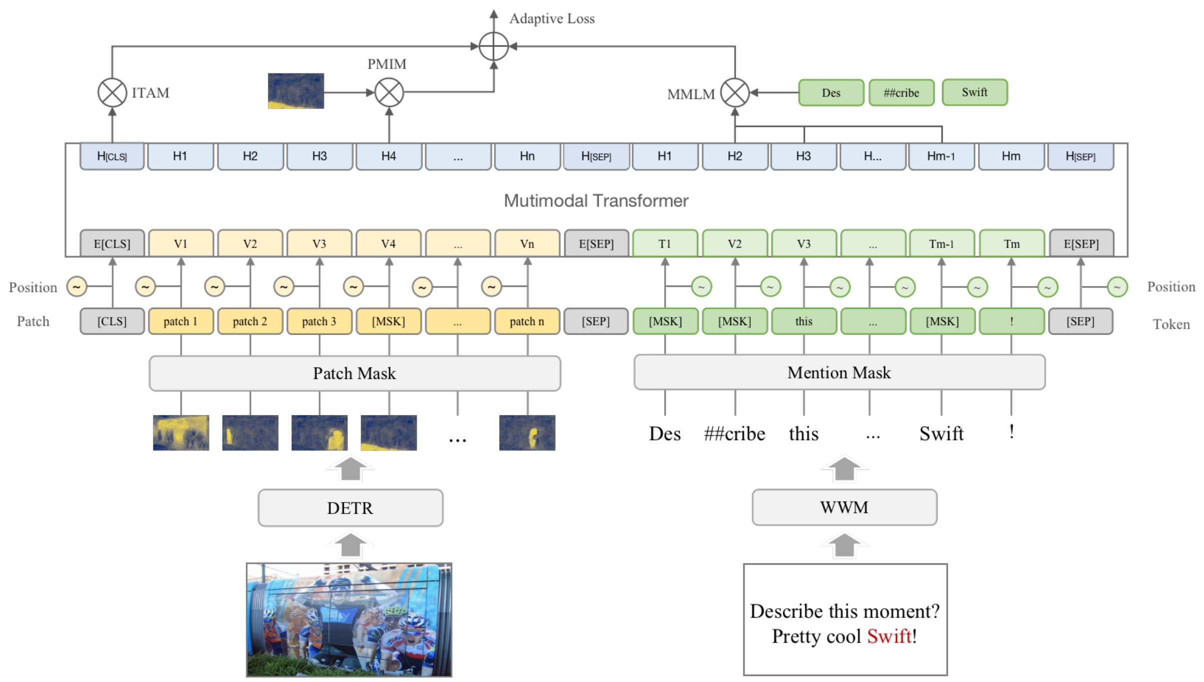

Figure 4: Our knowledge guided multimodal pre-training model. We cut the image into fixed-length patches with DETR, and concatenate textual tokens as the input sequence. Finally, the multimodal semantic representation is obtained through the transformer encoder.

A Implementation details

A.0.1 Pre-train Model Architecture

The overview of the pretrain model is illustrated in Figure 4. It is composed of four parts, textual representation, visual representation, transformer encoder and training with adaptive loss. The multimodal input is firstly tokenized into a token or patch sequence according to WordPieces and Object Detection. We use the standard BERT pre-process method to process the input sequence. And, the sum of the sequence embedding, position embedding and segmentation embedding is regarded as the text representation.

A.0.2 DETR Extractor

We use an end-to-end method to obtain the visual representation. DEtection TRansformer(DETR) approaches object detection as a direct set prediction problem. It consists of a set-based global loss, which forces unique predictions via bipartite matching, and a Transformer encoder-decoder architecture. Given a fixed small set of learned object queries, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions in parallel. Given an input image, we take the fixed-length vector sequence of the output layer of DETR decoder as the visual representation. Each of the vectors corresponds to one image patch, we regard each patch as an "patch token".

A.0.3 Negative Sampling in ITAM

For ITAM task, for one positive example in the train dataset, the text and image are extracted from the same post, while for one negative sample, the text and image are randomly selected from similar posts:

- 70% of the negative examples are randomly selected from the historical tweets posted by candidate entities of the mentions appearing in the article.
- 15% of the negative examples are randomly selected from the tweets with the same mention.
- 15% of the negative examples are randomly selected among the entire corpus.

B Experimental details

B.0.1 Dataset introduction

The entities in the corpus are composed of popular twitter users including people, companies, and organizations. For ground-truth entity generation, an important mechanism in Twitter communication is the usage of a user's screen name (@UserScreen-Name) in a tweet which helps to explicitly align mention with the ground-truth entity. Each tweet contains textual and visual content after a series of preprocessing including deleting single-modal, non-related and enumerated tweets. To sufficiently enrich the KB with ambiguous entities, thus make the MNED task challenging, a simple procedure was adopted to jointly generate ambiguous candidate entities and populate the KB. On the basic assumption that entities sharing the same last name or acronyms (when the Twitter user is an organization etc.) are potential candidate entities, entity generation can be achieved naturally by collecting entities sharing the same last name or acronyms. In the dataset, screen names in the original post were replaced with the last name or acronyms of the ground-truth entity as mentions.

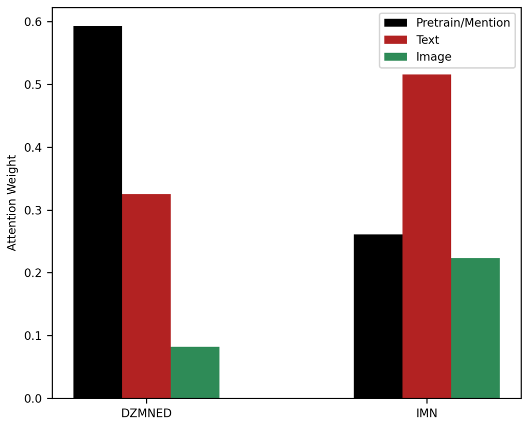

Figure 5: The average attention distribution on different modalities. The black column represents the average weight of joint embedding in IMN and mention embedding in DZMNED.

B.0.2 Attention Distribution

In order to evaluate the capability of extracting multimodal features, we output the final attention weight of each modality and make an average on the test set. Figure 5 shows the attention distribution over the joint/mention, text and image of the input posts. It is observed that the attention distribution of DZMNED is more imbalanced than ours. Specifically, the imbalance mainly lies on the average weight of images, which indicates that our model can extract visual features better than that of DZMNED. This can also support the good performance of our model on MNED.