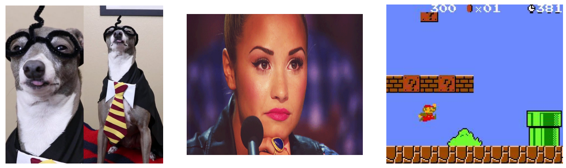
Multi-Granularity Contrastive Knowledge Distillation for Multimodal Named Entity Recognition

Anonymous ACL submission

Abstract

It is very valuable to recognize named entities from short and informal multimodal posts in this age of information explosion. Despite existing methods success in multi-modal named entity recognition (MNER), they rely on the well aligned text and image pairs, while a lot of noises exist in the datasets. And the representation of text and image with internal correlations is difficult to establish a deep connection, because of the mismatched semantic levels of the text encoder and image encoder. In this paper, we propose multi-granularity contrastive knowledge distillation (MGC) to build a unified joint representation space of two modalities. By leveraging multi-granularity contrastive loss, our approach pushes representations of matched image-text pairs or image-entity pairs together while pushing those unrelated image-text or image-entity pairs apart. By utilizing CLIP model for knowledge distillation, we can obtain a more fine-grained visual concept. Experimental results on two benchmark datasets prove the effectiveness of our method.



(a) Your world might be shaped like a big planet but mine is shaped like a tiny **[Harry Potter MISC]**

(b) When your family goes to **[Red Sox game MISC]** without you

(c) **[Kevin Durant PER]** putting up more bricks than **[Super Mario Bros MISC]**

Figure 1: Examples of different ways image content and named entity can be related in Multimodal Named Entity Recognition. The named entities and types are highlighted ("*MISC*" stands for other named entity and "*PER*" stands for person). (a) Image are significantly related to entity. (b) Image is hardly related to entity. (c) Image is partially related to entity.

1 Introduction

Named Entity Recognition (NER) is a crucial subtask of Information Extraction (IE), which aims to find and classify the type of named entities useful for downstream tasks. But in real scenarios (e.g., social media platforms), we are often exposed to limited and informal text, from which it is very difficult to identify named entities (Ritter et al., 2011). Some of the research on NER attempts to introduce multimodal information to help identify named entities in unstructured text (Zhang et al., 2018; Lu et al., 2018; Sui et al., 2021; Zhang et al., 2021). As shown in Figure 1(a), without the support of the image, it would be difficult to figure out that "*Harry Potter*" here refers to a dog, while easy to identify which as an actor or film title.

Multimodal Named Entity Recognition (MNER) has received increasing interest these years. Existing research focuses on how to fully exploit multimodal information (visual information) (Wu et al., 2020; Zhang et al., 2021), and how to fuse text and visual representation (Moon et al., 2018b; Yu et al., 2020; Chen et al., 2021). Despite their success, current MNER methods have two major limitations: **Firstly**, the current methods often relied on well aligned image and text pairs. But, in social media data, the relationship between image and entities is pluralistic (Hu et al., 2018; Vempala and Preotiuc-Pietro, 2019) and sometimes the images content may be unrelated to entities. Take Figure 1(b) as an example, the image is only used to express the mood of the uploader, which is unrelated to the entities in the text, and may even introduce undesirable noise. **Secondly**, the representation of text and images with internal correlations is difficult to establish a deep connection. Existing work often relies on language models pre-trained on massive raw data (e.g., BERT(Devlin et al., 2019), XLNET(Yang et al., 2019) and so on) and image classifiers pre-trained on large-scale annotated data such as Imagenet (Deng et al., 2009) and OpenImages (Kuznetsova et al., 2020). Such a pre-trained text encoder is knowledgeable. For example, in Figure 1(a), it could be found that "*Harry Potter*" could refer to a character, a novel or a series of films (Roberts et al., 2020; Petroni et al., 2019). But such a pre-trained image encoder is more concerned with low-level semantic information and relatively limited visual concepts. For instance, in Figure 1(a), it easily tells that the image consists of a dog, not a man, but hardly represents the dog dressed as "*Harry Potter*". Because it is difficult to learn fine-grained concepts on a standard image classification dataset. Furthermore, there is scarcely a one-to-one match between the image and the entity, but often an incomplete matching relationship. As Figure 1(c) illustrated, the image is a scene from "*Super Mario*", indicating that "*Super Mario*" is a game. But there is no direct match between the image and the entity "*Kevin Durant*". So there is no need to introduce image information as a distraction when classifying this entity. According to our statistics, incomplete matching exists in more than 31% of the image text pairs that contain more than one entity, in the Twitter-2017 dataset (Lu et al., 2018).

In this paper, to overcome above challenges, we propose **M**ulti-**G**ranularity **C**ontrastive Knowledge Distillation Learning (MGC) framework. We have constructed a joint representation space of text and image to learn the different relationship between images and texts or entities. In detail, in joint representation space, we leverage Global Contrastive loss to push embedding of matched image-text pairs together while pushing those unrelated image-text pairs apart. Besides, we leverage Local Contrastive loss to push embedding of matched image-entity pairs together while pushing those unrelated image-entity pairs apart. Moreover, in order to make the image encoder and the text encoder similar in capability and to bridge the two modality presentation better, we leverage Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021) as a teacher model. CLIP model is pre-trained on 400 million image-text pairs scraped from the website, which is able to express a much more fine-grained visual concepts in joint representation space and help to filter some unrelated image. As the framework is model-free, it could in theory also be used for many existing MNER methods.

Our contribution can be summarized as follows:

- We design a novel framework **MGC** (Multi-Granularity Contrastive Knowledge Distillation Learning) to align images and texts or entities. So more useful and fine-grained visual information can be used for NER.
- We propose an approach to build a joint representation space of image and text under the supervision of CLIP model and multi-granularity contrastive learning.
- We conduct extensive experiments on two public MNER datasets. Experimental results prove the effectiveness of our method. Our code has been uploaded as an attachment.

2 Methodology

In this section, we will introduce the details of **MGC** framework to multimodal named entity recognition. Before introducing our proposed approach, we first describe the task formalization of MNER.

Task Formalization: Given a piece of text X and an image V associated with the text. MNER aims to leverage multimodal information to classify and locate pre-defined types of entities from text X. As following most of studies about MNER, we have adopted the paradigm of sequence labeling. The input of MNER is a sequence of words $X = \{x_1, x_2, ..., x_n\}$, while the goal is to predict a sequence of label $Y = \{y_1, y_2, ..., y_n\}$, and that is to estimate $P(Y|X, V)$, where $y_i \in \mathbb{Y}$ and the $\mathbb{Y}$ is the pre-defined label set with the *BIO2* tagging schema (Sang and Veenstra, 1999).

As Figure 2 illustrating the overall architecture of our method, the key of our framework aims at how to build a unified joint representation space to help MNER. We introduce knowledge distillation from CLIP (Radford et al., 2021) and multi-granularity contrastive mechanism to bridge text and image modality. Consequently, we first introduce how to transform the input into the representation, and then describe knowledge distillation from CLIP, and multi-granularity contrastive mechanism. Finally, we elaborate how to fuse the two modality representation to cope with MNER task and the training process.

2.1 Instance Representation

First we need to obtain representations of the inputs from different modalities.

Text Encoder: To make better use of world knowledge, our text encoder employ BERT (Devlin et al., 2019). Give a batch of instances (text-image

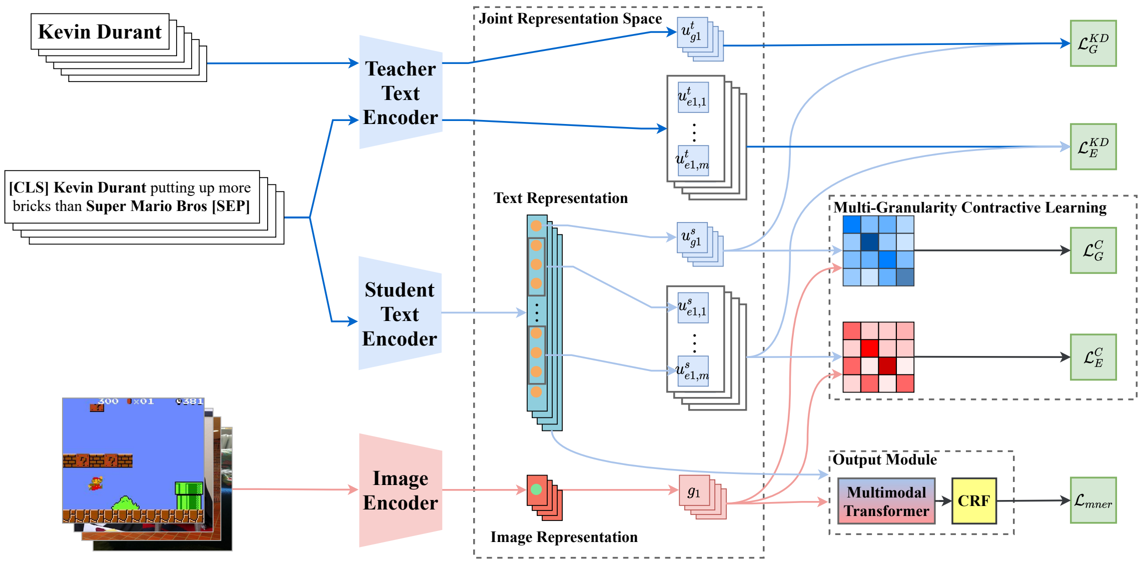


Figure 2: The overall architecture of our method in training. The top part illustrates unified text representation supervised by distilling from teacher text encoder output. The middle part shows that a multi-granularity contrastive mechanism is in charge of both entity-image and text-image matching. While the bottom part shows the text representation and image representation fuse to recognize named entities. The red line indicates the image data-flow and the blue line denotes the text data-flow.

pairs) $I = \{(X_1, V_1), (X_2, V_2), ..., (X_B, V_B)\}$, where X is the text, V denotes the image associated with the text, and each text contains some named entities $A_i = \{a_{i,k}\}_k^{N_a}$. We denote the text input as $X_i = \{[CLS], x_{i,1}, x_{i,2}, ..., x_{i,n}, [SEP]\}$, where $x_{i,j}$ is the word of text X_i, $[CLS], [SEP]$ are special tokens of BERT. We use BERT to obtain representation of each token in text $\mathbf{h}_{i,j}$, and representation of the whole text $\mathbf{h}_i^g$, which can be formulated as:

$$\mathbf{H}_i = \{\mathbf{h}_{i,j}\}_{j=1}^n = \mathbf{BERT}(X_i) \in \mathbb{R}^{n \times d_1}, \quad (1)$$

where d_1 stands for the hidden size of BERT. The representation of token $[CLS]$ stand by the whole text information denoted by $\mathbf{h}_i^g$. And then leverage a mapping function $\mathbf{E}(\cdot)$ to obtain unified text representation $\mathbf{u}_{gi}^s$ and $\mathbf{u}_{ei,k}^s$ in the joint representation space:

$$\mathbf{u}_{gi}^s = \mathbf{E}(\mathbf{h}_i^g) \in \mathbb{R}^{d_2}, \quad (2)$$

$$\mathbf{u}_{ei,k}^s = \mathbf{E}(\psi(\{\mathbf{h}_{i,j}\}_{x_{i,j} \in a_{i,k}})) \in \mathbb{R}^{d_2}, \quad (3)$$

where d_2 stands for the dimension of the joint representation space, $\psi(\cdot)$ is a pooling operation.

Image Encoder: To link text and images tightly together, we directly utlize the image encoder of the CLIP model (Radford et al., 2021) pre-trained on millions of image-text pairs, to extract image features. The image encoder is a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021), which encode each image to a vector g_i:

$$\mathbf{g}_i = \mathbf{ViT}(V_i) \in \mathbb{R}^{d_2}. \quad (4)$$

During the training process, the parameters of the image encoder are frozen.

2.2 Knowledge Distillation from CLIP

As we said in the introduction, in order to take advantage of fine-grained visual concept and bridge the two modality better, we take CLIP (Radford et al., 2021) model as teacher model. CLIP outperformance fully supervised ResNet (He et al., 2016) on a lot of image classification datasets such as Imagenet (Deng et al., 2009), under zero shot setting, which prove CLIP model has learned fine-grained image concept. And by introducing a priori knowledge of CLIP, some unrelated image-text pairs can be discarded. In previous section **MGC** leverage CLIP's Visual Transformer as image encoder. Proposed method adopts CLIP's text encoder, a pre-trained transformer, as the teacher text encoder, so as to allow the representation of the text linked the image's better. As Figure 2 top part illustrating, We constrain the text representation in terms of the whole sentence (global knowledge distillation) and each entity (local knowledge distillation).

Because CLIP's pre-training process only focuses on the overall representation of the text and does not learn the representation of each word, for each entity span in the text $A_i = \{a_{i,k}\}_k^{N_a}$, teacher text encoder encode them as a sentence. It can be formulated as:

$$\mathbf{u}_{gi}^t = \textbf{Transformer}_{CLIP}(X_i) \in \mathbb{R}^{d_2}, \quad (5)$$

$$\mathbf{u}_{ei,k}^t = \textbf{Transformer}_{CLIP}(a_{i,k}) \in \mathbb{R}^{d_2}. \quad (6)$$

The global knowledge distillation constrain the overall representation of the text, while the local knowledge distillation constrain the representation of the entity:

$$\mathcal{L}_G^{KD} = \sum_{i=1}^{B} ||\mathbf{u}_{gi}^t - \mathbf{u}_{gi}^s||_2 \quad (7)$$

$$\mathcal{L}_E^{KD} = \sum_{i=1}^{B} \sum_{k=1}^{N_a} ||\mathbf{u}_{ei,k}^t - \mathbf{u}_{ei,k}^s||_2 \quad (8)$$

In order to make the overall text representation and the entity text representation as similar as possible to the CLIP text encoder output, we minimized the Euclidean Distance between the two representation.

2.3 Multi-Granularity Contrastive Mechanism

To align the text and image, following recent studies on contrastive learning (Radford et al., 2021; Jia et al., 2021) , we propose global contrastive loss to push representation of matched image-text pairs together while pushing those unrelated image-text pairs apart. We assume that most of the image-text pairs in the dataset are related. As Figure 2 top part illustrating, we compute text to image similarity matrix:

$$A^G = \{a_{i,j}\} = \{g_i^\top u_{gj}^s\} \in \mathbb{R}^{B \times B}. \quad (9)$$

Before calculating the dot product we normalize the representation vectors from two modalities first. So the largest score value should be on the diagonal of the matrix:

$$\mathcal{L}_G^C = -\frac{1}{B} \sum_{i=1}^{B} log \frac{exp(a_{i,i}/\tau)}{\sum_{j=1}^{B} exp(a_{i,j}/\tau)}, \quad (10)$$

where τ denotes the temperature hyperparameter. Similarly, inspired by weakly supervised learning (Li et al., 2020; Wang et al., 2021), we propose local contrastive loss to push representation of matched image-entity pairs together while pushing those unrelated image-entity pairs apart. We assume that at least one of the entity should be related to the associated image.

$$u_{ei}^s = \underset{u_{ei,j}^s}{\textbf{argmax}} \{g_i^\top u_{ei,j}^s\} \quad (11)$$

$$A^E = \{a_{i,j}\} = \{g_i^\top u_{ej}^s\} \in \mathbb{R}^{B \times B}, \quad (12)$$

where u_{ei}^s stands for the most associated entity in text X_i with the image V_i. The local contrastive loss is calculated in the same way as global contrastive loss.

2.4 Output Module

The output module, illustrated as Figure 2 bottom part, aims at fusing the representations from the two modalities and predicting the label of each token. We leverage multimodal transformer proposed by Yu et al. (2020) to obtain the multimodal representation. First we get image-aware word representation by employ an m-head cross-modal attention, which treats visual representation $\mathbf{g}$ as query, text representation $\mathbf{H}$ as key and value:

$$\mathbf{C}_i(\mathbf{H}, \mathbf{g}) = softmax(\frac{[\mathbf{W_{qi}g}]^\top [\mathbf{W_{ki}H}]}{\sqrt{d/m}} [\mathbf{W_{vi}H}]^\top), \quad (13)$$

$$\mathbf{M}(\mathbf{H}, \mathbf{g}) = \mathbf{W}'[C_1(\mathbf{H}, \mathbf{g}), ..., C_m(\mathbf{H}, \mathbf{g})]^\top, \quad (14)$$

$$\widetilde{\mathbf{P}} = \text{LN}(\mathbf{g} + \mathbf{M}(\mathbf{H}, \mathbf{g})) \quad (15)$$

$$\mathbf{P} = \text{LN}(\widetilde{\mathbf{P}} + \textbf{FFN}(\widetilde{P})) \quad (16)$$

where $\mathbf{C}_i$ refers to the i-th head of cross-modal attention, $\{\mathbf{W_{qi}}, \mathbf{W_{ki}}, \mathbf{W_{vi}}\} \in \mathbb{R}^{d_1/m \times d_1}$ and $\mathbf{W}' \in \mathbb{R}^{d_1 \times d_1}$ are learnable parameters, FFN is the feed-forward network (Vaswani et al., 2017), LN is the layer normalization (Ba et al., 2016). And then, taking $\mathbf{P}$ as key and value, $\mathbf{H}$ as query, feed them into transformer layer to generate the final image-aware word representation $\mathbf{A} \in \mathbb{R}^{n \times d_1}$. Similarly, for word-aware visual representation, the fusion module adopt a cross-modal attention, which treats visual representation $\mathbf{g}$ as key and value, text representation $\mathbf{H}$ as query to get word-aware representation $\mathbf{Q} \in \mathbb{R}^n \times d_1$. To get the final representation, representation $\mathbf{Q}$ need to pass a visual gate, as follows:

$$\mathbf{c} = \sigma(\mathbf{W_a A} + \mathbf{W_q Q}) \in \mathbb{R}^n, \quad (17)$$

$$\mathbf{B} = \mathbf{c} \cdot \mathbf{Q} \in \mathbb{R}^{n \times d_1}, \quad (18)$$

where $\mathbf{W_a}, \mathbf{W_q} \in \mathbb{R}^{d_1 \times d_1}$ are learnable parameters. We can obtain the final output by concatenate representation two final representation from two modalities, which is $\mathbf{S} = [\mathbf{A}, \mathbf{B}] \in \mathbb{R}^{n \times 2d_1}$.

And to take advantage of the correlations between labels in neighbouring, we use Conditional Random Fields (CRF) (Lafferty et al., 2001) as decoder. The objective function of the MNER task is to maximum conditional likelihood estimation of CRF, as known as minimizing the log likelihood. Formally,

$$\mathcal{L}_{mner} = -\sum_i \log p(y|X). \tag{19}$$

2.5 Model Training

In the training process, our overall objective function is to minimum the combination of MNER task loss, contrastive loss and knowledge distillation loss. Our final loss function is given by

$$\mathcal{L} = \mathcal{L}_{mner} + \lambda(\mathcal{L}_G^C + \mathcal{L}_E^C) + \beta(\mathcal{L}_G^{KD} + \mathcal{L}_E^{KD}), \tag{20}$$

where λ and β are hyperparameters.

3 Experiments

This section will introduce the experiments we conduct to evaluate proposed method. The basic settings of the experiment will be described first. Then the performance results comparison with baseline methods will be introduce. Finally, the ablation study and case study will be elaborated.

3.1 Experimental Settings

Datasets: We take two public widely used Twitter datasets for MNER: **Twitter-2015** from Zhang et al. (2018) and **Twitter-2017** from Lu et al. (2018). The named entity types are *Person*, *Location*, *Organization* and *Misc*. We adopts the same configuration as Yu et al. (2020), in which 4,000/1,000/3,257 image-text pairs are used as **Twitter-2015** train/dev/test set, and 4,817/1,032/1,033 image-text pairs are used as **Twitter-2017** train/dev/test set.

Implementation Details: To ensure that the experiments are scientifically valid, our BERT based methods use same pretrained BERT(Devlin et al., 2019) (BERT-BASE-CASED)[1]. The maximum length of the sentence input and the batch size are set to 128 and 64 respectively. The Vision Transformer (ViT) is pretrained by CLIP (Radford et al., 2021) model, whose parameters are frozen during the training process. We adopt AdamW as optimizer(Loshchilov and Hutter, 2017), and the initial learning rate are set as 5e-5. The dimension of the joint representation space is set to 512. The head size of multi-head attention is set as 12. The hyperparameter τ is set to 0.05. Most of the other settings follow Devlin et al. (2019). All the nerual models are implemented with Pytorch, and all the experiments are conduct on NVIDIA RTX 3090 GPUs.

[1]https://github.com/google-research/bert

3.2 Baselines

We compared our approach with competitive text-based NER methods and multimodal-based NER methods. The results with the ♠ maker represent the methods we reproduce, which adopts the same hyperparameters as ours. For a fair comparison, other result of the baselines refer to Yu et al. (2020), Zhang et al. (2021) and Wu et al. (2020).

Text-based NER methods: (1) *BiLSTM-CRF* (Huang et al., 2015): First combine bidirectiional LSTM and CRF layer to solve sequence labeling problem. (2) *CNN-BiLSTM-CRF* (Ma and Hovy, 2016): A classical neural network model for NER, improve by introducing charater-level information. (3) *BERT* (Devlin et al., 2019): A sequence labeling model based on BERT, predict each word label by following a MLP layer. (4) *BERT-CRF*: A sequence labeling model based on BERT, predict each word label by following a CRF layer.

Multimodal-based NER methods: (1) *AdaCAN-CNN-BiLSTM-CRF* (Zhang et al., 2018): A sequence labeling model, which designs an adaptive co-attention network to learn word-aware visual representations from VGGNet (Simonyan and Zisserman, 2015) for each word. (2) *OCSGA* (Wu et al., 2020): A multimodal method adopts Mask-RCNN (He et al., 2020) to introduce object-level visual information to help recognize named entity. (3) *UMT* (Yu et al., 2020): A state-of-the-art approach for MNER, which proposes a multimodal transformer to fuse two modality representations from ResNet and BERT, and use auxiliary entity span detection task to help recognize named entity. (4) *UMT-ViT* (Yu et al., 2020): We use CLIP's Vision Transformer in place of ResNet in UMT. (5) *UMGF* (Zhang et al., 2021): Another state-of-the-art approach for MNER, which introduce visual object information and propose graph-based multimodal fusion to fuse two modality representations.

Modality	Methods	Twitter-2015							Twitter-2017						
		Single Type (F1)				Overall			Single Type (F1)				Overall		
		PER.	LOC.	ORG.	MISC.	P	R	F1	PER.	LOC.	ORG.	MISC.	P	R	F1
Text Only	BiLSTM-CRF	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
	CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
	BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
	BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
Multimodal	AdaCAN-CNN-BiLSTM-CRF	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
	OCSGA	84.68	79.95	56.64	39.47	**74.71**	71.12	72.92	-	-	-	-	-	-	-
	UMT	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	**70.10**	85.28	85.34	85.31
	UMT♠	84.95	81.97	61.15	40.38	70.98	75.36	73.11	90.51	84.09	82.08	64.29	83.79	84.53	84.16
	UMT-ViT♠	85.71	81.36	**63.64**	41.10	72.33	75.91	74.07	91.49	84.92	81.97	67.13	84.30	85.86	85.08
	UMGF	84.26	**83.17**	62.45	42.42	74.49	75.21	74.85	91.92	85.22	83.13	69.83	**86.54**	84.50	85.51
	MGC(Ours)	**85.76**	81.55	62.68	**42.94**	73.50	**76.66**	**75.05**	**92.38**	**85.39**	**83.84**	67.13	86.37	**85.86**	**86.12**

Table 1: Overall performance comparison in Twitter-2015 and Twitter-2017. The maker ♠ refers to the method reproduced by us and adopted same hyperparameters as ours.

3.3 Comparisons with SOTA methods

Table 1 reports the **F1** score (%) of each single named entity type, and overall **P** (Precision,%), **R** (Recall,%), **F1** (%) on two benchmark MNER datasets. From the table, we notice:

(1) Pre-trained based methods are knowledgeable. In text-based method, BERT-CRF outperforms CNN-BiLSTM-CRF of 4.66% and 4.07% F1 score on the two datasets. In multimodal-based methods using BERT as a language model, also outperform LSTM-based methods by a large margin. It is crucial for MNER to adopt a pre-trained language model.

(2) It is useful to introduce visual information in MNER. Compared with text-based methods, multimodal methods outperform them both in single type metric or overall performance. For example, UMT outperforms BERT-CRF 1.60% and 0.75% of F1 score on the two dataset. However, this improvement is not as enormous as adopting pre-trained language model.

(3) Introducing fine-grained visual concept is more helpful. Our approach (MGC) outperforms other multimodal methods on F1 score for two datasets. Besides, UMT are improved by replacing ResNet in UMT method with Vision Transformer pretrained in CLIP. These two phenomena prove that leveraging finer-grained visual concepts can help to take advantage of valid information from images. And it can be found that our method can recall more entities in dataset (1.45% of Recall score on Twitter-2015 and 1.36% of Recall score on Twitter-2017).

(4) Creating a joint representation space for MNER is beneficial. Our framework are based on UMT♠. By adopting our framework, the result are improve significantly (2.04% of F1 score on Twitter-2015 and 1.96% of F1 score on Twitter-2017). So it is of great benefit for MNER to build a joint representation to explore the relationship of image and text.

Method	Twitter-2017						
	Single Type (F1)				Overall		
	PER.	LOC.	ORG.	MISC.	P	R	F
MGC (Ours)	**92.38**	**85.39**	**83.84**	67.13	**86.37**	85.86	**86.12**
-KD	91.02	84.57	82.65	68.75	85.26	85.20	85.23
-Contra.	91.00	84.73	83.46	**68.87**	85.30	85.49	85.40
-Visual.	90.30	75.82	83.24	66.07	80.82	**85.95**	83.31

Table 2: Ablation study of MGC framework.

3.4 Ablation Study

To verify the effectiveness of each component of MGC, we conduct ablation studies on **Twitter-2017**. Here we consider three settings: (1) -KD: removing knowledge distillation from CLIP model. (2) -Contra.: removing multi-granularity contrastive constraint. (3) -Viisual.: removing entire vision-related modules (Such as Vision Transformer).

The results are shown in Table 2, and we can observe that: (1) Both the knowledge distillation from CLIP model and multi-granularity contrastive constraint are beneficial for MNER. In our analysis, this is due to the fact that both of these constraints are essential for building an unified joint representation space. Building such a space can filter the effects of noise images better and match finer-grained visual concepts with text. (2) Our approach also benefits from incorporating visual information to help recognize named entities. In Table 2, the Visual modules contributes +2.81% F1 score on Twitter 2017. And this improvement comes mainly from the fact that the model can predict labels more accurately, because we note that there are a significant drop in Precision score without using visual

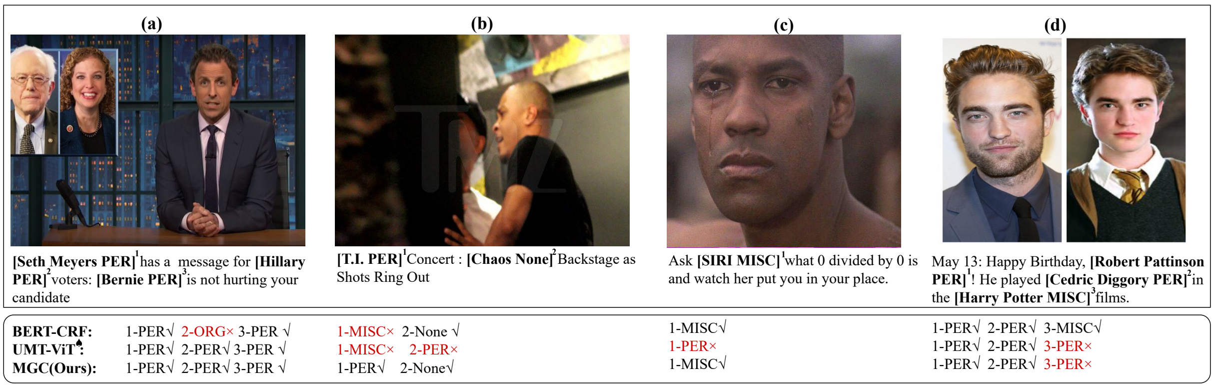


Figure 3: Four cases from different methods predictions in test set of two datasets. The top part shows the image-text pairs in the test set, and the named entities and their types annotated in the datasets are highlighted. The bottom part illustrates three methods predictions on these samples.

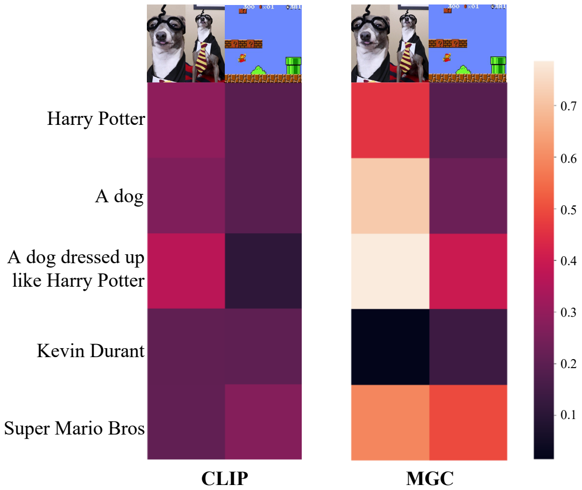


Figure 4: Cases for the cosine similarity of representations in joint representation space from two modalities.

information (5.55% in Precision score).

3.5 Case Study

In order to visualize the similarities and differences between our approach and other approaches and to investigate the unified joint representation space we have mentioned before, we choose some cases to explain.

The effectiveness and limitations of our framework. Figure 3 illustrates four examples of different predictions from three representative approaches (i.e, text-based baseline BERT-CRF, multimodal-based method UMT-ViT♠, and our approach MGC). We can see the Figure 3(a) that the image and the text are highly related. The text-based method without image prompts incorrectly labels "*Hillary*" as "*Location*". While, the multimodal-based method can leverage the visual information to label this case correctly. And Figure 3(b) illustrates the image are very hard to comprehend. If models only know the image contains a person, it is basically impossible to link "*T.I.*" with the image. For instance, BERT-CRF and UMT-ViT label the entity "*T.I.*" with a wrong type "*MISC*". The multimodal-based method UMT-ViT♠ consider that "*Chaos*" is a person, because of superfacial understanding of the image. While, our approach can acquire fine-grained visual concepts to predict correctly. Besides, as Figure 3(c) illustrating, the image content and text are hardly related. Over-consideration of the image may lead models to believe that the image is someone called "*Siri*". So the multimodal-based method UMT-ViT♠ makes a wrong prediction. But our framework can slightly resistant to this noise to keep the result same as the text-based method's. Nevertheless, our approach still has limitations. Since our method can be seen as a kind of weakly supervised approach, it is very difficult to ensure accurate correspondence between entities and images. As shown in Figure 3(d), multimodal-based UMT-ViT♠ and our approach misidentificate of "*Harry Potter*" as a character, while it refers to the film title here. The model that considers only text as input can make predictions correctly.

The joint representation space. The Figure 4 illustrates that our approach creates a unified joint representation space. We take the texts in the left part of the Figure as text encoder's input, and images in the top part of the Figure as image encoder's input, and normalize representations from two modalities. And then, we compute the cosine similarity scores of two modalities' representations. We can figure out: (1) Our proposed method can leverage fine-grained visual concepts. The similarity score of representation of "*Harry Potter*" and "*A dog dressed uplike Harry Potte*" with representation of the first image are high, which illustrates our approach can leverage the concept "Harry Potter" not just a dog. (2) Our proposed method can push related entity-image pairs together while push unrelated entity-image pairs away. In example "*Kevin Durant putting up more bricks than Super Mario Bros.*", there are two entities "*Kevin Durant*" and "*Super Mario Bros*". But only entity "*Super Mario Bros*" is related to the image. Our proposed method draw a large margin between the similarity score of related image-entity and unrelated image-entity (0.3625 in cosine similarity score). (3) Our approach is better suited to the current data set. In the case, we can observe that our approaches similarity score distribution is much sharper than CLIP (Radford et al., 2021). In other words, our method can get a higher similarity score in related image-text pairs and more lower similarity score in unrelated pairs. On the other hand, it is also a limitation to think that our method over-fits the current dataset.

4 Related Work

In this section, we review the related work of our method from: multimodal named entity recognition (MNER) and multimodal representation learning.

4.1 Multimodal Named Entity Recognition

With the popularity of social media, billions of image-text pairs posts are produced everyday. Some study begin to leverage visual information to help recognize named entity (Zhang et al., 2021; Lu et al., 2018; Moon et al., 2018b) or disambiguate named entity (Moon et al., 2018a). MNER has received increasing interest these years, where a lot of approaches has been proposed.

From the perspective of multimodal fusion. Some studies (Zhang et al., 2021; Lu et al., 2018; Moon et al., 2018b) are attention-guided method, and they try to adopt visual information by attention mechanism (Bahdanau et al., 2015). Yu et al. (2020) proposes multimdal transformer which extends multimodal interaction between two modalities in traditional Transformer (Vaswani et al., 2017). Zhang et al. (2021) proposes to leverage a multimodal graph to fuse the representation from two modalities. However, these methods rely on the image and text in the dataset are well aligned. And their methods are always adopt mismatched visual encoder and text encoder, by which it is hard to bridge the information from two modalities.

From the perspective of visual information. Some studies (Zhang et al., 2021; Lu et al., 2018; Moon et al., 2018b; Yu et al., 2020; Zhang et al., 2021) attempt to use general information, such as ResNet features (He et al., 2016), VGG features (Simonyan and Zisserman, 2015). Another studies (Wu et al., 2020; Zhang et al., 2021) try to fuse object positions information to the MNER task. In addition, Chen et al. (2021) try to use image caption generated by model to improve performance. Unlike them our approach attempts to leverage finer-grained visual information, and try to build a unified joint representation space for two modalities to model correspondence better.

4.2 Multimodal Representation Learning

Multimodal representation learning is a fundamental problem in multimodal machine learning, which aims at exploiting complementarity and redundancy of multiple modalities (Baltrusaitis et al., 2019). Good representations are crucial for the performance of machine learning systems, as evidenced behind the recent leaps in performance of natrual language processing (Bengio et al., 2013) and visual object classification (Krizhevsky et al., 2012) systems. The multimodal representation learning methods can be divided into two categories: joint and coordinated. For joint representation, different features from various modalities are represented in the same vector space. While in coordinated representation, each modality has a corresponding projection function that maps it into a coordinated multimodal space. Our MGC framework try to build a joint representation space by using multi-granularity contrastive loss and knowledge from CLIP model, a model pretrained on millions of image-texts pairs.

5 Conclusions

In this paper, we proposed a new framework Multi-Granularity Contrastive Knowledge Distillation (MGC) for multimodal named entity recognition (MNER). We have built a joint representation space by introducing multi-granularity contrastive loss and leveraging the knowledge guidance of CLIP model. We conduct extensive experiments on two benchmark datasets. The experimental results prove the effectiveness of our approach. In the future, we will further explore how to establish a more generalised approach.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Can images help recognize entities? A study of the role of images for multimodal NER. In *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT 2021, Online, November 11, 2021*, pages 87–96. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2020. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397.

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.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. 2018. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Trans. Multim.*, 20(4):927–938.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2557–2568. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne,* *Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1990–1999. Association for Computational Linguistics.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018a. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2000–2008. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018b. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 852–860. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1524–1534. ACL.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 173–179. The Association for Computer Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale chinese multimodal NER dataset with speech clues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2807–2818. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2830–2840. Association for Computational Linguistics.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. 2021. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14090–14100. Computer Vision Foundation / IEEE.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1038–1046. ACM.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3342–3352. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14347–14355. AAAI Press.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. 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), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5674–5681. AAAI Press.