Reducing the Bias of Visual Objects in Multimodal Named Entity Recognition

Xin Zhang
School of Computer Science and Artificial Intelligence,
Wuhan University of Technology
Wuhan, Hubei, China
xinz@whut.edu.cn

Jingling Yuan
School of Computer Science and Artificial Intelligence,
Wuhan University of Technology
Engineering Research Center of Digital Publishing
Intelligent Service Technology, Ministry of Education
Wuhan, Hubei, China
yjl@whut.edu.cn

Lin Li
School of Computer Science and Artificial Intelligence,
Wuhan University of Technology
Engineering Research Center of Digital Publishing
Intelligent Service Technology, Ministry of Education
Wuhan, Hubei, China
cathylilin@whut.edu.cn

Jianquan Liu
NEC Corporation
Tokyo, Japan
jqliu@nec.com

ABSTRACT

Visual information shows to empower accurately named entity recognition in short texts, such as posts from social media. Previous work on multimodal named entity recognition (MNER) often regards an image as a set of visual objects, trying to explicitly align visual objects and entities. However, these methods may suffer the bias introduced by visual objects when they are not identical to entities in quantity and entity type. Different from this kind of explicit alignment, we argue that implicit alignment is effective in optimizing the shared semantic space learning between text and image for improving MNER. To this end, we propose a de-bias contrastive learning based approach for MNER, which studies modality alignment enhanced by cross-modal contrastive learning. Specifically, our contrastive learning adopts a hard sample mining strategy and a debiased contrastive loss to alleviate the bias of quantity and entity type, respectively, which globally learns to align the feature spaces from text and image. Finally, the learned semantic space works with a NER decoder to recognize entities in text. Conducted on two benchmark datasets, experimental results show that our approach outperforms the current state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

multimodal named entity recognition; contrastive learning

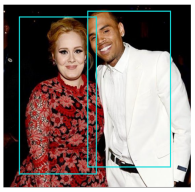
WSDM '23, February27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570485

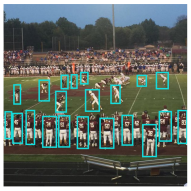
ACM Reference Format:
Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. Reducing the Bias of Visual Objects in Multimodal Named Entity Recognition. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570485

1 INTRODUCTION

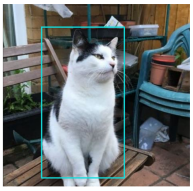
As an important research direction of NER, multimodal named entity recognition (MNER) has attracted more and more attention, due to its research significance in multimodal deep learning and wide applications, such as social media posts. It significantly extends the conventional text-based NER by taking images as additional inputs[2, 14, 23, 25–28]. The assumption is that visual information in images can help classify the entity types in text, especially when text semantics are ambiguous.



(a). Happy Birthday to [**Chris Brown PER**] and [**Adele PER**]



(b). [**Bishop Chatard ORG**] VS. [**Lawrence Central ORG**]



(c). [**Jeremy Corbyn PER**]'s cat El Gato just threw some glorious shade at [**David Cameron PER**]. Yes, really

Figure 1: Three Examples of Multimodal Named Entity Recognition in Social Media.

As a key of MNER, it is generally believed that explicit alignment can unearth the fine-grained correspondence between text and image. As shown in Fig. 1.a, by observing the image containing two people (visual objects), it is easy to classify the types of "Chris Brown" and " Adele" (entities) in the text as "PER". However, such explicit alignment will inevitably have problems when visual objects and entities are inconsistent in quantity or type. For example,

in Fig. 1.b, there are many detected objects (n visual objects) in the image, which makes it difficult to explicitly align them with the "Bishop Chatard" and "Lawrence Central" (2 entities) in the text. And in Fig. 1.c, it is expected to find some "people" ("PER" type) in the image to align with "Jeremy Corbyn" or "David Cameron" ("PER" type) in the text, but there is a "cat" ("MISC" type) in the image. Just like these examples, when there is no such precise correspondence between text and image, it will bring difficulties for the graph-based method by establishing the relationship between entities and visual objects [26] , and the method of taking visual objects as a semantic representation of image [2, 23, 28].

As we can see, recent studies in MNER mainly focus on capturing various semantic correspondences between multimodal semantic units (entities and visual objects). They consider leveraging detected visual objects to help identify entities in the text by explicitly aligning them. In such a case, these methods can work well when the detected visual objects exactly correspond to the entities. As a result, the bias caused by visual objects may mislead the recognition of entities when the detected visual objects and entities are inconsistent in quantity or type. According to our data analysis, the inconsistency of quantity alone accounts for 85.58% and 84.04% in two MNER datasets. It is necessary to propose a new multimodal alignment method to address these biases to alleviate the bias from visual objects.

Contrastive learning, as a recent popular self-supervised method, has been widely used in many fields[3][4][13]. It can effectively exploit the natural pairing relationship between text and image, which can be used to enhance the multimodal latent semantic space learning. The reasoning behind this is that contrastive learning can narrow the semantic distance between positive samples and widen the semantic distance between negative samples by constructing positive and negative samples with different feature distributions. Inspired by this inference, we try to optimize the learning of the text-image shared latent semantic space by combining MNER with a de-bias contrastive learning, in this way to effectively optimize implicit alignment and alleviate the bias from visual objects for better NER performance.

In this paper, we propose a novel de-bias contrastive learning based approach, which combines MNER with cross-modal contrastive learning to alleviate the bias from visual objects for MNER in social media posts. Specifically, we first introduce a multimodal interaction module consisting of multilayer self-attention and cross-modal attention to learn text-image shared latent semantic space. To effectively alleviate the bias in quantity, we propose a visual object density guided hard sample mining strategy to select text-image pairs with high visual object density as hard samples. To effectively alleviate the bias in entity type, we adopt a debiased contrastive loss to replace standard contrastive loss, which can alleviate the bias caused by negative samples with wrong types. Then, we combine MNER with the above de-bias contrastive learning to optimize the learning of the latent semantic space between visual and textual representations. Finally, we exploit the textual representation in semantic space with a NER decoder to perform entity labeling. Compared with previous models, ours can effectively alleviate the bias caused by visual objects in multimodal alignment from multiple perspectives (quantity and entity type).

Our main contributions can be summarized as follows:

- We propose a de-bias contrastive learning based modal for the task of MNER, which achieves multimodal implicit alignment by optimizing the learning of text-image shared latent semantic space.
- We propose a novel de-bias contrastive learning, which combines a hard sample mining strategy and adopts a debiased contrastive loss. It aims at alleviating the bias caused by visual objects in quantity and entity type.
- Conducted on the Twitter 2015 and 2017 datasets, the experimental results demonstrate that our proposed model outperforms state-of-the-art methods.

2 RELATED WORK

2.1 Multimodal NER

As multimodal data become increasingly popular on social media platforms, NER in the social media domain has raised broad concerns. The multimodal NER task was first explored by Zhang et al.[27], Moon et al. [17], and Lu et al.[14] in the same period. They take the approach of encoding the entire image, which implicitly interacts the information of two modalities. As a typical implicit alignment method, Yu et al.[25] propose a multimodal interaction module to capture the inter-modality dynamics between words and images. However, simple multimodal interaction will lead to poor semantic alignment and failure to find latent semantic correspondences. Thus, recent work began to study explicit alignment methods in MNER. Chen et al. [2] introduced image attributes and knowledge to help improve named entity extraction. Wu et al.[23] and Zheng et al. [28] proposed to mine relations between fine-grained visual objects and entities to predict. Similarly, Zhang et al.[26] proposed using a unified multimodal graph to capture various semantic relationships between words and visual objects. However, these methods all focus on the visual objects in the image and attempt to achieve explicit alignment by learning association weights between entities and visual objects. Different from the above methods, ours attempts to combine MNER with contrastive learning to achieve implicit alignment by optimizing the learning of text-image shared latent semantic space.

2.2 Contrastive Learning

Contrastive learning has become a rising domain because of its significant success in various CV and NLP tasks. Several researchers (Chen et al.[3]; Kim et al.[10]; Misra and Maaten[16]) proposed to make the representations of the different augmentation of an image agree with each other and showed positive results. The main difference between these works is their various definition of image augmentation. At the same time, researchers in the NLP domain have also started to work on finding suitable augmentation for text (Giorgi et al.[7]; Wu et al.[22]; Yang et al.[24]). However, a major limitation of the above methods is that they are only uni-modal contrastive learning. Recently, with the rise of multimodal pre-trained models, many studies have incorporated the multimodal contrastive learning in their methods (Radford et al.[18]; Li et al.[13]; Li et al.[12]). However, most of them directly use standard cross-modal contrastive learning based on random samples or only perform data augmentation based on text, which does not consider optimizing

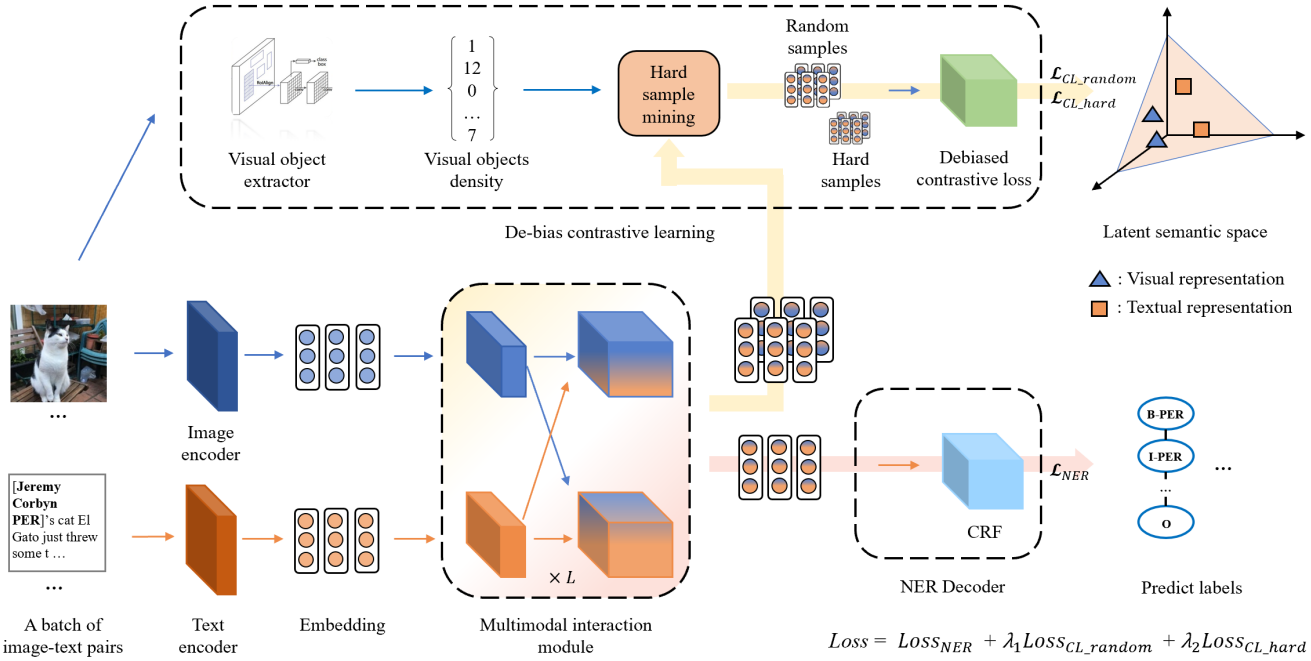


Figure 2: The Overall Architecture of Our De-bias Contrastive Learning based Approach. (Our approach achieves implicit alignment and alleviates the bias of visual objects by combining MNER with a de-bias contrastive learning to optimize the learning of the text-image shared latent semantic space)

from the visual objects, but this is exactly the difficulty we face in MNER. Therefore, we propose a hard samples mining strategy based on visual object density in the image and adopt a debiased contrastive loss, so as to address the bias caused by visual objects in multimodal alignment.

3 THE PROPOSED METHOD

In this section, we present a novel de-bias contrastive learning based approach. Given text-image pairs, we aim to use visual information in images to help recognize the words as predefined entity types. Compared with the classic two-stream model in the multimodal task, our model adds a de-bias contrastive learning module. The overall architecture is shown in Figure 2. We will first introduce the process of text and image embedding, and then detail the components of the proposed de-bias contrastive learning. Finally, we give a technical explanation of our decoder for NER.

3.1 Text and Image Embedding

3.1.1 Text Embedding. Due to the capability of giving different representations for the same word in different contexts, same as previous work[25][26], BERT[6] is chosen as our textual feature extractor to obtain contextualized representation. Then, to project the textual representation into the same semantic space as the visual representation, a projection head composed of linear transformation layer, Relu activation function layer and dropout layer is added after BERT. Formally, let $S = (s_0, s_1, ..., s_{n-1})$ be the input text. As shown in Figure 2, S is fed to the text encoder consisting of BERT and projection head to obtain the embedding of textual representation $\mathbf{W} = (\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_{n-1})$, where $\mathbf{w}_i \in R^d$ is the contextualized representation for s_i.

3.1.2 Image Embedding. As one of the state-of-the-art CNN models for image recognition, Residual Network (ResNet)[9] has shown its capability to extract meaningful representations of input images. Therefore, same as previous work[25][26], ResNet is chosen as our visual feature extractor to obtain contextualized representation, which will split each input image into 7×7 = 49 visual blocks with the same size. Then, like textual feature extractor, a projection head is used to project the visual representation into the same semantic space as the textual representation. Specifically, the input image V will be resized to 224 × 224 pixels and then input to the visual encoder consisting of ResNet and projection head to obtain the embedding of visual representation $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{49})$, where $\mathbf{v}_i$ is the visual representation for the i-th visual block.

3.1.3 Multimodal Interaction (MMI) Module. To effectively learn the textual representation that incorporates visual information and the visual representation that incorporates textual information, a multimodal interaction module is proposed, which stacks L multimodal fusion layers to encode the input text-image pairs. At each fusion layer, intra- and inter-modal fusions are sequentially conducted to update the visual and textual representations. This way, the final visual and textual representations simultaneously encode the context within the same modality and the cross-modal semantic information.

Specifically, in the l-th fusion layer, both updates of textual representation $\mathbf{H}_w^{(l)}$ and visual representation $\mathbf{H}_v^{(l)}$ mainly involve the following steps:

Intra-modal Fusion. Self-attention[21] is employed to generate the contextual representation of each modal. Formally, the textual contextual representation $\mathbf{C}_w^{(l)}$ is calculated as follows:

$$\mathbf{C}_w^{(l)} = SA(\mathbf{H}_w^{(l-1)}, \mathbf{H}_w^{(l-1)}, \mathbf{H}_w^{(l-1)}) \tag{1}$$

where $SA(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is a multi-head self-attention function taking a query matrix $\mathbf{Q}$, a key matrix $\mathbf{K}$ and a value matrix $\mathbf{V}$ as inputs. Similarly, the visual contextual representation $\mathbf{C}_v^{(l)}$ is generated as:

$$\mathbf{C}_v^{(l)} = SA(\mathbf{H}_v^{(l-1)}, \mathbf{H}_v^{(l-1)}, \mathbf{H}_v^{(l-1)}) \tag{2}$$

Inter-modal Fusion. Cross-attention[20] is applied to gather the cross-modal semantic information between two modalities. Formally, the fused representation $\mathbf{R}_w^{(l)}$ of text is generated as:

$$\mathbf{R}_w^{(l)} = CA(\mathbf{C}_w^{(l-1)}, \mathbf{C}_v^{(l-1)}, \mathbf{C}_v^{(l-1)}) \tag{3}$$

where $CA(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is a multi-head cross-attention function. Similarly, the fused representation $\mathbf{R}_v^{(l)}$ of visual is generated as:

$$\mathbf{R}_v^{(l)} = CA(\mathbf{C}_v^{(l-1)}, \mathbf{C}_w^{(l-1)}, \mathbf{C}_w^{(l-1)}) \tag{4}$$

For simplicity, the descriptions of layer normalization[1] and residual connection[9] are omitted in the above description.

Then, the position-wise feed forward networks (FFN)[21] are adopted to generate the final textual representation $\mathbf{H}_w^{(l)}$ and final visual representation $\mathbf{H}_v^{(l)}$ as the output of l-th fusion layer :

$$\mathbf{H}_w^{(l)} = FFN(\mathbf{R}_w^{(l)}) \tag{5}$$

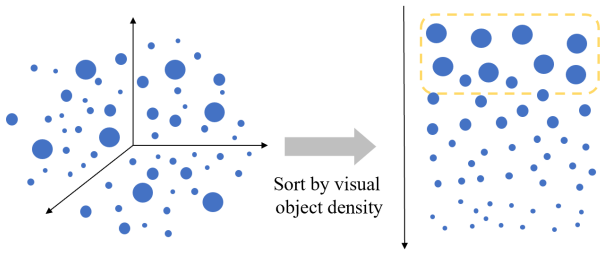
$$\mathbf{H}_v^{(l)} = FFN(\mathbf{R}_v^{(l)}) \tag{6}$$

Meanwhile, $\mathbf{H}_w^{(l)}$ and $\mathbf{H}_v^{(l)}$ also will be the input of the $(l+1)$-th fusion layer. By this layer-by-layer iteration, MMI gradually learns the accurate image and text representations in the text-image shared latent semantic space.

3.2 De-bias Contrastive Learning

3.2.1 Hard Sample Mining Strategy. As we know, contrastive learning can effectively narrow the distance between positive samples and widen the distance between negative samples. Thus, the construction of negative samples will directly affect the performance of contrastive learning.

At the same time, we notice a high quantity inconsistency between visual objects and entities in MNER datasets. According to our data analysis, 65.73% and 65.01% text-image pairs have more visual objects than entities in the two datasets, and we observe that visual object densities tend to be higher in these data. When too many visual objects in the image are not conducive to the precise alignment between text and image. Based on this observation, to effectively alleviate the bias in quantity, we consider whether these text-image pairs with high visual object density can be used as negative samples in contrastive learning. However, the negative samples of standard contrastive learning are constructed by combining images and texts with different pairing relationships in the same batch. Since the model also needs to be trained for the NER task, we cannot purposefully select batch data with a specific feature distribution. Therefore, different from existing contrastive learning methods, we consider selecting some samples with high visual object density from random samples as hard samples. This way, we can get the hard negative samples for contrastive learning to alleviate the quantity bias.



(1). Random samples (2). High visual object density hard samples

Figure 3: Random Samples and Hard Samples for Contrastive Learning. (The size of the dot represents the density of visual objects.)

Specifically, we take the text-image pairs fused by the MMI module as the input for de-bias contrastive learning. Since these text-image pairs in each batch are randomly sampled, we call these text-image pairs random samples, denoted as $(\mathbf{R}_w^r, \mathbf{R}_v^r)$.

After this, first, we detect the visual objects in each input image via the pre-trained object detection model[8], and calculate visual object density based on the number n of detected objects and the size p of images, denoted as $D = (d_0, d_1, ..., d_{n-1})$, where $d = n/p$. As shown in Figure 3, each dot represents an image, and its corresponding text is omitted. The size of the dot represents the visual object density of an image. Thus, the larger the dot, the higher the visual object density. Then, to select hard samples with high visual object density, we sort the dots from largest to smallest, and take the largest $\mathcal{N}$ of them as our hard samples $\mathbf{R}_w^h, \mathbf{R}_v^h$.

Algorithm 1: Hard sample mining strategy

Input: random samples: $(\mathbf{R}_w^r, \mathbf{R}_v^r)$
Output: hard samples: $(\mathbf{R}_w^h, \mathbf{R}_v^h)$

$(\mathbf{R}_w^h, \mathbf{R}_v^h) \leftarrow \{\}$;
$D \leftarrow \{\}$;
for $(\mathbf{R}_w^r, \mathbf{R}_v^r)_i$ ***in*** $(\mathbf{R}_w^r, \mathbf{R}_v^r)$ **do**
  $\{obj\} \leftarrow ObjectDetection((\mathbf{R}_v^r)_i)$;
  $n \leftarrow Count(\{obj\})$;
  $p \leftarrow GetSize((\mathbf{R}_v^r)_i))$;
  $d \leftarrow n/p$;
  $D_i \leftarrow d$;
end
$(\mathbf{R}_w^r, \mathbf{R}_v^r)_{sorted} \leftarrow DescendingSort((\mathbf{R}_w^r, \mathbf{R}_v^r), D)$;
$(\mathbf{R}_w^h, \mathbf{R}_v^h) \leftarrow FirstN(\mathbf{R}_w^r, \mathbf{R}_v^r)_{sorted})$;
return $(\mathbf{R}_w^h, \mathbf{R}_v^h)$

The main process of hard sample mining is shown in Algorithm 1, where *GetSize* means to return the input image size, *DescendingSort* means to sort the former according to the latter in descending order, *FirstN* means to return the first N values from the input.

Based on the above hard sample mining strategy, we obtain the hard samples, an extension of random samples. Since the hard samples are selected from random samples, in essence, the weight of contrastive learning on these selected samples (text-image pairs with high visual object density) is enhanced.

Finally, these hard samples will be input into the contrastive learning module with random samples to get their contrastive learning loss. In this way, the MMI module is optimized to learn a better text-image shared latent semantic space, and the bias caused by quantity inconsistency can also be alleviated.

3.2.2 De-bias Contrastive Learning Loss. By observing the features of text-image pairs in the MNER dataset, we found that many texts often contain multiple different entity types. In contrast, the visual objects in corresponding images may only contain one or part of these entity types, and other images in the same batch may contain the remaining types. In such cases, some text-image pairs that are actually positive samples are wrongly regarded as negative samples, which will inevitably negatively impact contrastive learning. We call this the bias in entity type.

To effectively alleviate this bias, we notice that constructing specified hard samples for contrastive learning is unsuitable because it is difficult to obtain accurate image type labels. Therefore, we consider addressing the bias from the perspective of the loss function. Specifically, our proposed de-bias contrastive learning adopts a debiased contrastive loss [5] to replace the standard contrastive loss(NTXent)[3]. The key idea of the debiased contrastive loss is to indirectly approximate the distribution of negative samples, which corrects the sampling bias of negative samples even without knowing the true type labels.

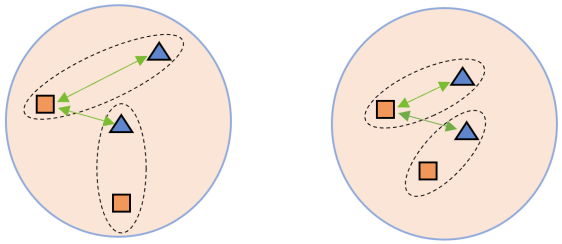
During each training step, we construct mini-samples of size N, and take text-image pairs in samples as negative examples of each other. In this way, we can create large volumes of positive examples $\mathcal{X}^+$ and negative examples $\mathcal{X}^-$ for each text-image pair. Each data point is trained to find its counterpart among $(2N-2)$ in-batch negative samples. Then the loss function for a positive pair of examples is defined as:

$$\mathcal{L}_{(\mathcal{X}^+,\mathcal{X}^-)} = -\log \frac{\sum_{x\in\mathcal{X}^{\{+\}}} f(x)}{\sum_{x\in\mathcal{X}^{\{+\}}} f(x) + \frac{Q}{N}\sum_{x\in\mathcal{X}^{\{-\}}} f(x)} \tag{7}$$

where $f(x)$ means $exp(d(x)/\tau)$ and $d(\cdot)$ indicates the cosine similarity function, τ denotes the temperature parameter. Q is the weighting parameter, and N is the number of negative samples.

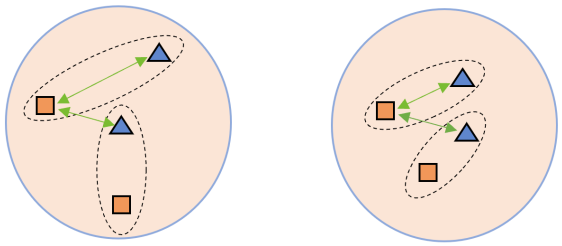
Finally, we average all N in-batch classification losses to obtain the final contrastive loss.

As shown in Figure 4, in the case of an accurate selection of negative samples, compared with the latent semantic space before contrastive learning, the distance between the positive pairs is closer, and the distance between the negative pairs is relatively farther. However, when the selection of negative samples is inaccurate, standard contrastive learning tends to pull them away, as shown in Figure 4. In contrast, contrastive learning based on the debiased contrastive loss can be closer to the real situation without pulling the wrong negative samples away, as shown in Figure 5. It effectively alleviates the bias caused by unreasonable negative samples, that is, the bias in entity type mentioned above.



(1). Latent semantic space before contrastive learning (2). Latent semantic space after contrastive learning

Figure 4: Contrastive learning when negative samples are accurate. (Green means semantically relevant and red means semantically irrelevant.)



(1). Latent semantic space before contrastive learning (2). Latent semantic space after contrastive learning

Figure 5: Contrastive learning when negative samples is inaccurate. (Green means semantically relevant and red means semantically irrelevant.)

Based on the two sets of samples $(\mathbf{R}^r_w, \mathbf{R}^r_v)$ and $(\mathbf{R}^h_w, \mathbf{R}^h_v)$ by randomly sampling and visual object density guided hard sample mining in the previous section, as described above, we construct two sets of positive and negative examples, denote as $(\mathcal{X}^+_r, \mathcal{X}^-_r)$ and $(\mathcal{X}^+_h, \mathcal{X}^-_h)$, and calculate their contrastive learning loss separately:

$$\mathcal{L}_{cl\_r} = \mathcal{L}_{(\mathcal{X}^+_r,\mathcal{X}^-_r)} \tag{8}$$

$$\mathcal{L}_{cl\_h} = \mathcal{L}_{(\mathcal{X}^+_h,\mathcal{X}^-_h)} \tag{9}$$

To sum up, with the proposed hard sample mining strategy and the adopted debiased contrastive learning loss, we can accurately optimize the learning of latent semantic space and alleviate the bias in quantity and entity type to improve the implicit alignment.

3.3 NER Decoder

Since visual information has been incorporated into textual representation via the MMI module enhanced by contrastive learning, we introduce a decoder to perform conditional sequence labeling on textual representation.

It has been shown that Conditional Random Fields (CRF) have the ability to mine information from semantic space for sequence labeling and have played a good role in many MNER tasks[14, 23, 25–27]. Therefore, it is considered as our NER decoder and computes the prediction loss simultaneously with the debiased contrastive loss. Before decoding, we additionally introduce the image types identified by the pre-trained model[18] to optimize the final prediction.

The final loss of the DebiasCL can be expressed as the weighted sum of $Loss_{NER}$, $Loss_{CL\_random}$ and $Loss_{CL\_hard}$:

$$Loss = Loss_{NER} + \lambda_1 Loss_{CL\_random} + \lambda_2 Loss_{CL\_hard} \quad (10)$$

4 EXPERIMENT SETTING

4.1 Dataset

Following previous works in MNER[14, 23, 25–27], we take *Pre.*, *Rec.* and F_1 as our main evaluation metric and conduct experiments on the two MNER datasets (i.e., Twitter-2015 and Twitter-2017), which are respectively provided by Zhang et al.[27] and Lu et al.[14]. The entity types include "Person", "Location", "Organization" and "Misc". The tagging schema is BIO[19]. Moreover, a default image is leveraged to replace the missing images in Twitter 2017, like Zhang et al.[26]. Two datasets are divided into training, development and testing parts following the same setting as Yu et al.[25]. Table 1 shows the number of entities for each type and the counts of multimodal tweets in detail.

Table 1: The Statistics Summary of Two Twitter Datasets.

Entity Type	Twitter-2015			Twitter-2017		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	928	247	839	1674	375	395
Miscellaneous	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Num of Tweets	4000	1000	3257	3373	723	723

4.2 Baseline Methods

Different from Yu et al.[25] and Zhang et al.[26], we mainly compare three groups of baseline systems with our approach.

The first group contains several representative text-based NER approaches:

- **CNN-BiLSTM-CRF**[15]: It is a classical neural network for NER based on CNN and LSTM.
- **HBiLSTM-CRF**[11]: It is an improvement of CNN-BiLSTM-CRF, replacing the bottom CNN layer with the LSTM layer.
- **BERT**[6]: It is a BERT-based model with a softmax layer for entity prediction.
- **BERT-CRF**: It is a variant of BERT replacing the softmax layer with a CRF layer.

The second group contains several competitive implicit alignment-based multimodal approaches for MNER:

- **VG**[14]: It utilizes a visual attention and a gate mechanism to mine implicit information from the global image to enrich word representation based on HBiLSTM-CRF.
- **ACoA**[27]: It designs an adaptive co-attention network for learning the shared implicit semantics between text and global image based on CNN-BiLSTM-CRF.
- **UMT**[25]: It extends Transformer[21] to obtain both image-aware word representations and word-aware visual representations, and incorporates an auxiliary entity span detection module to alleviate visual bias.

The three group contains several competitive explicit alignment-based multimodal approaches for MNER:

- **OCSGA**[23]: It incorporates object-level visual information with textual representations for explicit alignment.
- **AGBAN**[28]: It explicitly extracts entity-related features from both visual objects and text, and combines adversarial training to fuse two different representations.
- **IAIK**[2]: It explicitly introduces image attributes and knowledge to help improve named entity extraction.
- **UMGF**[26]: It is the state-of-the-art approach for MNER, which exploits the explicit semantic correspondences by a unified text-image graph that takes visual objects and words as nodes.

4.3 Parameter Settings

Our model is implemented by the PyTorch framework. We set the maximum length of text input and batch size to 128 and 16. In our approach, the word embeddings are initialized by the uncased $BERT_{base}$[6] with a dimension of 768. The visual embeddings are initialized by $ResNet152$ with a dimension of 2048. After the projection head, the dimension of each modality is transformed into 200. To better learn the text-image shared latent semantic space, like UMGF[26], our heads of multi-head attention and number of layers in the MMI module are set to 8 and 12, the latter is double that in UMGF.

Based on best-performed development results, The learning rate of the BERT, the MMI module and other parts are respectively set to 5e-5, 1e-4, and 0.1. The temperature parameter of our de-bias contrastive learning is 10. The ratio between $Loss_{NER}$, $Loss_{CL\_random}$, $Loss_{CL\_hard}$ is 1:1:2. There are only two parameters that differ on two datasets. The dropout rates of the MMI module are 0.2 and 0.25, and the numbers of hard samples N in Section 3.2.1 are 4 and 5. Other parameters are the same and set by the development. The source code of this paper can be found in https://github.com/xinzcode/DebiasCL.

5 RESULTS AND DISCUSSION

Table 2: The Proportion of Inconsistent Data in Two Twitter Datasets.

Data	Twitter-2015 (%)			Twitter-2017 (%)		
	$N_o < N_e$	$N_o = N_e$	$N_o > N_e$	$N_o < N_e$	$N_o = N_e$	$N_o > N_e$
Train	19.15	13.51	67.34	18.98	16.90	61.12
Dev	21.05	14.14	64.81	17.30	14.50	68.20
Test	19.32	15.60	65.08	20.81	16.49	62.70
Total	19.84	14.42	65.74	19.03	15.96	65.01
Overall	85.58 ($N_o \neq N_e$)			84.04 ($N_o \neq N_e$)		

Table 3: Performance Comparison on Two TWITTER Datasets.

Modality	Methods	Twitter-2015							Twitter-2017						
		Single Type (F_1)				Overall			Single Type (F1)				Overall		
		PER.	**LOC.**	**ORG.**	**MISC.**	*Pre.*	*Rec.*	F_1	**PER.**	**LOC.**	**ORG.**	**MISC.**	*Pre.*	*Rec.*	F_1
Text	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
	HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	78.16	80.37
	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
Text+Image	VG	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
	ACoA	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
	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
	OCSGA	84.68	79.95	56.64	39.47	74.71	71.21	72.92	-	-	-	-	-	-	-
	Object-AGBAN	84.75	79.41	58.31	40.72	74.13	72.39	73.25	-	-	-	-	-	-	-
	IAIK	84.28	79.42	58.97	41.47	**74.78**	71.82	73.27	-	-	-	-	-	-	-
	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
	DebiasCL(Ours)	**85.97**	81.84	**64.02**	**43.38**	74.45	**76.13**	**75.28**	**93.46**	84.15	**84.42**	67.88	**87.59**	**86.11**	**86.84**

5.1 Data Inconsistency Analysis

Since we cannot obtain accurate entity type information for each image, we only analyze the quantity inconsistency proposed in Section 1 in our experiments. Specifically, we obtain the number of visual objects in the image and the number of entities in the text by object detection and text labels. The condition for judging the quantity inconsistency is that: the number of visual objects (N_o) in the image is not the same as the number of entities (N_e) in the text.

As shown in Table 3, we calculated the proportion of N_o <, = and > N_e. We can see that the proportion of inconsistent data in the two datasets reaches 85.58% and 84.04%. Therefore, it is necessary to alleviate the bias caused by the inconsistency of quantity. Moreover, we can see that the data of $N_o > N_e$ accounts for most of them. Therefore, handling the data with high visual object densities is necessary.

5.2 Overall Experimental Results

We mainly report the metric *Pre.*, *Rec.* and F_1 for every single type and overall on two benchmark MNER datasets. Table 3 shows the performance comparison of different competitive uni-modal and multimodal approaches. From this table, we can see that:

1) For the uni-modal approaches, BERT-based approaches perform better than the CNN and LSTM apparently in *Pre.*, *Rec.* and F_1. It indicates the obvious advantages of BERT as a text encoder in NER. Regarding the single type and overall results of both datasets, BERT-CRF with CRF decoding performs better than BERT except for the metric *Rec.*. It shows the effectiveness of CRF as the NER decoder.

2) Compared with uni-modal approaches, multimodal approaches generally achieve better performance, proving that visual information is helpful for entity recognition. The most recent approach UMGF performs much better than all multimodal implicit and explicit alignment approaches. The performance gains mainly come from the following reasons: First, UMGF, a recent representative of explicit alignment approaches, utilizes a graph to represent words and visual objects to model explicit alignment relationships between them. Then, UMGF leverages a graph-based multimodal fusion module to mine the semantic correspondence for final entity recognition, which helps to improve entity recognition performance.

3) Different from the explicit alignment approaches such as UMGF, our proposed DebiasCL does not need to model the correspondence between entities and visual objects explicitly. DebiasCL combines MNER with de-bias contrastive learning to fully capture global implicit semantic interaction between text and image, which effectively alleviates the bias caused by visual objects. Compared to the sota model UMGF which is a representative of explicit alignment, DebiasCL has an improvement of 0.43% and 1.33% overall on two datasets, which proves that DebiasCL is effective for improving the performance of MNER as an implicit alignment approach.

5.3 Ablation Study

To investigate the effectiveness of de-bias contrastive learning in our DebiasCL architecture, we perform a comparison between the full DebiasCL and its ablations concerning the hard samples (w/o Hard CL), the entire de-bias contrastive learning module(w/o CL) and the debiased contrastive loss(w/o Debias-loss)

Table 4 shows the results of DebiasCL and its ablated approaches. First, we remove the hard samples in the de-bias contrastive learning module. We can see that performance drops on data where $N_o = N_e$ and $N_o > N_e$ in both datasets, especially in data where $N_o > N_e$, however, rises on the data where $N_o < N_e$. We speculate this is because, after the introduction of hard samples, contrastive learning is more fully optimized for the data with high visual object density, and affects the performance of the data with low visual object density. On this observation, when we further remove the entire de-bias contrastive learning module, the performance all drops significantly, indicating the usefulness of the proposed de-bias contrastive learning for improving MNER performance. In addition, we replace the debiased contrastive loss with the standard contrastive loss in full DebiasCL, and we found that both F_1 decreased, which proves the help of introducing the debiased contrastive loss to alleviate samples bias.

Table 4: Ablation Study of DebiasCL on Different Data.

Methods	Twitter-2015 (F_1)			Twitter-2017 (F_1)		
	$N_o < N_e$	$N_o = N_e$	$N_o > N_e$	$N_o < N_e$	$N_o = N_e$	$N_o > N_e$
DebiasCL	74.45	**75.49**	**74.98**	84.47	**88.76**	**86.31**
DebiasCL w/o Hard CL	**74.47**	75.12	74.16	**84.95**	86.67	85.12
DebiasCL w/o CL	74.18	74.84	73.87	82.11	86.03	84.53
DebiasCL w/o Debias-loss	74.36	75.29	74.38	83.78	87.83	85.69

5.4 Parameter Sensitivity Study

In this section, we evaluate our model on different parameter settings. We have mentioned in Section 3 that we pick up N pairs in batch-size random samples as hard samples. Since we select the N as the batch size of hard samples according to the visual object density from high to low, the value of N directly affects the average visual object density in hard samples. Therefore the value of N is noteworthy because it is crucial for alleviating bias.

Table 5 describes the results of our model influenced by different values of N. The results show that when the N starts to increase from a small size, the performance is poor, which indicates that When there are too few negative examples, it is difficult to learn useful features but noise. When the batch is set larger, there are more hard negative samples, and the effect of learning improves. We can see that DebiasCL achieved the best performance on both datasets when the N was set to 4 and 5. As N becomes larger, the proportion of visual objects in negative samples may decrease. This results in that when N exceeds a specific value, the effect of learning will decrease. On the whole, The results above prove that a proper batch size(N) of hard samples can effectively help de-bias contrastive learning to improve MNER performance.

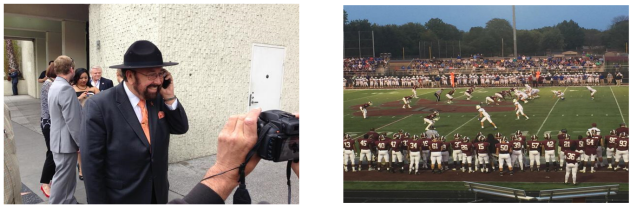
Table 5: The Performance of DebiasCL When Hard Sample Mining Takes Different N.

N	Twitter-2015			Twitter-2017		
	Pre.	*Rec.*	F_1	*Pre.*	*Rec.*	F_1
2	71.92	75.44	73.64	85.78	85.10	85.44
3	72.29	75.29	74.78	86.75	85.31	86.02
4	**74.45**	**76.13**	**75.28**	86.67	**86.18**	86.43
5	74.28	75.15	74.71	**87.59**	86.11	**86.84**
6	74.20	74.44	74.32	86.35	84.91	85.62

5.5 Case Study

Figure 6 shows the case study comparing our method with the BERT-CRF[6] and UMGF[26]. Our method performs better in all the cases due to the enhancement of de-bias contrastive learning.

First, from Figure 6 (a), we can see that according to the textual modality only, BERT-CRF can correctly predict the entity type due to its strong contextual learning. However, UMGF gives a wrong identification of "Hollywood Walk of Fame". We speculate that this may be influenced by the type of visual objects (people) in the image. The graph-based UMGF mistakenly uses the visual objects as graph nodes to change the type of entity in the text.



(a). With @ShotgunTomkelly about to get a star on the [**Hollywood Walk of Fame LOC**][1]

(b).[**Bishop Chatard ORG**][1] VS. [**Lawrence Central ORG**][2]

BERT-CRF:	1-LOC √	1-PER ×, 2-ORG √
UMGF:	1-MISC ×	1-PER ×, 2-ORG √
DebiasCL:	1-LOC √	1-ORG √, 2-ORG √

Figure 6: The Results of DebiasCL Compared with BERT-CRF and UMGF.

Then, as shown in Figure 6 (b), both BERT-CRF and UMGF fail to recognize "Bishop Chatard" in the text. BERT-CRF cannot predict the type of "Bishop Chatard" based on the text alone, misjudging it as PER. For the multimodal method UMGF, it cannot explicitly align the "Bishop Chatard" and "Lawrence Central" with numerous visual objects in the image. Different from the above two approaches, DebiasCL successfully identified "Bishop Chatard" based on implicit alignment, which effectively alleviates the bias in quantity.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel de-bias contrastive learning based approach, which combines MNER with a cross-modal contrastive learning to learn a text-image shared latent semantic space for implicit alignment between text and visual representation. The vision and language are bridged by the latent semantic correspondence. To effectively alleviate the bias caused by visual objects in quantity and entity types, we further propose a hard sample mining strategy guided by visual object density, which can effectively deal with the bias of quantity, and introduce a debiased contrastive loss to alleviate the bias of entity types. Conducted on two benchmark datasets, experimental results demonstrate that our proposed DebiasCL outperforms state-of-the-art methods.

For future work, we plan to investigate the data augmentation method to obtain hard negative samples, which is another way to mine implicit correspondence between vision and language.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China No.62276196 and the Hubei Key Laboratory of Big Data in Science and Technology (Wuhan Library of Chinese Academy of Science) No.20211h0437.

REFERENCES

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *CoRR* abs/1607.06450 (2016).

[2] Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal Named Entity Recognition with Image Attributes and Image Knowledge. In *Database Systems for Advanced Applications - 26th International Conference, DASFAA*. 186–201.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML*. 1597–1607.

[4] Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: Contrastive Instance Learning Framework for Distantly Supervised Relation Extraction. 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*. 6191–6200.

[5] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased Contrastive Learning. In *Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*.

[6] 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*. 4171–4186.

[7] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. 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*.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision, ICCV*. 2980–2988.

[9] 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*. 770–778.

[10] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. 2021. Self-supervised Pre-training and Contrastive Representation Learning for Multiple-choice Video QA. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*. 13171–13179.

[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT*. 260–270.

[12] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS*. 9694–9705.

[13] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. 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*. 2592–2607.

[14] 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*. 1990–1999.

[15] 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*.

[16] Ishan Misra and Laurens van der Maaten. 2020. Self-Supervised Learning of Pretext-Invariant Representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. 6706–6716.

[17] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. 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*. 852–860.

[18] 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*. 8748–8763.

[19] Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing Text Chunks. In *9th Conference of the European Chapter of the Association for Computational Linguistics, EACL*. 173–179.

[20] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. 6558–6569.

[21] 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 *Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS*. 5998–6008.

[22] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive Learning for Sentence Representation. *CoRR* abs/2012.15466.

[23] 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 *The 28th ACM International Conference on Multimedia, MM*. 1038–1046.

[24] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. 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*. 5065–5075.

[25] 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*. 3342–3352.

[26] 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*. 14347–14355.

[27] 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*. 5674–5681.

[28] Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2021. Object-Aware Multimodal Named Entity Recognition in Social Media Posts With Adversarial Learning. *IEEE Trans. Multim.* 23 (2021), 2520–2532.