

# Named Entity and Relation Extraction with Multi-Modal Retrieval

Xinyu Wang<sup>◇‡</sup>, Jiong Cai<sup>◇‡</sup>, Yong Jiang<sup>\*</sup>, Pengjun Xie, Kewei Tu<sup>◇\*</sup>, and Wei Lu<sup>♣</sup>

<sup>◇</sup>School of Information Science and Technology, ShanghaiTech University  
Shanghai Engineering Research Center of Intelligent Vision and Imaging  
Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences  
University of Chinese Academy of Sciences

<sup>♣</sup>StatNLP Research Group, Singapore University of Technology and Design  
{wangxy1,caijiong,tukw}@shanghaitech.edu.cn  
{jiangyong.ml,xpjandy}@gmail.com  
luwei@sutd.edu.sg

## Abstract

Multi-modal named entity recognition (NER) and relation extraction (RE) aim to leverage relevant image information to improve the performance of NER and RE. Most existing efforts largely focused on directly extracting potentially useful information from images (such as pixel-level features, identified objects, and associated captions). However, such extraction processes may not be knowledge aware, resulting in information that may not be highly relevant. In this paper, we propose a novel Multi-modal Retrieval based framework (MoRe). MoRe contains a text retrieval module and an image-based retrieval module, which retrieve related knowledge of the input text and image in the knowledge corpus respectively. Next, the retrieval results are sent to the textual and visual models respectively for predictions. Finally, a Mixture of Experts (MoE) module combines the predictions from the two models to make the final decision. Our experiments show that both our textual model and visual model can achieve state-of-the-art performance on four multi-modal NER datasets and one multi-modal RE dataset. With MoE, the model performance can be further improved and our analysis demonstrates the benefits of integrating both textual and visual cues for such tasks.<sup>1</sup>

## 1 Introduction

Utilizing images to improve the performance of Named Entity Recognition (NER) and Relation Extraction (RE) has attracted increasing attentions in natural language processing. Image information can be utilized in various domains such as social

media (Zhang et al., 2018b; Moon et al., 2018; Lu et al., 2018; Zheng et al., 2021b), movie reviews (Gan et al., 2021) and news (Wang et al., 2022d). Most of the previous approaches utilize images by extracting information such as feature representations (Moon et al., 2018; Yu et al., 2020), object tags (Wu et al., 2020; Zheng et al., 2021a) and image captions (Wang et al., 2022b) to improve the model performance. However, most of image feature, object tag and caption extractors are trained based on datasets such as ImageNet (Deng et al., 2009) and Visual Genome (Krishna et al., 2016), which mainly contain common nouns<sup>2</sup> instead of named entities. As a result, the extractors (especially those of object tags and image captions) often output information about common nouns, which may not be helpful for entity-based task models. Recently, pretrained vision-language models (Tan and Bansal, 2019b; Chen et al., 2020; Li et al., 2020) have improved the performance of cross-modal tasks such as VQA (Agrawal et al., 2015), NLVR (Young et al., 2014) and image-text retrieval (Suhr et al., 2019) significantly. However, suffering from the same problem, the pretrained vision-language models do not achieve better performance than textual models in multi-modal NER (Sun et al., 2021; Wang et al., 2022b).

Recently, approaches based on text retrieval have shown their effectiveness over question answering (Liu et al., 2020; Xu et al., 2022; Wang et al., 2022a), machine translation (Gu et al., 2018; Zhang et al., 2018a; Xu et al., 2020), language modeling (Guu et al., 2020; Borgeaud et al., 2021), NER (Wang et al., 2021, 2022c; Zhang et al., 2022b) and entity linking (Zhang et al., 2022a; Huang et al., 2022). The approaches use the input texts

<sup>\*</sup> Yong Jiang and Kewei Tu are the corresponding authors.  
<sup>‡</sup>: equal contributions. This work was done when Xinyu Wang was visiting StatNLP Research Group at SUTD.

<sup>1</sup>Our code is publicly available at <http://github.com/modelscope/adaseq/examples/MoRe>.

<sup>2</sup>[https://visualgenome.org/data\\_analysis/statistics](https://visualgenome.org/data_analysis/statistics)

as the search query to retrieve the related knowledge in the knowledge corpus (KC), which is a key-value structured memory built from the knowledge source. Besides, humans can recognize the entities (such as famous persons and locations) in the image based on their learned knowledge in practice. When they are not sure about the entities in the image, they can even use image-based retrieval in the search engine to get the related knowledge about the image. Inspired by that, for multi-modal NER and RE models, we believe retrieving the related knowledge of the image can be utilized to help the task models to disambiguate the named entities as well. In this paper, we propose Multi-modal Retrieval based framework (MoRe), which explores the knowledge behind the input image and text pairs for multi-modal NER and RE. MoRe retrieves related knowledge for the input text and image using the textual retriever and image retriever respectively. The text retriever retrieves the most related paragraphs in the KC and the image retriever finds the documents containing the most related images. The retrieval results of each modality are sent to the textual and visual models respectively and used for training on NER and RE tasks. After both of the models are trained, the Mixture of Experts (MoE) module is trained to learn how to combine the model predictions from the two models.

The contributions of MoRe can be summarized in four aspects:

1. We propose a simple and effective way to inject knowledge-aware information into multi-modal NER and RE tasks using multi-modal retrieval, which is rarely introduced on multi-modal NER and RE tasks in previous work.
2. We empirically show that the knowledge from our text retrieval and image-based retrieval modules can significantly improve the performance of multi-modal NER and RE tasks.
3. We further propose MoE for multi-modal NER and RE, which can combine the knowledge from the image and text retrieval modules well. We show MoE can further improve the performance and achieve state-of-the-art accuracy.
4. We conduct detailed analyses that compare the advantage of the text retrieval module and image-based retrieval module. We show the MoE module can correctly take the advantage of the knowledge from each modal.

System	Query	Key	Value	Alg.
Text	$x$	Sentence	Appeared Paragraph	BM25
Image	$I$	Img. Feat.	Introduction Section	$k$ -NN

Table 1: A comparison of text retrieval and image-based retrieval system. Img. Feat.: Image feature, Alg.: Search algorithm,  $k$ -NN:  $k$ -nearest-neighbors.

## 2 MoRe

Given an input text and image pair  $(x, I)$ , where  $x = \{x_1, \dots, x_n\}$ , MoRe aims to predict the outputs  $y$ .  $y$  can be the label sequence or the relations for the NER and RE task respectively. MoRe feeds  $(x, I)$  into the multi-modal retrieval module, which returns retrieved texts  $z_T$  and  $z_I$  from a text retrieval module and a image-based retrieval module respectively. The textual task model and visual task model concatenate the input sentence  $x$  with the retrieved knowledge  $z_T$  and  $z_I$  from KC and predict the output distribution  $P(y|x, I, z_T)$  and  $P(y|x, I, z_I)$  respectively. Finally, the MoE module fuses the predictions from the models based on each modality and makes the final prediction  $P(y|x, I)$ . The architecture of our framework is shown in Figure 1.

### 2.1 Multi-modal Retrieval Module

Text retrieval has been shown to be effective for NER (Wang et al., 2021), as the retrieval result can provide information for disambiguation of complex entities. In order to retrieve related knowledge of the text and image, we design a multi-modal retrieval module. We choose Wikipedia, the largest online encyclopedia as our knowledge source because it has a rich collection of articles describing entities and the articles should provide used clues for entity-related tasks. Considering the difficulty of pairing text and image in Wikipedia, we constructed two separate retrieval systems in the module for text and image respectively.

A retrieval system has two components: KC and knowledge retriever. The KC is denoted by  $\{(k, v)\}$ . The knowledge retriever calculates the relevance between an input query  $q$  and the keys in the KC. The retriever then returns the values corresponding to top- $k$  keys that are most relevant to the query. We denote the retrieved result as  $\{t_1, \dots, t_k\}$ . A summarization of the two retrieval modules is shown in Table 1.

**Textual Retrieval System** We retrieve the related knowledge of input text  $x$  with the textual

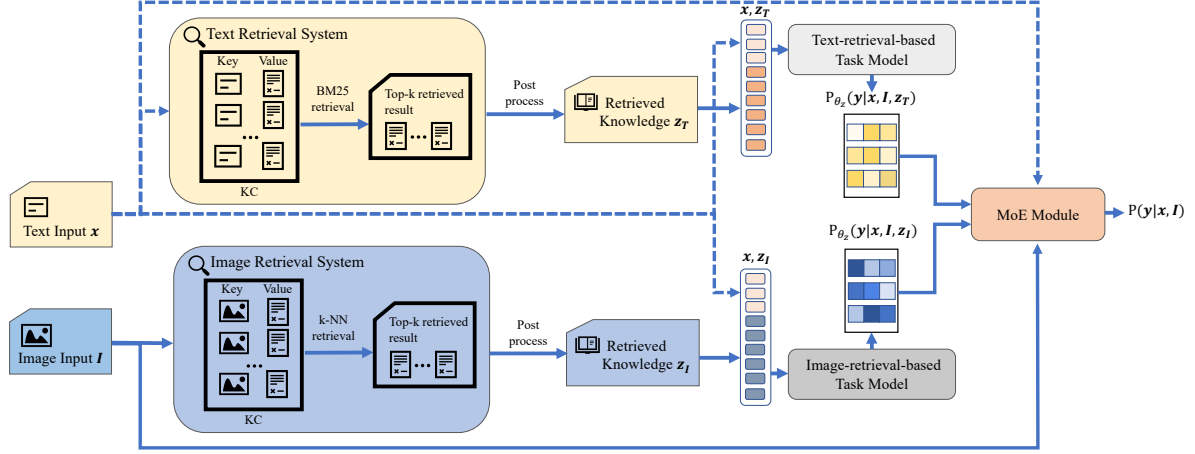


Figure 1: The architecture of MoRe.

retrieval system. We build the KC from the articles in Wikipedia. In the KC, each key is the sentence in Wikipedia and the corresponding value is the paragraph where the sentence appears. Considering the retrieval efficiency over a KC with 200 million entries, we choose to use a term-based text retriever. The retriever use BM25 (Robertson et al., 1995) to calculate the relevance score between key  $k$  and query  $x$ .

**Image-base Retrieval System** According to the style manual of Wikipedia<sup>3</sup>, the introduction section of an article is the summary of the most important contents and an image in an article is an important illustrative aid to understanding. Given an input image  $I$ , we search the related images in Wikipedia to collect the knowledge of related entities. Each key in KC is an image from a Wikipedia article and the corresponding value is the concatenation of the article title and introduction section of this article. To find the related images, we use an image encoder to encode the images into feature vectors. For each query  $I$ , it retrieves its  $k$ -nearest-neighbors with the inner-product metric.

**Context Processing** The top- $k$  results from the KC are concatenated into  $z = \{[X], t_1, \dots, t_k\}$ , where  $[X]$  is a special mark indicating the following sequence to be the retrieved texts. Given a certain transformer-based embedding model, we chunk the retrieved knowledge  $z$  so that the total subtoken number of  $x$  and  $z$  does not exceed the embedding’s subtoken number limits. Since the retrieved texts are usually very long, it is hard to combine the retrieval results from two modalities

<sup>3</sup>[https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style)

together as a single  $z$ . As a result, we feed the retrieved knowledge  $z_T$  and  $z_I$  to separate models to let the model fully utilize each kind of information.

## 2.2 Task Model

Given the retrieved knowledge  $z$  (which is either  $z_T$  or  $z_I$ ), the task model predicts the probability distribution  $P(y|x, I, z)$  of the task.

**Named Entity Recognition** We take the NER task as a sequence labeling problem, which predicts a label sequence  $\mathbf{y} = \{y_1, \dots, y_n\}$  at each position. The NER model feeds the concatenated text  $[x; z]$  into the transformer-based encoder and gets the token representations  $\{r_1, \dots, r_n\}$  corresponding to  $x$ :

$$\{r_1, \dots, r_n, \dots, r_{n+m}\} = \text{embed}([x; z])$$

where  $m$  is the length of  $z$ . With the attention module in the transformer-based encoder, the token representations  $\{r_1, \dots, r_n\}$  contain the retrieved knowledge  $z$ . MoRe then feeds the representations into a linear-chain CRF layer to predict the probability distribution  $P(y|x, I, z)$  of the label sequence:

$$\psi(y', y, r_i) = \exp(\mathbf{W}_y^T r_i + \mathbf{b}_{y', y})$$

$$P(\mathbf{y}|x, I, z) = \frac{\prod_{i=1}^n \psi(y_{i-1}, y_i, r_i)}{\sum_{\mathbf{y}' \in \mathcal{Y}(x)} \prod_{i=1}^n \psi(y'_{i-1}, y'_i, r_i)}$$

where  $\psi$  is the potential function. In the potential function,  $\mathbf{W} \in \mathbb{R}^{d \times t}$  and  $\mathbf{b} \in \mathbb{R}^{t \times t}$  are parameters for calculating emission and transition scores respectively.  $d$  is the hidden size of  $r$  and  $t$  is the

label set size.  $\mathcal{Y}(\mathbf{x})$  denotes the set of all possible label sequences given  $\mathbf{x}$ .

**Relation Extraction** In RE, the model aims to predict a relation label  $P(y|\mathbf{x}, \mathbf{I}, \mathbf{z})$  given the subject entity  $\{x_{start_s}, \dots, x_{end_s}\}$  and object entity  $\{x_{start_o}, \dots, x_{end_o}\}$ . We follow PURE (Zhong and Chen, 2021), which adds special markers in the input  $\mathbf{x}$  to indicate the named entities:

$$\mathbf{x}' = \{x_1, \dots, \langle S \rangle, x_{start_s}, \dots, x_{end_s}, \langle /S \rangle, \dots, \langle O \rangle, x_{start_o}, \dots, x_{end_o}, \langle /O \rangle, x_n\}$$

In the equation,  $\langle S \rangle$  and  $\langle /S \rangle$  indicate the start and end of the subject entity while  $\langle O \rangle$  and  $\langle /O \rangle$  indicate the start and end of the object entity. Similar to the NER model, the RE model feeds the concatenated text  $[\mathbf{x}'; \mathbf{z}]$  into the transformer based encoder and gets the token representations of  $\langle S \rangle$  and  $\langle O \rangle$ , denoted by  $\mathbf{r}_s$  and  $\mathbf{r}_o$  respectively:

$$\{\mathbf{r}_1, \dots, \mathbf{r}_s, \mathbf{r}_{start_s}, \dots, \mathbf{r}_o, \mathbf{r}_{start_o}, \dots, \mathbf{r}_{n+m+4}\} = \text{embed}([\mathbf{x}'; \mathbf{z}])$$

The probability distribution  $P(\mathbf{y}|\mathbf{x}, \mathbf{I}, \mathbf{z})$  for relation extraction is then given by:

$$\begin{aligned} \psi'(y, \mathbf{r}_s, \mathbf{r}_o) &= \exp(\mathbf{M}_y^T [\mathbf{r}_s; \mathbf{r}_o] + \mathbf{b}'_y) \\ P(y|\mathbf{x}, \mathbf{I}, \mathbf{z}) &= \frac{\psi'(y, \mathbf{r}_s, \mathbf{r}_o)}{\sum_{y' \in \mathcal{Y}'} \psi'(y', \mathbf{r}_s, \mathbf{r}_o)} \end{aligned}$$

where  $\mathbf{M} \in \mathbb{R}^{2d \times t'}$  and  $\mathbf{b}' \in \mathbb{R}^{t' \times 1}$  are the parameters for RE.  $t'$  is the label set size.  $\mathcal{Y}'$  denotes the relation label set.

### 2.3 Mixture of Experts

As we mentioned in Section 2.1, we use a separate retrieval module for each modality. The retrieval scores of retrieved texts from each modal are therefore not comparable. As a result, it is hard to determine a priori which retrieved knowledge is more helpful to the model performance. We use MoE to alleviate this problem. The MoE module aims to fuse the probability distributions from the textual model and visual model to get better model performance. To obtain the overall probability distribution of generating  $\mathbf{y}$ , we treat  $e$  as a latent variable and calculate the marginal distribution over  $e$ , which is:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{I}) = \sum_{e \in \{T, I\}} P_{\theta_e}(\mathbf{y}|\mathbf{x}, \mathbf{I}, \mathbf{z}_e) P_{\theta_c}(e|\mathbf{x}, \mathbf{I})$$

where  $\theta_e$  is the task model parameters trained with the retrieved knowledge  $\mathbf{z}_e$  and  $\theta_c$  is the model parameters of MoE. To calculate  $P_{\theta_c}(e|\mathbf{x}, \mathbf{I})$ , we use a text encoder and an image encoder to extract the representation of  $\mathbf{x}$  and  $\mathbf{I}$  respectively:

$$\begin{aligned} \mathbf{r}_T &= \text{TextEncoder}(\mathbf{x}); \mathbf{r}_I = \text{ImageEncoder}(\mathbf{I}) \\ P_{\theta_c}(T|\mathbf{x}, \mathbf{I}) &= \sigma(\mathbf{U}_y^T [\mathbf{r}_T; \mathbf{r}_I] + \mathbf{b}^*) \\ P_{\theta_c}(I|\mathbf{x}, \mathbf{I}) &= 1 - P_{\theta_c}(T|\mathbf{x}, \mathbf{I}) \end{aligned}$$

where  $\mathbf{U} \in \mathbb{R}^{(d_T+d_I) \times t}$  and  $\mathbf{b}^* \in \mathbb{R}^{t \times 1}$ .  $d_T$  and  $d_I$  are the dimension of  $\mathbf{r}_T$  and  $\mathbf{r}_I$  respectively. The final prediction from MoE module is given by:

$$\begin{aligned} \hat{\mathbf{y}} &= \arg \max_{\mathbf{y} \in \hat{\mathcal{Y}}} P(\mathbf{y}|\mathbf{x}, \mathbf{I}) \\ &= \arg \max_{\mathbf{y} \in \hat{\mathcal{Y}}} \sum_{e \in \{T, I\}} P_{\theta_e}(\mathbf{y}|\mathbf{x}, \mathbf{I}, \mathbf{z}_e) P_{\theta_c}(e|\mathbf{x}, \mathbf{I}) \end{aligned}$$

where  $\hat{\mathcal{Y}}$  can be  $\mathcal{Y}(\mathbf{x})$  for NER or  $\mathcal{Y}'$  for RE. In RE,  $\hat{\mathbf{y}}$  can be easily calculated by finding the largest probability among all possible relation label  $y$ . However, for NER with a linear-chain CRF layer,  $\mathbf{y}$  represents the corresponding label sequence of the input  $\mathbf{x}$ . The possible label sequence set  $\mathcal{Y}(\mathbf{x})$  is exponential in size. As a result, it is difficult to calculate the weighted summation of two probability distributions (i.e.  $P_{\theta_T}(\mathbf{y}|\mathbf{x}, \mathbf{I}, \mathbf{z}_T)$  and  $P_{\theta_I}(\mathbf{y}|\mathbf{x}, \mathbf{I}, \mathbf{z}_I)$ ) with an exponential number of possible label sequence. Instead of directly calculating the equation, we approximate this process by assuming the label at each position can be independently determined:

$$\begin{aligned} P_{\theta_e}(\mathbf{y}|\mathbf{x}, \mathbf{I}) &\approx \prod_{i=1}^n P_{\theta_e}(y_i|\mathbf{x}, \mathbf{I}) \\ \hat{\mathbf{y}} &\approx \arg \max_{y_1, \dots, y_n \in \mathcal{Y}^*} \sum_{e \in \{T, I\}} \prod_{i=1}^n P_{\theta_e}(y_i|\mathbf{x}, \mathbf{I}) P_{\theta_c}(e|\mathbf{x}, \mathbf{I}) \\ \hat{y}_i &= \arg \max_{y_i \in \mathcal{Y}^*} \sum_{e \in \{T, I\}} P_{\theta_e}(y_i|\mathbf{x}, \mathbf{I}) P_{\theta_c}(e|\mathbf{x}, \mathbf{I}) \end{aligned}$$

where  $\mathcal{Y}^*$  is the NER label set. We use forward-backward algorithm to calculate the marginalized probability distribution  $P_{\theta_e}(y_i|\mathbf{x}, \mathbf{I})$ :

$$\begin{aligned} \alpha(y_i) &= \sum_{\{y_0, \dots, y_{i-1}\}} \prod_{k=1}^i \psi(y_{k-1}, y_k, \mathbf{r}_k) \\ \beta(y_i) &= \sum_{\{y_{i+1}, \dots, y_n\}} \prod_{k=i+1}^n \psi(y_{k-1}, y_k, \mathbf{r}_k) \\ P_{\theta_e}(y_i|\mathbf{x}, \mathbf{I}) &\propto \alpha(y_i) \times \beta(y_i) \end{aligned}$$



## 2.4 Training

**Named Entity Recognition** We use the negative log-likelihood (NLL) as the training loss for the input sequence with gold labels  $\mathbf{y}^*$ :

$$\mathcal{L}_{\text{NER-NLL}}(\theta_e) = -\log P_{\theta_e}(\mathbf{y}^*|\mathbf{x}, \mathbf{I}, \mathbf{z}_e) \quad (1)$$

**Relation Extraction** Similar to NER, we calculate the NLL loss with the gold label  $y^*$ :

$$\mathcal{L}_{\text{RE-NLL}}(\theta_e) = -\log P_{\theta_e}(y^*|\mathbf{x}, \mathbf{I}, \mathbf{z}_e) \quad (2)$$

**Mixture of Experts** Given the trained task models with parameters  $\theta_T$  and  $\theta_I$ , the MoE model is trained with NLL loss with  $P_{\theta_c}(\mathbf{y}|\mathbf{x}, \mathbf{I})$ . The parameters of trained task model  $\theta_T$  and  $\theta_I$  are fixed during the training of MoE.

## 3 Experiments

### 3.1 Settings

**Retrieval System Configuration** For the retrieval systems, we build the KCs using the English Wikipedia dumps. We convert the dumps into plain text and download the images appearing in the articles. To take advantage of the rich anchors in Wikipedia, we mark them with a special tag. For example, the anchor of “*Alan Turing*” is tagged and the text “*Alan Turing published an article ...*” is transformed into “*<e: Alan\_Turing> Alan Turing </e> published an article ...*”. There are about 200 million entries in the KC for textual retrieval and 4 million entries in the one for image-based retrieval. We build the term-based textual retriever with the search engine ElasticSearch<sup>4</sup>. We use ViT-B/32 in CLIP to encode images in the feature-based image retriever and use Faiss (Johnson et al., 2019) for efficient search. For both retrieval modules, we use the top-10 retrieval candidates.

**Datasets** For NER, we show the effectiveness of MoRe on Twitter-15, Twitter-17, SNAP and WikiDiverse datasets<sup>5</sup> containing 4,000/1,000/3,257, 3,373/723/723, 4,290/1,432/1,459, 6,312/755/757 sentences in train/development/test split respectively. The Twitter-15 dataset is from Zhang et al. (2018b). The SNAP dataset is from Lu et al. (2018). The Twitter-17 dataset is a filtered version of SNAP constructed by Yu et al. (2020). These 3 datasets are from the social media domain and

<sup>4</sup><https://www.elastic.co/>

<sup>5</sup>The datasets are available at <https://github.com/jefferyYu/UMT>, <https://github.com/Multimodal-NER/RpBERT> and <https://github.com/wangxw5/wikiDiverse>.

are the most commonly used datasets for multi-modal NER. The WikiDiverse dataset is a very recent multi-modal entity linking dataset constructed by Wang et al. (2022d) based on Wikinews. The dataset has annotations of entity spans and entity labels. We convert the multi-modal entity linking dataset into a multi-modal NER dataset to further show the effectiveness of MoRe on the news domain. For RE, we use MNRE dataset<sup>6</sup>. The dataset is constructed by Zheng et al. (2021b) based on user tweets and contains 12,247/1,624/1,614 sentences in train/development/test split. MNRE dataset is also from the social media domain.

**Model Configuration** In all experiments of MoRe, we use XLM-RoBERTa large (XLMR; Conneau et al., 2020) model as the encoder of task models, which has a strong ability for modeling the contexts. For the text and image encoder in MoE, we use the same CLIP model as it in the retrieval system to extract the text and image features.

**Training Configuration** During training, we finetune the models by AdamW (Loshchilov and Hutter, 2018) optimizer. In all experiments, we use the grid search to find the learning rate for the embeddings within  $[1 \times 10^{-6}, 5 \times 10^{-5}]$ . We use a learning rate of  $5 \times 10^{-6}$  and a batch size of 4 for task model training. Following ITA (Wang et al., 2022b), we use the cross-view alignment loss to minimize the KL divergence between the output distributions of retrieval based input and original input. For MoE, we use the same learning rate and a batch size of 64 instead. The task models are trained for 10 epochs and the MoE models are trained for 50 epochs. All of the results are averaged from 3 runs with different random seeds.

### 3.2 Results

We compare MoRe with our baseline and previous state-of-the-art approaches on multi-modal NER and RE. Our baseline is the model without any retrieval module and MoE module. To fully show the effectiveness of our approach, we rerun ITA on WikiDiverse for NER and MNRE for RE. ITA is one of the very recent state-of-the-art approaches to the multi-modal NER, which extracts the image captions, object tags and OCR texts in the image to help NER predictions. For MoRe, we also show the model performance only with the

<sup>6</sup><https://github.com/thecharm/MNRE>

	T-15	T-17	SNAP	Wiki	MNRE
Wu et al. (2020)	72.92	-	-	-	-
Yu et al. (2020)	73.41	85.31	-	-	-
Sun et al. (2020)	73.80	-	86.80	-	-
Sun et al. (2021)	74.90	-	87.80	-	-
Zhang et al. (2021a)	74.85	85.51	-	-	-
Zheng et al. (2021b)	-	-	-	-	65.56
Zheng et al. (2021a)	-	-	-	-	66.41
<b>ITA</b>	78.03	89.75	90.15	76.87	66.89
Ours: <b>Baseline</b>	77.29	88.68	89.35	76.01	65.77
<b>MoRe</b> <sub>Text</sub>	77.91	89.50	90.09	77.97	66.62
<b>MoRe</b> <sub>Image</sub>	78.13	89.82	90.20	77.46	67.24
<b>MoRe</b> <sub>MoE</sub>	<b>79.21</b>	<b>90.67</b>	<b>91.10</b>	<b>79.33</b>	<b>68.60</b>

Table 2: A comparison of our approaches and state-of-the-art approaches on multi-modal NER and RE. **T-15**: Twitter-15, **T-17**: Twitter-17, **Wiki**: WikiDiverse. The results of ITA on WikiDiverse and MNRE datasets are reproduced by us.

text retrieval module<sup>7</sup> and the performance only with the image-based retrieval module to show the strength of the retrieval module. The results in Table 2 show that MoRe outperforms all of the previous state-of-the-art approaches<sup>8</sup>. Only with the text retrieval module or the image-based retrieval module, our model performance is competitive and even outperforms ITA. On WikiDiverse dataset, our models have more improvements compared with ITA. The possible reason is that our approach can retrieve more helpful information from KC in the news domain while the caption and object extractors do not perform well in this domain. This shows the performance of ITA may be limited for certain task domains. Comparing our models with text retrieval, image-based retrieval and our baseline, our retrieval approaches are significantly stronger than our baseline (with Student’s t-test with  $p < 0.05$ ). In most of the cases, models with the image-based retrieval module perform better than models with a text retrieval module except on WikiDiverse dataset. The possible reason is the knowledge from the text retrieval is more critical in the news domain.

## 4 Analysis

### 4.1 Comparison with Other Variants of MoE

To further show the advantage of our MoE module over text and image models, we compare several variants in Table 3. In this analysis, we mix two tex-

<sup>7</sup>The model is similar to the model of Wang et al. (2021). Our textual retrieval is based on Wikipedia, while the textual retrieval in Wang et al. (2021) is based on Google. Our local retrieval module is much faster and more practical.

<sup>8</sup>Note that all of the previous approaches do not use any retrieval techniques for the tasks.

	T-15	T-17	SNAP	Wiki	MNRE
<b>Avg. Pool.</b> <sub>Text</sub>	78.37	89.70	90.72	78.25	66.91
<b>Avg. Pool.</b> <sub>Image</sub>	78.81	89.47	90.55	78.16	68.07
<b>Avg. Pool.</b> <sub>Text+Image</sub>	79.00	89.86	91.02	78.82	68.29
<b>MoE</b> <sub>Text</sub>	78.62	89.86	90.80	78.45	68.22
<b>MoE</b> <sub>Image</sub>	78.88	90.32	90.78	78.24	68.44
<b>MoE</b> <sub>Text+Image</sub>	<b>79.21</b>	<b>90.67</b>	<b>91.10</b>	<b>79.33</b>	<b>68.60</b>

Table 3: A comparison of MoE in MoRe and other variants of MoE. **Text/Image**: a mixture of two text retrieval/image-based retrieval based models with two random seeds, **Text+Image**: a mixture of a text retrieval based model and a image-based retrieval based model.

tual models and two image models with different random seeds for comparison. Firstly, we compare our MoE approach with average pooling, which averages the probability distributions over two models. The results show that our MoE approach outperforms all the average pooling approaches significantly (with student’s T test with  $p < 0.05$ ) except on the SNAP dataset, which shows the effectiveness of MoE. Comparing among the three average pooling approaches, we can find that averaging the probability distributions over the text and image models performs better than averaging the probability distributions of two models from the same modality. Our MoE is also significantly stronger than the single modality MoE approaches on all the datasets (with  $p < 0.05$ ). Moreover, we find the relative improvements of MoE depend on each specific dataset. For example, the improvements are relatively smaller in T-15 and SNAP compared with the MoE and average pooling while the improvements on T-17 and Wiki datasets are relatively larger. A possible reason is that the importance of text retrieval and image retrieval is almost equal in most of the samples in the T-15 and SNAP datasets. Similarly, the advantage of multi-modal MoE over single-modal MoE depends on the dataset as well. When the retrieved knowledge from images and that from texts are more complementary, the relative improvements will be much higher (e.g. Wiki).

### 4.2 How Text Retrieval and Image-based Retrieval Affect Model Prediction

To further show the advantage of text retrieval and image-based retrieval module in MoRe, we compare the label-wise F1 in Table 4. For the diversity of domains and tasks, we choose SNAP, WikiDiverse and MNRE as the representative datasets to show the label-wise F1 score. In the WikiDiverse dataset, there are 13 entity types. We select loca-

	SNAP				WikiDiverse				MNRE			
	LOC	ORG	OTHER	PER	LOC	ORG	OTHER	PER	LOC	ORG	OTHER	PER
<b>Baseline</b>	86.45	89.71	76.82	93.70	78.16	76.63	62.40	89.42	71.83	62.70	61.92	62.94
<b>MoRe<sub>Text</sub></b>	<b>87.90</b>	89.85	<b>79.65</b>	93.79	<b>79.97</b>	76.64	<b>66.32</b>	90.68	71.85	64.67	<b>66.37</b>	63.85
<b>MoRe<sub>Image</sub></b>	87.72	<b>89.88</b>	79.00	<b>94.34</b>	78.30	<b>78.53</b>	64.88	<b>91.41</b>	<b>74.74</b>	<b>65.86</b>	65.38	<b>64.24</b>

Table 4: Label-wise F1 score on SNAP, WikiDiverse and MNRE datasets.

	T-15	T-17	SNAP	Wiki	MNRE
<b>Baseline</b>	77.29	88.68	89.35	76.01	65.77
<b>Random<sub>Text</sub></b>	77.03	88.53	89.21	75.88	65.34
<b>Random<sub>Image</sub></b>	77.27	88.61	89.29	75.96	65.28

Table 5: A comparison of our baseline and the model with random retrieval results.

tion, organization, others and person as the representative labels, which are the most common labels in the dataset and are consistent with the entity label set in SNAP and MNRE datasets. For MNRE, we calculate the entity-label-based F1 score, which calculates the relation F1 score for each entity type. For example, if a relation of two entities is predicted as “/per/org/member\_of” (which means the subject is person type, the object is organization type and the relationship between them is “member\_of”), the relation will be counted into the relation F1 score for both “per” and “org” entities. We use this way to calculate the relation F1 score to analyze how the retrieval system affects each entity label in RE. From the results in Table 4, we can observe that 1) the models with the retrieval module outperform our baselines over all the labels; 2) the image-based retrieval module in MoRe is much helpful for recognizing person and organization entities; 3) the text retrieval module in MoRe is helpful for recognizing other entities; 4) for location entities, the text retrieval module has an advantage in NER while the image-based retrieval module has an advantage in RE. The possible reason is that the image-based retrieval can easily capture the person and organization entities since people and organization usually appear in the image. However, other entities such as the entity name of creative works and festivals are hard to be presented in the images. The related knowledge of such kinds of entities can be easily found through text retrieval.

### 4.3 How the Knowledge Quality Affects Performance

We analyze how the task model will perform when the quality of retrieved knowledge drops. We ran-

	T-15	T-17	SNAP	Wiki
<b>MoRe<sub>Text</sub></b>	77.91	89.50	90.09	77.97
<b>MoRe<sub>Text-Marg.</sub></b>	77.85	89.40	90.00	77.92
<b>MoRe<sub>Image</sub></b>	78.13	89.82	90.20	77.46
<b>MoRe<sub>Image-Marg.</sub></b>	78.07	89.73	90.15	77.39

Table 6: A comparison of the performance of the MoRe NER models and the performance of their marginal distributions (labeled with Marg.).

domly select the retrieved knowledge from the KCs of text retrieval module and image-based retrieval module respectively and train the models based on the random knowledge. The results in Table 5 shows that in both of the conditions, the model performance drops moderately compared with our baseline that is trained without retrieval results. The observation shows that using the random retrieved knowledge can introduce noises to the model. The improvements of the task models come from the related knowledge provided by our designed retrieval module rather than the extended input sequence length.

### 4.4 How Approximation Affects Performance

In Section 2.3, we propose to calculate the marginal probability distribution of the CRF layer for NER to approximately calculate the MoE target function. To show how the approximation may affect the model performance, we compute the prediction of our NER model by calculating  $\arg \max_{y_i \in \mathcal{Y}^*} P_{\theta_e}(y_i | \mathbf{x}, \mathbf{I})$  at each position. The results of our task models with the text retrieval and image-based retrieval modules are shown in Table 6. The results show that the approximation only drops the model performance by no more than 0.1 F1 score. Therefore we can use the approximated probability distribution to calculate the MoE target function, which can be much easier than the original function for the linear-chain CRF layer.

### 4.5 Speed Comparison

In Table 7, we compare the speed of each module in MoRe on a single Tesla V100 GPU with 16GB

Module	Sentences/Second
ITA Feature Extraction	0.7
CLIP Feature Extraction	6.4
Text Retrieval	64.6
Image Retrieval	650.1
Model Prediction	8.2

Table 7: Speed of ITA feature extraction module and each module of MoRe. Note that the CLIP feature extraction includes the text and image feature extraction.

memory with a batch size of 1. To further show the advantage of MoRe, we also calculate the speed of feature extraction parts (i.e. object and caption extractors based on VinVL (Zhang et al., 2021c)) of ITA<sup>9</sup>. We can observe that the bottleneck of MoRe is the CLIP feature extraction part, but the speed is much faster than the feature extraction module in ITA. The observation shows the speed advantage of MoRe over ITA.

#### 4.6 Case Study

In the case study, we show the importance of knowledge from text retrieval and image-based retrieval. In Figure 2 (a), the text input talks about the festival at “Cannes” while the image input shows the beaches at the place. The text retrieval results are mainly talking about the Cannes Film Festival while the image-based retrieval results are mainly talking about the similar beaches in the world. For the entity “Cannes”, the text retrieval results are much more helpful to the disambiguation of the entity since the text retrieval results mention the location “Cannes” multiple times. In Figure 2 (b), it is hard to recognize the named entity “Kolo” is the dog’s name only given the input text. The text retrieval also fails to find the related information to the sentence. However, the image-based retrieval returns knowledge about similar kinds of the dog in the input image, which helps the model to recognize the entity “Kolo” is possibly the dog’s name instead of a person’s name.

## 5 Related Work

**Introducing Visual Information to Improve Multi-modal NLP Tasks** In the natural language processing community, improving NLP tasks by introducing visual information becomes a hotspot of recent studies. In many scenarios, there is a plenty of visual information for a lot of NLP tasks such as

<sup>9</sup>The prediction speed of ITA is the same as that of MoRe since the input sequence lengths of the models are similar.

NER (Zhang et al., 2018b; Moon et al., 2018; Lu et al., 2018), RE (Zheng et al., 2021a,b), keyphrase prediction (Wang et al., 2020) and entity linking (Gan et al., 2021; Zhang et al., 2021b; Wang et al., 2022d). Most of the approaches propose to introduce a special attention mechanism to model the interaction between the representations of objects in the image and the input text (Zhang et al., 2018b; Yu et al., 2020; Sun et al., 2021; Wang et al., 2020; Zheng et al., 2021a). Wang et al. (2020) and Wang et al. (2022b) additionally introduce OCR texts and image captions to the tasks for further improvements. Recently, Wang et al. (2022b) suggests that the representations of images and texts are trained separately and the representations are not aligned. It is hard for the newly introduced attention mechanism to model the interaction. They propose to convert an image into the text to ease the alignment problem between text and image. They convert the image into object tags, image captions, and OCR texts for the model. However, the approach may be limited to the training domain of the image information extractor. In comparison, we explore the related knowledge of an image instead of the surface information of an image. The KC can be much easier to build for the specific domains since building it only requires a large scale of domain-specific unlabeled data.

Pretrained vision-language models such as LXMERT (Tan and Bansal, 2019a), UNITER (Chen et al., 2020), Oscar (Li et al., 2020), E2E-VLP (Xu et al., 2021) and mPLUG (Li et al., 2022) are trained on image-text pairs and achieve significant improvement on tasks like captioning, VQA and image-text retrieval. The pretraining targets at aligning the image and text features into the same space so that the performance of multi-modal tasks can be improved. However, the text representations in pretrained vision-language models are usually not as strong as the pretrained language models. As a result, some of the recent work (Sun et al., 2021; Wang et al., 2022b) find that the pretrained vision-language models do not perform well on multi-modal NLP tasks such as NER.

**Retrieval-based NLP** For knowledge-intensive NLP tasks, retrieval is an effective methods to utilize external knowledge. The knowledge retrieval has been applied to a lot of NLP tasks such as question answering (Liu et al., 2020; Karpukhin et al., 2020; Lewis et al., 2020; Xu et al., 2022; Izacard and Grave, 2021), machine translation (Gu et al.,









(a) Importance of Textual Knowledge	(b) Importance of Visual Knowledge
<p><b>Input</b></p>  <p>Tech Companies Preparing to Take Over the Beaches of Cannes at This Year's Festival.</p>	<p><b>Input</b></p>  <p>RT @KerriFar : RT @mscator : Kolo loves the sun and is so pretty , too . #grilling</p>
<p><b>Text Retrieval Result</b></p> <ol style="list-style-type: none"> <li>The film was based on the Mishima Incident, which was a failed coup d'etat attempt led by Yukio Mishima in 1970. The film competed in the Un Certain Regard section at the 2012 Cannes Film Festival.</li> <li>12 Storeys (十二樓 or Shí'èr lóu in Mandarin) is a 1997 Singaporean drama film written and directed by Eric Khoo. It features an ensemble cast of Jack Neo, Koh Boon Pin and Quan Yi Fong. It was screened in the Un Certain Regard section at the 1997 Cannes Film Festival.</li> </ol>	<p><b>Text Retrieval Result</b></p> <ol style="list-style-type: none"> <li>RT-1556, RT-1794, RT-1824, RT-1851, RT-1851A, and RT-1939, RT-1939A, RT-1990, RT-1990A and the RT-2036. The earliest model covered 30-400MHz. The fifth generation RT-1939 is one of the first military radios to have software-programmable encryption under the National Security Agency's (NSA) Cryptographic Modernization Initiative. Its frequency range is extended and includes the following bands ...</li> <li>.....</li> </ol>
<p><b>Image Retrieval Result</b></p>   <ol style="list-style-type: none"> <li>Vallauris is a commune in the Alpes-Maritimes department in the Provence-Alpes-Côte d'Azur region in southeastern France. It is located in the metropolitan area, and is today effectively an extension of the town of Antibes, bordering it on its west side.</li> <li>Palamós is a town and municipality in the Mediterranean Costa Brava, located in the comarca of Baix Empordà, in the province of Girona, Catalonia, Spain.</li> </ol>	<p><b>Image Retrieval Result</b></p>   <ol style="list-style-type: none"> <li>The Tyrolean Hound is a breed of dog that originated in Tyrol also called the Tyroler Bracke or Tyroler Bracke. They are scent hounds that descended from the celtic hounds in the late 1800s, mainly for their hunting skills. ....</li> <li>The Segugio Maremmano is an Italian breed of scent hound from the coastal plains of the Maremma, in Tuscany. It is mainly used for hunting wild boar, but may also be used to hunt hare and other mammals. They may be either smooth-haired or rough-haired.</li> </ol>
<p><b>Label</b></p> <p>Gold: S-LOC  Baseline: S-MISC  MoRe<sub>Text</sub>: S-LOC  MoRe<sub>Image</sub>: S-MISC</p>	<p><b>Label</b></p> <p>Gold: S-MISC  Baseline: S-PER  MoRe<sub>Text</sub>: S-PER  MoRe<sub>Image</sub>: S-MISC</p>

Figure 2: Two case studies of how the text retrieval and image-based retrieval help model predictions.

2018; Zhang et al., 2018a; Xu et al., 2020), NER (Wang et al., 2021, 2022c; Zhang et al., 2022b) and entity linking (Zhang et al., 2022a; Huang et al., 2022). Compared with these work, our work novelly introduces an image-based retrieval module, which retrieves the knowledge behind the image to improve multi-modal NER and RE tasks. Recently, some of the work introduces the knowledge retrieval to language model pretraining. REALM (Guu et al., 2020) trains the latent knowledge-retriever and knowledge-augmented encoder in an end-to-end manner during the pretraining and fine-tuning. The generative process in REALM is decomposed into retrieving and predicting. The retrieved knowledge is treated as a latent variable and marginalized. Inspired by the generative process, our MoE module treats whether the retrieved knowledge is from text or image as the latent variable. While REALM aggregates the top- $k$  retrieved knowledge from text with the latent variable, we use it to aggregate the knowledge retrieved from different modalities. Since the retriever is trainable, REALM needs to asynchronously re-embedding and re-indexing all documents during the training. In order to scale with a larger database size, RETRO (Borgeaud et al., 2021) freezes the retriever and applies a chunked cross-attention mech-

anism to make use of databases of trillion tokens. For efficiency consideration, we also freeze the retriever module in MoRe.

## 6 Conclusion

In this paper, we introduce a novel Multi-modal Retrieval based framework that utilizes the knowledge behind the multi-modal inputs. MoRe first retrieves related knowledge of input text and image from a text retrieval module and an image-based retrieval module. MoRe then feeds the retrieved knowledge from the text retrieval module and the image-based retrieval module into the textual and visual task models respectively to make predictions. Given the predictions from the task models of each modality, MoRe combines the prediction by a Mixture of Experts (MoE) module. The MoE module takes the features of each input text and image into consideration and makes the final decision. In our experiments, we show that both our textual model and visual model can achieve state-of-the-art performance on four multi-modal NER datasets and one multi-modal RE dataset. With MoE, the model performance can be further improved. In analysis, we demonstrate the advantage of integrating both textual and visual cues for such tasks over different types of labels.

## Limitations

In this paper, MoRe requires a textual and a visual KC for the task. We build the KCs based on Wikipedia. However, in some of the scenarios, the KC needs domain-specific unlabeled data for these scenarios. In these cases, the unlabeled data, especially the data with images, should be collected with effort. Moreover, the input length of MoRe is significantly longer than the original input texts since the new inputs contain the retrieved knowledge. As a result, the inference speed should be significantly slower than the speed with the original input texts. Therefore, MoRe may not satisfy some of the time-critical scenarios. However, we can use the techniques such as knowledge distillation (Hinton et al., 2015) to distill the knowledge from MoRe to smaller models for faster model speed.

## Ethics Statement

In this paper, we use the publicly available datasets for experiments. For the KCs, we build them based on Wikipedia, which is one of the largest online encyclopedia and is publicly available. Therefore, we believe we do not use any personal data that invades users' privacy.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976139).

## References

- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. L. Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. [Improving language models by retrieving from trillions of tokens](#). *CoRR*, abs/2112.04426.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. [Multimodal Entity Linking: A New Dataset and A Baseline](#), page 993–1001. Association for Computing Machinery, New York, NY, USA.
- Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. [Distilling the knowledge in a neural network](#). In *NIPS Deep Learning and Representation Learning Workshop*.
- Shen Huang, Yuchen Zhai, Xinwei Long, Yong Jiang, Xiaobin Wang, Yin Zhang, and Pengjun Xie. 2022. DAMO-NLP at NLPCC-2022 task 2: Knowledge enhanced robust ner for speech entity linking. In *Natural Language Processing and Chinese Computing*, pages 284–293, Cham. Springer Nature Switzerland.
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. [Visual](#)

- genome: Connecting language and vision using crowdsourced dense image annotations.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. 2022. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of EMNLP*.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- 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 (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.
- 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, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.
- Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.
- Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *ACL*.
- Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. In *AAAI*.
- Hao Tan and Mohit Bansal. 2019a. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Hao Hao Tan and M. Bansal. 2019b. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.
- Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022a. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179, Dublin, Ireland. Association for Computational Linguistics.
- Xinyu Wang, min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2022b. ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.
- Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.
- Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022c. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022d. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *ACL*.



- Yue Wang, Jing Li, Michael Lyu, and Irwin King. 2020. [Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3311–3324, Online. Association for Computational Linguistics.
- 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. *ACM MM*.
- Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. [E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, Online. Association for Computational Linguistics.
- Jitao Xu, Josep Crego, and Jean Senellart. 2020. [Boosting neural machine translation with similar translations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.
- Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. Human parity on commonsenseqa: Augmenting self-attention with external attention. In *IJCAI*.
- Peter Young, Alice Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- 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*, pages 3342–3352, Online. Association for Computational Linguistics.
- Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. 2018a. [Guiding neural machine translation with retrieved translation pieces](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.
- Li Zhang, Zhixu Li, and Qiang Yang. 2021b. Attention-based multimodal entity linking with high-quality images. In *Database Systems for Advanced Applications*, pages 533–548, Cham. Springer International Publishing.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021c. [Vinvl: Revisiting visual representations in vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.
- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018b. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022a. [Entqa: Entity linking as question answering](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Xin Zhang, Yong Jiang, Xiaobin Wang, Xuming Hu, Yueheng Sun, Pengjun Xie, and Meishan Zhang. 2022b. [Domain-specific NER via retrieving correlated samples](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2398–2404, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306.
- Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.
- Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.