MCIL: Multimodal Counterfactual Instance Learning for Low-resource Entity-based Multimodal Information Extraction

Baohang Zhou[1], **Ying Zhang**[1]*, **Kehui Song**[1], **Hongru Wang**[2],
Yu Zhao[1], **Xuhui Sui**[1], **Xiaojie Yuan**[1]
[1] College of Computer Science, VCIP, TMCC, TBI Center, Nankai University, China
[2] The Chinese University of Hong Kong, China
{zhoubaohang, songkehui, zhaoyu, suixuhui}@dbis.nankai.edu.cn
{yingzhang, yuanxj}@nankai.edu.cn, hrwang@se.cuhk.edu.hk

Abstract

Multimodal information extraction (MIE) is a challenging task which aims to extract the structural information in free text coupled with the image for constructing the multimodal knowledge graph. The entity-based MIE tasks are based on the entity information to complete the specific tasks. However, the existing methods only investigated the entity-based MIE tasks under supervised learning with adequate labeled data. In the real-world scenario, collecting enough data and annotating the entity-based samples are time-consuming and impractical. Therefore, we propose to investigate the entity-based MIE tasks under the low-resource settings. The conventional models are prone to overfitting on limited labeled data, which can result in poor performance. This is because the models tend to learn the bias existing in the limited samples, which can lead them to model the spurious correlations between multimodal features and task labels. To provide a more comprehensive understanding of the bias inherent in multimodal features of MIE samples, we decompose the features into image, entity, and context factors. Furthermore, we investigate the causal relationships between these factors and model performance, leveraging the structural causal model to delve into the correlations between the input features and output labels. Based on this, we propose the multimodal counterfactual instance learning framework to generate the counterfactual instances by the interventions on the limited observational samples. In the framework, we analyze the causal effect of the counterfactual instances and exploit it as a supervisory signal to maximize the effect for reducing the bias and improving the generalization of the model. Empirically, we evaluate the proposed method on the two public MIE benchmark datasets and the experimental results verify the effectiveness of it.

Keywords: Causal Learning, Multimodal Information Extraction, Low-resource Scenario

1. Introduction

Multimodal information extraction (MIE) is a series of fundamental tasks to extract the structural information from the free text coupled with the image for constructing the multimodal knowledge graph (Zhu et al., 2022). Among of the MIE tasks, the entity-based ones include multimodal relation extraction (Zheng et al., 2021) and named entity typing (Wang et al., 2022). They are defined to identify the relation or fine-grained types of the entities as shown in Figure 1. The existing MIE models are designed with the various multimodal fusion strategies to improve the performance under the supervised learning (Hu et al., 2023). However, the supervised learning methods necessitate sufficient labeled data, which requires labor-intensive annotation. Considering to reduce the labeling-cost, some researchers proposed the semi-supervised learning method to utilize the less labeled data and more unlabeled one to train the MIE models and gain the better performance (Zhou et al., 2022).

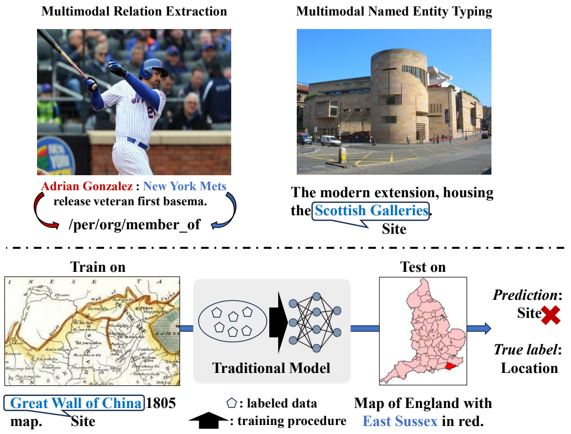

Multimodal Relation Extraction
Adrian Gonzalez : New York Mets
release veteran first basema.
/per/org/member_of
Multimodal Named Entity Typing
The modern extension, housing the Scottish Galleries.
Site
Train on
Great Wall of China 1805 map.
Site
Traditional Model
⬠: labeled data
: training procedure
Test on
Map of England with East Sussex in red.
Prediction: Site ✗
True label: Location

Figure 1: The examples of the entity-based multimodal information extraction tasks and the example of the bias influencing the model performance.

In real-world scenarios, collecting enough data is time-consuming and impractical. Therefore, we propose to investigate the entity-based MIE tasks under the low-resource scenarios. Under low-resource scenarios, MIE tasks face the additional challenge of training models with limited labeled data, making them more difficult com-

[*] Corresponding author.

LREC-COLING 2024, pages 11101–11110
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

pared to tasks based on supervised and semi-supervised learning. The existing studies focus on the low-resource text-based information extraction tasks like: relation extraction (Yu et al., 2022; Xu et al., 2022, 2023) and named entity recognition (Nguyen et al., 2023). The above methods use the data augmentation or distant supervision to expand the limited labeled data for training (Wang et al., 2023).

However, the low-resource MIE tasks are faced with a more challenging situation because of the complexity of multimodal data. The traditional low-resource methods enrich the limited text data by augmentation and train the models with the self-training. But the original distribution of limited labeled samples implies the bias which influences the generalization of the trained models (Wang et al., 2023). The bias of the limited samples is reflected as the spurious correlations that the models learn by the incomplete features. In the case of MIE tasks with low-resources, the models' performance is also negatively impacted by the bias resulting from the limited multimodal samples. This is because text and image data contain rich semantic correlations, and deep neural networks trained on limited samples tend to capture spurious correlations between input features and output labels (Volodin et al., 2020).

To overcome the above disadvantage, we provide a theoretical foundation to analyze the correlation between the features and the model performance from the causal perspective. For a multimodal sample, we decompose its features into three parts: **image**, **entity** and **context**. For example, given the sentence "*Map of England with East Sussex in red.*" and the coupled image is a map of England, the fine-grained type of "*East Sussex*" is *Location*. The model, as demonstrated in Figure 1, tends to predict "*Site*" when trained with bias. Because it primarily focuses on **context** features that contain spurious correlations learned from the limited training data while disregarding **entity** and **image** features. Based on the above analysis, we propose the multimodal counterfactual instance learning framework to tackle the low-resource entity-based MIE tasks. In the framework, we exploit the structural causal model to contruct the counterfactual instances by the interventions on the limited observational samples. To reduce the bias, we analyze the causal effect of the counterfactual instances and exploit it as a supervisory signal to maximize the effect for improving the generalization of the model under limited labeled samples. The experimental results demonstrate that our method can gain the significant improvement of the performances on the different MIE benchmark datasets under the low-resource scenarios. In summary, the contributions of this manuscript can be summarized as follows:

1. To the best of our knowledge, we are the first one to investigate the entity-based multimodal information extraction tasks under the low-resource settings. We provide the theoretical foundation from the causal perspective to analyze the bias that exists in the limited observational samples and explore the spurious correlations between multimodal representations and labels.

2. Based on the above analysis, we propose the multimodal counterfactual instance learning framework to improve the generalization of the low-resource MIE models. In the framework, the causal effect of the counterfactual instances generated by the interventions on the observational samples is exploited as the supervisory signal to train the model.

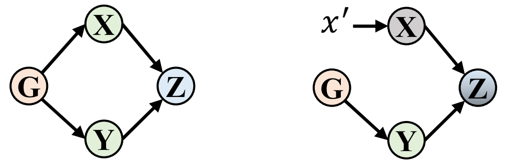
3. We conduct the experiments with the public benchmark datasets on the two entity-based MIE tasks including: multimodal relation extraction and named entity typing. And the corresponding results verify the effectiveness of the proposed framework on the above tasks.

2. Related Work

2.1. Multimodal Information Extraction

Multimodal information extraction (MIE) is a series of tasks to extract the structural information from the free text coupled with the image. The MIE tasks mainly include multimodal named entity recognition (MNER) (Moon et al., 2018), multimodal relation extraction (MRE) (Zheng et al., 2021), multimodal named entity typing (MNET) and multimodal entity linking (MEL) (Wang et al., 2022). Among of the above tasks, MNER is defined as the pre-processing task to extract the named entities based on the text and image. And the other tasks are the entity-based MIE tasks because they depend on the results from the MNRE task. To control the fusion of the text and image representations in MNER dynamically, Zhang et al. (2018) proposed an adaptive co-attention network and Yu et al. (2020) designed the uniform multimodal transformer. MRE task targets to identify the relation between the entities and (Zheng et al., 2021) proposed to fuse the multimodal representations with the graph alignment module. Considering to map the entity to standard knowledge base, Wang et al. (2022) proposed a public dataset WikiDiverse with fine-grained entity types from Wikinews on different topics and utilized multimodal pre-trained models to learn representations of texts and images consistently.

The above methods are based on the supervised learning which requires the adequate labeled data.



(a) An example of the structural causal model (factual world).

(b) An example of the counterfactual world corresponding to Figure 2a.

Figure 2: The structural causal model (SCM) that represents the mechanism of the discriminative models. (a) Complete SCM without interventions. G: confounder variable, X and Y: input features (i.e. text and image features), Z: task label. (b) Counterfactual world with the intervention on the variable X by the reference value x'.

To reduce the labeling costs, Zhou et al. (2022) proposed the semi-supervised MNER model to make use of the limited labeled and unlabeled samples. Compared with the previous works, we are the first one to investigate the MIE tasks under the low-resource scenarios which reduce the procedures to collect data and annotate them heavily.

2.2. Low-resource Information Extraction

Traditional information extraction (IE) tasks including: named entity recognition (Chieu and Ng, 2002) and relation extraction (Kumar, 2017) are focused on the text modality. However, the procedures of collecting data and annotating them are time-consuming and labor-intensive. Therefore, the low-resource IE models are proposed to make full use of the limited labeled samples. Zeng et al. (2020) proposed the weakly-supervised model for named entity recognition which generates the counterfactual examples to expand the original limited data. For the low-resource relation extraction, Xu et al. (2022) summarized the different strategies for training the model with the limited samples including: data augmentation, and balancing methods. In summary, the low-resource information extraction tasks focus on alleviating the long-tailed distribution and the number of original limited samples. But the existing methods are not efficient to the low-resource MIE because the complexity of the multimodal data implies the bias for training the specific models. Therefore, we propose to investigate the low-resource MIE tasks that has not been explored before.

3. Preliminary

3.1. Structural Causal Model

The mechanism of the discriminative models involves the different factors. To investigate the causal relationship between the data and models in a theoretical way, we exploit the structural causal model (SCM) (Pearl, 2009) to analyze the factors. SCM can be expressed as a directed acyclic graph where the nodes represent the random variables and the edges represent the direct causal correlations between two variables. As shown in Figure 2, the variables X and Y denote the input features (i.e. text and image features), Z is the task label and $X \to Z$ represents the causation from variable X to variable Z. Empirically, G is the confounder variable that influences the generation of both variables X and Y which implies the semantic correlation between them.

3.2. Counterfactual Reasoning and Causal Effect

To deduce outcomes under hypothetical circumstances that diverge from the factual world, Pearl (2009) proposed the counterfactual reasoning as a statistical inference method. For example, Figure 2a is a factual world where the calculation of Z is formulated as $Z_{x,y} = Z(X = x, Y = y)$. To estimate the causal effect of a treatment variable X on a response one Z, we conduct the counterfactual reasoning on the factual world. As shown in Figure 2b, we construct a counterfactual world by the intervention on the variable X. Formally, the intervention operation is denoted as $do(\cdot)$ in causal inference literature. And $do(X = x')$ in the counterfactual world means that we cut-off the link $G \to X$ to force the variable X to not be caused by its causal parent G by fixing the variable X to the specific value x'. The causal effect (ce) (Rao et al., 2021) of the variable X on the prediction task label Z can be calculated by the difference between the observational prediction $Z(X = x, Y = y)$ and its counterfactual alternative $Z(do(X = x'), Y = y)$:

$$\begin{aligned} Z_{ce}^{X} &= Z_{x,y} - Z_{x',y} \\ &= Z(X = x, Y = y) - Z(do(X = x'), Y = y) \end{aligned} \tag{1}$$

where we denote the causal effect of variable X on the prediction variable Z as Z_{ce}^{X}. Similarly, we can also utilize the above way to investigate the causal effect Z_{ce}^{Y} of the variable Y on the variable Z by the intervention $do(Y = y')$.

4. Model

Before getting into the details of the proposed framework, we formalize the problem of the entity-based multimodal relation extraction tasks. The training and test sets are denoted as D_{tr} and D_{te} respectively. The training dataset with $|D_{tr}|$ samples is formulated as $\{(S_i, V_i, e_i, y_i) | i = 1, 2, \ldots, |D_{tr}|\}$ where S_i and V_i are the sentence and image of i-th sample. Further more, e_i is defined as the entity information including: the entity location and text in the sentence S_i, and $y_i \in Y$ is the task label such as: the relation between entities or fine-grained entity type. Under the requirement of low-resource scenario, the number of samples in training set should be much smaller than that in test set $|D_{tr}| << |D_{te}|$.

The **m**ultimodal **c**ounterfactual **i**nstance **l**earning (**MCIL**[1]) framework for the entity-based multimodal information extraction tasks is illustrated in Figure 3. In addition to language and vision pretrained models for multimodal input representations, MCIL is comprised of three main components: the multimodal representation fusion module, counterfactual instance generation module, and causal effect learning module. These components work together to model the semantic correlation between multimodal data and task labels while minimizing the bias from limited data.

4.1. Counterfactual Instance Generation

As shown in Figure 3, we construct the structural causal model (SCM) for the entity-based multimodal information extraction (MIE) tasks including: multimodal relation extraction and named entity typing. We decompose the multimodal data into three parts and denote the image as variable V, the entity as variable E and the context as variable C which determine the label of the sample in SCM. Considering that the limited multimodal data imply the bias that influences the performance of the MIE models, we need to investigate the causal effect of the three treatment variables on the response one. Based on the counterfactual reasoning, we construct the counterfactual worlds by the interventions on the variables of multimodal data respectively.

Given the observational sample $f_i = (S_i, V_i, e_i, y_i)$ of MIE tasks, we need to generate the counterfactual instances of it by the interventions. As shown in Figure 3, we randomly select the reference sample $f_j = (S_j, V_j, e_j, y_j)$ where $i \neq j$ from the limited training set. We do the intervention $do(V = v')$ on the variable V to cut-off the link $G \to V$. The confounder variable G determines the generation of the multimodal data and semantic correlation between the three parts. Empirically, the intervention $do(V = v')$ can be operated by replacing the image V_i of the sample f_i with the image V_j of the reference sample f_j. Therefore, we generate the vision-based counterfactual instance $\hat{f}_i^V$ of the sample f_i by the intervention $do(V = v')$. Besides, the entity information is also important to the entity-based MIE tasks. We generate the entity-based counterfactual instance $\hat{f}_i^E$ by the intervention $do(E = e')$ which replaces the entity information e_i with that e_j. Similarly, the context of the sentence implies the rich semantic information that we should take account. We denote the context as $o_i = S_i/e_i$ and the operation $/$ represents the part of the sentence excludes the entity. And the context-based counterfactual instance $\hat{f}_i^C$ is generated by the intervention $do(C = c')$ which replaces the context o_i with that o_j of the reference sample f_j. Through the above generation process, we acquire the counterfactual instance set $\hat{f}_i = \{\hat{f}_i^V, \hat{f}_i^E, \hat{f}_i^C\}$ of the sample f_i.

4.2. Multimodal Task Prediction

Given the multimodal data including the image and sentence, we need to map them into the dense representations for training neural networks. For the visual information, we take advantage of ViT (Dosovitskiy et al., 2021) to extract the fine-grained representations of images. Compared with convolution neural networks based pretrained model like ResNet (He et al., 2016), ViT splits the image into small patches and uses transformer modules to keep the local vision information in the high-level representation. And the regional representations of the image are denoted as $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{|V|}\}$ where $\mathbf{v}_i \in \mathbb{R}^d$ and $|V|$ is the number of feature vectors extracted from ViT.

As for the texual information, We denote the input sentence with $|S|$ words as $S = \{w_1, w_2, \ldots, w_{|S|}\}$. To make use of pre-trained language models with impressive performance, we use BERT (Devlin et al., 2019) as the sentence encoder to map the discrete words into the dense representations. Each sentence need inserting special tokens `[CLS]` and `[SEP]` into the start and end of it before fed into BERT. And considering that the entity information is critical to the entity-based MIE tasks, we need to mark the begin and end of the entities in the sentence. For the entity-based MIE tasks including: multimodal relation extraction and named entity typing, we utilize the different feature extraction procedure respectively.

Multimodal Relation Extraction (MRE). Formally, the extended sentence of the task is denoted as $\hat{S}$. The sentence representation calculation process can be formulated as $\mathbf{T} = \text{BERT}(\hat{S}) \in \mathbb{R}^{d\times(|S|+6)}$. To combine the visual and textual representations, we utilize the multimodal representation fusion module which can be implemented with the architectures of different multimodal methods such as: UMT (Yu et al., 2021) or MKGFormer (Chen et al., 2022). The fusion module consists of the cross-modal transformer and various mechanisms like: attention and gate layer for aligning and combining the multimodal representations. Therefore, we can utilize the acquire the multimodal representation as $\mathbf{U} = \text{M}(\mathbf{V}, \mathbf{T}; \theta_1)$ where $\mathbf{U} \in \mathbb{R}^{u\times(|S|+6)}$ and $\text{M}(\cdot)$ denotes the fusion module. Among the multimodal represen-

[1] https://github.com/ZovanZhou/MCIL

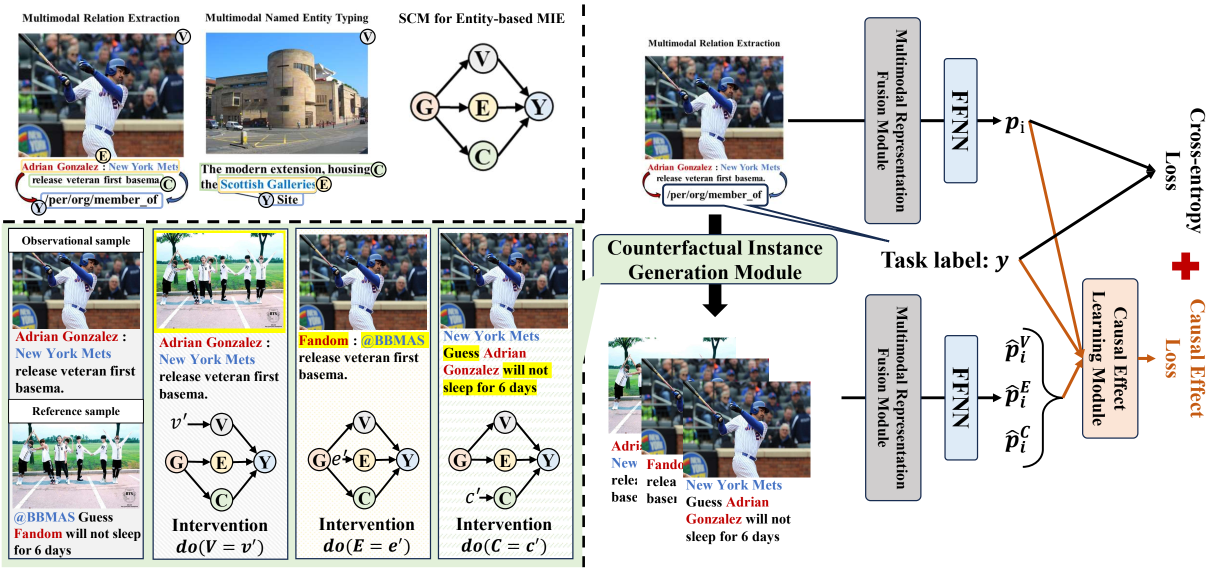

Multimodal Relation Extraction
Multimodal Named Entity Typing
SCM for Entity-based MIE
V
E
C
Y
G
Adrian Gonzalez : New York Mets
release veteran first basema
/per/org/member_of
The modern extension, housing
the Scottish Galleries
Site
Observational sample
Adrian Gonzalez : New York Mets release veteran first basema.
Reference sample
@BBMAS Guess Fandom will not sleep for 6 days
Adrian Gonzalez : New York Mets release veteran first basema.
v'
Intervention $do(V = v')$
Fandom : @BBMAS release veteran first basema.
e'
Intervention $do(E = e')$
New York Mets Guess Adrian Gonzalez will not sleep for 6 days
c'
Intervention $do(C = c')$
Counterfactual Instance Generation Module
Multimodal Representation Fusion Module
FFNN
p_i
Task label: y
Cross-entropy Loss
$\hat{p}_i^V$
$\hat{p}_i^E$
$\hat{p}_i^C$
Causal Effect Learning Module
Causal Effect Loss

Figure 3: The multimodal counterfactual instance learning framework for low-resource entity-based multimodal information extraction tasks. In the counterfactual instance generation module, the replaced features of counterfactual instances are marked with background color.

tations, $\mathbf{u}_{[CLS]}$, $\mathbf{u}_{[E1]}$ and $\mathbf{u}_{[E2]}$ denote the features of the special token [CLS] and start tokens [E1] and [E2] of head and tail entities respectively. We combine them with the concatenation as $\hat{\mathbf{U}} = [\mathbf{u}_{[CLS]}, \mathbf{u}_{[E1]}, \mathbf{u}_{[E2]}] \in \mathbb{R}^{3u}$. Eventually, we feed the combination feature into the feed-forward neural network (FFNN) to get the prediction probability as $\mathbf{p} = \text{FFNN}(\hat{\mathbf{U}}; \theta_2) \in \mathbb{R}^{|Y|}$.

Multimodal Named Entity Typing (MNET). Unlike that there are the head and tail entities in the sentence of MRE, the sentence of MNET only refers one entity. Therefore, the sentence representation of MNET can be formulated as $\mathbf{T} \in \mathbb{R}^{d\times(|S|+4)}$. With the same multimodal representation fusion module, we can acquire the multimodal feature as $\mathbf{U} \in \mathbf{R}^{u\times(|S|+4)}$. And the combination feature of the sample is denoted as $\hat{\mathbf{U}} = [\mathbf{u}_{[CLS]}, \mathbf{u}_{[E1]}] \in \mathbb{R}^{2u}$. Similarly, we use the FFNN to calculate the prediction probability $\mathbf{p}$.

4.3. Causal Effect Learning

Given the multimodal sample f_i and its counterfactual instance set $\hat{f}_i = \{\hat{f}_i^V, \hat{f}_i^E, \hat{f}_i^C\}$, we can calculate the prediction probabilities of them as $\mathbf{p}_i$ and $\hat{\mathbf{p}}_i \in \mathbb{R}^{3\times|Y|}$ through the multimodal task prediction procedure. Considering that the various samples are sensitive to the different the counterfactual instances, we design the gated weight mechanism to control the casual effect learning of the three kinds of interventions. The combination features of the counterfactual instances are denoted as $\{\hat{\mathbf{U}}^k | k \in \{V, E, C\}\}$. The weight scores are calculated as follows:

$$\alpha^k = \frac{\exp(s^k)}{\sum_{j\in\{V,E,C\}} \exp(s^j)} \tag{2}$$

where $s^k = \sigma(\mathbf{W}\hat{\mathbf{U}}^k + \mathbf{b})$, σ is the sigmoid function, and $\mathbf{W}$ and $\mathbf{b}$ are the trainable parameters. The causal effect between the factual sample and counterfactual ones can be calculated as follows:

$$Z_{ce}^V = Z_{v,e,c} - Z_{v',e,c} = \mathbf{p}_i - \hat{\mathbf{p}}_i^V, \tag{3a}$$

$$Z_{ce}^E = Z_{v,e,c} - Z_{v,e',c} = \mathbf{p}_i - \hat{\mathbf{p}}_i^E, \tag{3b}$$

$$Z_{ce}^C = Z_{v,e,c} - Z_{v,e,c'} = \mathbf{p}_i - \hat{\mathbf{p}}_i^C. \tag{3c}$$

Furthermore, we exploit the causal effect as the supervisory signal to maximize the effect for improving the generalization of the model. And the causal effect loss is defined as follows:

$$\mathcal{L}_{ce} = -\sum_{k\in\{V,E,C\}} \alpha^k \cdot y \cdot \log(\text{SoftMax}(Z_{ce}^k)). \tag{4}$$

4.4. Training Procedure

For predicting the task label, we regard the factual samples as the useful supervisory signals to train the model. And the corresponding cross-entropy loss is denoted as follows:

$$\mathcal{L}_{task} = -y \cdot \log(\text{SoftMax}(\mathbf{p})). \tag{5}$$

Considering to train the model while reducing the bias of the limited mulitmodal data under low-resource scenario, we sum the above losses of Equation 4 and Equation 5 in an overall loss as:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{task}. \tag{6}$$

To train the weights of the model, we use the stochastic gradient descent (SGD) methods to update them after feeding multimodal data into the model and calculating the loss by Equation 6.

Item	MRE	MNET
train set	12,247	13,205
valid set	1,624	1,552
test set	1,614	1,570
# labels	23	13
# words	30,124	18,431
# entities	30,970	16,327
# sentences	15,485	7,824
Average length	16.7	10.2

Table 1: The information of the datasets for entity-based multimodal information extraction tasks including: multimodal relation extraction (MRE) and multimodal named entity typing (MNET).

Multimodal Relation Extraction						
N	UMT			MKGFormer		
	NoAug.	CG	MCIL	NoAug.	CG	MCIL
100	$19.9_{2.9}$	$24.0_{5.7}$	$\mathbf{27.0}_{3.1}$	$20.9_{4.7}$	$21.6_{6.8}$	$\mathbf{24.5}_{5.3}$
200	$29.5_{2.4}$	$32.4_{2.7}$	$\mathbf{35.1}_{2.4}$	$30.4_{2.8}$	$30.1_{4.5}$	$\mathbf{35.0}_{1.9}$
300	$34.7_{3.1}$	$36.4_{2.6}$	$\mathbf{39.9}_{2.9}$	$34.5_{2.8}$	$36.4_{2.2}$	$\mathbf{38.7}_{2.9}$
400	$37.6_{2.2}$	$38.1_{4.0}$	$\mathbf{42.4}_{1.2}$	$40.9_{1.5}$	$39.9_{1.6}$	$\mathbf{42.6}_{1.2}$
500	$40.9_{1.7}$	$39.6_{4.4}$	$\mathbf{44.3}_{1.9}$	$42.0_{2.3}$	$41.2_{1.8}$	$\mathbf{42.5}_{1.6}$

Multimodal Named Entity Typing						
N	UMT			MKGFormer		
	NoAug.	CG	MCIL	NoAug.	CG	MCIL
100	$48.5_{1.5}$	$55.6_{1.7}$	$\mathbf{56.1}_{1.2}$	$47.4_{3.0}$	$49.4_{3.3}$	$\mathbf{50.5}_{1.2}$
200	$59.5_{1.5}$	$\mathbf{70.3}_{2.2}$	$69.3_{2.7}$	$58.5_{1.9}$	$63.2_{3.3}$	$\mathbf{64.4}_{3.0}$
300	$66.8_{1.7}$	$73.5_{2.1}$	$\mathbf{76.1}_{1.8}$	$66.4_{1.7}$	$\mathbf{73.0}_{1.5}$	$72.8_{1.9}$
400	$69.5_{3.6}$	$76.5_{1.6}$	$\mathbf{78.9}_{1.0}$	$69.6_{2.4}$	$76.0_{1.1}$	$\mathbf{76.9}_{0.8}$
500	$74.6_{1.8}$	$76.6_{1.1}$	$\mathbf{79.4}_{0.7}$	$74.6_{0.9}$	$70.3_{16.6}$	$\mathbf{79.0}_{1.5}$

Table 2: Performance comparison of different low-resource entity-based multimodal information extraction methods on the benchmark datasets. The performance of the models is presented as mean_{std}. "NoAug." means that the multimodal models are only trained with the limited samples and not with the data augmentation methods.

5. Experiments

5.1. Datasets and Experiment Settings

We investigate the entity-based multimodal information extraction tasks including: multimodal relation extraction (MRE) and multimodal named entity typing (MNET) under the low-resource scenario. For the above two tasks, we conduct the experiments on the corresponding benchmark datasets. (Zheng et al., 2021) proposed the MRE dataset which is built on the posts from Twitter and the samples were randomly selected by annotators with various topics. And there are 15,484 samples with 23 relation categories in the MRE dataset. Besides, we utilize the WikiDiverse (Wang et al., 2022) dataset as the benchmark one for the MNET. Although the WikiDiverse is proposed for multimodal entity linking, each sample in the dataset is collected as a text-image pair from Wikinews, and the entity mention is manually annotated with 13 fine-grained types in the sentence. The detailed statistical information of the benchmark datasets are shown in Table 1. To compare our model with the baselines under the low-resource settings, we mimic this scenario by randomly selecting N samples from the original training set with the balanced label distribution. And we keep the validation and test sets unchanged for evaluating the models in all experiments. Eventually, we repeat the experiments 5 times with different seeds and report the evaluation metrics including: the mean and standard deviation of the models on the samples of test set.

For the proposed framework and baseline models, we use the base version of pre-trained language model BERT (Devlin et al., 2019) and that of vision model ViT (Dosovitskiy et al., 2021) to extract multimodal input representations. The size of hidden layers is set to 768, and the learning rate and batch size are set to 1e-5 and 4 respectively. During the training procedure, we firstly train the model with the training set for specific epochs at most and test it on the validation set to select the best model. And we set the training epochs to 30 and 15 for the MRE and MNET respectively because of the category numbers. All experiments are accelerated by NVIDIA GTX A6000 devices.

5.2. Compared Methods

Considering that there are no existing studies on the low-resource entity-based multimodal information extraction (MIE), we compare the proposed framework with the previous text-based low-resource information extraction method. Zeng et al. (2020) proposed the counterfactual generator (CG) for the weakly-supervised named entity recognition. The CG method is a kind of data augmentation way to construct the counterfactual samples by replacing the entity in one sentence with that in another sentence which is the same type. And it is an effective and model-independent baseline for the low-resource MIE. Besides, the low-resource MIE is involved with the multimodal representation fusion module which is based on multimodal base architectures. To investigate the influence of different multimodal architectures, we utilize the architectures of UMT Yu et al. (2020) and MKGFormer Chen et al. (2022) as the multimodal representation fusion modules respectively.

5.3. Experimental Results

We compare MCIL with the baseline model on the two entity-based MIE tasks under low-resource scenarios, and report the metrics of micro-averaged F1 for multimodal relation extraction (MRE) and multimodal named entity typing (MNET). The detailed experimental results are shown in Table 2. Our model can achieve the best

Method	MRE		MNET	
	UMT	MKGFormer	UMT	MKGFormer
MCIL	27.0	26.5	56.1	50.5
w/o VCI	24.4	25.5	51.4	49.5
w/o ECI	20.8	23.7	53.0	48.9
w/o CCI	22.4	23.0	55.0	50.2

Table 3: The ablation study for multimodel counterfactual instance learning framework (MCIL) on the benchmark datasets under the low-resource scenario where $N = 100$. "VCI", "ECI" and "CCI" represent the vision-, entity-, context-based counterfactual instance in the MCIL.

results on the two benchmarks, and the micro-averaged F1 scores on MNET and MRE of the proposed framework increase 3.4% and 10.6% over the baseline respectively. Compared with the "NoAug." model, the CG method can gain the improvements on the two benchmarks 6.7% and 11.1% with UMT, and 0.7% and 5.1% with MKGFormer. And the proposed MCIL method increases 18.0% and 13.1% with UMT, and 9.8% and 8.5% with MKGFormer on the two datasets over the "NoAug." model. Therefore, our method is a more effective way to utilize the limited samples on the entity-based MIE tasks under the low-resource scenarios.

Besides, the low-resource methods are sensitive to the different entity-based MIE tasks. The CG method can gain the significant improvements on the MNET but not on the MRE because its strategy only focus on the entity information and ignore the rich semantic information of context and image. And our MCIL framework can always increase the results effectively on the two entity-based MIE tasks because we construct the whole structural causal model to investigate the each part of mulitmodal data for reducing the bias of limited samples and improving the generalization of the model. Moreover, the low-resource methods gain the various improvements with the different multimodal architectures like: UMT and MKGFormer. The methods with UMT can achieve the better results that those with MKGFormer and the improvements of the results on UMT are more significant than those on MKGFormer. In summary, the proposed MCIL framework takes advantage of the multimodal counterfactual instances for learning the causal effect to reduce the bias of the limited samples and achieve the significant improvements on the two entity-based MIE tasks with the benchmark datasets.

5.4. Further Discussion

To investigate the model further, we conduct the detailed analysis for presenting it in different aspects. The effectiveness of different counterfactual instances proposed in MCIL is verified by the ablation study. We analyze the causal effects of different parts including: image, entity and context of multimodal data to present the importance of them for the process of the entity-based MIE models inference. Besides, we conduct the case study to discuss the influence of multimodal counterfactual instances to the entity-based MIE tasks and apply the visualization analysis on the multimodal representation to present the usefulness of the proposed framework.

Multimodal Relation Extraction

ACE	UMT			MKGFormer		
	NoAug.	CG	MCIL	NoAug.	CG	MCIL
$do(V=0)$	-6.4	-3.2	-3.5	-3.7	-9.7	-8.2
$do(E=0)$	-2.6	-7.3	-4.4	-6.2	-3.0	-5.2
$do(C=0)$	-9.9	-11.3	-12.4	-9.7	-8.5	-15.6

Multimodal Named Entity Typing

ACE	UMT			MKGFormer		
	NoAug.	CG	MCIL	NoAug.	CG	MCIL
$do(V=0)$	-1.4	-0.9	-1.2	-4.1	-7.3	-1.5
$do(E=0)$	-8.3	-23.8	-10.5	-12.1	-11.7	-15.3
$do(C=0)$	-6.0	-10.6	-9.0	-5.9	-9.2	-7.2

Table 4: Causal effect of image V, entity E and context C on the model's performance. We compare the F1 scores of the original model and the ones that we intervene on the above three variables by the operation $do(\cdot = 0)$. The lower value represents the higher importance of the factor to the model during the inference procedure.

5.4.1. Ablation Study

To fully understand the effectiveness of the different multimodal counterfactual instances in MCIL, we conducted an ablation study as shown in Table 3. The results demonstrate the significant impact of each component on the overall performance of the model. Firstly, we observe that removing the entity-based counterfactual instance (ECI) leads to a significant decrease in performance, indicating the importance of entity information to the entity-based MIE tasks. This is because the MIE models rely on the entity information which implies the bias of limited samples. Considering that there are various forms of entity mentions, the bias will force the model to learn the simple features like: the similar context or image to identify the target label. By introducing the ECI into MCIL, the model can capture the non-spurious correlation between the entity information and the label, leading to improved performance.

Secondly, we examine the contribution of vision-based counterfactual instance (VCI) and context-based one (CCI) to the final results. The results of MRE reveal that both VCI and CCI make significant contributions to the final performance of the task, albeit in different degrees. But the MNET

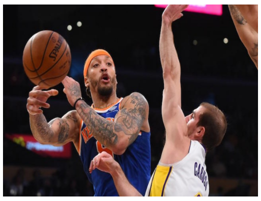
Multimodal Relation Extraction		Multimodal Named Entity Typing	
			
A. Lakers officially announced the signing of Michael Beasley.		B. [Gonzalez] wrestled for the World Wrestling Federation in 1993.	
GT:	/per/org/member_of	GT:	People
NoAug.:	/per/per/alternate_names ✗	NoAug.:	Organization ✗
CG:	/per/per/alternate_names ✗	CG:	People ✓
MCIL:	/per/org/member_of ✓	MCIL:	People ✓

Table 5: The two cases in the test sets of the entity-based multimodal information extraction datasets, and the prediction results of different low-resource methods on these test samples. "GT" is short for the ground truth.

task depends on the VCI more than the CCI which demonstrates the importance of the rich semantic information from image data to the task. Besides, the different multimodal fusion modules like: UMT and MKGFormer are sensitive to the specific type counterfactual instance because of their unique fusion mechanisms and architectures.

5.4.2. Causal Effect Analysis

As shown in Table 4, we conduct the causal effect analysis of image V, entity E and context C on the model's performance. We intervene on the above three variables by the operation $do(\cdot = 0)$ and it means that we replace the features of the specific variable with the zero vectors during the inference procedure. The difference between the F1 scores of the original model and the ones where we apply the intervention is denoted as average causal effect (ACE) (Zeng et al., 2020). And the lower value indicates that the variable is more important to the model during inference. We can observe that the methods are sensitive to the different variables. For the MRE, the context variable is the most important to the performance of the model and the proposed framework enhances the context information which plays more vital role to the model. Besides, MCIL can re-weight the variables to reduce the bias of the limited samples and imporve the performance compared with the baselines. And for the MNET, the entity variable is the most important to the model and MCIL can capture the reasonable combination of the multimodal features to improve the performance by not only focusing on the specific variables.

5.4.3. Case Study

As shown in Table 5, we present a case study to investigate the difference of the prediction results from low-resource methods. The Table 5.A shows the results of the multimodal relation extraction and the "NoAug." and CG methods acquired the wrong results. Because they do not capture the reasonable combination of the multimodal representations and MCIL can re-weight the different features to improve the prediction results. Besides, the Table 5.B shows the results of the multimodal named entity typing and the CG and MCIL methods predicted the correct results. Because the CG and MCIL methods pay more attention to the entity representation and the "NoAug." method only learn the bias from the multimodal representation because of the limited samples. In summary, the proposed MCIL framework can enhance the multimodal representations by reducing the bias of the limited samples.

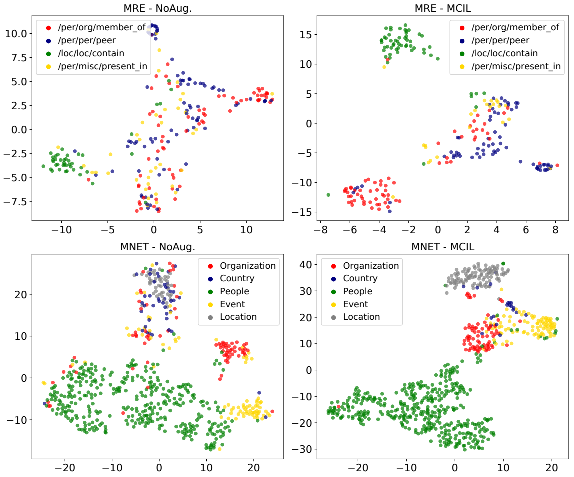


Figure 4: The t-SNE visualization results of the multimodal representations with the specific types extracted from UMT combined with the original and MCIL low-resource methods respectively.

5.4.4. Visualization Analysis

To assess the efficacy of multimodal representations for entity-based MIE tasks, we visualize the features extracted from UMT combined with different low-resource methods in Figure 4. We select the samples with specific labels from the test sets of the two benchmark datasets and visualize the learned representations of "NoAug." and MCIL methods respectively. We then reduce the dimensionality of the representations to two using t-SNE. Our results reveal that the representations output from "NoAug." do not exhibit clustering clearly, indicating that the model has not learned discriminative features for each sample category. In contrast, the representations of MCIL are more densely clustered within each category, suggesting that the proposed framework is more effective in capturing the nuanced consistency between multimodal samples of the same type.

6. Conclusion

In this paper, we propose the effective low-resource method named multimodal counterfactual instance learning (MCIL) framework for the entity-based multimodal information extraction (MIE). We provide the theoretical foundation by the use of the structural casual model to explore the correlation between the different features and the output labels. In the MCIL, we generate the counterfactual instances by the interventions on the limited samples and exploit the causal effect as a supervisory signal to maximize the effect for improving the generalization of the model. The experimental results and detailed analysis show the effectiveness of MCIL to the entity-based MIE tasks in various aspects.

7. Acknowledgements

We thank the anonymous reviewers for the valuable comments on our manuscript. This research is supported by the National Natural Science Foundation of China (No. 62272250, 62302243, U22B2048), and the Natural Science Foundation of Tianjin, China (No. 22JCJQJC00150).

8. Bibliographical References

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 904–915, New York, NY, USA. Association for Computing Machinery.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: A maximum entropy approach using global information. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S. Yu. 2023. Multimodal relation extraction with cross-modal retrieval and synthesis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 303–311, Toronto, Canada. Association for Computational Linguistics.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *CoRR*, abs/1705.03645.

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.

Ngoc Dang Nguyen, Wei Tan, Lan Du, Wray L. Buntine, Richard Beare, and Changyou Chen. 2023. AUC maximization for low-resource named entity recognition. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13389–13399. AAAI Press.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1005–1014. IEEE.

Sergei Volodin, Nevan Wichers, and Jeremy Nixon. 2020. Resolving spurious correlations in causal models of environments via interventions. *CoRR*, abs/2002.05217.

Peng Wang, Tong Shao, Ke Ji, Guozheng Li, and Wenjun Ke. 2023. fmlre: A low-resource relation extraction model based on feature mapping similarity calculation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13700–13708. AAAI Press.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada. Association for Computational Linguistics.

Xin Xu, Xiang Chen, Ningyu Zhang, Xin Xie, Xi Chen, and Huajun Chen. 2022. Towards realistic low-resource relation extraction: A benchmark with empirical baseline study. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 413–427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Junjie Yu, Xing Wang, Jiangjiang Zhao, Chunjie Yang, and Wenliang Chen. 2022. STAD: Self-training with ambiguous data for low-resource relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2044–2054, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Weijiang Yu, Yingpeng Wen, Fudan Zheng, and Nong Xiao. 2021. Improving math word problems with pre-trained knowledge and hierarchical reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3384–3394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5674–5681. AAAI Press.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5298–5306, New York, NY, USA. Association for Computing Machinery.

Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6293–6302, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

9. Language Resource References