



# Unsupervised multimodal learning for image-text relation classification in tweets

Lin Sun<sup>1</sup> · Qingyuan Li<sup>1,2</sup> · Long Liu<sup>3</sup> · Yindu Su<sup>1,2</sup>

Received: 15 May 2023 / Accepted: 6 September 2023 / Published online: 10 October 2023  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

## Abstract

Recent studies show that the use of multimodality can effectively enhance the understanding of social media content. The relations between texts and images become an important basis for developing multimodal data and models. Some studies have attempted to label image-text relation (ITR) and build supervised learning models. However, manually labeling ITR is a challenging task and incurs many controversial labels because of disagreements among the annotators. In this paper, we present a novel unsupervised multimodal method called ITR pseudo-labeling (ITRp) that learns multimodal representations for various ITR types using different finetuning strategies. Our ITRp method generates pseudo-labels by clustering and uses them as supervision to train the classifier and encoders. We evaluate the ITRp method on the ITR dataset and the effects of the samples with incorrect labels on both the supervised and unsupervised models. The code and data are available on the website <https://github.com/SuYindu/ITRp>.

**Keywords** Image-text relation · Unsupervised learning · Deep clustering · Multimodal learning

## 1 Introduction

Research on image-text relation (ITR) in the web is important because text and images are the two main elements of content on a website, and they have to work effectively together in order to create engaging and effective content. Understanding the relationship between text and images on a website can lead to improved website design, user

experience, and communication [1]. In addition, search engines like Google have algorithms that analyze text and image content on web pages in order to rank them on search engine results pages.

Research on the topic can also help businesses and organizations create more targeted and effective marketing strategies by understanding how to use text and images together to communicate messages more effectively. Besides, ITR can be leveraged to assist downstream tasks, e.g., multimodal named entity recognition [2], multimodal aspect-level sentiment analysis [3], and multimodal disaster classification [4].

In order to formalize and comprehend the relationship between text and images, Vempala and Preotiuc-Pietro's research identified four types of semantic image-text relations that are commonly observed in tweets [5]. These relations are used to classify the relationship between text and images in a practical manner and are divided into two tasks: the *text task* and the *image task*. The text task focuses on identifying if there is a semantic overlap between the content of the text and the image, and further divides the relations into “text is represented” or “text is not represented”. The image task focuses on identifying if an image's content contributes additional information to the meaning of the tweet beyond the text, and divides the relation into “image add” or “image does not add”. By combining the labels of the two

---

✉ Lin Sun  
sunl@hzcu.edu.cn  
Qingyuan Li  
liqingyuan@zju.edu.cn  
Long Liu  
liulong\_9@163.com  
Yindu Su  
yindusu@zju.edu.cn

<sup>1</sup> Department of Computer Science, Hangzhou City University, 51 Huzhou Street, Hangzhou 310015, Zhejiang, China

<sup>2</sup> College of Computer Science and Technology, Zhejiang University, 38 Zheda Road, Hangzhou 310027, Zhejiang, China

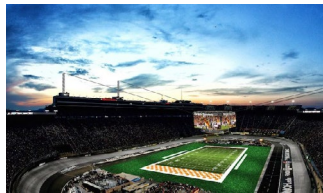
<sup>3</sup> Zhejiang Development and Planning Institute, 598 Gudun Road, Hangzhou 310012, Zhejiang, China

**Fig. 1** Examples of four types of ITR [5]. **a** Text is represented & image adds, **b** Text is represented & image does not add, **c** Text is not represented & image adds, and **d** Text is not represented & image does not add



**Fig. 2** Two examples with incorrect labels in the ITR dataset [5]

LOOK: A giant videoboard will hang over the field for UT-VT at Bristol



(a) Labeled as “text is not represented”, but ‘field’ and ‘videoboard’ are in the image.

I NEED THIS



(b) Labeled as “image does not add”, but ‘this’ refers to the sport shoe in the image.

binary tasks, four types of ITRs are formed, called *image-text task*, as similar as the main types in [6] which are “complementary”, “illustration”, “anchorage”, and “independent”. See Fig. 1 for examples of four types of image-text relations.

However, labeling enough ITR samples for supervised learning is a tough task. Vempala and Preotiuc-Pietro [5] manually annotated ITRs on tweets and the inter-annotator agreement measured using Krippendorff’s  $\alpha$ . The strength of agreement of  $\alpha$  is listed as follows [7]:

- Fair agreement— $0.2 < \alpha < 0.4$
- Moderate agreement— $0.4 < \alpha < 0.6$
- Substantial agreement— $0.6 < \alpha < 0.8$
- Almost agreement— $0.8 < \alpha < 1.0$

Most computational linguistics researchers followed the more stringent conventions [8, 9], i.e.,  $\alpha > 0.8$  as good reliability. The  $\alpha$  of the text task is only 0.46 and that of the image task is 0.71, which means that it does not represent good reliability, especially in the text task because the labeling results strongly rely on the understanding of individuals.

The main objective of this study was to address the challenge of manually labeling image-text relation, which is a time-consuming and challenging task due to disagreements among annotators. Through careful review of the ITR dataset, we found that many samples had incorrect labels, which could adversely affect the performance of any supervised learning model trained on this dataset.

Figure 2 shows two examples with incorrect labels. To overcome this challenge, we proposed a novel unsupervised multimodal learning method for image-text relation classification. The proposed method leverages the knowledge contained in pre-trained models (PTMs) to classify ITRs with pseudo-labels generated by clustering. By using this unsupervised approach, we aim to improve the accuracy and reliability of ITR classification in tweets without the need for time-consuming manual annotation. The contributions of this paper are as follows:

- This paper addresses the challenge of manually labeling image-text relation (ITR), which is a time-consuming and challenging task due to disagreements among annotators. The study reviewed the ITR dataset and found wrongly labeled samples. This is a novel contribution to the field of image-text relation classification, as it questions the reliability of existing labeled datasets, which will help to better study the image-text relation in the future.
- The proposed method, ITR pseudo-labeling (ITRp), is a novel unsupervised multimodal learning approach for image-text relation classification. The study utilizes a simple and fast clustering algorithm to achieve state-of-the-art performance through iterative learning with fewer computing resources. To better match various image-text relations, we propose a finetuning strategy on multimodal PTMs, which can effectively align the semantics between texts and images for the text the image tasks separately.

- We have demonstrated the effectiveness of our proposed method on the ITR dataset, achieving competitive results compared to existing supervised learning approaches while avoiding the need for manual labeling. The results also indicate that the proposed method can effectively address the challenge of incorrect labels in the ITR dataset, leading to more reliable and accurate image-text relation classification.

## 2 Related work

**Image-text relation** Early studies of ITRs focused on illustrations in web pages [10], advertisements and drawings [6]. The authors categorized and analyzed ITRs based on conceptual closeness and subordinate directions for potential applications, such as extracting keywords from texts for precise retrieval and extending hypermedia links between texts and images. Wang et al. [11] introduced image-text association to build a probabilistic model to discover topics from microblogs. However, none of these studies tested the performance of machine learning methods for categorizing ITRs.

Some researchers used machine learning methods to train ITR classification models using labeled data. Chen et al. [12] annotated ITRs as “visually-relevant” and “non visually-relevant” and built a naive Bayes model to predict the relationship type. Later, Chen et al. [13] added an emotionally relevant factor to relate tweet’s image and text besides visually-relevant and proposed Visual-Emotional LDA for ITR classification. Zhang et al. [14] collected an advertisement ITR dataset for whether the image and slogan form a “parallel” or “non-parallel” relationship, and trained SVM based classifiers on it.

Recently, deep neural networks were used as encoders for ITR classification. Henning and Ewerth [15] proposed two semantic metrics for ITR and a LSTM+Inception-v3 model to automatically predict them. Based on this, Otto et al. [1] presented a categorization of eight semantic ITR types and derived how they can systematically be characterized by a set of three metrics. They also update the encoders to GRU+Inception-v4. Kruk et al. [16] introduced a multi-modal dataset of Instagram posts labeled for inferring author intents from image-caption relations, the encoders they used for classification are ELMo+ResNet. Furthermore, Vempala and Preotiuc-Pietro [5] built a dataset of tweets annotated with four ITR types that express whether images or texts provide additional or overlapping information to the other modality, based on which they trained multiple supervised models using various features.

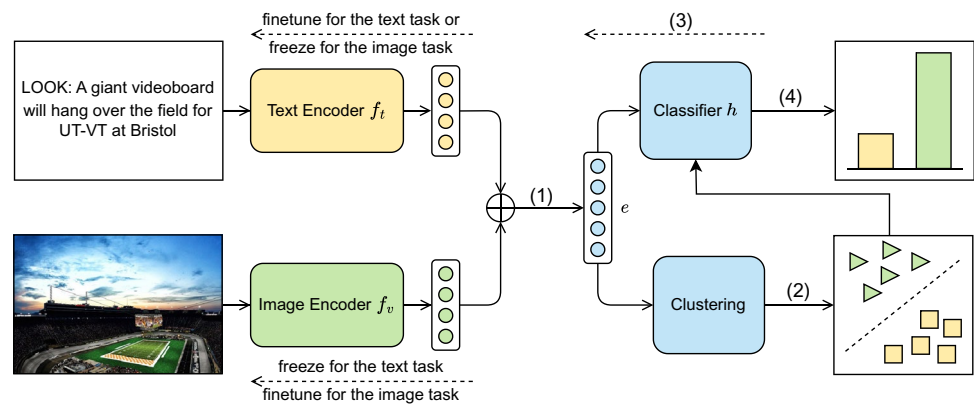
All of the above image-text classification models were trained using labeled data. In this work, we seek to learn classification models in an unsupervised manner based on a large number of tweets. To verify the effectiveness of our model, we test it on the dataset of [5] because it is with clear semantics, publicly available, and the size is relatively large compared with the other datasets [1, 12–14, 16].

**Unsupervised clustering learning** Deep clustering, as a technique that jointly optimizes clustering and representation learning, has attracted growing attention. The iterative deep clustering conducts the interaction between clustering and representation learning by iterating two steps: (1) calculating clustering results given current representations and (2) updating representations given current clustering results. Caron et al. [17] proposed deep clustering for the unsupervised training of CNN, it iterates between applying k-means to the representation from CNN and updating the backbone along with the classifier by utilizing cluster assignments as pseudo-labels. Alwassel et al. [18] extended to audio-video multi-modal setting using clusters learned from one modality as pseudo-labels to refine the representation of the other modality. Instead of applying k-means to feature vectors extracted by the neural network, Asano et al. [19] solved an optimal transport problem to obtain the pseudo-labels and Caron et al. [20] used a swapping prediction problem to predict the online clustering assignment. Li and Tang [21] introduced a weakly-supervised deep matrix factorization framework [22] to learn latent image and tag representations, followed by the development of collaborative factor analysis [23] and hashing code learning incorporated methods [24] in later works.

## 3 Method

An overview of the ITRp model is illustrated in Fig. 3. Our model mainly consists of three parts: (1) multimodal feature extraction (see Sect. 3.1); (2) clustering for pseudo-label generation (see Sect. 3.2); (3) classifier training with pseudo-labels (see Sect. 3.3). Given an unlabeled training set  $X = \{x_i\}_{i=1}^n$  of  $n$  image-text pairs consisting of a text  $x_i^t$  and an image  $x_i^v$ , encoders  $f_t(\cdot)$  and  $f_v(\cdot)$  are initialized with pre-trained language and vision models, respectively. We start by extracting the text and image features and concatenate them to generate a multimodal feature. Next, we produce pseudo-labels using clustering. Based on the pseudo-labels, we train a binary classifier and finetune the encoders. Clustering and pseudo-label-based classification training are performed in an iterative manner for continued performance improvement.

**Fig. 3** Illustration of the ITRp model. The model clusters (1) multimodal representations and then uses (2) cluster assignments as pseudo-labels to (3) train the neural classifiers  $h$ . The model adapts to the text task by only finetuning the text encoder and the image task by only finetuning the image encoder. Finally, we (4) predict relation types by classifiers  $h$



(see Sect. 3.4). Finally, we infer the relations between texts and images via the binary classifier (see Sect. 3.5).

### 3.1 Multimodal feature extraction

We employ pre-trained language models such as BERT [25] as a text encoder and denote the output as  $f_t(x^t) \in \mathbb{R}^{d_t}$ . We use pre-trained vision models such as ResNet [26] as an image encoder and denote the output as  $f_v(x^v) \in \mathbb{R}^{d_v}$ . The text embeddings are obtained by taking the output of the [CLS] token or averaging all token embeddings from the last layer of the pre-trained models, while the image embeddings are obtained by taking the output of the global average pooling layer of the pre-trained model. We then concatenate the outputs of two encoders and generate a multimodal embedding  $e \in \mathbb{R}^{d_t+d_v}$ , which are used as input for the clustering and classification tasks in the ITRp method. Formally, a sample  $x$  is mapped to a multimodal representation  $e$  as follows:

$$e = f_t(x^t) \oplus f_v(x^v). \quad (1)$$

### 3.2 Pseudo-label generation

The pseudo-label generation process in the ITRp method involves clustering the multimodal embeddings obtained through the feature extraction process. Specifically, k-means clustering is used to group the embeddings into  $k$  clusters, where  $k$  is a hyperparameter that is set based on the number of ITR types to be classified. The centroid of each cluster is then used as a representative pseudo-label for that group of embeddings. These pseudo-labels are then used to train both the classifier and the encoder in a self-supervised manner, without the need for manual labeling of the data. To train image-text relation classifiers for the text and image tasks, we froze the respective encoders to enable better semantic

alignment for different ITR tasks. During training, we set  $k$  to 2, which corresponded to ‘text presented’ vs. ‘text not presented’, and ‘image adds’ vs. ‘image not add’. Although  $k$  can also be set to 4 to classify the data into four types of ITR, we discover that this approach does not achieve optimal performance because it becomes more challenging to cluster into 4 distinct groups, and there is a lack of aligned semantic learning for the text and image tasks, each of which requires separate treatment.

The k-means algorithm takes a set of vectors as input and clusters them into  $k$  distinct groups based on pairwise Euclidean distance. More precisely, it jointly learns a centroid matrix  $C \in \mathbb{R}^{d \times k}$  and the cluster assignments  $y_i \in \{0, 1\}^k$  s.t.  $\|y_i\| = 1$  of each sample  $x_i$  by solving the following problem:

$$\min_C \frac{1}{n} \sum_{i=1}^n \min_{y_i} \|e_i - Cy_i\|_2^2. \quad (2)$$

The samples that belong to the same cluster are assigned the same pseudo-label.

### 3.3 Training with pseudo-labels

**Training set sampling** To balance the number of samples in different clusters, we adopt three balanced sampling strategy: *Over-sampling* [27] randomly duplicates samples from the minority class to make two sets equal, and *under-sampling* [28] randomly deletes samples from the majority class; *Combination sampling* [29] randomly selects samples from each cluster to create a balanced dataset with a size of  $\frac{n}{k}$ , involving over-sampling the minority class to  $\frac{n}{k}$  samples and under-sampling the majority class to  $\frac{n}{k}$  samples. We refer to the balanced dataset with size  $n'$  as  $X'$  and use it for pseudo-label-based classification training in the next paragraph.

**Classifier training** A randomly initialized classifier head  $h \in \mathbb{R}^{(d_t+d_v) \times 2}$  which is a fully connected layer on top of the multimodal representation  $e$ , as well as an encoder, is trained

using the pseudo-labeled samples. The parameters  $W$  of classifier  $h$  and  $\theta$  of an encoder are jointly learned by minimizing the following objective:

$$\min_{\theta, W} \frac{1}{n'} \sum_{i=1}^{n'} \mathcal{L}(h(e), y_i), \quad (3)$$

where  $\mathcal{L}$  is cross-entropy loss function.

### 3.4 Iterative training algorithm

For the complete training, the iterative training procedure is shown in Algorithm 1. We divide the training set  $X$  into subsets  $X_1, X_2, \dots, X_{n/m}$  of size  $m$ . At each iteration, we choose  $m$  data for clustering because using all data could increase the difficulty of clustering. This also facilitates unsupervised learning by building more dynamic and up-to-date pseudo-labels [30].

**Table 1** The statistics of the datasets

Dataset	Train set size	Test set size
Twitter100k dataset	100k divided into 20 subsets	–
ITR dataset	3,576 (663/785/918/1,210)	895 (171/198/231/295)

The size of types ‘Text represented &Image adds’ / ‘Text represented &Image not add’ / ‘Text not represented &Image adds’ / ‘Text not represented &Image not add’ in the ITR dataset are also listed

**Table 2** Training settings for ITR tasks

Settings	Text task	Image task
BERT	Finetuning	Frozen
ResNet	Frozen	Finetuning
Sampling	Under-sampling	Over-sampling

**Algorithm 1:** The ITRp training algorithm

**Input:** Training set  $X = \{x_i\}_{i=1}^n$ .

**Output:** Parameters  $\theta_t$ ,  $\theta_v$ , and  $W$  of the text encoder, image encoder, and classifier  $h$ , respectively.

Load text and image encoders;

**while**  $epoch \leq num\_epochs$  **do**

**for** all subsets  $\{X_k\}_{k=1}^{n/m}$  **do**

        Compute embedding  $e$  for  $x_i \in X_k$ ;

        Generate pseudo-labels;

        Sample  $X_k$  to balanced  $X'_k$ ;

        Randomly initialize  $W$ ;

**for** all batches in  $X'_k$  **do**

            Update  $W$  and  $\theta_t$  for the text task or update  $W$  and  $\theta_v$  for the image task;

**end**

**end**

**end**

### 3.5 ITR type prediction

We adopt two distinct different training strategies (as summarized in Table 2) to train ITRp models for the text and image tasks, resulting in two classifiers, with one being for the text task inferring and the other for the image task inferring; the first strategy involves finetuning the text encoder and freezing the image encoder for the text task, while the

second strategy involves freezing the text encoder and finetuning the image encoder for the image task.

During the testing stage, we feed image-text pairs into two classifiers to predict the corresponding types for both the text and image tasks. Combining the labels of two tasks gives rise to four types of ITRs. We do category mapping [31] after the ITRp iterative training, therefore there is no annotated data used in the classifier training.



## 4 Experimental settings

In this section, we will discuss the experimental datasets used, the parameters and settings employed in the model implementation, and the comparison models used for evaluation.

### 4.1 Datasets

We use the unlabeled Twitter100k [32] dataset for unsupervised training, and the labeled ITR dataset [5] for validation and evaluation. These two datasets were independently collected from different users. Table 1 shows the statistics of two datasets and we briefly introduce them as follows:

**Twitter100k dataset.**<sup>1</sup> This dataset is comprised 100,000 image-text pairs collected from Twitter. It covers a wide range of domains, such as sports, architecture, food, animals, news, plants, people, posters, and others. We divide it into  $m = 20$  subsets for each clustering.

**ITR dataset.**<sup>2</sup> In this dataset, the authors annotated tweets with two tasks and four image-text relation types. We follow the same split of 3,576:895 to form the train/test sets as in the original paper. We use the train set to train the supervised model and use the test set for comparisons of ITRp with supervised models.

In the ITR dataset, the authors stated that there were relatively low agreement scores between annotators, therefore we believe that there must be a lot of samples with controversial labels. To better compare the performance of classification models, we checked the test set. We employed five annotators who have 3 years of experience in social media research to identify the wrongly labeled samples in the ITR dataset. We provided them with the annotation guidelines defined by Vempala and Preotiuc-Pietro [5] and asked them to review the dataset and identify any samples with controversial labels. For each sample with a controversial label, the annotators provided reasons for why they thought the label was incorrect. After reviewing the reasons, the annotators discussed among themselves to reach a consensus on whether the label was correct or incorrect. Once a consensus was reached, we considered the labels with a majority agreement to be incorrect and selected them for relabeling. It is important to note that our relabeling process was not based solely on the votes by the annotators but also on a discussion among the annotators to ensure a fair and consistent evaluation. By using this approach, we identified a large number of wrongly labeled samples in the ITR dataset, which could adversely affect the performance of supervised learning models trained on this dataset. 218 wrongly labeled samples were selected, where 124 samples are in the text

task only, 61 samples are in the image task only, and 33 samples are in both tasks. We have uploaded the incorrectly labeled sample set in the file “controversial\_samples.txt” of the public GitHub repository.<sup>3</sup>

To verify the improvement of data quality after relabeling, we use another powerful text-image semantic model, CLIP [33], to compute the correlation between the CLIP similarity score and ITR category score on the test set. Specifically, CLIP similarity score is the cosine similarity between text and image embeddings from the CLIP model, and the categories of “text is not represented & image does not add”, “text is not represented & image adds”, “text is represented & image does not add” and “text is represented & image adds” are assigned to a respective score of 1, 2, 3 and 4, according to the strength of association between text and image. As result, Spearman’s rank correlation coefficient increases from 0.09 to 0.28 on the corrected labels and is 0.38 on all data of the test set, thus indicating an enhancement in the quality of sample labeling.

### 4.2 Baselines

Previous work on the ITR dataset mainly employed multimodal features and a linear neural network for classification:

**LSTM+InceptionNet** [5]. This model concatenated the features from the final layers of LSTM and InceptionNet [34] and passed them through a multilayer perceptron with one hidden layer.

**RoBERTa+EfficientNet** [35]. This linear model is trained over the concatenation of static image and text representations from EfficientNet [36] and RoBERTa [37].

**LXMERT** [35]. LXMERT [38] is a multimodal Transformer pre-trained with five tasks: masked language modeling, RoI feature regression, detected label classification, cross-modality matching, and image question answering. Hessel and Lee [35] fine-tuned LXMERT for the integrated image-text task, achieving state-of-the-art performance.

**CLIP** [33]. CLIP was pre-trained by contrastive learning on image-text pairs and performed excellent zero-shot transfer to downstream tasks. The concatenation of embeddings from the CLIP text encoder and image encoder was used as a classification feature.

**CMA-CLIP** [39]. CMA-CLIP unified two types of cross-modality attention, sequence-wise attention, and modality-wise attention, and is capable of performing multi-task classification with multi-modalities.

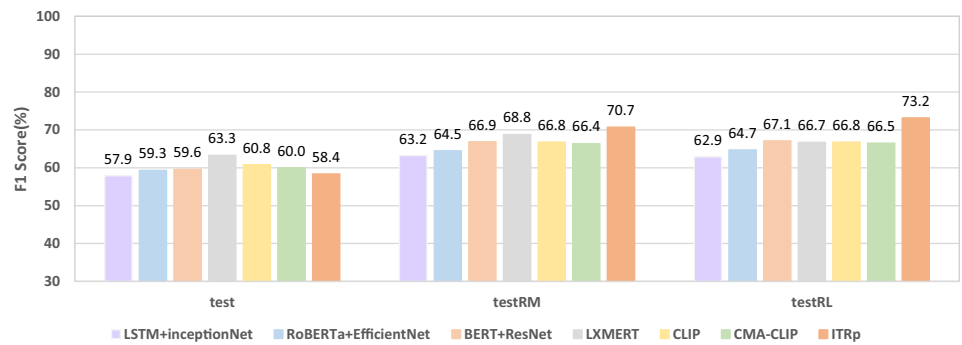
**BERT+ResNet**. This model uses the same encoders as ITRp and adapts a similar architecture as

<sup>1</sup> <https://github.com/huyt16/Twitter100k>.

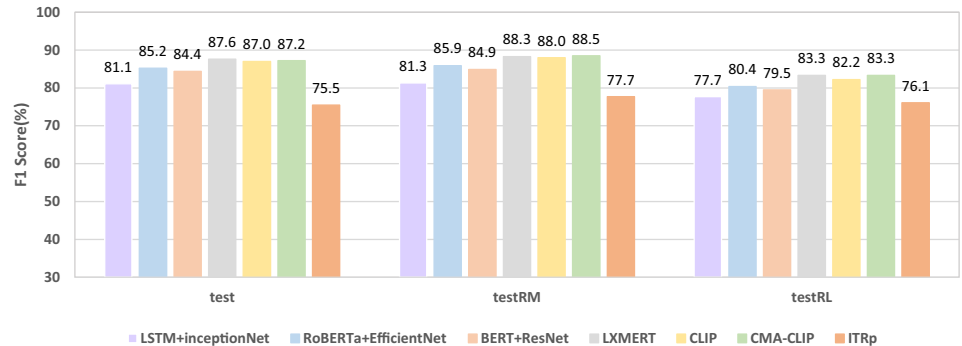
<sup>2</sup> <https://github.com/danielpreotiuc/text-image-relationship>.

<sup>3</sup> <https://github.com/SuYindu/ITRp>.

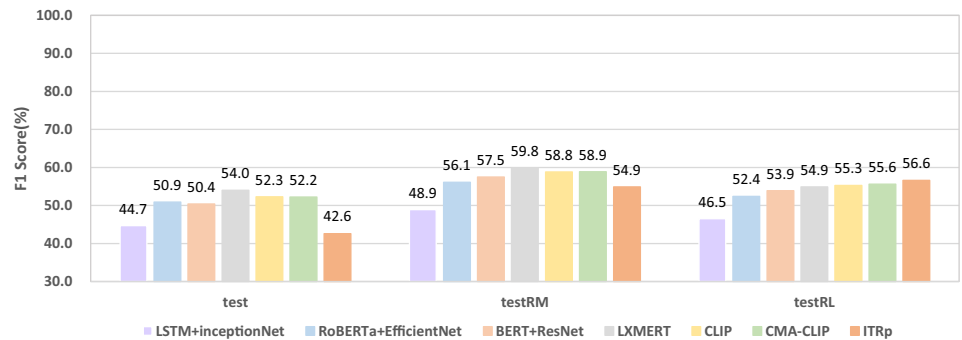
**Fig. 4** Performance comparisons between supervised baselines and our ITRp method in ITR tasks, test<sub>RM</sub> refers to the test set where 218 wrongly labeled samples are removed and test<sub>RL</sub> refers to the test set where 218 wrongly labeled samples are corrected. The results are the averages over 6 runs



(a) Text task.



(b) Image task.



(c) Image-text task.

RoBERTa+EfficientNet, with RoBERTa and EfficientNet replaced by the BERT-Base and ResNet-101.

All parameters in the aforementioned models are finetuned in supervised learning to achieve better performance.

### 4.3 Implementation details

We leverage BERT-Base to encode the text part of tweets, and ResNet-101 for the image part. Specifically, we use the last layer's sequence of hidden states in BERT to obtain an overall representation of the sentence, which was then averaged to generate a 768-dimensional text embedding. Similarly, the feature map from ResNet's last convolution layer, with shape  $7 \times 7 \times 2048$  for  $224 \times 224$  input image,

and mean-pool it to generate a 2048-dimensional image embedding.

To fuse the text and image embeddings, we experimented with different fusion modules, including concatenation, element-wise multiplication, and element-wise addition. We found that concatenation, a simple yet effective parameter-free operation, worked best and used it as the fusion module  $g$ . As for the classifier  $h$ , a fully connected layer followed by a softmax layer is employed. We use Adam [40] with a learning rate of  $2e-5$  to finetune pretrained models, and a learning rate of  $1e-4$  to train the classifier  $h$ .

To ensure a uniform framework for different ITR tasks, we design different finetuning settings for each task, as summarized in Table 2. Specifically, we freeze the ResNet model to perform the text task learning and freeze the

**Table 3** Performance comparisons with vanilla clustering algorithms in F1 score (%) on the test<sub>RL</sub> set

Method	Text task	Image task	Image-text
GMM	55.3	69.9	39.1
ITRp (GMM)	72.4	75.2	55.6
k-means	59.6	70.5	44.4
ITRp (k-means)	<b>73.2</b>	<b>76.1</b>	<b>56.6</b>

“Image-text” is short for image-text task

The best scores are indicated in bold

BERT model to perform the image task learning. These settings are based on fixing the encoder representation of one modality and finetuning the encoder of the other modality to better align the semantic relations for different ITR tasks. By doing so, we aim to achieve a fine-tuned joint representation of both modalities that captures the complementary and overlapping information between them. In Sect. 5.2, we show the performance comparisons in different finetuning settings.

To balance the trade-off between computational efficiency and model performance, we set the subset size  $m$  to 5k and batch size to 32. We train the model on a machine with NVIDIA A100 40GB GPU and Hygon C86 32-core CPU. Each epoch takes approximately 40 min, with the k-means clustering accounting for 62% of the time because the clustering algorithm is computed on the CPU. The overall training takes 6 h on one GPU. We use the weighted F1 score as an evaluation metric, which is the same as in [5].

## 5 Results

### 5.1 Performance comparisons

Figure 4 shows the ITR task results of supervised baselines and the ITRp method on the test set of the ITR dataset. Using powerful PTMs, the supervised RoBERTa+EfficientNet, LXMERT, CLIP and CMA-CLIP models achieve F1 scores of 50.9%, 54.0%, 52.3%, and 52.2%, where large margins (+6.2%, +9.3%, +7.6%, and +7.5%, respectively) are obtained compared to LSTM+InceptionNet. It is worth noting that the performance of the text task is approximately 25% absolute lower than that of the image task. Because of the low Krippendorff's  $\alpha = 0.46$  in the text task, the unreliability of annotation makes supervised learning intractable.

We provide new evidence of performance on the test<sub>RM</sub> set and test<sub>RL</sub> set when wrongly labeled samples are concerned. From Fig. 4, we gain the following insights:

**Table 4** Comparison of different PTM finetuning strategies in F1 score (%) on the test<sub>RL</sub> set

BERT	ResNet	Text task	Image task
Frozen	Frozen	54.7	67.7
Finetuning	Frozen	<b>73.2</b>	74.1
Frozen	Finetuning	67.8	<b>76.1</b>
Finetuning	Finetuning	60.3	70.6

The results are the averages over 6 runs

The best scores are indicated in bold

**Table 5** Comparison of different sampling strategies in F1 score (%) on the test<sub>RL</sub> set

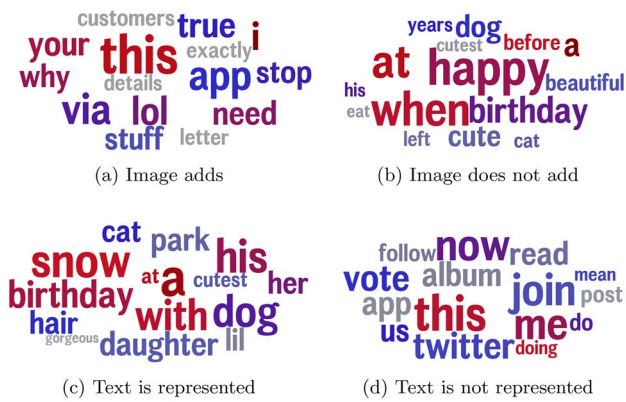
Sampling strategy	Text task	Image task
Random sampling	71.0	69.8
Under-sampling	<b>73.2</b>	75.4
Over-sampling	67.8	<b>76.1</b>
Combination sampling	69.9	76.0

The results are the averages over 6 runs

The best scores are indicated in bold

- Removing the wrongly labeled data, the ITR task performance of all models is significantly improved. This indicates the validity of our selected wrong labels. The smaller Krippendorff's  $\alpha$ , the greater performance impact of the wrongly labeled samples. The increases of F1 score for the text task (+5.2% for RoBERTa+EfficientNet, +5.5% for LXMERT, +6.0% for CLIP, +6.4% for CMA-CLIP, and +12.3% for ITRp) are significantly larger than those of the image task (+0.7% for RoBERTa+EfficientNet, +0.7% for LXMERT, +1.0% for CLIP, +1.3% for CMA-CLIP, and +2.2% for ITRp).
- The wrongly labeled data are harder to be classified for supervised models. F1 scores of the image task on the test<sub>RL</sub> set decrease by −4.8% for RoBERTa+EfficientNet, −4.3% for LXMERT, −5.8% for CLIP, and −5.2% for CMA-CLIP, while the F1 score of our unsupervised ITRp increases slightly.
- The advantage of ITRp is that the performance of text and image tasks is close on the test<sub>RL</sub> set: 73.2% for the text task and 76.1% for the image task. In contrast, the differences between the F1 scores of text and image tasks on the test<sub>RL</sub> set are still large for supervised models, i.e., 15.7% for RoBERTa+EfficientNet, 16.6% for LXMERT, 15.4% for CLIP, and 16.8% for CMA-CLIP.
- In the overall image-text task, ITRp obtains an absolute increase of 2.7% compared to BERT+ResNet and 1.7% to previous state-of-the-art LXMERT on the test<sub>RL</sub> set.





**Fig. 5** Word clouds regarding four ITR types using the ITRp's results



**Fig. 6** Word clouds regarding four ITR types using human annotations

In addition, we compare the performance of vanilla clustering algorithms such as k-means [41] and GMM [42] in Table 3, these methods perform clustering on the static multimodal features concatenated from BERT and ResNet. Other clustering algorithms such as DBSCAN [43] or spectral clustering [44] cannot control the number of clusters or the training cost is much higher than k-means and GMM on high-dimensional features. Due to clustering being performed on the CPU and taking up a significant portion of the training time, the advantage of using the k-means method in this model is that it can complete the model training in a relatively short amount of time. The ITRp method outperforms vanilla clustering algorithms with an increase in +16.5% for GMM and +12.2% for k-means in F1 score, showing that our pseudo-label training with deep neural networks can learn better multimodal representations for ITR tasks.

## 5.2 Discussions


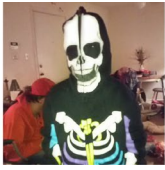






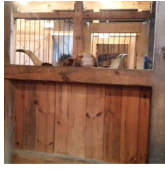



**PTM finetuning strategies.** We experiment on how the optimizing settings of PTMs affect the performance of the ITRp method and show the results in Table 4. First, we take the setting of frozen BERT and frozen ResNet as a baseline model. The results show that the baseline setting is better adapted to the image task, achieving 67.7% in the F1 score. Second, we finetune the text encoder BERT only and obtain an increase in +18.5% for the text task. We finetune the image encoder ResNet only and obtain an increase of +8.4% for the image task. The ITRp model optimizes respective language and vision PTMs for the text and image tasks and achieves the best scores, which also verifies the rationality and validity of our model.

**Sampling strategies** In this section, we show the reason for choosing sampling strategies in the ITRp method. The data distribution of different image-text relations is not balanced. For example, 40% of the ITR dataset are labeled as “text is represented” and 60% are labeled as “text is not represented”, and 44% are labeled as “image adds” and 56% are labeled as “image does not add”. On the unlabeled dataset, the pseudo-labels generated by clustering could become more imbalanced for classifier training. The imbalanced classes could have a negative impact on the learning results [17]. We test the three balanced sampling methods mentioned in Sect. 3.3 and compare them in Table 5. The results show that under-sampling is the best for the text task because the text task is much harder than the image task and the clusters of the text task are not well separated in the feature space. Therefore, under-sampling can improve discrimination by removing the instances from the majority cluster. The balanced sampling methods are better than random imbalanced sampling in the image task and over-sampling is the best.

**Text analysis** Here we show the text analysis of ITRp compared to that of human annotation. We follow the same settings (i.e., *unigram* feature) as in [5] and draw word clouds using Differential Language Analysis ToolKit [45]. Figure 5 illustrates the words correlated with each ITR category performed by ITRp. The results of human annotation are directly copied from Figures 3 and 4 of [5] and shown in Figure 6. The color represents the word's frequency (gray to red for infrequent to frequent), and the size represents the correlation strength.

For “image adds”, compared to human annotation, the ITRp adds several valid words such as ‘*app*’ for images that show what the app looks like, ‘*details*’ for images that show detailed content, and ‘*why*’ for images that refer to the reason. It also removes words such as ‘*would*’ and ‘*not*’ that seem unrelated to images. For “image does not add”,

**Table 6** Example comparisons between our ITRp and human annotation (color table online)

	LOOK: A giant videoboard will hang over the field for UT-VT at Bristol	My 8year old dressing like Tyler ! #num-beronefan	One week! @estfesto-hio	@Fothron Flies	Time
					
Human	text is not represented	text is not represented	text is represented	text is represented	
ITRp	text is represented	text is represented	text is not represented	text is not represented	
	I NEED THIS	Am I ready to go back to school?	This mass of humanity" A heart that loves is always young."	#quotes #motivation #inspiration	
					
Human	image does not add	image does not add	image does not add	image adds	
ITRp	image adds	image adds	image adds	image does not add	
	This or (Wyatt) really wanted my attention... lol	RT @ChiIIVibes: Time lapse of 100 sunsets	This is how I feel when I see @hiromiishima post on twitter:	Good morning	
					
Human	text is not represented & image does not add	text is not represented & image adds	text is represented & image does not add	text is represented & image adds	
ITRp	text is represented & image adds	text is represented & image does not add	text is not represented image adds	text is not represented & image does not add	

The results are best viewed in color

in ITRp, the results are similar to those of human annotation, such as objects ('dog', 'cat', 'birthday') and feelings ('happy' and 'cute'). Other words such as prepositions ('at', 'before'), articles ('a'), and conjunctions ('when') cannot be given extra meaning by images.

For "text is presented", the results of the ITRp and human annotation are quite different. The ITRp mainly adds nouns and pronouns ('snow', 'daughter', 'dog', 'hair', 'park', 'birthday', 'lil',<sup>4</sup> 'her', 'his') about objects or persons presented in images. According to the definition of the text task, these words correctly reflect this kind of image-text relationship. In contrast, in human annotation, the words ('the', 'free', 'without', 'away', 'new') can hardly be represented in images. For "text is

not presented", most words (e.g., 'this', 'join', 'now') are not content words and hardly appear in images.

### 5.3 Case study

Table 6 lists several wrongly labeled examples and the classification results of the ITRp method and human annotation. The 1st row shows four wrongly labeled examples in the text task, the 2nd row shows four wrongly labeled examples in the image task, and the 3rd row shows four wrongly labeled examples in both tasks. We use the ITR type definition in [5] to identify these examples.

In the 1st row, "videoboard" and "the field" are represented in column 1 as well as the "dressing like Tyler" in column 2 (highlighted in yellow background). While

<sup>4</sup> Lil Wayne, an American rapper.

examples in columns 3 and 4 are “only a comment about the content of the image”.

In the 2nd row, “image depicts something that adds information to the text” such “THIS” in column 1, “Am I ready” in column 2, and “This mass of humanity” in column 3, while the image in column 4 “does not add additional content”.

In the 3rd row, “ox” and “sunsets” are represented in the image of columns 1 and 2 respectively, while none of the content words in columns 3 and 4 are displayed. On the other hand, the image in columns 1 and 3 “depicts something that adds information to the text” (how the ox want my attention and how I feel), which is not true for images in columns 2 and 4.

## 6 Conclusion and future works

In this paper, we investigate the ITR dataset because of the low inter-annotator agreement in labeling and propose an unsupervised method ITRp for image-text relation classification. We have carefully checked the samples with incorrect labels in the test set of the ITR dataset and found 216 or 24.3% incorrect samples. The experimental results show that for supervised learning models, the main reason for poor performance on the text task is unreliable annotations. We correct the annotations and re-evaluate the models, and obtain an increase of 5.5% in F1 score on the corrected test set. Our proposed ITRp model achieves an increase of 2.7% in F1 score compared to the vanilla vision and language models and outperforms the supervised models overall on the corrected test set. These results demonstrate the effectiveness of our unsupervised approach in handling the ITR task, especially in scenarios where reliable annotations are limited or costly to obtain.

Although the classification performance of our proposed method was gradually improved through iterations, the model performance was still constrained by the unsupervised clustering performance. Despite the fact that both results exceed the baseline algorithms, the scores of the overall text-image task remain relatively low. In future work, we plan to explore the impact of noisy data on the clustering performance and how it may affect the results. Existing image-text relation datasets are relatively small, and it would be beneficial to develop larger and more diverse datasets for evaluating the robustness and generalizability of models. In addition, the four types of image-text relation identified in the paper are a good starting point, but there may be other types of relation that are not captured by these categories. Further research could

explore new types of relation and develop methods for identifying them.

**Data availability** The datasets generated during and/or analyzed during the current study are available in the GitHub repository, <https://github.com/SuYindu/ITRp>.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Otto C, Springstein M, Anand A (2020) Ewerth R Characterization and classification of semantic image-text relations. *Int J Multimed Inf Retrieval* 9:31–45
2. Sun L, Wang J, Zhang K, Su Y, Weng F (2021) Rpbert: A text-image relation propagation-based BERT model for multimodal NER. In: *AAAI*, pp 13860–13868
3. Ju X, Zhang D, Xiao R, Li J, Li S, Zhang M, Zhou G (2021) Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In: *EMNLP*, pp 4395–4405
4. Sosea T, Sirbu I, Caragea C, Caragea D, Rebedea T (2021) Using the image-text relationship to improve multimodal disaster tweet classification. In: *ISCRAM 2021 conference proceedings—18th international conference on information systems for crisis response and management*, pp 691–704
5. Vempala A, Preotiuc-Pietro D (2019) Categorizing and inferring the relationship between the text and image of twitter posts. In: *Annual meeting of the association for computational linguistics*
6. Martinec R, Salway A (2005) A system for image-text relations in new (and old) media. *Vis Commun* 4(3):337–371
7. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 159–174
8. Carletta J, Isard A, Isard S, Kowtko JC, Doherty-Sneddon G, Anderson AH (1997) The reliability of a dialogue structure coding scheme. *COLING* 23(1):13–31
9. Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. *COLING* 34(4):555–596
10. Marsh EE, White MD (2003) A taxonomy of relationships between images and text. *J Document* 59(6):647–672
11. Wang Z, Cui P, Xie L, Zhu W, Rui Y, Yang S (2014) Bilateral correspondence model for words-and-pictures association in multimedia-rich microblogs. *ACM Trans Multim Comput Commun Appl* 10(4):34–13421
12. Chen T, Lu D, Kan MY, Cui P (2013) Understanding and classifying image tweets
13. Chen T, SalahEldeen H, He X, Kan MY, Lu D (2015) Velda: relating an image tweet’s text and images. In: *AAAI conference on artificial intelligence*
14. Zhang M, Hwa R, Kovashka A (2018) Equal but not the same: understanding the implicit relationship between persuasive images and text. In: *British machine vision conference*
15. Henning CA, Ewerth R (2017) Estimating the information gap between textual and visual representations. *Int J Multimed Inf Retrieval* 7:43–56

16. Kruk J, Lubin J, Sikka K, Lin X, Jurafsky D, Divakaran A (2019) Integrating text and image: Determining multimodal document intent in instagram posts. In: Conference on empirical methods in natural language processing
17. Caron M, Bojanowski P, Joulin A, Douze M (2018) Deep clustering for unsupervised learning of visual features. In: European conference on computer vision
18. Alwassel H, Mahajan D, Korbar B, Torresani L, Ghanem B, Tran D (2020) Self-supervised learning by cross-modal audio-video clustering. In: Advances in neural information processing systems, vol 33, pp 9758–9770
19. Asano YM, Rupprecht C, Vedaldi A (2020) Self-labelling via simultaneous clustering and representation learning. In: International conference on learning representations
20. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A (2020) Unsupervised learning of visual features by contrasting cluster assignments. In: Neural information processing systems
21. Li Z, Tang J (2016) Weakly supervised deep matrix factorization for social image understanding. *IEEE Trans Image Process* 26(1):276–288
22. Li Z, Liu J, Tang J, Lu H (2015) Robust structured subspace learning for data representation. *IEEE Trans Pattern Anal Mach Intell* 37(10):2085–2098
23. Li Z, Tang J, Mei T (2019) Deep collaborative embedding for social image understanding. *IEEE Trans Pattern Anal Mach Intell* 41(9):2070–2083
24. Li Z, Tang J, Zhang L, Yang J (2020) Weakly-supervised semantic guided hashing for social image retrieval. *Int J Comput Vision* 128:2265–2278
25. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp 4171–4186
26. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
28. Liu XY, Wu J, Zhou ZH (2008) Exploratory undersampling for class-imbalance learning. *IEEE Trans Syst Man Cyber B* 39(2):539–550
29. Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *JMLR* 18(17):1–5
30. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: CVPR, pp 9726–9735
31. Xie J, Girshick RB, Farhadi A (2016) Unsupervised deep embedding for clustering analysis. In: Balcan M, Weinberger KQ (eds) ICML, pp 478–487
32. Hu Y, Zheng L, Yang Y, Huang Y (2018) Twitter100k: a real-world dataset for weakly supervised cross-media retrieval. *IEEE TMM* 20(4):927–938
33. Radford A, Kim J.W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: ICML, pp 8748–8763
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: CVPR, pp 1–9
35. Hessel J, Lee L (2020) Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In: EMNLP, pp 861–877
36. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML, pp 6105–6114
37. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized BERT pretraining approach. *arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692)*
38. Tan H, Bansal M (2019) LXMERT: Learning cross-modality encoder representations from transformers. In: EMNLP, pp. 5100–5111
39. Fu J, Xu S, Liu H, Liu Y, Xie N, Wang CC, Liu J, Sun Y, Wang B (2022) Cma-clip: Cross-modality attention clip for text-image classification. In: 2022 IEEE international conference on image processing (ICIP), pp 2846–2850
40. Kingma D.P, Ba J (2015) Adam: A method for stochastic optimization. In: ICLR
41. MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth berkeley symposium on mathematical statistics and probability, pp 281–297
42. Bishop CM (2007) Pattern recognition and machine learning, 5th Edition. In: Information science and statistics
43. Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, pp 226–231
44. Von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17:395–416
45. Schwartz H.A, Giorgi S, Sap M, Crutchley P, Eichstaedt J, Ungar L (2017) Dlatk: differential language analysis toolkit. In: EMNLP, pp 55–60

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.