Relational Distant Supervision for Image Captioning without Image-Text Pairs

Yayun Qi[1], Wentian Zhao[1], Xinxiao Wu[1,2*]

[1]Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China

[2] Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China

{qiyayun,wentian_zhao,wuxinxiao}@bit.edu.cn

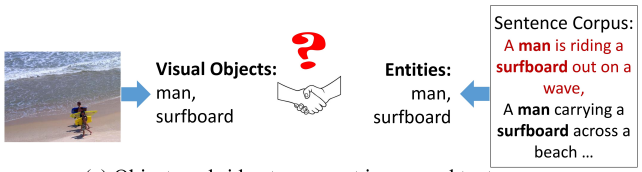
Abstract

Unsupervised image captioning aims to generate descriptions of images without relying on any image-sentence pairs for training. Most existing works use detected visual objects or concepts as bridge to connect images and texts. Considering that the relationship between objects carries more information, we use the object relationship as a more accurate connection between images and texts. In this paper, we adapt the idea of distant supervision that extracts the knowledge about object relationships from an external corpus and imparts them to images to facilitate inferring visual object relationships, without introducing any extra pre-trained relationship detectors. Based on these learned informative relationships, we construct pseudo image-sentence pairs for captioning model training. Specifically, our method consists of three modules: (i) a relationship learning module that learns to infer relationships from images under the distant supervision; (ii) a relationship-to-sentence module that transforms the inferred relationships into sentences to generate pseudo image-sentence pairs; (iii) an image captioning module that is trained by using the generated image-sentence pairs. Promising results on three datasets show that our method outperforms the state-of-the-art methods of unsupervised image captioning.

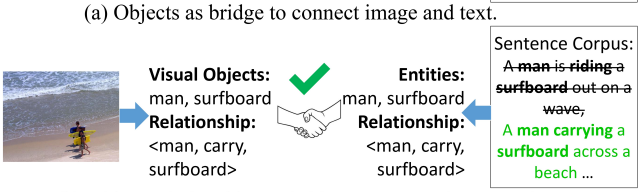
Introduction

Unsupervised image captioning has aroused growing interest from researchers in recent years, as it does not require large-scale high-quality paired images and sentences for training. In existing settings (Feng et al. 2019; Laina, Rupprecht, and Navab 2019; Guo et al. 2020; Meng et al. 2022), only image data, sentence corpus, and an off-the-shelf object detector are needed to train a captioning model. The sentence corpus usually comes from external resources, e.g., sentences crawled from real-world websites (Feng et al. 2019).

The main challenge of unsupervised image captioning task lies in how to build the connection between images and sentences from the external corpus. To address this challenge, with the help of object detector, previous methods regard detected visual objects and extracted entities as shared elements to connect images and sentences as illustrated in Figure 1(a), such as guiding the reconstruction between two modalities (Feng et al. 2019) or constraining the learned manifolds (Laina, Rupprecht, and Navab 2019). Several other methods use the detected objects as representations of visual content, in the form of labels (Guo et al. 2020) or regions (Meng et al. 2022).



(a) Objects as bridge to connect image and text.



(b) Object relationships as bridge to connect image and text.

Figure 1: Illustration of using objects (a) and their relationships (b) as the bridge to connect images and texts in the unsupervised image captioning task.

Since objects are not independent and their relationships carry more important contextual information, we address the unsupervised image captioning task by introducing object relationships to bridge images and sentences, as illustrated in Figure 1(b). An intuitive solution is to apply an off-the-shelf relationship detector to detect the object relationships from images. However, this solution suffers from a semantic gap between the relationships detected from images and those expressed in sentences. For example, many possessive or geometric relationships between objects, like ⟨dog, have, head⟩ and ⟨man, wear, shoes⟩, are easily detected from images, but they are seldom described in sentences.

In this paper, we exploit knowledge about object relationships extracted from external corpus to infer visual object relationships in images. To enable the knowledge imparting from texts to images, we propose a novel relational distant supervision method, which first infers visual object relationships through scene-level alignment between images

*Corresponding author.

and sentences, and then converts the inferred relationships into sentences to generate pseudo image-sentence pairs to train an image captioning model. The relationships inferred by our method are more similar to those described in sentences than those detected by off-the-shelf detectors. This brings the benefit of narrowing the semantic gap between relationships in images and texts, resulting in generating better captions.

Specifically, our method consists of a relationship learning module, a relationship-to-sentence module, and an image captioning module. The relationship learning module infers object relationships with images and object labels as input, whose training data are image-relationship pairs provided by distant supervision. The distant supervision is built on the assumption that if a sentence contains the same entities as an image, then the sentence and the image describe the same scene. With this assumption, we align images with sentences from the corpus that depict the same scene, and the entity relationships parsed from the paired sentences are treated as the relationship labels in images. The relationship-to-sentence module converts the object relationships inferred by the relationship learning module into sentences, and is trained by reconstructing sentences from the parsed entity relationships in the original sentences. The converted sentences by the relationship-to-sentence module and the corresponding images are served as pseudo image-sentence pairs. The image captioning module generates captions to describe a given image, and is trained by using the pseudo image-sentence pairs generated by the relationship-to-sentence module.

The main contributions of this paper are three-fold:

- We address the unsupervised image captioning task by exploiting the knowledge of object relationships extracted from the external corpus to connect images and texts.
- We propose a relational distant supervision method that infers visual object relationships by imparting the object relationship knowledge from texts to images, without introducing any extra relationship detectors.
- Extensive experiments on three datasets demonstrate that our method achieves better results than the state-of-the-art methods of unsupervised image captioning.

Related Work

Unsupervised Image Captioning

Unsupervised image captioning has attracted increasing attention in recent years, since it alleviates the heavy reliance on large-scale high-quality image-sentence pairs. This task is first proposed by Feng et al. (2019), where a GAN-based bidirectional reconstruction model is designed to learn the correlation between images and texts. The detected objects are used to initialize the model and calculate a special object-based reward. Laina, Rupprecht, and Navab (2019) map images and sentences into a shared manifold and decode captions from embeddings, where the objects are used as the anchors to find an accurate space.

In addition to using objects as rewards or constraints in this task, there have been other attempts to use them directly as visual representations. Guo et al. (2020) propose a concepts-to-sentence memory translator that takes object labels detected from images as the visual information. Meng et al. (2022) collect a set of object regions from images and train a Transformer based model to generate captions based on the input regions.

Different from these methods that use objects as the bridge to connect images and texts, we impart the knowledge of object relationships from texts to images, and then use those relationships as more accurate connections between images and sentences.

Distant Supervision

Distant supervision is a paradigm that uses external knowledge bases to provide labels for unlabeled datasets. It is introduced by Mintz et al. (2009) into the relation extraction task in the field of natural language processing, which aligns a given knowledge base with text to train a relation extractor. The alignment process is guided by the assumption:"If two entities participate in a relation, then all sentences mentioning those two entities express that relation". There have emerged plenty of works following this thought and achieving satisfying results without labeled data for their specific tasks, such as relation extraction (Mintz et al. 2009; Zeng et al. 2015; Ji et al. 2017), named entity recognition (Meng et al. 2021; Shang et al. 2018; Liang et al. 2020; Hedderich, Lange, and Klakow 2021; Peng et al. 2019), part-of-speech tagging task (Fang and Cohn 2016; Plank and Agić 2018), and sentiment classification (Go, Bhayani, and Huang 2009; Purver and Battersby 2012; Sahni et al. 2017).

In recent years, the idea of distant supervision has been successfully applied to computer vision tasks. Yao et al. (2021) propose visual distant supervision to train a scene graph generator without labeled data. In this paper, we adapt the idea of distant supervision to unsupervised image captioning, by using sentence corpus as the knowledge base to provide useful information for images through scene-level alignment, so as to establish the connection between images and texts to train the image captioning model.

Our Method

Our relational distant supervision method includes three modules: a relationship learning module, a relationship-to-sentence module, and an image captioning module. The relationship learning module infers object relationships from images and object labels, and is trained under relational distant supervision. The relational distant supervision is represented by the image-relationship pairs generated by aligning images with extracted knowledge about entities and relationships from sentences. The relationship-to-sentence module converts the input object relationships into sentences, and is trained by reconstructing sentences from the parsed entity relationships of the original sentences. The image captioning module generates captions of input images, and is trained using pseudo image-sentence pairs. The pseudo image-sentence pairs are collected by first inferring relationships in images through the relationship learning module and then converting them into sentences through the relationship-to-sentence module. The overview of our method is illustrated

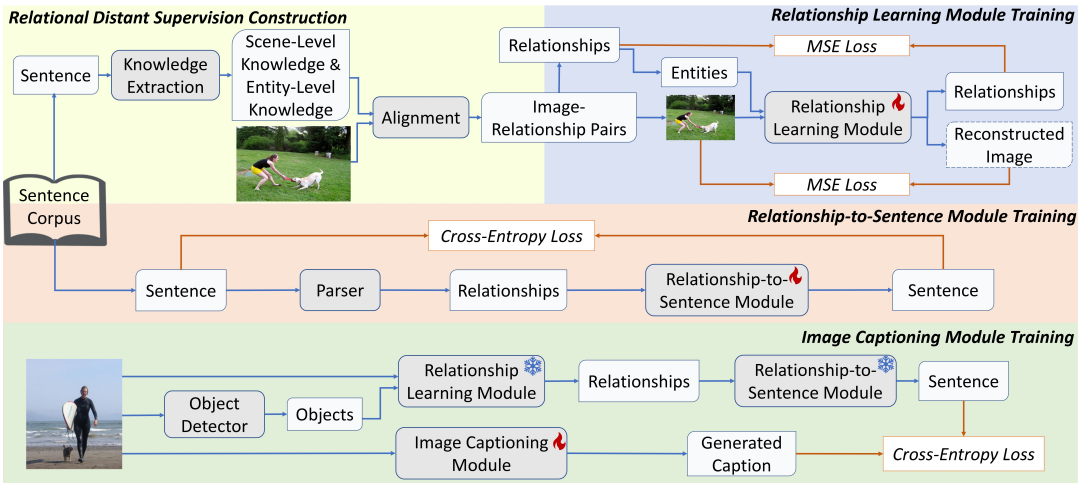


Figure 2: Overview of our method. The relationship learning module is trained on image-relationship pairs using MSE losses, where the data pairs come from the relational distant supervision constructed by an alignment using extracted scene-level and entity-level knowledge. The relationship-to-sentence module is trained by reconstructing sentences from the parsed entity relationships in the original sentences with a cross-entropy loss. The image captioning module is trained on image-sentence pairs using a cross-entropy loss, where the data pairs are generated by first inferring visual object relationships from images using the relationship learning module and then converting them into sentences using the relationship-to-sentence module.

in Figure 2, where the training processes of three modules are illustrated with different background colors.

Relationship Learning Module

We address the unsupervised image captioning task by introducing object relationships as informative connections between images and sentences, where we impart the knowledge about object relationships from text to image in order to infer relationships in images. This process is achieved by a relationship learning module, whose training data comes from the constructed relational distant supervision. We will introduce the training data collection and the training procedure for this module in the following sections.

Relational Distant Supervision Construction. The main challenge of connecting images and sentences using object relationships lies in inferring visual object relationships that have a small semantic gap to entity relationships in the textual modality. To tackle this issue, we exploit object relationship knowledge carried by the sentence corpus at two levels, *e.g.* scene-level and entity-level, and impart it from texts to images to establish connections between images and sentences. Specifically, the scene-level knowledge is used for constructing relational distant supervision to provide relationship labels for images, which is achieved by a scene-level alignment. The entity-level knowledge is used as the commonsense knowledge to remove unreasonable relationships. Since the size of the corpus is usually in millions, almost all possible scenes or entity relationships are included in the corpus.

First, we use the Stanford CoreNLP toolkit (Manning et al. 2014) and the NLTK toolkit (Bird, Klein, and Loper 2009) to parse entity relationships and extract entities from sentences. Then, for the scene-level knowledge, we consider the content of sentences as a whole and collect it in the form of tuples $< \mathcal{E}, \mathcal{S} >$, where $\mathcal{E}$ represents the entities extracted from sentences and $\mathcal{S}$ represents sentences containing these entities. These tuples capture textual descriptions of scenes that contain the same entities, and implicitly embody the relationships in sentences that are more likely to be used to describe the scenes. For the entity-level knowledge, we collect it from the parsing results by considering a pair of entities as a whole, and integrating the relationships depicted by these two entities in all sentences as their relationship candidate set. These sets provide instructive knowledge about reasonably existing relationships between two entities, enabling noise mitigation.

Based on the scene-level knowledge, we employ the idea of distant supervision to generate image-relationship pairs in order to obtain visual relationships similar to those in the textual modality. Specifically, instead of directly applying the original assumption of distant supervision (Mintz et al. 2009), we propose a new assumption that is more suitable for our purpose: "If a sentence contains the same entities as an image, then they describe the same scene. The more entities in the scene, the higher the confidence of this scene". This assumption takes all entities/objects in a scene as a whole into account for the relationships they describe. Because for the image captioning task, a generated sentence describes the visual content of an entire image, not just a single relationship between objects.

On this basis, we align images with sentences satisfying the above assumption according to the detected objects and extracted entities. These aligned image-sentence pairs serve as training data for an image captioning module to comprehensively explore the alignment results for the entire image set, where we use an off-the-shelf method (Anderson et al. 2018) as the captioning module. The mod-

ule then infers the descriptions of images, from which the relationships are parsed by the NLP toolkits as the labels for the corresponding images. Since these relationship labels are parsed from sentences that have a small semantic gap to textual modality, they achieve the purpose of imparting object relationship knowledge from texts to images. The relational distant supervision is represented by $(I_i, \langle e_1, p_1, e_2\rangle, \cdots, \langle e_m, p_j, e_n\rangle, \cdots)$, where I_i indicates the image, e_m, e_n and p_j indicate the entity as subject, the entity as object and the predicate from parsed entity relationships, respectively. Meanwhile, to mitigate the impact of noisy labels, we only keep inferred captions including the detected objects, and also remove the parsed relationships not covered by the entity-level knowledge.

Relationship Learning Module Training. The relationship learning module is responsible for inferring object relationships in images, and its training data comes from the relational distant supervision. It takes the image feature $\boldsymbol{f}_i$ of image I_i and the entities $\mathcal{E}_i$ in the paired relationship labels as input, and infers relationships between any entities. Specifically, this module infers the embeddings of predicates in relationships instead of predicting certain predicate labels, which makes the learning procedure easier and more stable for subsequent procedures. To handle the case where no relationship exists between two entities, we additionally add a "none" predicate into the categories. The training procedure is as follows.

First, we take the embedding vectors of Glove (Pennington, Socher, and Manning 2014) as the word embeddings for the entities and the predicates, denoted as $\boldsymbol{E}_e$ and $\boldsymbol{E}_p$, respectively. Then, we organize the entities into pairs and represent concatenated embeddings of the entity pairs as $\boldsymbol{E}_{bina}$. The formulation of the learning procedure is given by

$$\boldsymbol{E}'_p = \mathrm{MLP}([\boldsymbol{f}_i, \overline{\boldsymbol{E}}_e, \boldsymbol{E}_{bina}]), \tag{1}$$

where $\boldsymbol{E}'_p$ represents the inferred predicate embeddings, $[\cdot]$ denotes the concatenate operation, $\overline{\boldsymbol{E}}_e$ denotes the mean of $\boldsymbol{E}_e$, and $\mathrm{MLP}(\cdot)$ denotes a multi-layer perceptron. Afterward, reconstructed image features are computed from the inferred relationship embeddings by a Transformer encoder, which constrains the inferred relationships corresponding to the original visual information. The encoder has multiple encoder layers stacked together, with the same structure as a standard Transformer (Vaswani et al. 2017). The formulation of this reconstruction is given by

$$\boldsymbol{H}_i = \begin{cases} \mathrm{EncoderLayer}_i([\boldsymbol{E}_a, \boldsymbol{E}_r]), i = 0, \\ \mathrm{EncoderLayer}_i(\boldsymbol{H}_{i-1}), \quad i > 0, \end{cases} \tag{2}$$

$$\boldsymbol{f}'_i = \mathrm{FC}(\boldsymbol{E}'_a), \tag{3}$$

where $\boldsymbol{E}_r = [\boldsymbol{E}_{bina}, \boldsymbol{E}'_p]$ indicates the concatenated embeddings of the entity pairs and inferred predicates. $\boldsymbol{E}_a$ indicates additionally added features initialized with zero, which has the same dimension as $\boldsymbol{E}_r$. $\boldsymbol{E}'_a$ is the updated features of $\boldsymbol{E}_a$ through N_{rec} encoder layers, which is extracted from $\boldsymbol{H}_{N_{rec}}$ according to the initial location of $\boldsymbol{E}_a$ in $[\boldsymbol{E}_a, \boldsymbol{E}_r]$. $\mathrm{FC}(\cdot)$ denotes a fully connected layer, which projects $\boldsymbol{E}'_a$ into the reconstructed image feature $\boldsymbol{f}'_i$.

The relationship learning module is trained by an MSE loss between inferred predicate embeddings and the ground-truth, and an MSE loss between reconstructed image features and the original image features in N images, formulated as

$$\mathcal{L}_{rel} = \frac{1}{N}\sum_{i=1}^{N}((\boldsymbol{E}^i_p - \boldsymbol{E}^{i\prime}_p)^2 + (\boldsymbol{f}_i - \boldsymbol{f}'_i)^2). \tag{4}$$

Relationship-to-Sentence Module

The relationship-to-sentence module is responsible for generating sentences to describe the input relationships, which first updates the embeddings of relationships through a Graph Convolutional Network (GCN) (Johnson, Gupta, and Fei-Fei 2018) and then decodes them into word sequences via a Transformer. It is trained on the sentence corpus by reconstructing sentences from their corresponding parsed entity relationships.

First, the "none" predicate is added to the parsing result for entity pairs that have no relationship. Then, Glove vectors are used to embed the entities and predicates, *i.e.* $\boldsymbol{E}_e, \boldsymbol{E}_p$, similar to those in the relationship learning module. The information propagation through these relationships is given by

$$\begin{aligned} \boldsymbol{v}_i, \boldsymbol{v}_p, \boldsymbol{v}_j &= g_p(\boldsymbol{E}^i_e, \boldsymbol{E}_p, \boldsymbol{E}^j_e), \\ \boldsymbol{v}^s_i &= g_s(\boldsymbol{v}_i, \boldsymbol{v}_p, \boldsymbol{v}_j), \\ \boldsymbol{v}^o_i &= g_o(\boldsymbol{v}_j, \boldsymbol{v}_p, \boldsymbol{v}_i), \\ \boldsymbol{v}'_i &= h(\boldsymbol{V}^s_i, \boldsymbol{V}^o_i), \end{aligned} \tag{5}$$

where $g_p(\cdot)$, $g_s(\cdot)$, $g_o(\cdot)$ are responsible for the calculation of predicates, entities as subjects, and entities as objects, respectively. All of these three functions are MLP-based functions that generate updated features of the input embeddings. The propagated feature $\boldsymbol{v}'_i$ of the entity is the pooled result of $\boldsymbol{v}^s_i$ and $\boldsymbol{v}^o_i$, where $h(\cdot)$ is the pooling function. $\boldsymbol{V}^s_i$ and $\boldsymbol{V}^o_i$ are the sets of $\boldsymbol{v}^s_i$ and $\boldsymbol{v}^o_i$, respectively.

After the information propagation, the updated features for predicates and entities are re-organized as a set of relationship features $\boldsymbol{V}_r$ by collecting all concatenated features of each entity pair $(\boldsymbol{v}'_i, \boldsymbol{v}'_j)$ and its predicate $\boldsymbol{v}_p$, *i.e.* $[\boldsymbol{v}'_i, \boldsymbol{v}_p, \boldsymbol{v}'_j]$. Then, $\boldsymbol{V}_r$ is used as the representation of the input relationships and further decoded by an encoder-decoder structured Transformer. The calculation of the encoder is similar to Eq.(2), where only the input data in the case of $i = 0$ becomes $[\overline{\boldsymbol{V}}_r, \boldsymbol{V}_r]$. $\overline{\boldsymbol{V}}_r$ represents the average pooling relationship features. The calculation of the decoder follows the standard Transformer, with the final output of the encoder and the embeddings of the word sequence at each time-step as input. N_{enc} encoder layers and N_{dec} decoder layers are stacked, where a subsequent fully connected network and softmax operation further translate the final output from the decoder to the probability distribution over words.

The relationship-to-sentence module is trained by a standard cross-entropy loss calculated between the ground-truth sentences and the generated ones:

$$\mathcal{L}_{XE} = -\frac{1}{T}\sum_{t=1}^{T}\log p_\theta(\overline{y}^*_t | y^*_{1:t-1}). \tag{6}$$

Dataset	Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
COCO-Shutterstock	UIC (Feng et al. 2019)	41.0	22.5	11.2	5.6	12.4	28.7	28.6
	R^2M (Guo et al. 2020)	44.0	25.4	12.7	6.4	13.0	31.3	29.0
	OCM (Meng et al. 2022)	30.7	14.1	5.9	2.5	13.4	24.1	15.0
	Fast RF-UIC (Yang et al. 2023)	43.3	24.2	11.7	6.1	12.5	30.2	25.9
	Ours	**51.1**	**28.9**	**15.4**	**8.0**	**16.2**	**33.7**	**38.5**
Flickr30K-COCO	SME (Laina, Rupprecht, and Navab 2019)	—	—	—	7.9	13.0	32.8	9.9
	R^2M (Guo et al. 2020)	53.1	**32.8**	**19.2**	**11.7**	13.7	35.9	18.1
	OCM (Meng et al. 2022)	45.1	26.9	15.2	9.0	13.3	34.1	11.0
	Ours	**53.7**	32.3	18.1	10.5	**13.8**	**36.4**	**19.2**
COCO-GCC	SME (Laina, Rupprecht, and Navab 2019)	—	—	—	6.5	12.9	35.1	22.7
	R^2M (Guo et al. 2020)	51.2	29.5	15.4	8.3	14.0	35.0	29.3
	OCM (Meng et al. 2022)	42.0	23.3	11.4	5.8	14.9	33.6	17.7
	Ours	**53.3**	**31.7**	**17.8**	**10.1**	**15.3**	**36.8**	**36.3**

Table 1: Comparison results with the state-of-the-art unsupervised image captioning methods on three datasets.

Image Captioning Module

The image captioning module is responsible for describing the visual content of an image with a sentence, which takes images as input and decodes word sequences. In this work, we employ an off-the-shelf image captioning model BUTD (Anderson et al. 2018) to achieve this functionality.

At this point, the relationship learning module and the relationship-to-sentence module have been trained and their parameters are frozen. Given an image, the relationship learning module first uses the knowledge learned from the relational distant supervision to infer visual object relationships, whose input data are images and detected object labels. It then concatenates the inferred predicate embeddings with the Glove embeddings of the detected object labels to form complete inferred relationship embeddings. Afterward, the relationship-to-sentence module converts the inferred visual object relationship embeddings into sentences. Subsequently, the generated sentences are combined with the input images to generate pseudo image-sentence pairs. Finally, the image captioning module is trained using a standard cross-entropy loss based on these pseudo image-sentence pairs and is able to generate a description for the input image.

Discussion of the Combination with LLMs

In our method, the relational distant supervision is provided by knowledge in the external corpus. With the rise of large-scale pre-training paradigm, more and more large language models (LLMs) can be regarded as powerful black-box language knowledge bases. Therefore, our method also has the potential to be combined with LLMs, from which the relational distant supervision can be obtained by constructing prompts using objects in a scene. LLMs generate descriptions of certain scenes based on their knowledge learned during pre-training. In this way, our unsupervised image captioning method can further remove the reliance on the external corpus, allowing it to be extended to an image-only training setting. In addition, the language styles of captions are no longer limited to the styles of the corpus, but can be flexibly obtained from the LLMs by tuning the prompts.

Experiments

Datasets

To evaluate the effectiveness of our method, we conduct experiments on three different datasets: (1) COCO-Shutterstock with images from COCO (Lin et al. 2014) and sentence corpus from Shutterstock (Feng et al. 2019); (2) Flickr30K-COCO with images from Flickr30K (Plummer et al. 2015) and sentence corpus from COCO; (3) COCO-GCC with images from COCO and sentence corpus from GCC (Sharma et al. 2018).

Implementation Details

The visual features are extracted by ResNet-101 (He et al. 2016). A Faster R-CNN detector (Huang et al. 2017) trained on Visual Genome (Krishna et al. 2017) is applied to detect objects in images. Note that we only use the predicted object category labels and do not rely on the detected regions to maintain the minimal supervision. We remove the object labels that occur less than 500 times in the corpus.

In the relationship learning module, the hidden layer dimensions of MLP and the Transformer encoder are set to 1024, and the layer number N_{rec} of the Transformer encoder is set to 2. In the relationship-to-sentence module, the hidden layer dimensions of GCN and Transformer are set to 512. The layer number N_{enc} and N_{dec} of the Transformer encoder and decoder are both set to 6. In the image captioning module, the word embedding dimension and the hidden layer dimension are both set to 512. The Adam optimizer (Kingma and Ba 2014) is adopted. The learning rates are set to 5×10^{-5}, 1×10^{-4} and 1×10^{-4} for the relationship learning module, the relationship-to-sentence module, and the image captioning module, respectively.

The widely used evaluation metrics, *i.e.*, BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), are used for performance evaluation.

Comparison with State-of-the-Art Methods

We compare our method with several state-of-the-art unsupervised image captioning methods, including UIC (Feng

Dataset	Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
COCO-Shutterstock	w/o relationship learning	44.6	23.8	11.8	5.8	12.9	31.2	27.2
	w/o scene-level knowledge	47.8	26.1	13.5	6.8	14.6	31.2	32.0
	w/o entity-level knowledge	50.7	28.4	15.0	7.7	15.8	33.4	37.3
	w/o reconstruction loss	49.9	27.8	14.5	7.5	16.1	33.2	37.7
	Ours	**51.1**	**28.9**	**15.4**	**8.0**	**16.2**	**33.7**	**38.5**
Flickr30K-COCO	w/o relationship learning	51.2	29.7	16.0	9.1	12.9	34.3	14.9
	w/o scene-level knowledge	50.0	27.9	14.2	7.3	12.6	33.9	15.7
	w/o entity-level knowledge	**53.9**	32.1	17.8	10.3	13.6	36.0	18.2
	w/o reconstruction loss	53.5	32.1	17.7	10.1	**13.8**	36.2	18.6
	Ours	53.7	**32.3**	**18.1**	**10.5**	**13.8**	**36.4**	**19.2**
COCO-GCC	w/o relationship learning	47.2	27.6	14.4	7.6	13.2	35.7	27.0
	w/o scene-level knowledge	50.7	29.8	16.3	8.9	14.3	35.3	31.4
	w/o entity-level knowledge	**53.7**	31.6	17.4	9.5	15.2	36.6	35.9
	w/o reconstruction loss	**53.7**	**31.7**	17.3	9.4	15.2	36.5	35.6
	Ours	53.3	**31.7**	**17.8**	**10.1**	**15.3**	**36.8**	**36.3**

Table 2: Results of ablation studies in three datasets.

Dataset	Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
COCO-Shutterstock	relationship detector	43.3	23.0	11.0	5.4	14.6	32.1	24.1
	Ours	**51.1**	**28.9**	**15.4**	**8.0**	**16.2**	**33.7**	**38.5**
Flickr30K-COCO	relationship detector	43.5	23.8	11.4	5.9	12.7	32.1	11.6
	Ours	**53.7**	**32.3**	**18.1**	**10.5**	**13.8**	**36.4**	**19.2**
COCO-GCC	relationship detector	43.7	23.6	10.9	5.3	13.7	32.6	21.7
	Ours	**53.3**	**31.7**	**17.8**	**10.1**	**15.3**	**36.8**	**36.3**

Table 3: Comparison results between applying the relationship detector and our method to infer the visual object relationships.

et al. 2019), SME (Laina, Rupprecht, and Navab 2019), R^2M (Guo et al. 2020), OCM (Meng et al. 2022) and Fast RF-UIC (Yang et al. 2023). For UIC, SME, R^2M and Fast RF-UIC, we directly report the results from their original papers. For OCM, we use their official code to evaluate it with the same settings as other methods, where the training procedure remains the same as the original OCM method.

The comparison results are shown in Table 1. It is obvious that our method achieves the best performance on all evaluation metrics on COCO-Shutterstock and COCO-GCC, and comparable performance on Flickr30k-COCO. It is also interesting that when the corpus is collected by crawling from websites (*i.e.*, GCC and Shutterstock), our method outperforms existing methods by a large margin. For example, our method improves the CIDEr scores by 32.8% on COCO-Shutterstock. These promising results demonstrate the effectiveness of our method, especially the superiority on exploiting the relationship knowledge carried by the corpus.

Ablation Study

To further evaluate the effectiveness of individual components, we introduce four variants of our method for comparison: (1) **w/o relationship learning**: we remove the relationship learning module and the relationship-to-sentence module, and only keep the process of aligning images with sentences for distant supervision construction. This variant shows the performance of purely applying distant supervision, in order to evaluate the effectiveness of subsequent relationship learning; (2) **w/o scene-level knowledge**: we remove the scene-level knowledge provided by the distant supervision, in order to evaluate the effectiveness of bridging images and texts via relationships. Specifically, we no longer take into account the relationships between objects or entities, where both the inferred predicates for visual objects and the parsed predicates between entities are set to "none"; (3) **w/o entity-level knowledge**: during the construction of distant supervision data, we remove the suggested relationship candidates provided by the entity-level knowledge to evaluate its effectiveness; (4) **w/o reconstruction loss**: we remove the reconstruction loss in Eq. (4) to evaluate its effectiveness.

The results of ablation studies are shown in Table 2. We have the following observations: (1) Our method substantially outperforms "w/o relationship learning", which indicates that the following relationship learning process is essential after the construction of distant supervision. For example, a gain of 9.3 on CIDEr is achieved on COCO-GCC. (2) With the removal of scene-level knowledge for relationship learning, the performance of all metrics drops significantly. This demonstrates the importance of considering relationships in images and sentences as a bridge connecting them and the effectiveness of the relational distant supervision in our method. (3) Likewise, both the removal of entity-level knowledge and reconstruction loss degrade the performance, which verifies their effectiveness.

Figure 3: Example captions generated by our method and its two variants. Ground-truth captions of images are shown in "GT".

Method		
w/o scene-level knowledge	**Objects:** snow, skis, helmet, man,... grass **Sentence:** A man with a helmet of skis walks on the snow covered ground.	**Objects:** beach, ocean, man,..., surfboard **Sentence:** Man with surfboard on beach near ocean and ocean.
relationship detector	**Relationships:** <hand, hold, wire>, <man, has, head >, ... ,<man, has, leg> **Sentence:** A man holding up a stick to the other one of the other as he holds a line through the.	**Relationships:** <man, has, head>, <man, watch, man >, <man, ride, wave>, <man, carry, surfboard>, ..., <man, walk on, beach> **Sentence:** a man walks on the beach and rides a surfboard under his arm.
Ours	**Sentence:** Man in helmet riding on snow on winter day on grass with snowboard.	**Sentence:** Man holding a surfboard on the beach with blue sky and ocean background.

Figure 4: Examples of pseudo image-sentence pairs generated by our method and its two variants.

Comparison with Relationship Detector

To evaluate the negative impact caused by the semantic gap between the relationships detected by off-the-shelf detectors from images and those described in sentences, we conduct more experiments by replacing our relationship learning module with an off-the-shelf relationship detector.

Specifically, we apply the unbiased scene graph generator (Tang et al. 2020) based on Neural motifs (Zellers et al. 2018) that is trained on Visual Genome to detect object relationships in images. Then, we directly take these detected relationships as the input of the relationship-to-sentence module when generating the training data for the image captioning module, while keeping other procedures unchanged. We denote this variant **"relationship detector"**. It can be observed from Table 3 that directly using visual relationships detected by an off-the-shelf relationship detector as modality connections do not perform well, which shows the benefit of inferring visual object relationships by imparting the object relationship knowledge from texts to images.

Qualitative Results

To further demonstrate the effectiveness of the proposed relational distant supervision, we show several examples of pseudo image-sentence pairs generated by our method and other two variants ("w/o scene-level knowledge" and "relationship detector") in Figure 4. It is interesting to observe that: (1) Compared with "w/o scene-level knowledge" that ignores object relationships, our method is able to generate pseudo sentences that more accurately describe images, like "man riding on snow ... with snowboard" rather than "man walks on the snow". This verifies the importance of introducing object relationships to bridge images and sentences. (2) Compared with "relationship detector", our method generates pseudo-sentences with higher relevance of visual content, such as "man holding a surfboard on the beach" rather than "man rides a surfboard". (3) There are many possessive relationships in the detection results, like ⟨man, has, leg⟩. These relationships are obvious and scarcely described in sentences, which leads to poor performance of "relationship detector" in generating pseudo sentences for images. These results validate the efficacy of imparting knowledge from external corpus to infer visual object relationships.

Figure 3 shows more examples of captions generated by our method and other two variants. Compared to "relationship detector", our method generates more fluent captions, *e.g.* "dog standing in the grass with a frisbee" instead of "dog behind a tree" in the first example. Compared with "w/o scene-level knowledge", our method generates more accurate captions, *e.g.* "riding on skateboard" rather than "sitting on a wooden fence" in the second example.

Conclusion

We have presented a relational distant supervision method for unsupervised image captioning, which successfully relieves the dependency on large-scale image-sentence data for training. By imparting the extracted knowledge of object relationships from the external corpus to images, our method can infer the visual object relationships in images without any extra relationship detectors, and thus succeeds in constructing pseudo image-sentence pairs to train an image captioning model. Extensive experiments on three datasets demonstrate that our method achieves the best results compared with the state-of-the-art methods. In the future, we are going to extend the proposed relational distant supervision to the image-only setting with the combination of LLMs, and also apply it to unsupervised video captioning with more complex relationships.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No 62072041.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (ACLW)*, 65–72.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Fang, M.; and Cohn, T. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. *arXiv*, abs/1607.01133.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4125–4134.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12).

Guo, D.; Wang, Y.; Song, P.; and Wang, M. 2020. Recurrent relational memory network for unsupervised image captioning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 920–926.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hedderich, M. A.; Lange, L.; and Klakow, D. 2021. ANEA: Distant supervision for low-resource named entity recognition. *arXiv*, abs/2102.13129.

Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; and Murphy, K. 2017. Speed/Accuracy trade-Offs for modern convolutional object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3296–3297.

Ji, G.; Liu, K.; He, S.; and Zhao, J. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 3060–3066.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1219–1228.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv*, abs/1412.6980.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision (IJCV)*, 123: 32–73.

Laina, I.; Rupprecht, C.; and Navab, N. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7414–7424.

Liang, C.; Yu, Y.; Jiang, H.; Er, S.; Wang, R.; Zhao, T.; and Zhang, C. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery & data mining (KDD)*, 1054–1064.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 55–60.

Meng, Y.; Zhang, Y.; Huang, J.; Wang, X.; Zhang, Y.; Ji, H.; and Han, J. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. *arXiv*, abs/2109.05003.

Meng, Z.; Yang, D.; Cao, X.; Shah, A.; and Lim, S.-N. 2022. Object-Centric unsupervised image captioning. In *European Conference on Computer Vision (ECCV)*, 219–235.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJNLP)*, 1003–1011.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Peng, M.; Xing, X.; Zhang, Q.; Fu, J.; and Huang, X. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. *arXiv*, abs/1906.01378.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Plank, B.; and Agić, Ž. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. *arXiv*, abs/1808.09733.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer

image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2641–2649.

Purver, M.; and Battersby, S. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 482–491.

Sahni, T.; Chandak, C.; Chedeti, N. R.; and Singh, M. 2017. Efficient Twitter sentiment classification using subjective distant supervision. In *International Conference on Communication Systems and Networks (COMSNETS)*, 548–553.

Shang, J.; Liu, L.; Ren, X.; Gu, X.; Ren, T.; and Han, J. 2018. Learning named entity tagger using domain-specific dictionary. *arXiv*, abs/1809.03599.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2556–2565.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3716–3725.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

Yang, R.; Cui, X.; Qin, Q.; Deng, Z.; Lan, R.; and Luo, X. 2023. Fast RF-UIC: A fast unsupervised image captioning model. *Displays*, 79: 102490.

Yao, Y.; Zhang, A.; Han, X.; Li, M.; Weber, C.; Liu, Z.; Wermter, S.; and Sun, M. 2021. Visual distant supervision for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15816–15826.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5831–5840.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1753–1762.