Learning Cross-modal Representations with Multi-relations for Image Captioning

Peng Cheng, Tung Le[a], Teeradaj Racharak[b], Cao Yiming, Kong Weikun and Minh Le Nguyen[c]
School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan

Keywords: Cross-modality, Multi-relational Semantics, Object Semantics, Vision-and-Language, Pre-training.

Abstract: Image captioning is a cross-domain study that generates image description sentences based on a given image. Recently, (Li et al., 2020b) shows that concatenating sentences, object tags, and region features as a unified representation enables to overcome state-of-the-art works in different vision-and-language-related tasks. Such results have inspired us to investigate and propose two new learning methods that exploit the relation representation in the model and improve the model's generation results in this paper. To the best of our knowledge, we are the first that exploit both relations extracted from text and images for image captioning. Our idea is motivated by the phenomenon that humans can correct other people's descriptions by knowing the relationship between objects in an image while observing the same image. We conduct experiments based on the MS COCO dataset (Lin et al., 2014) and show that our method can yield the higher SPICE score than the baseline.

1 INTRODUCTION

Image captioning is the task of generating a natural language sentence describing the content of an image, placing itself in the intersection of two important AI areas: natural language processing (NLP) and computer vision (CV). As NLP and CV are highly active in research and development, especially various available pre-trained models, recent advancement on image captioning can be highly expected from the results of these two fields. Apropos to the CV side, recent state-of-the-art object detection architectures, e.g., Faster R-CNN (Ren et al., 2015), have shown to improve the performance of image captioning. Similarly, sophisticated Transformer-based architectures in NLP have shown to enhance caption's generation (Vaswani et al., 2017).

Inspired by neural machine translation, most image captioning systems (cf. (Vinyals et al., 2015; Anderson et al., 2018)) employ an encoder-decoder architecture, i.e., an input image is encoded into an intermediate representation of the information contained in an image and is subsequently decoded into a descriptive sentence as an output of the systems. More recently, studies on vision-language pre-training (VLP) (cf. (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2019; Li et al., 2020a; Zhou et al., 2020)) revealed that vision-language pre-training model can effectively lead to generic representations from massive image-text pairs and is able to fine-tune with task-specific data, such as image captioning, with state-of-the-art results.

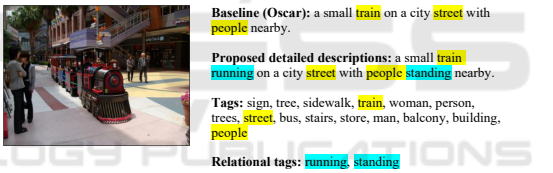

Figure 1: Examples of the image descriptions that exploit the relations extracted from both image (object tags) and text (relational labels); Objects tags and relational label are highlighted by yellow and blue, respectively.

Object-semantics aligned pre-training for vision-language tasks (Oscar) (Li et al., 2020b) is one of the VLP approaches that improve learning of cross-modal representations by utilizing object tags detected in images as anchor points to better learn the semantic alignments between images and texts. However, current VLP approaches do not utilize well the semantic relationship between objects in both images and texts, causing the generation of unnatural sentences as relations between objects are not considered (cf. the baseline's captions in Figure 1).

[a] https://orcid.org/0000-0002-9900-7047
[b] https://orcid.org/0000-0002-8823-2361
[c] https://orcid.org/0000-0002-2265-1010

Cheng, P., Le, T., Racharak, T., Yiming, C., Weikun, K. and Nguyen, M.
Learning Cross-modal Representations with Multi-relations for Image Captioning.
DOI: 10.5220/0010915100003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 346-353
ISBN: 978-989-758-549-4; ISSN: 2184-4313

To overcome this challenge, we demonstrate that learning of cross-modal representations can be further improved from the baseline by utilizing the relationship between semantics of visual and linguistic perspectives. Our motivation comes from an observation that when humans look at an image, they usually pay attention to the identification of the relationship between the objects in the image before they give a description on them. Following the motivation, we propose a novel framework extended from Oscar that enables the utilization of potential relational labels between objects in an image. In our work, objects in an image are indicated by the object detector (e.g. Faster R-CNN) and the relation labels between them (e.g. verb, preposition, and conjunction) are indicated by the ground-truth sentences.

Figure 1 shows an intuitive comparison between the proposed framework and the baseline. In the first example, it shows that the baseline (Oscar) outputs 'a small train on a city street with people nearby'. While this sentence looks natural and reasonable, it will be better if the model can utilize well possible relations ('running' and 'standing') between objects ('train' and 'people') and generate a sentence: 'a small train running on a city street with people standing nearby'. This example shows that an image's caption can be enhanced by adding more information related to the objects, particularly their relations.

To the best of our knowledge, we are the first time that exploits relational information of objects in the VLP architecture. We have analyzed the state-of-the-art models and found that they only consider image regions, object tags, and captions in the training; thus, they omit information describing the relationship between objects from both visual and linguistic perspectives. We fill this gap by proposing to incorporate this information for training the image captioning.

Furthermore, we extend the original MS COCO dataset (Lin et al., 2014) to include linguistic and positional relations for training our proposed model. For this purpose, our modified COCO dataset contains two forms of linguistic relational data: NLTK POS tagging and OpenNRE; both are extracted from the captions in the original COCO. Also, our modified dataset contains one form of positional data that captures potential links between objects in an image.

In sum, the contributions of this work are twofold: (i) we develop a new dataset by extending MS COCO to include linguistic and positional features and (ii) we propose novel training and inference methods that utilize well semantic relations on the image-text pairs. The experimental results on our extended dataset show that the proposed methods can improve SPICE score based on Oscar's pre-training and also improve SPICE 1.00 in CIDEr optimization (Rennie et al., 2017). Indeed, the results show that our methods do not only outperform the baseline but also become the state-of-the-art under the same experimental setting.

2 BACKGROUND

The training data for image captioning generally consists of image-text pairs. Notationally, a training dataset of size N is denoted by $\mathcal{D} := \{(\boldsymbol{I}_i, \boldsymbol{s}_i)\}_{i=1}^{N}$ with image $\boldsymbol{I}$ and text sequence $\boldsymbol{s}$. The goal of such model architectures is to train a caption generation for each image in a self-supervised manner.

Most VLP architectures including Oscar (Li et al., 2020b) employ multi-layer self-attention Transformers to learn cross-modal contextualized representations. The success of these model architectures is basically based on the quality of each modality's embedding. Notationally, for each input pair, an image $\boldsymbol{I}$ is represented by region features $\boldsymbol{v} = \{v_1, \ldots, v_K\}$ and a sequence $\boldsymbol{s}$ is represented by word embeddings $\boldsymbol{w} = \{w_1, \ldots, w_K\}$. Unlike traditional VLP architectures, Oscar additionally use object tags $\boldsymbol{q}$ to improve semantic alignments between images and texts. Since our proposed model is extended from Oscar, the following briefly provides its detail for self-containment.

Figure 2 (left) illustrates the input form of Oscar, showing that each image-text pair is represented as a word-tag-image triple $(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v})$. Here, $\boldsymbol{w}$ is a vector representing word embeddings of a text sequence, $\boldsymbol{q}$ is a vector representing word embeddings of the object tags detected on an image, and $\boldsymbol{v}$ is a vector representing a set of image regions detected on an image. Regarding the initialization, each image-text pair proceeds with the following steps as suggested in (Li et al., 2020b):

1. A pre-trained BERT (Devlin et al., 2018) is used to initialize $\boldsymbol{w}$;
2. A Faster R-CNN (Ren et al., 2015) is applied on each image to indicate semantic regions as (v', z) in which $v' \in \mathbb{R}^S$ represents region features and $z \in \mathbb{R}^T$ denotes region positions. The values of S and T are the same as in (Li et al., 2020b);
3. We concatenate v' and z to form a position-sensitive region feature vector and transform into $\boldsymbol{v}$ with a linear projection using matrix $\boldsymbol{W}$ in order to ensure that the dimension of $\boldsymbol{v}$ is the same as the dimension of $\boldsymbol{w}$;
4. The same Faster R-CNN is used to detect a set of high precision object tags and the pre-trained BERT is used to yield a sequence $\boldsymbol{q}$ of word embeddings for that tags.

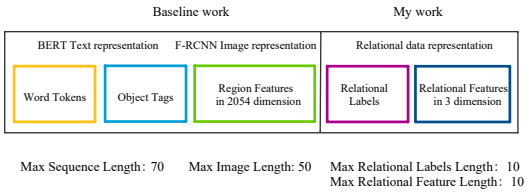

Figure 2: A comparison between the baseline modality representation and the proposed modality representation.

An example of real test dataset:
A cat eating a banana from someone's hand.

NLTK POS tagging:
Object tags: ('cat', 'NN') ('banana', NN)
Relational label: ('eating', 'VBG ')

OpenNRE relation extraction:
Relational label: ('part of ')

Figure 3: An image-caption example pair from our test dataset. Objects tags and relational labels are listed.

3 PROPOSED METHOD: OBJECT-SEMANTICS ALIGNED PRE-TRAINING WITH MULTI-RELATIONS

Motivated by an observation that humans usually pay attention on objects in an image and the relationship between them simultaneously to describe any images, we originally introduce to use linguistic features $\boldsymbol{l}$ and positional features $\boldsymbol{f}$ for image captioning. Formally, each training input i in $\mathcal{D}$ is defined as follows:

$$(\underbrace{\boldsymbol{w}_i, \boldsymbol{q}_i, \boldsymbol{l}_i}_{\text{language}}, \underbrace{\boldsymbol{v}_i, \boldsymbol{f}_i}_{\text{image}}) \quad (1)$$

The above equation shows that our proposed input form is conservatively extended from the Oscar input (cf. Section 2). The main objective of our proposed input is to integrate the pre-trained BERT models with relational linguistic features and positional features to achieve better performance on image captioning. We aim to address three research questions as follows:

1. What kinds of relations can enhance image captioning?
2. How can we train the proposed input form for image captioning?
3. Given an image and its relation features, how can we perform inference to yield a caption?

We provide the answers to all of the above questions in the following subsections sequentially.

3.1 Relational Linguistic Label and Positional Feature Extraction

We investigate two kinds of linguistic features: NLTK POS tagging and OpenNRE's relations, and one kind of image-based features using positional dimensions.

NLTK POS Tagging

We use word tokens directly from the image's caption in the dataset by adapting NLTK POS tagging. NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, as well as text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

We implement BERT tokenizer to tokenize a sentence of each caption into tokens. Then, the NLTK library is used for part-of-speech (POS) tagging to analyze tokens and labels from the MS-COCO image captioning dataset, as well as to find out those sentences containing the labels in POS's noun phrases. We filter out verb tokens or POS's preposition phrases from those sentences to make relational labels.

To illustrate the above method, let us consider a sentence taken from a test set (shown in Figure 3):

A cat eating a banana from someone's hand.

Here, the NLTK object tags ('cat', 'NN') and ('banana', 'NN') are POS's noun phrase (namely, 'NN', 'NNS', 'NNP', 'NNPS', 'PRP', 'PRP$'), while ('eating', 'VBG') is a relational word corresponding to the verb phrase (namely, 'VBG', 'VBZ', 'VB', 'VBN', 'VBD', 'VBP'). Thus, we take the word 'eating' from the dataset and annotate it as a relational label.

OpenNRE on Relational Label Extraction

Next, we consider the OpenNRE (Han et al., 2019) relation extraction. Note that OpenNRE is an Open and Extensible Toolkit for Neural Relation Extraction.

To illustrate OpenNRE's relations, let us consider the following sentence taken from OpenNRE:

He was the son of Máel Dúin mac Máele Fithrich, and grandson of the high king Áed Uaridnach (died 612).

Here, the sentence has two entities 'Máel Dúin mac Máele Fithrich' and 'Áed Uaridnach', as well as a relationship ('directed_by'). In other words, with OpenNRE, one can a triple of (entity, relation, entity) from any sentence to obtain structured information.

We apply OpenNRE's wiki80-based relationship extraction to analyze the relationship between noun

phrase label, and used the analyzed data as a relational label. The OpenNRE is an open-source and extensible toolkit, which provides a standard API to inference existing models for Relation Extraction (RE).

Consider the same sentence of the test data:

A cat eating a banana from someone's hand.

Here, the object tags ('cat', 'NN') and ('banana', NN) are still POS's noun phrase. Using the OpenNRE's wiki80-based relationship extraction model, the relationship ('part of') can be inferred and is thus regarded as a relational label in our proposed method. Note that the relational labels may not contain appropriate meaning literally while analyzing the COCO dataset image caption as they are inferred by the OpenNRE model. However, we hypothesize that the OpenNRE's relational labels can still represent the relationship of a sentence, which also reflects from our experimental results.

Relational Features Extraction

Finally, we investigate the positional dimension on images to define relational features. While relational labels (NLTK/OpenNRE-based labels) are extracted to express relationships from text, relational features are extracted to express positional relationships between objects in an image. We design a relational features' extraction simply but efficiently, which consider an angle ratio, an area ratio, a center point connection divided by the length of the area. The parameter setting is illustrated in Figure 4.

To efficiently represent the relational positions between objects, we introduce three relational position parameters (P_1, P_2, P_3) to describe the relation on bounding box i which is generated by object detection from an image. For each bounding box, the width $boxw_i$, the height $boxh_i$, the starting coordinates $(boxx_i, boxy_i)$ are indicated. The three parameters can be defined as follows.

Firstly, P_1 represents a relative position between two objects using an angle. It is defined as a gradient between two centered points $(cenx_1, ceny_1)$ and $(cenx_2, ceny_2)$, where i denotes a bounding box.

$$cenx_i := boxx_i + (boxw_i/2) \tag{2}$$

$$ceny_i := boxy_i + (boxh_i/2) \tag{3}$$

$$P_1 := \text{sigmoid}\left(\frac{ceny_2 - ceny_1}{cenx_2 - cenx_1}\right) \tag{4}$$

Secondly, P_2 represents how big a bounding box is relative to another based on the ratio of an area between two bounding boxes.

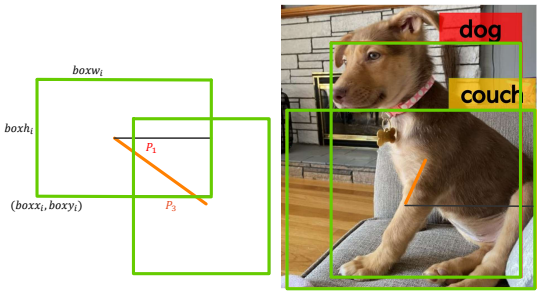

Figure 4: Two bounding boxes from object detection in an image with their intersections. P_1 indicates the angle of the two bounding boxes, P_3 indicates the relative distance of the two bounding boxes.

$$area_i := boxw_i \times boxh_i \tag{5}$$

$$P_2 := \frac{area_1}{area_1 + area_2} \tag{6}$$

Lastly, P_3 computes the relative distance between two bounding boxes. This distance is defined as the ratio between two centered points and re-scaled with respect to the relative area (the green line in Figure 4).

$$P_3 := \sqrt{\frac{(cenx_2 - cenx_1)^2 + (ceny_2 - ceny_1)^2}{area_1 + area_2}} \tag{7}$$

After calculation above, all the parameter P_1, P_2, P_3 expand to the range between 0 to 10 due to data scale of region feature.

3.2 Proposed Training: Multi-relations Aligned Vision-language Pre-training

To train the proposed multi-relations aligned vision-language model, we introduce a training scheme as shown in Figure 5. Firstly, we adapt relational labels and relational features as relational data into the embeddings of the baseline model. Then, we adapt the baseline method using relational data embedding as an extension, relational labels are input as extra text sequence into BERT, while relational features are input as extra visual sequence into Transformer.

The goal of training has two parts. The first part *Masked Token Loss* is to calculated the cross entropy of output probability of discrete token sequence h_i, which defined as $[\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{l}]$, based on their surrounding tokens sequence $h_{\setminus i}$ and image features $\boldsymbol{v}$ and relational features $\boldsymbol{f}$ by minimizing the negative log-likelihood:

$$L_{MTL} = -\mathbb{E}_{(v,h)\sim D} \log p(h_i | h_{\setminus i}, v, f) \tag{8}$$

The second part is *Contrastive Loss*, which is for image representation and relational data representation, as shown in Equation 9. The discrete input sequence h' is define as $[\boldsymbol{q}, \boldsymbol{v}, \boldsymbol{l}, \boldsymbol{f}]$. A set of 'polluted' images is created, then replacing q with probability 50% with a different tag sequence randomly sampled from the dataset D. We apply a fully-connected (FC) layer on the top of it as a binary classifier $f_{bin}(.)$ to predict whether the pair contains the original image representation (y = 1) or any polluted ones (y = 0).

$$L_C = -\mathbb{E}_{(h',w)\sim D} \log p(y \mid f_{bin}(h', w)) \tag{9}$$

The full objective of proposed method is:

$$L_{Pre-training} = L_{MTL} + L_C \tag{10}$$

We use the cross-entropy as the loss function. The sequence length of discrete tokens h and region features v are 35 and 50, respectively. The length of relational labels l and relational features f are 10 and 10, respectively. We pre-train model variants, linear project relational labels and relational features sequence dimension to the BERT hidden size of 768 as word tokens, Object Tags & region features dimension. In sentence generation, we apply Beam search and set the number of output sequence to 5.

3.3 Proposed Inference

Since our model extends from the baseline (Oscar) by exploiting both relation labels and relational features for image captioning, this work also introduces a new inference scheme consisting of the following steps:

1. Given an image, we first predict a caption from the baseline model. Although Oscar is used in this work, it can be changed. We also plan to investigate and experiment with other models in future,
2. The relational labels of the predicted caption will be extracted based on our proposed method,
3. The relational features are extracted from the bounding boxes of the image, generated by the object detection system (Faster R-CNN),
4. We apply the extracted relational labels, the extracted relational features, object tags, relational features for inferring the image's caption.

The overall process of the proposed inference is shown in Figure 6.

3.4 Fine-tuning: CIDEr Optimization based on SCST

SCST is a reinforce learning algorithm for solving the exposure bias problem in mainstream MLE training methods. It is an improvement of the reinforcement learning-based method called Mixed Incremental Cross-Entropy Reinforce (MIXER) (Ranzato et al., 2015). Using SCST, attempting to estimate the reward signal, as actor-critic methods must do, and estimating normalization, as REINFORCE algorithms must do, is avoided, while at the same time harmonizing the model with respect to its test-time inference procedure. We further fine-tune Oscar with self-critical sequence training (SCST) using CIDEr Optimization to improve sequence-level learning.

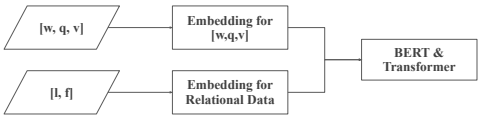

Figure 5: The proposed training process, where a square box represents model's process, a diamond box represents data, and $[\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v}]$ denotes the baseline input of word tokens $\boldsymbol{w}$, object tags $\boldsymbol{q}$ and region features $\boldsymbol{v}$. Also, $[\boldsymbol{l}, \boldsymbol{f}]$ represents the relational labels $\boldsymbol{l}$ and relational features $\boldsymbol{f}$.

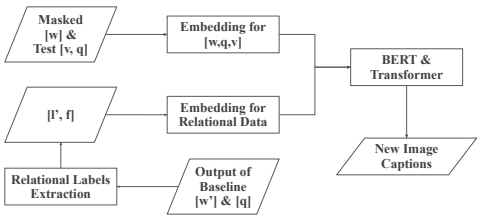

Figure 6: The proposed inference process, where a square box represents model's process, a diamond box represents data, and $[\boldsymbol{l'}, \boldsymbol{f}]$ denotes the relational labels $\boldsymbol{l'}$ extracted from baseline model prediction $\boldsymbol{w'}$. Also, $\boldsymbol{q}$, $\boldsymbol{f}$ denote object tags and relational features extracted from object detection bounding box, respectively.

4 EXPERIMENT

4.1 Dataset

We evaluate our approach on Image Captioning using the MS-COCO Image Caption dataset (Lin et al., 2014), which provides resources for training, validation, and testing. The dataset's statistical information is shown on Table 1. Currently, the MS COCO 2014 dataset contains six million captions and over 160,000 images. Each set contains an image with at least five sentences. The ratios that an image and its paired text is shared at least one, two, three objects are 49.7%, 22.2%, 12.9%, respectively. Each image is encoded to region feature with 2054 dimension by Faster-RCNN.

Table 1: The quantity of Image Captions, Relational Data (R. Data), average Data per captions of training set (Train), validation set (Val), and test set (Test) in NLTK relational label, OpenNRE relational label and relational feature (per one feature is shown in parenthesis).

	Captions	R. Data	Data Per Captions
NLTK Relational Label			
Train	566732	311419	0.45
Val	102604	62285	0.60
Test	102550	84125	0.82
Open-NRE Relational Label			
Train	566732	791238	1.40
Val	102604	142310	0.72
Test	102550	182280	1.81
Relational Feature			
Train	566732	3105670	5.49 (1.83)
Val	102604	573806	5.58 (1.86)
Test	102550	719050	7.02 (2.34)

4.2 Evaluation Metrics & Baseline

In this study, we implement four evaluation metrics in total, and focus on two main evaluation metrics:

- SPICE measures how effectively image captions recover objects, attributes, and the relations between them, by comparing consistency using a parse tree constructed with a ground truth sentence.
- BLEU is a popular machine translation metric that analyzes the co-occurrences of n-grams between the candidate and reference sentences, comparing consistency by each ground truth sentence. This work uses $n = 4$ (a.k.a. BLEU-4).

The SPICE and BLEU-4 describe the consistency of one sentence in ground truth sentences based on parse tree and n-gram, Thus SPICE and BLEU-4 are selected as this study's main evaluation metrics.

The CIDEr enables an objective comparison of machine generation approaches based on their "human-likeness", without having to make arbitrary calls on weighing content, grammar, saliency, etc. with respect to each other. The METEOR is calculated by generating an alignment between the words in the candidate and reference sentences, with an aim of 1:1 correspondence. The CIDEr and METEOR evaluation are reference evaluation metrics to provide the auxiliary result verification.

The baseline is set to the output result of the original Oscar model. However, the baseline score of pre-training is set to the baseline model executed from our environment, and the baseline score of CIDEr Optimization is set to the output result of the original Oscar model.

4.3 Experimental Setting

The relational labels consist of two forms of data: NLTK and OpenNRE labels. The statistical information of each label is as shown on Table 1. In total, there are 0.45 NLTK relational labels per each caption sentences extracted from the training dataset, while there are on average 1.40 OpenNRE relational labels extracted. A potential reason is that not many captions contain POS's verb phrase. The table also shows that on average 5.49 relational features are extracted from training dataset and each proposed parameter of P_i has 1.83 features on average per sentence.

The quantity of relational features are more than the quantity of relational labels. This is because we can extract positional relational features as long as the object tags pair is found in captions, and also because there are multiple bounding boxes (e.g. objects) for one object tags in an image. However, when using predicted captions of baseline, we adapt the generated top-5 captions to achieve the same standard of ground truth test set.

We compare our method against the best performance of baseline models of similar size to the BERT base. We train our method for at least 900k steps, with learning rate of $1e^{-5}$ and a batch size of 64 samples from the baseline work. The AdamW Optimizer is used.

We conducted two experiments, including pre-training and CIDEr Optimization. To verify the impact of the two different relational labels and relational features, we firstly compare the relational label by NLTK with relational features and the relational label by OpenNRE with relational features, compared to baseline respectively. After adapting the pre-training, we choose the best pre-training result of the relational label by OpenNRE to conduct CIDEr Optimization fine-tuning based on SCST.

4.4 Experimental Results & Analysis

The experimental results on pre-training are presented in upper part of Table 2, while the lower part also summarizes the experiment results of fine-tuning, which conducts CIDEr Optimization by adapting relational label by OpenNRE, compared to the baseline.

As shown in Table 2, there are two main findings. The better SPICE score is achieved by the representation of OpenNRE relational labels during pre-training. However, in fine-tuning of SCST, a great impact is made on SPICE score during CIDEr Optimization. The representation of relational features affects differently in BLEU-4 score on different relational labels.

Table 2: A summary of all compared with the baseline. The upper part and lower part show the results of pre-training and CIDEr optimization, respectively. In this table, there are four experimental setting of relational labels and relational features are implemented, and one experimental settings of relational labels are adapted. **Δ** represents the difference between the best performance and the baseline's result.

Method	SPICE	BLEU-4	METEOR	CIDer
Baseline	23.04	35.97	30.1	122.26
Proposed relational labels (NLTK)	22.90	35.11	29.80	120.03
Proposed relational labels (NLTK) & relational feature	22.31	34.39	28.93	115.80
Proposed relational labels (OpenNRE)	**23.09**	35.68	30.06	121.75
Proposed relational labels (OpenNRE) & relational feature	23.01	35.92	30.02	121.88
Δ (Pre-training process)	**0.05** ↑	0.29 ↓	0.04 ↓	0.51 ↓
Baseline	22.8	40.5	29.7	137.6
Proposed relational labels by OpenNRE	**23.8**	39.2	**29.8**	132.1
Δ (CIDEr Optimization process)	**1.0** ↑	1.3 ↓	**0.1** ↑	5.5 ↓

OpenNRE relational labels improve SPICE score rather than BLEU-4 score, indicating that OpenNRE relational labels can influence the model in the accurate single caption rather than imitating validation dataset. Comparing to the NLTK relational labels, OpenNRE relational labels can represent the relation more accurately between entities than a single keyword in captions. However, the great increasing SPICE score in CIDEr Optimization reveals the importance of the proposed OpenNRE relational labels, which influence the limit by our objective function but also influence greatly by the objective function of the CIDEr Optimization based on SCST.

Relational features affect differently on the results of relational labels of NLTK and OpenNRE, demonstrating that the relational features have potential auxiliary effectiveness on positional relation due to different relational labels.

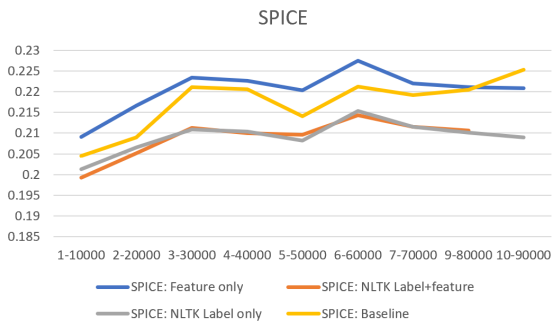

Figure 7: Analyzing the influence of relational label and relational feature on the proposed method, while pre-training relational label and relational feature separately.

4.5 Ablation Study

For further studying on relational features, we also implement ablation study about the single effect of relational features on our proposed method. We implement single relational features on our proposed method, which is shown in Figure 7. We find that, even though the SPICE score of NLTK labels falls in the training process, the relational features cause increasing SPICE score in the first 10 epochs. This may mean that more accurate relational features on positional relation will have a chance to make an improvement to the results.

5 RELATED WORK

The usage of pre-training generic models which solve a variety of V+L problems is becoming significant, such as visual question answering (VQA), image-text retrieval and image captioning etc. The existing methods (Sun et al., 2019; Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2020; Zhou et al., 2020; Su et al., 2019; Li et al., 2020a; Hao et al., 2020) employ BERT-series model (Devlin et al., 2018) to adapt cross-domain representations from a concatenated-embedding of image features and sentence tokens. For example, a two-stream and three-stream Transformer-based framework with co-attention to fuse the two modalities are proposed by early research (Tan and Bansal, 2019; Lu et al., 2019). Chen et al. (Chen et al., 2020) conduct comprehensive studies on the effects of different pre-training objectives for the learned generic representations. Zhou et al. (Zhou et al., 2020) propose the first unified model to deal with both understanding and generation tasks, using only VQA and image captioning as the downstream tasks. The Oscar models have been applied to a wider range of downstream tasks, including both understanding and generation tasks.

Comparing with existing VLP methods, the most salient difference of this research is the use of relational data for revealing the relationship between objects in image, and adapt different inference process.

6 CONCLUSION & FUTURE WORK

In this paper, we have presented a new method of using relational data based on relation extraction to enhance image captioning. This method uses two different relational labels and the relational features, aligning the image and language modalities to enhance the shared semantic space. We validate the schema by pre-training Oscar models with relation extraction on a public corpus with 6.5 million text-image pairs. The VLP models based on our proposed method archive new results on image captioning.

In the future, the sequence of visual region features will be a breakthrough, due to the more accurate semantic representation of image could lead to more accurate alignment for image and text representation. The lasted research Oscar+(Zhang et al., 2021) could also bring more potential improvement. In the relation extraction perspective, more standard relation extraction for image caption will be needed, which may influence key components proposed in this paper.

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 conference on computer vision and pattern recognition*, pages 6077–6086.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., and Sun, M. (2019). OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pages 169–174.

Hao, W., Li, C., Li, X., Carin, L., and Gao, J. (2020). Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. (2020a). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020b). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

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 *European conference on computer vision*, pages 740–755. Springer.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2015). Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019). Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. *arXiv preprint arXiv:2101.00529*.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.