OCL-NOC: Object-level Contrastive Learning in Novel Object Captioning

Anonymous ICCV submission

Paper ID 1258

Abstract

We present in this paper a novel object-level contrastive learning approach (OCL-NOC) to novel object captioning. The proposed approach aims at learning visual and textual modality alignment by maximizing the compatibility between regions and object tags in a contrastive manner. To set up a proper contrastive objective, for each image and its paired object tags, we first augment object tags by taking into account both the image and tag context. Then we utilize the rank of each augmented object tag in a list as a relative relevance label to contrast each top ranked tag with a set of lower ranked tags. Such a learning objective encourages the top ranked tags to be more compatible with their image and text context than lower ranked tags, hence effectively improving the discriminative ability of the learned multi-modality representation. The experimental results on two datasets show that our proposed OCL-NOC approach outperforms state-of-the-art methods by large margins, demonstrating the effectiveness of OCL-NOC in improving the object-level vision-language representation for novel object captioning.

1. Introduction

Describing novel objects unseen in training data is a highly desired capability for a real-world image captioning model. Conventional image captioning models [3, 26, 18] often fail to describe novel objects because they only cover limited visual concepts and generalize poorly to wild images [24]. To overcome this limitation, approaches that rely on object detection as the external resource[25, 18, 31, 7, 23, 6, 16, 12, 33] have been widely explored and demonstrated breakthroughs in vision-language (VL) understanding.

Although object detection models (*e.g.*, Faster RCNN [21]) have been improved to recognize a wide range of objects including novel ones like zero-shot object detection [13], using object detection in novel captioning models brings a new challenge. VIVO [12] leverages extra object tags to pre-train a visual vocabulary and help image captioning generalize to new categories. NOC-REK [27] tries to augment object tags in the training stage based on the similarity between region and word. Such methods simply concatenate regions, object tags, and caption features as input to a Transformer-based model and use masked token reconstruction to implicitly learn an alignment between vision and language. Although such an alignment can help bring regions and words of the same concept closer, it lacks an effective mechanism to push away unrelated concepts. As a result, it still makes mistakes on some confusing concepts, as shown in Figure 1.

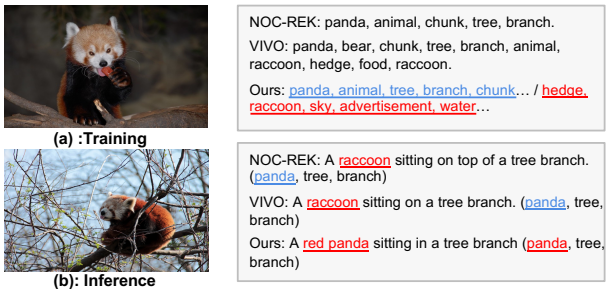

Figure 1: An illustrative example of our method to leverage object-level relative semantic relevance to achieve modality alignment. Our method could give accurate captions `"A cute red panda sitting in a tree branch"` conditioned on the objects `"red panda"` while VINVL+VIVO generates the wrong caption `"A cute raccoon sitting in a tree branch"`. This inference result shows that our method can differentiate some confusing objects and generate accurate captions for novel objects when the object detection is well aligned with other modalities.

Different from previous methods, this study aims to learn visual and textual modality alignment in a contrastive manner. Gupta *et al*. [9] showed that modality alignment could be achieved by maximizing the information lower bound between image and object tags. That is, given pairs of images and object tags, we can maximize the compatibility between tags and their attention-weighted region representations, compared to regions and non-corresponding tags. However, there

are two critical problems that need to be further addressed: 1) how to effectively generate contrastive tags that are closely relevant to an image, and 2) how to design a proper contrastive learning objective that allows the model to effectively leverage the augmented tags to align visual and textual modality.

To tackle the first problem, a straightforward solution is to construct some positive/negative tags given the context of an image and its caption. The main difficulty of constructing such tags is the requirement of a multi-modality semantic understanding of both image and caption. Gupta *et al*. [9] developed a solution to find hard negative words that can replace a word in a caption, but without considering its paired image, which is crucial for modality alignment. To consider both the vision and language context, for each tag, we utilize a pre-trained multi-modality model in VIVO[12] to generate a list of contrastive tags that are closely related to both image and object tags. Such a simple approach can help find more useful tags which are crucial for contrastive learning.

To address the second problem, a proper contrastive learning objective needs to be explored. The major challenge here is that the augmented contrastive tags are inaccurate and inevitably noisy. We cannot simply treat one tag as positive and others as negative to perform contrastive learning, because the augmented tags might be highly correlated or similar to each other. To tackle this problem, we leverage the relative relevance, rather than the absolute relevance, of the augmented tags, which is more robust to data noise.

Specifically, given an image, for each of its labeled object tags, we generate a ranked list of contrastive tags using a VIVO pre-trained model. We regard the rank of each augmented tag as its relative semantic relevance with the image. In general, the top-ranked tags are assumed to be more relevant to the image than the lower-ranked tags. We divide the list into two parts: the (relatively) relevant part with higher rank and the (relatively) irrelevant part with lower rank (Figure 1 (a)). In our proposed objective, we treat each tag in the relevant part as positive and all tags in the irrelevant part as negative to perform contrastive learning. Note that we do not let tags both in the relevant part contrast with each other as they might be highly correlated concepts. In this way, our approach weakens the strict assumption of contrastive learning in previous works and exploits the relative ranking in a loose form to achieve modality alignment.

To fully utilize object-level representation, we also propose an object-level contrastive learning (OCL) loss to enhance the alignment for both cross-modality (region-tag) and inner-modality (caption-tag). As a result, our method (OCL-NOC) not only enhances visual concepts for better region-tag alignment, but also learns caption-tag semantics to connect related caption tokens for better region-caption interaction. Moreover, to further mitigate the negative effect brought by noisy contrastive tags, we develop an uncertainty-aware selection and reweighting (UASR) strategy to let the learning algorithm focus on highly reliable samples and demote false positives and false negatives.

We conduct experiments on the Nocaps and Held-Out COCO datasets to demonstrate the effectiveness of OCL-NOC. Our contributions can be summarized as follows:

- We propose to learn the relative semantic relevance in a loose form by maximizing the compatibility between regions and their relevant tags compared with regions and irrelevant tags to achieve vision and language alignment and improve the discriminative ability of the multi-modality representation.
- An object-level contrastive learning loss function is proposed to enhance alignment for both inner-modality and cross-modality, with an uncertainty-aware selection and reweighting strategy to further mitigate the negative effect brought by noisy contrastive samples.
- We validate the proposed method on the Nocaps and Held-Out COCO benchmarks, which outperforms other state-of-the-art methods by a large margin.

2. Related work

Novel object captioning aims to describe images with objects that are unseen in the training stage (we define these objects as novel objects) where many methods. [25, 18, 31, 7, 6, 16, 12, 33, 20, 29] have been proposed. Early works such as Hendricks *et al*. [11] and Venugopalan *et al*. [25] utilize unpaired labeled image and sentence data to enhance semantically visual concepts. Recent studies propose to explicitly leverage the object detection results for NOC, Lu *et al*. [18], Wu *et al*. [31], and Demirel *et al*. [7] fill the generated template sentence with objects detected by object/novel object detectors. Chen *et al*. [6] combine object detector and human attention to identify novel objects. In addition, Li *et al*. are the first to utilize object-level semantics in VLP tasks, which are further extended by Zhang *et al*. [33]. Hu *et al*. [12] built upon [33] and propose to leverage extra region-tag pairs to conduct pretraining. Duc *et al*. [27] tried to augment object tags in the training stage based on the similarity of regions and objects.

However, most aforementioned methods for NOC ignore the misalignment problem of object tags, thereby failing to fully exploit the object-level semantic relationship between vision and language (Figure 1), which we argue is crucial to the quality of generated captions. In this paper, we propose a simple but effective object-level contrastive learning objective to learn the relative semantic relevance in a loose form where the object tags could be explicitly aligned with their corresponding image feature representations in a semantic space.

Contrastive Learning aims to learn discriminative representations to distinguish an image from oth-

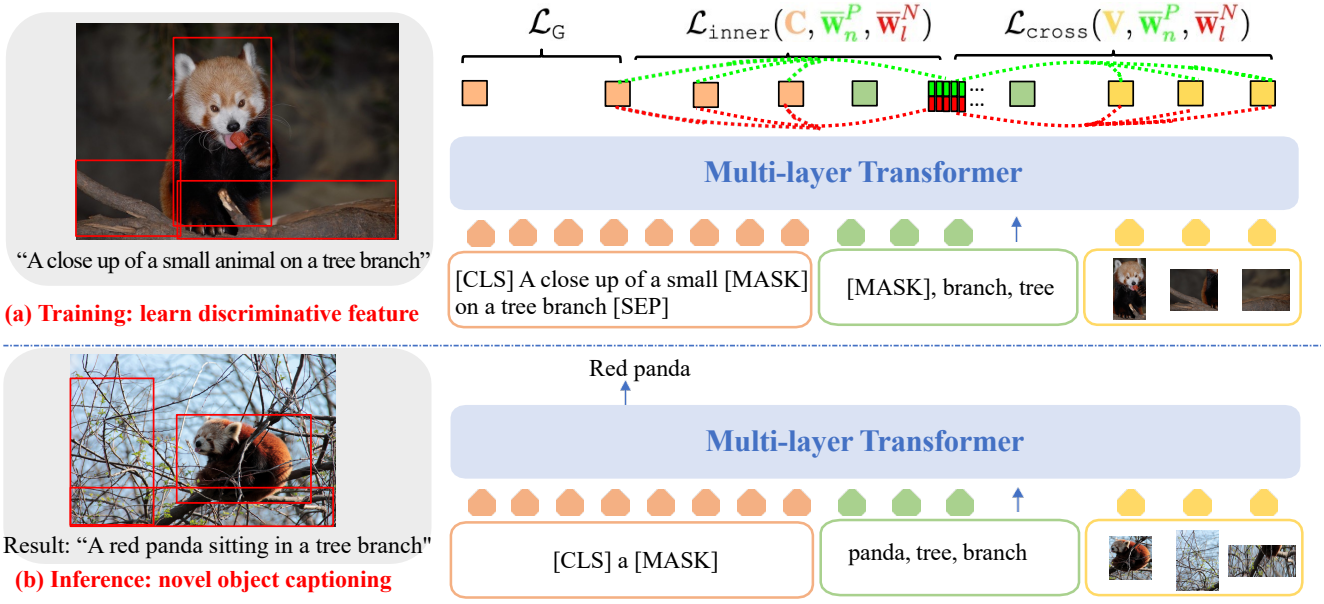

Figure 2: The main pipeline of our proposed OCL-NOC, including training and inference: In addition to the normal caption loss L_G in the training process, we also compute the cross-modality and inner-modality loss L_{cross} and L_{inner} in Eq. 15 and Eq. 16 by contrasting the positive tags against the negative tags. The corresponding relationship is denoted in green and red line separately. During inference, our model could accurately generate the caption that contains novel objects.

ers. Many methods [5, 10, 30, 14, 9, 20, 15] have shown their effectiveness. For example, Chen *et al.* [5] proposed to learn visual representations by maximizing agreement between differently augmented views of the same image via a contrastive loss. He *et al.* [10] proposed Momentum Contrast (MoCo) for unsupervised visual representation learning. Wei *et al.* [30] utilized contrastive learning in the detection task. Li *et al.* [14] focused on the problem of contextual outpainting. Gupta *et al.* [9] proposed a method CPG[1] to find hard negative words by replacing a word in a caption.

Other data-driven studies such as CLIP [20] and ALIGN [15] focus on learning a corresponding relative relationship from massive noisy web data. CLIP [20] predicts which text goes with which image and learns a relative image-text corresponding relationship from broad web data with noisy supervision. ALIGN [15] further scales up CLIP by leveraging a noisy dataset that covers more than one billion image-text pairs. These methods achieve remarkable results but require massive fully-annotated data, which are difficult to obtain.

3. Proposed Methods

By maximizing the compatibility between regions and their relevant tags compared with regions and irrelevant tags, we leverage the object tags' relative semantic relevance to learn vision-language alignment and improve the discriminative ability of the multi-modality representation. Specifically, we use a two-stage scheme that consists of a contrastive sample generation stage and a training stage with contrastive learning. Fig. 2 and Fig. 3 illustrate our approach.

First, in the contrastive sample generation stage (Fig. 3), for each image and its paired object tags, we generate a ranked list of augmented tags using a pre-trained multi-modality model [12]. In particular, the input to the pre-trained multi-modality model includes region features and object tags. The output includes two sets, including positive tags and negative tags. In the training stage (Fig. 2(a)), given image-caption pairs, we utilize the rank of each augmented object tag in a list as a relative relevance label to contrast each top-ranked tag (*e.g.*, positives: `"panda"` and `"tree"`) with a set of lower ranked tags (*e.g.*, negatives: `"raccoon"` and `"hedge"`). Such a learning objective encourages the top-ranked tags to be more compatible with their image and text context than lower-ranked tags. Thus, our model could achieve modality alignment and discriminative multi-modality representation learning, allowing the model to differentiate confusing object classes while composing familiar constituents.

In addition to the cross-modality (region-tag) contrastive loss L_{cross} (Eq. 3.2) and the conventional caption loss L_G [27], we also compute inner-modality contrastive loss L_{inner} (Eq. 7) for caption-tag pairs

[1]For better illustration, we use the abbreviation CPG to denote the paper "Contrastive Learning for Weakly Supervised Phrase Grounding" [9]

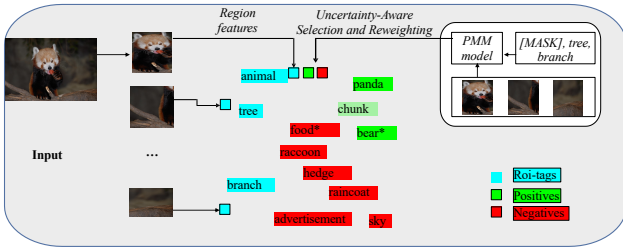

Figure 3: An example of contrastive sample generation. The cyan, green, and red colors denote object tags, positive tags, and negative tags, respectively. $*$ denotes false positives/negatives. PMM denotes a Pre-trained Multi-Modality model.

(Fig. 2(a)). This encourages the model to also align caption words with object tags in the different context and hence generate captions which are more consistent with the given object tags. Additionally, to reduce the impact of noisy contrastive tags, Uncertainty-Aware Selection and Reweighting (UASR) is involved in the training process to let the learning algorithm focus on highly reliable samples and demote false positives and false negatives. In the following, we describe our methods step by step.

3.1. Contrastive Samples Generation

To fully leverage the object tags' relative semantic relevance, we focus on contrasting relevant tags against relatively irrelevant tags. However, extracting such contrastive tags that are related to the corresponding image is challenging as it requires the multi-modality semantic understanding of both the image and caption. To consider both the vision and language context, we leverage a pre-trained multi-modality model in VIVO[12] to get a list of contrastive tags that are closely related to both image and object tags. We adopt VIVO for its rich object-level multi-modality semantics. Other multi-modality models could also be used, such as CLIP [20] and SimVLM [29]. The comparison results are shown in the ablation study.

Specifically, for an image, given its region features $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^{K}$ and object tags $T' = \{t_j\}_{j=1}^{T}$, for its m-th tag in T', we define a contrastive tag sample as one which has the same context $\mathbf{V}$ and $T'_c = \{t_j\}_{j=1,j\neq m}^{T}$ but is different from t_m with the following properties: a positive tag should be (relatively) plausible for the given context and a negative tag should be (relatively) implausible for the given context. As shown in Fig 3, consider the detected tags `"animal, tree, branch"` where t'_m is chosen as `animal` and tag context is defined by `"[MASK], tree, branch"` where `[MASK]` is the token that denotes a missing tag. Potential augmented candidates for a substituted tag might be `"panda, tree, branch"` where the augmented positive tag is `panda`. `sky` and `raincoat` are negative tags which are relatively implausible given the context.

Constructing contrastive samples. In particular, for the m-th tag in T', we feed the corresponding region features $\mathbf{V}$ and object tags T'_c context into the pre-trained multi-modality model to extract M augmented tags $T = \{t_l\}_{l=1}^{M}$ using probabilities $p(t_m|context)$ predicted by a pre-trained multi-modality model. For better illustration, we use the abbreviation $p(t_m)$ to denote $p(t_m|context)$ in the following sections. Then we utilize the rank of each augmented object tag in T as a relative relevance label. In general, the top-ranked tags are assumed to be more relevant to the image than the lower-ranked tags. We divide T into positive and negative tags: positive tags being the (relatively) relevant tags with higher rank $T^P = \{t_n^P\}_{n=1}^{K} = \{t_l\}_{l=1}^{K}$ and negative tags being the (relatively) irrelevant tags with lower rank $T^N = \{t_l^N\}_{l=1}^{K} = \{t_l\}_{l=K+1}^{2K}$ $(M = 2K)$.

We empirically find that such a learning objective is effective and easy to optimize by relatively encouraging the top ranked tags to be more compatible with their image and text context than lower ranked ones. Fig 3 shows a contrastive sample generation result with corresponding region features and tags. The green and red tags denote the positive and negative tags extracted from the pre-trained multi-modality model. $*$ denotes the corresponding false positives/negatives, which we will discuss in the following.

3.2. Cross and Inner modality Alignment Module

InfoNCE [19] is a typical type of contrastive loss function used for self-supervised learning that contrasts one positive sample against a set of negative samples. CPG [9] builds upon InfoNCE and proposes a compatibility function for regions and a caption word. Both InfoNCE and CPG utilize the absolute relevance to optimize the contrastive loss.

However, such a learning objective is hard to optimize for object tags since they may be highly correlated to each other. In OCL-NOC, we focus on the relative semantic relevance and utilize the attention-based compatibility function to measure the relativity information for regions and a set of object tags, which is more robust to data noise.

Specifically, we first compute the dot product for a region-tag pair.

$$s_{jk} = \mathbf{w}_j^T \mathbf{v}_k / \sqrt{d}, \tag{1}$$

where $\mathbf{w}_j$ and $\mathbf{v}_k$ refer to the corresponding embeddings for tag and region, and d is the feature dimension. Here, s_{jk} represents the similarity between the j-th tag and the k-th region. To find a contextualized region representation for the j-th tag, we define $\mathbf{a}_j^v$ as follows.

$$\mathbf{a}_j^v = \sum_{k=1}^{K} \alpha_{jk} \mathbf{v}_k, \tag{2}$$

where

$$\alpha_{jk} = \frac{e^{s_{jk}}}{\sum_{k'=1}^{K} e^{s_{jk'}}}. \tag{3}$$

To measure the compatibility between a tag $\mathbf{w}_j$ and its contextualized region representation $\mathbf{a}_j^v$, we define

$$\phi(\mathbf{V}, \mathbf{w}_j) = \mathbf{w}_j^T \mathbf{a}_j^v. \tag{4}$$

In this way, we can derive our cross-modality contrastive loss function L_{cross}.

$$\mathcal{L}_{\text{cross}}(\mathbf{V}, \mathbf{W}^P, \mathbf{W}^N) = -\frac{1}{K}\sum_{n=1}^{K} \log\left(\frac{e^{\phi(\mathbf{V}, \mathbf{w}_n^P)}}{e^{\phi(\mathbf{V}, \mathbf{w}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{V}, \mathbf{w}_l^N)}}\right), \tag{5}$$

where $\mathbf{W}^P = \{\mathbf{w}_l^P\}_{l=1}^K$ and $\mathbf{W}^N = \{\mathbf{w}_n^N\}_{n=1}^K$ are the word embeddings for positives/negatives T^P and T^N, extracted from a pre-trained multi-modality model. $\phi()$ is the compatibility function defined in Eq. 4.

This equation encourages each positive tag $\mathbf{w}_n^P$ to be more compatible with image regions $\mathbf{V}$ than all negative tags in $\mathbf{W}^N$. Essentially, if tag t_j is not relevant to the image (negative), its representation $\mathbf{w}_j$ should not be similar to its contextualized region representation $\mathbf{a}_j^v$ since it would not be able to collect good information while computing $\mathbf{a}_j^v$. Note that we do not let two positive tags (top-ranked tags) contrast with each other as it is hard to tell which one is more relevant.

The formulation for inner-modality contrastive loss (tag-caption) is similar but slightly different. We enforce contrastive learning not to be between tags and all caption words, but just between tags and noun caption words. This is because a caption may have many unrelated tokens, such as "the, a, is", which will harm our alignment process and cause redundant computational costs.

Finally, we can derive our inner-modality contrastive loss function L_{inner}.

$$\phi(\mathbf{C}, \mathbf{w}_l) = \mathbf{w}_l^T \mathbf{a}_l^c \tag{6}$$

$$\mathcal{L}_{\text{inner}}(\mathbf{C}, \mathbf{W}^P, \mathbf{W}^N) = -\frac{1}{K}\sum_{n=1}^{K} \log\left(\frac{e^{\phi(\mathbf{C}, \mathbf{w}_n^P)}}{e^{\phi(\mathbf{C}, \mathbf{w}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{C}, \mathbf{w}_l^N)}}\right), \tag{7}$$

where this time we take noun caption words $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^P$ in Eq. 6 as input. By further considering inner-modality alignment, we not only learn caption-tag semantics to connect related caption tokens, but also enable region-caption interactions.

3.3. Uncertainty-Aware Selection and Re-weighting

The augmented contrastive tags are often noisy and would harm the discriminative ability and robustness of the contrastive learning process. Therefore, we design a sample selection strategy to deal with noisy contrastive tags (*e.g.*, false positives/negatives).

Uncertainty-Aware Selection. Specifically, we use the cosine similarity to calculate the correlation between a region and a tag and filter false positives/negatives. The stronger the correlation, the more reliable a positive tag is and the more uncertain a negative tag.

$$\mathrm{u}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\| \, \|\mathbf{w}\|}, \tag{8}$$

Then we retrieve L tags and their corresponding representations $\mathbf{H} = \{\mathbf{w}_j\}_{j=1}^L$ for the given L region features $\{\mathbf{v}_i\}_{i=1}^L$. We use $\arg\max$ to choose the most correlated tag t_j for region $\mathbf{v}_i$.

$$j = \arg\max_k(\mathrm{u}(\mathbf{v}_i, \mathbf{w}_k)) \tag{9}$$

According to the score in Eq. 8, the corresponding top-L tags will be listed as confusion samples for negatives (actually positive), and these confusion samples will be removed from the original contrastive tags. Similarly, we will also choose the true positives based on these top-L tags since tags with low-similarity scores should be regarded as outliers to the positives and they are too sparse to have a positive influence on the shape of embedding space. In this way, we can derive Eq. 10, where / is the removing set operation, $\cap$ is the intersection set operation. $\overline{\mathbf{W}}^P = \{\bar{\mathbf{w}}_l^P\}_{l=1}^K$ and $\overline{\mathbf{W}}^N = \{\bar{\mathbf{w}}_n^N\}_{n=1}^K$ is the final positive/negative tags' embedding sets. Note that if the set length of $\overline{\mathbf{W}}^P/\overline{\mathbf{W}}^N$ is less than M, we will over-sample $\overline{\mathbf{W}}^P/\overline{\mathbf{W}}^N$ until the set length is equal to M.

$$\overline{\mathbf{W}}^P = \mathbf{W}^P \cap \mathbf{H}, \overline{\mathbf{W}}^N = \mathbf{W}^N/\mathbf{H} \tag{10}$$

Uncertainty-Aware Re-weighting. To further mitigate the negative effect of noisy tags, we use the correlation function defined above to enhance reliable samples and reduce the influence of high-uncertainty samples. The information from more reliable samples should have a larger impact on the shape of the embedding and vice versa. Specifically, we first introduce the weight in Eq. 11.

$$q_1(\bar{\mathbf{w}}_n^P) = \exp\left(\mathrm{u}(\mathbf{v}_k, \bar{\mathbf{w}}_n^P)\right) \tag{11}$$

$$j = \arg\max_k(\mathrm{u}(\mathbf{v}_i, \bar{\mathbf{w}}_k^P)) \tag{12}$$

$$q_2(\bar{\mathbf{w}}_n^P) = p(t_n) \tag{13}$$

where $q_1(\bar{\mathbf{w}}_n^P)$ in Eq. 11 is the uncertainty score computed by the most relevant region with the corresponding tag. Moreover, we further consider how certain each tag is and combine the corresponding score $q_2(\bar{\mathbf{w}}_n^P)$ in Eq. 13 with $q_1(\bar{\mathbf{w}}_n^P)$ to derive the final uncertainty score $q(\bar{\mathbf{w}}_n^P)$ in Eq. 14. Here $p(t_n)$ in Eq. 13 is the corresponding classification probabilities for tag t_n.

$$q(\bar{\mathbf{w}}_n^P) = q_1(\bar{\mathbf{w}}_n^P) * q_2(\bar{\mathbf{w}}_n^P) \tag{14}$$

Thus, our method could mitigate the negative effect of noisy contrastive tags by considering both the confidence of each tag itself and the multi-modality supervision signal from the region-tag relationship. Finally,

we could rewrite our contrastive loss function in Eq. 15, Eq. 16.

$$\mathcal{L}_{\text{cross}}(\mathbf{V}, \overline{\mathbf{W}}^P, \overline{\mathbf{W}}^N) = -\frac{1}{K}\sum_{n=1}^{K} -q(\bar{\mathbf{w}}_n^P) \log\left(\frac{e^{\phi(\mathbf{V}, \bar{\mathbf{w}}_n^P)}}{e^{\phi(\mathbf{V}, \bar{\mathbf{w}}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{V}, \bar{\mathbf{w}}_l^N)}}\right) \quad (15)$$

$$\mathcal{L}_{\text{inner}}(\mathbf{C}, \overline{\mathbf{W}}^P, \overline{\mathbf{W}}^N) = -\frac{1}{K}\sum_{n=1}^{K} -q(\bar{\mathbf{w}}_n^P) \log\left(\frac{e^{\phi(\mathbf{C}, \bar{\mathbf{w}}_n^P)}}{e^{\phi(\mathbf{C}, \bar{\mathbf{w}}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{C}, \bar{\mathbf{w}}_l^N)}}\right) \quad (16)$$

4. Experiments

4.1. Experimental Setup

Datasets. Our main experiments and ablation studies are based on the Nocaps[1] dataset. Following VIVO [12], we use the Open Images V5 challenge training set to implement the pre-trained multi-modality model in our paper, and no ground-truth tags are used on the Nocaps validation and test sets. For the training stage, we use the COCO training set which consists of 118K images, each with 5 captions. We evaluate our model on the validation and test sets of the Nocaps dataset, which consist of 4.5K and 10.6K images from the Open Images validation and test sets, respectively. Additionally, we test the proposed method on the Held-Out COCO [11], which is a subset of MS COCO [17] where the following eight object categories are excluded from the training set: bottle, bus, couch, microwave, pizza, racket, suitcase, and zebra. We randomly split the COCO validation set and use half of it for validation and the other half for testing, each with 20,252 images. In addition, we empirically set $M = 50$ to generate the original contrastive samples for the best performance.

Evaluation. We use the object detector from VinVL [33] to extract region features (2048D for the visual features and 6D for the bounding box encoding including top-left and bottom-up corners as well as the box's width and height) and tags (600 objects in the challenge set). In the training stage, the model is trained for 30 epochs with a batch size of 512 and a learning rate of 10^{-4}, optimized using the cross-entropy loss and our contrastive loss. We set the maximum caption length to 40 and the maximum tag length to 30. To further boost the performance, we also perform the SCST optimization (Rennie *et al.* [22]) with a learning rate of 1.4×10^{-6} for 10 epochs. During inference, we use greedy decoding to generate image captions with a maximum length of 20. Our model is trained with 8 A100 GPUS and takes 2 days to train.

4.2. Quantitative evaluation.

Results on Nocaps dataset. In this section, we extensively compare our frameworks with previous methods on the Nocaps benchmark. The other compared state-of-art results are from NOC-REK [27]. Note that all the methods use the BERT-base[8] for a fair comparison. We compare our method with UpDown [4, 1], OSCAR [16], VinVL [33], VIVO [12], and NOC-REK [27], which hold the state-of-the-art result on the Nocaps benchmark. The training data for the baselines is the COCO dataset. Following prior settings, we also report the results after the model is optimized using SCST [22]. Our generated captions also adopt Constrained Beam Search (CBS) following [2].

In Table 1, we present our results of SPICE and CIDErscores on the Nocaps validation and test sets. We also show the results of our different training stages in the middle of the table, where OCL-NOC** denotes our first training stage result and OCL-NOC is our final result after CIDEr optimization. Note that we do not compare with the data-driven methods[15, 28, 32] which use massive out-of-domain corpus and fail to follow the rule of Nocaps. By leveraging relative semantic relevance in contrastive learning, our method has achieved a significant improvement compared to all prior works. It is worth noting that our first stage training could generate a comparable result with [12, 33, 16]. After the CIDEr optimization process, our method could outperform the recent state-of-art method NOC-REK [27] with a large margin, which is 5.1 and 0.5 (CIDEr and SPICE) better than [27] on the Nocaps Test set. This suggests that our model is more capable of generating captions with novel objects.

To quantitatively evaluate how well the model can describe novel objects, we also calculate the F1-score between our generated and the ground-truth tags on the validation set. Table 2 shows the comparison with VIVO and NOC-REK on the Nocaps validation set. We see that OCL-NOC improves NOC-REK and VIVO on F1-score substantially, especially for out-of-domain objects. This again verifies the effectiveness of OCL-NOC's discriminative ability to describe and distinguish novel objects.

Results on the Held-Out COCO dataset. To further prove the generalization ability of our method, we also conduct experiments on the Held-Out COCO dataset. As we can see from Table 3, our method consistently beats the baseline VinVL+VIVO with a large margin. The improvements are 7.7, 2.5, 5.3, and 10.0 with BERT-base on the F1, SPICE, METEOR, and CIDEr metrics. Additionally, our method is superior to all other state-of-art methods. In particular, compared with the recent state-of-art method NOC-REK [27], the recent state-of-art on the Held-Out COCO dataset, OCL-NOC achieves a 4.4 improvement on CIDEr (from 138.4 to 142.8) and 3.1 improvement on METEOR (from 32.8 to 35.9). The 3.2 improvement on the F1 score also shows that OCL-NOC performs better in describing novel objects.

Table 1: Our evaluation results using SPICE and CIDEr on the Nocaps validation and test sets. We achieve the best scores for in-domain, near-domain, out-domain and Overall. Notably, the captions generated by our method are better than those by human in most cases. We note that our results on the test set are better than those by other methods which are publicly submitted to Nocaps leader-board[b]. Higher score is better.

Method	Validation set								Test set							
	in-domain		near-domain		out-domain		Overall		in-domain		near-domain		out-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
UpDown [3]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1	76.0	11.8	74.2	11.5	66.7	9.7	73.1	11.2
Oscar [16]	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7	81.3	11.9	79.6	11.9	73.6	10.6	78.8	11.7
VIVO [12]	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL [33]	96.8	13.5	90.7	13.1	87.4	11.6	90.9	12.8	93.8	13.3	89.0	12.8	66.1	10.9	85.5	12.5
VinVL + VIVO [33, 12]	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	**78.0**	11.5	92.5	13.1
NOC-REK [27]	**104.7**	**14.8**	**100.2**	**14.1**	**100.7**	**13.0**	**100.9**	**14.0**	**100.0**	**14.1**	**95.7**	**13.6**	77.4	**11.6**	**93.0**	**13.4**
OCL-NOC**	95.0	13.6	91.3	13.1	93.2	12.1	92.2	13.0	87.7	13.3	88.1	12.9	77.9	11.4	86.3	12.7
OCL-NOC	**109.7**	**14.7**	**100.9**	**14.2**	**100.8**	**13.2**	**102.5**	**14.1**	**101.5**	**14.4**	**100.2**	**14.2**	**87.5**	**12.5**	**98.1**	**13.9**
Δ	5.0↑	0.1↓	9.6↑	0.1↑	0.1↑	1.2↑	1.6↑	1.6↑	1.5↑	0.3↑	4.5↑	0.6↑	9.5↑	0.9↑	5.1↑	0.5↑
Human [1]	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6

4.3. Ablation study.

In this subsection, we will discuss every component's contribution to our framework. If not mentioned by purpose, all methods are conducted on the Nocaps validation set with BERT-base. (we cannot use the Nocaps test set because only 5 times submissions are allowed for the test set).

Effectiveness of multi-modality semantics. By utilizing the multi-modality semantics from the pre-trained model VIVO [12], we can introduce more novel information to our framework. To further prove the generalization ability of our method, we change this pre-trained model for contrastive sample generation and add comparison experiments using CPG [9], CLIP[20] to generate contrastive tags.

Table 2: Comparison of F1-scores (in %) on object classes of Open Images, evaluated on the Nocaps validation set. There are 504 classes in total. 80 of them are in-domain, which are common classes from COCO. The remaining 424 classes are the out-of-domain objects.

model	in-domain	out-of-domain	entire
VIVO	39.2	29.1	30.4
NOC-REK	45.3	30.5	32.8
Ours	58.6	39.2	43.1

Table 3: Evaluation results on the Held-Out COCO test set. The best results are highlighted in red.

Method	Avg. F1-score	SPICE	Meteor	CIDEr
NOC	48.8	–	21.4	–
NBT	48.5	15.7	22.8	77.0
DNOC	57.9	–	21.6	–
ZSC	29.8	14.2	21.9	–
ANOC	64.3	18.2	25.2	94.7
VinVL+VIVO	71.8	24.5	30.6	132.8
NOC-REK	**76.3**	**26.9**	**32.8**	**138.4**
OCL-NOC	**79.5**	**27.0**	**35.9**	**142.8**

Table 4: Effectiveness of the multi-modality semantics from pre-trained model. Changes in the type of pre-trained model result in changes in performance. Higher score is better.

Method	in-domain		near-domain		out-of-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
N.A.	103.2	13.8	95.4	13.5	83.4	12.0	94.0	13.2
CPG	104.5	14.0	96.8	13.6	84.9	12.1	95.4	13.3
CLIP	106.8	14.2	99.6	13.8	90.1	12.6	98.5	13.6
OCL-NOC	109.6	14.7	100.8	14.2	100.9	13.2	102.5	14.1

Table 5: Effectiveness of the CIAM. CAM: Cross-modality Alignment Module, IAM: Inner-modality Alignment Module, CIAM: Cross-modality and Inner-modality Alignment Module.

Method	in-domain		near-domain		out-of-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
N.A.	103.2	13.8	95.4	13.5	83.4	12.0	94.0	13.2
CAM	105.2	14.4	98.5	13.9	96.3	12.5	98.4	13.6
IAM	103.9	14.3	97.1	13.7	95.3	12.7	97.7	13.5
CIAM	109.7	14.7	100.86	14.2	100.7	13.2	102.5	14.1

The result shows that using CPG [9] to generate contrastive tags could only slightly improve our framework, which brings 1.4 and 0.1 (CIDEr and SPICE) gains compared with the baseline. This may be caused by the limited linguistic knowledge defined in the language model BERT [8], which may only cover limited novel categories and lack the crucial multi-modality semantic understanding. For CLIP[20], we feed the whole image into the model to retrieve contrastive tags based on its multi-modality semantic understanding. CLIP could bring 4.5 and 0.4 (CIDEr and SPICE) extra boosts compared with the baseline since it is pre-trained on massive data that contains rich information for novel categories. Finally, VIVO [12] performs the best for our approach, which brings 8.5 and 0.9 (CIDEr and SPICE) gains compared with the baseline. This can be attributed to the rich object-level semantics from VIVO, while CLIP is mainly designed for learning the image-level corresponding relationship.

Effectiveness of the CIAM. We conduct ablation experiments for each component of the Cross-modality and Inner-modality Alignment Module (CIAM) step by step. Table 5 shows that directly using cross-modality

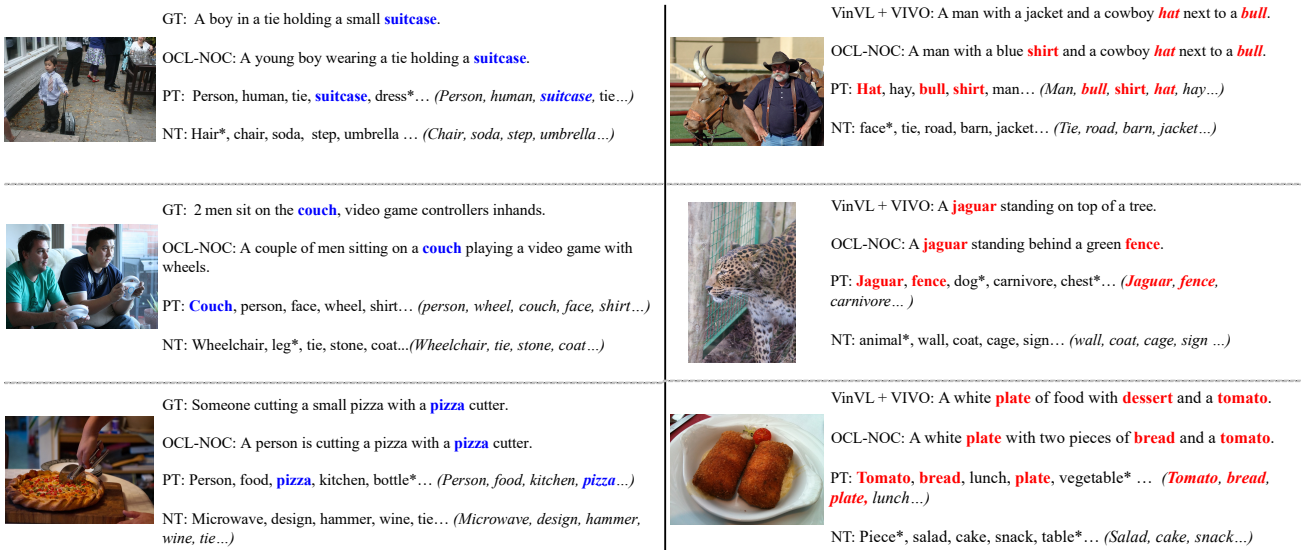

Figure 4: Examples of generated captions and contrastive tags by compared methods on Held-Out COCO (left) and Nocaps (right). We show the ground-truth captions (GT) on Held-Out COCO for reference. PT/NT denotes the positive/negative tags before and after Uncertainty-Aware Selection and Reweighting (UASR). **Blue**/**Red** text indicates novel objects in Held-Out COCO/Nocaps, and $*$ indicates the false positives/negatives.

Table 6: Effectiveness of the UASR. UAR: Uncertainty-Aware Re-weight, UAS: Uncertainty-Aware Selection, UAS^N: Uncertainty-Aware Selection for Negatives), UASR: Uncertainty-Aware Selection and Re-weight.

Method	in-domain		near-domain		out-of-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
N.A.	103.2	13.8	95.4	13.5	83.4	12.0	94.0	13.2
UAS^N	105.2	14.2	96.1	13.7	90.4	12.6	97.2	13.7
UAS	107.8	14.3	98.0	13.8	96.6	13.0	98.9	13.9
UAR	105.3	14.1	97.0	13.6	90.1	12.3	97.1	13.5
UASR	109.7	14.7	100.86	14.2	100.7	13.2	102.5	14.1

contrastive loss could bring 4.4 and 0.4 boosts on CIDEr and SPICE by aligning region-tag features. On top of the Cross-modality Alignment Module (CAM), further extending our method into CIAM yields 4.1 and 0.5 (CIDEr and SPICE) extra gains, which may be attributed to inner-modality alignment for caption-tag features. Finally, our method improves the performance by 8.5 and 0.9 (CIDEr and SPICE) by combining the Inner-modality Alignment Module (IAM) and CAM. This is because CIAM could not only achieve region-tag and caption-tag alignment but also enable region-caption interaction.

Effectiveness of the UASR. We investigate the performance of our Uncertainty-Aware Selection and Reweighting (UASR) on the Nocaps validation set in this part. Table 6 shows that Uncertainty-Aware Reweighting (UAR) could effectively enhance the reliable contrastive samples and bring 3.1 and 0.3 boosts on CIDEr and SPICE. In addition, the false positives/negatives severely harm our training process. By further incorporating Uncertainty-Aware Selection for Negatives (UAS^N) into our method and filtering false negatives, our model's performance is boosted to 97.2 and 13.7 on CIDEr and SPICE. When filtering false positives/negatives simultaneously, we could reach 98.9 and 13.9 on CIDEr and SPICE with Uncertainty-Aware Selection (UAS). Finally, our method improves the performance by 8.5 and 0.9 (CIDEr and SPICE) by combining UAR and UAS, which illustrates the proposed UASR is a more powerful tool for tackling noisy contrastive samples in NOC.

4.4. Qualitative Results

In Figure 4, we display some qualitative results on Held-Out COCO and Nocaps, and all the approaches are based on BERT-base. On Held-Out COCO, the result compared with GT shows that our proposed method could effectively generate accurate and precise captions with novel objects by leveraging relative semantic relevance into training. On Nocaps, VinVL+VIVO [33, 12] sometimes cannot include the desired novel objects in their captions (first and third examples) or generate wrong captions (second example). Our OCL-NOC, on the other hand, successfully generates correct and coherent captions via differentiating confusing classes and aligning object detection with other modalities. For better illustration, we show top-5 Positive and Negative Tag results extracted from the pre-trained multi-modality model.

5. Conclusion

In this paper, we have proposed a novel Object-level Contrastive Learning in Novel Object Captioning called OCL-NOC to learn visual and textual modality alignment by maximizing the compatibility between re-

gions and object tags in a contrastive manner. Specifically, for each image and its paired object tags, we first augment object tags by taking into account both the image and tag context. Then we utilize the rank of each augmented object tag in a list as a relative relevance label to contrast each top ranked tag with a set of lower ranked tags. We empirically find that such a learning objective is effective and easy to optimize by encouraging the top ranked tags to be more compatible with their image and text context than lower ranked tags, hence improving the discriminative ability of the learned multi-modality representation. We prove the effectiveness of our paradigm in novel object caption, with the spotlight on Nocaps and Held-Out COCO benchmark.

The effectiveness of object tags has already been proved in recent works. However, there are few studies to further explore why such tags could perform well and the problem of object tags' misalignment is often ignored. So how to further exploit use tags explicitly while having a deep understanding of their roles is meaningful in the future.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 6, 7

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017. 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 7

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[6] Xianyu Chen, Ming Jiang, and Qi Zhao. Leveraging human attention in novel object captioning. In *IJCAI*, 2021. 1, 2

[7] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Image captioning with unseen objects. In *BMVC*, 2019. 1, 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 6, 7

[9] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 2, 3, 4, 7

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[11] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 2, 6

[12] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1575–1583, 2021. 1, 2, 3, 4, 6, 7, 8

[13] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1

[14] Jiacheng Li, Chang Chen, and Zhiwei Xiong. Contextual outpainting with object-level contrastive learning supplementary material. 3

[15] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming

Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 3, 6

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2, 6, 7

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[18] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 1, 2

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3, 4, 7

[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[22] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 6

[23] Mikihiro Tanaka and Tatsuya Harada. Captioning images with novel objects via online vocabulary expansion. In *ECCV*, 2020. 1

[24] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich image captioning in the wild. In *CVPR Workshop on DeepVision*, 2016. 1

[25] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017. 1, 2

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1

[27] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge, 2022. 1, 2, 3, 6, 7

[28] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 6

[29] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2, 4

[30] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 3

[31] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1029–1037, 2018. 1, 2

[32] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 6

[33] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1, 2, 6, 7, 8