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

Anonymous CVPR submission

Paper ID 5616

Abstract

We present in this paper a novel Object-level Contrastive Learning in Novel Object Captioning and offer a deeper understanding of object representation. Although existing methods that heavily rely on an object detection model have shown the success to describe objects absent from training data in NOC, we show that this dominant paradigm is still limited by the cross-modality and inter-modality discrepancy for object-level feature learning. To address the issues, we propose focusing to enhance the discriminative object-level feature learning with two simple yet effective means from two core aspects (1) Symmetric triplet loss to guide the training procedures by simultaneously considering the cross-modality and inter-modality discrepancy. (2) Sample selection and re-weight for object vocabulary to improve the object-level features with more discriminability and robustness. Additionally, a reference model with external knowledge is adopted to generate the object vocabulary for contrative samples. The experimental results on two datasets show that our proposed OCL-NOC approach distinctly outperforms state-of-the-art methods by large margins, demonstrating the effectiveness of our OCL-NOC to enhance the discriminative object-level feature learning for NOC.

1. Introduction

Image captioning is an important and hot topic in the multi-modal learning field [**?**, 31, 45], it is the task of describing the content of an image in words. As a fundamental problem, image captioning requires both visual understanding and recognition and natural language processing [20, 46]. In the past few years, image captioning methods based on deep neural networks (DNNs) have made tremendous progress, such as [3, 24, 38]. However, most of these methods cover only limited visual concepts and generalize poorly to in-the-wild images [35].

To effectively describe objects absent from training data and enrich visual concepts, the approaches that rely on object detection model [7,8,18,22,24,34,37,44,48] have been widely proposed, demonstrating breakthroughs in vision-language tasks. It concatenates image-text-tag features as a whole input and utilizes the object tags as external resources to alleviate the burden of the annotated data. Most of the methods enjoy the merit of modern object detection models and improve them by designing specific architectures as well as training strategies [18, 22, 48]. [22] is the first work to adopt the tags as an extra input to achieve vision-language alignment. [48] utilizes an object detection model pre-trained on much larger training corpora to broaden the knowledge with novel categories. [18] enrich novel categories information via resorting to alternative data sources without caption supervision and using image-tag pairs to conduct vision language pretraining.

Figure 1. An example of caption results on Nocaps val set with VIVO and ours. The result show that our proposed method can differentiate between hard-to-distinguish object classes and accurately generate the novel caption given novel objects.

However, these data-driven methods still depend on the large scale and the high quality of the human-annotated dataset, which becomes a burden to apply in the real world. Regarding limited human annotations, the network cannot discern the various appearance within the category and is easily over-fitted to the restricted samples, which generates a novel but not accurate or completely irrelevant description for some confusing categories. Approaches like contrastive learning could be an alternative way to achieve better feature alignment and enhance discriminative feature learning. Given an anchor and its corresponding positive and negative samples, when projecting them into the joint space, it is hoped that the distance between the anchor and positive samples can be similar, and anchor and negative samples

can be far away from each other. Many implementations of this principle have been proposed: max-margin loss [11], triplet loss [16, 40, 42], and InfoNCE [36]. Usually positive pairs are defined as synthetic, spatial [5, 29], or temporal variations of an instance [29]. Instance discrimination has also been applied to cross-modal tasks, where positive pairs (or a set of positives) are sampled from the same time window (MILNCE [26], AVSA [27], CM-ACC [25]).

Despite all the progress, few works pay attention to the the discriminability of object-level descriptions in novel object captioning tasks. In this study, we find that the conventional NOC (Novel Object Captioning) models tend to generate over-generic captions or even identical captions when input objects are actually different. As shown in figure 1) (b), given two similar but different images, VINVL [48]+VIVO [18] generates the wrong caption "`A cute raccoon sitting in a tree branch`" result though conditioned on the right objects hint "`red panda`" while our method could give accurate descriptions "`A cute red panda sitting in a tree branch`". Obviously, previous works fail to align the different modalities and capture the object-level discriminability of the target image. To address the problems, we present in this paper a novel OCL-NOC to enhance the object-level representations by exploiting better feature alignment and increasing the object-level feature discriminative ability by contrasting positive tag pairs from sets of negative tags. In addition, we leverage [18] as our reference model to extract the multi-modal semantic relationship. Given region features and object-descriptions context, our reference model will find the most likely and unlikely objects based on multi-modal semantics.

To fully enjoy the merit of object-level representation, we propose the symmetric module CIAM (Cross and Inner modality Alignment Module) based on infoNCE [28]. The intuition is to enforce the feature compactness within the class and increase the discriminative across classes, similar to [10], but the target scope is different. [10] do not take the inner-modality interactions into account while our method leverages both inner-modality and cross-modality alignment. The lack of inner-modality modality will lead to inefficient embeddings, as the caption may also attend to object-level descriptions. Therefore, our method could not only enhance visual concepts by alignment region-tag but also learn caption-tag semantics to connect related caption tokens and enable region-caption interaction. Furthermore, to mitigate the negative effect brought by noisy and biased contrastive samples, we propose the UASR (Uncertainty-Aware Selection and Re-weighting) to focus on the high-reliability samples and avoid issues with false positives/negatives.

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

- We present in this paper a novel Object-level Contrastive Learning in Novel Object Captioning (OCL-NOC) to enhance the feature discriminative ability by contrasting positive tag pairs from sets of negative tags.
- A symmetric module CIAM (Cross-modality and Inner modality Alignment Module) is proposed to reduce the feature discrepancy and enhance interactions for inner-modality and cross-modality. We propose UASR (Uncertainty-Aware Selection and Rweighting) to filter false positives/negatives and enhance the reliable contrastive samples via modeling uncertainty.
- We validate the proposed method on Nocaps and Held-Out COCO benchmarks, which outperforms other state-of-art methods and boosts performance over baseline by a large margin.

2. Related work

Novel object captioning. This task aims to describe images with objects that are unseen in the training stage (we define these objects as novel objects) where many methods. [7, 8, 18, 22, 24, 30, 37, 43, 44, 48] have been proposed. The early works such as Hendricks et al. (DCC) [14] and Venugopalan et al. (NOC) [37] 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. (NBT) [24], Wu et al. (DNOC) [44], and Demirel et al. (ZSC) [8] fill the generated template sentence with objects detected by object/novel object detectors. Chen et al. (ANOC) [7] combine object detector and human attention to identify novel objects. In addition, Li et al. (Oscar) is the first to utilize object-level semantics in VLP tasks and further extended by Zhang et al. (VinVL) [48]. Hu et al. (VIVO) [18] builds upon [48] and proposes to leverage region-tag pairs to conduct pretraining.

Other data-driven studies Radford et al. (CLIP) [30] and Wang et al. (CLIP) [43] that don't depend on the object detection results also show strong performance in many VLP tasks. [30] conducts pretraining and utilizes massive data to align the vision and language feature in a single-step manner. [43] is trained with large-scale weak supervision and a single prefix language modeling objective. However, we find the large-scale data-driven method tends to reach its performance bottleneck in relatively low-data regimes while the model that relies on object detection results could still demonstrate good performance. The potential of the object-level semantics hasn't been fully exploited.

Contrastive Learning Contrastive learning could be an alternative way to enhance object-level feature learning where many methods have shown their effectiveness. The early work such as Oord et al. [28] proposes Contrastive Predictive Coding (CPC) that learns representations for se-

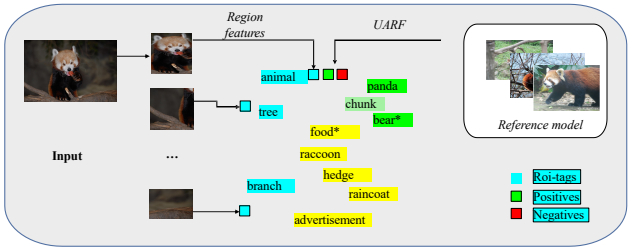

Figure 2. An example of contrastive samples generation process. The cyan, red, and green denote ROI-tags, positive tags, and negative tags separately. $*$ denotes the corresponding false positives/negatives.

quential data. Hjelm et al. [15] presents Deep InfoMax for unsupervised representation learning. More recently, Chen et al. [6] proposes to learn visual representations by maximizing agreement between differently augmented views of the same image via a contrastive loss. He et al. [13] proposes Momentum Contrast (MoCo) for unsupervised visual representation learning. Wei et al. [**?**] utilizes object-level contrastive learning in detection task. Li et al. [21] focuses on the problem of contextual outpainting. However, few works pay attention to object-level feature learning in VL tasks, especially for the NOC task, which we argue is crucial to the quality of generated captions.

In addition, sample selection also plays an important role in contrastive learning. Han et al. [12] introduces a co-training method to mine hard positive samples by using other complementary views of the data for video representation learning. Recent works Kalantidis et al. and Robinson et al. [19,33] explore informative (hard) negatives to facilitate better and faster contrastive learning. Zolfaghari et al. [49] focuses on selecting reliable negative samples in the multi-video representation task. Our method also benefits from sample selection. In fact, the ROI-tags extracted from the object detection model often contain many unrelated or even wrong descriptions, our contrastive samples also have object-level noises. Thus, we propose UASR and focus on filtering false positives/negatives and mitigating the negative effect of unreliable contrastive samples via modeling uncertainty.

3. Proposed Methods

Image captioning models have achieved impressive results in offering novel descriptions via utilizing the object detection as external source. But they lack the cruicial discriminability to differentiate between hard-to-distinguish object classes while composing familiar constituents in the wild. For example, the models trained on COCO Captions can accurately describe images containing objects such as “raccoon”, “branch”, or “tree”, but fail to generate a reasonable caption for any image containing the hard-to-distinguish classes like “red panda” and ”raccoon” due to

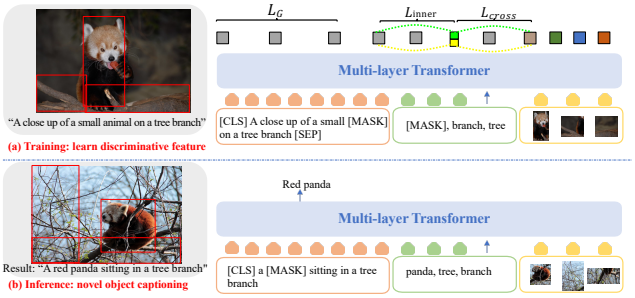

Figure 3. The main pipeline of our proposed OCL-NOC, including training and inference: In addition to the normal caption loss L_G in training process, we compute the cross-modality and inner-modality loss L_{cross} and L_{inner} by contrasting the positive pairs against the negative pairs. The corresponding relationship is denoted in green and red line separately. During inference, our model could accurately generate the caption that contains novel objects.

the lack of object-level discriminability for wild images.

To address the problem, we propose a novel contrastive learning paradigm to enhance the discriminative object-level feature learning and explicitly align different modalities. Our approach uses a two-stage scheme that consists of contrastive sample generation and training with contrastive learning. Fig. 2 and Fig. 3 illustrates our approach using an example. First, in the generation stage (Fig. 2), a reference model with external knowledge is adopted to retrieve the positive tags and negative tags given a correspoding image, where the objects ”panda”, and ”tree” are included. Then in the training stage (Fig. 3(a)), given image-caption pairs, contrastive samples (e.g., positives “panda” and “tee”, negatives “raccoon” and “hedge”) from stage1, the model will pull in the distance with positive tags and image push away the negative tags. e.g., for “[A] on a tree branch ...”, [A] could attend to object tags and the object tags may refer to novel visual objects that are unseen in image-caption pairs.

Thus, our model could achieve modality alignment and discriminability, allowing the model to differentiate between hard-to-distinguish object classes while composing familiar constituents. As shown in (Fig. 3(b)), at inference time the model can effectively utilize the detected novel object (e.g., “red panda”) and form an accurate caption ”A red panda sitting in a tree branch”. In addition, as shown in the figure (Fig. 3(a)), the cross-modality contrastive loss L_{cross} and inner-modality contrastive loss L_{inner} is added after getting caption loss L_G. Moreover, CIAM (Cross-modality and Inner modality Alignment Module) could guide the training procedures by simultaneously considering the cross-modality and inter-modality discrepancy since both the visual concepts and word representation may attend to object tags. Another module, UASR (Uncertainty-Aware Selection and Re-weight) could filter false positives/negatives while simultaneously mitigating the effect of unreliable samples. In what follows, we describe our meth-

ods step by step.

3.1. Contrastive Samples Generation

In this paper, we focus on contrasting the positive pairs against negative pairs whose tags are replaced with plausible negative tags. However, extracting contrastive tag samples that are related to the corresponding image is challenging as it requires the semantic understanding of tags in multi-modal context. Here, we leverage [18] as our reference model to perform contrastive samples generation (we adopt [18] due to its rich object-level semantics and this reference model could also [30,43], the comparison results are shown in ablation study)

Given a tag g and corresponding context image region features $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^{K}$ and tag tokens $\mathbf{T}' = \{\mathbf{t_j}\}_{\mathbf{j=1,j\neq m}}^{\mathbf{T}}$, we define contrastive tag sample as one which has the same context but a different substituted tag g' with the following properties: the positive tag should be plausible in the context and the negative tag should be untrue for the given context. As shown in Fig 2, consider the detected tags "animal, tree, branch" where g is chosen as "animal" and tag context is defined by "[MASK], tree, branch" where [MASK] is the token that denotes a missing tag. A potential candidate for a substituted tag might be "panda, tree, branch" where g' is "panda". Where "hedge" and "raincoat" are negative samples which are not plausible given the context.

Constructing contrastive samples. Our approach for generating such tags can be divided into two steps: First, we feed the image region features and tag tokens context into the reference model to extract 50 most related result $\{g'_l\}_{l=1}^{50}$ using probabilities $p(g'|context)$ predicted by reference (for better illustration, we use the abbreviation$p(g')$ to denote $p(g'|context)$ in the following section. Intuitively, these tags with high-rank are corresponding to the context according to the reference model while the low-rank tags can be regarded as outliers. Thus, in the second step, we divide the original $\{g'_l\}_{l=1}^{50}$ as positive tags $\{w_n\}_{n=1}^{25} = \{g'_l\}_{l=1}^{50}$ and negative tags $\{s_l\}_{l=1}^{25} = \{g'_l\}_{l=26}^{50}$.

We empirically find that the proposed approach is effective in extracting positive and negative tags given multi-modal context. Fig 2 shows a contrastive sample result with corresponding image features and tags. The green and red tags denote the positive and negative tags extracted from the reference model. $*$ denotes the corresponding false positives-negatives, which we will discuss in the following section.

3.2. Cross and Inner modality Alignment Module

The idea of utilizing object-level semantics in VL tasks is firstly introduced in OSCAR [22] and further extended by VinVL [48]. Both OSCAR and VinVL utilize detected object tags and concatenate image-text-tags features as input to the model to achieve alignment. In OCL-NOC, region-tag alignment is learned in an explicit manner, we further learn caption-tag semantics to connect related caption tokens and enable region-caption interaction.

Prior to introducing our specific loss function, we will first introduce a compatibility function to measure the similarity and take the region-tag feature pairs for example. We define the compatibility function as $\phi_\theta(\mathbf{V}, t_j) = V_k^T(t_j)V_{att}(\mathbf{V}, t_j)$ and V_{att} is the region representation with attention scores defined in Eq. 1.

$$V_{att}(\mathbf{V}, t_j) = \sum_{k=1}^{K} a(v_k, t_j)v_k. \tag{1}$$

Here v_k is the region embedding extracted from an embedding layer. The attention scores $a(v_k, t_j)$ are used as a soft selection mechanism to compute a tag-specific region representation using a linear combination of the region. The definitions are as the following.

$$a(v_k, t_j) = \frac{e^{s(v_k, t_j)}}{\sum_{k=1}^{K} e^{s(v_{k'}, t_j)}}, \tag{2}$$

Where $s(v_k, t_j) = t_j^T v_k/\sqrt{d}$, t_j and v_k refer to the corresponding embedding for tag and region, d is the feature dimension. Given the contrastive samples $\{w_n\}_{n=1}^{25}$ and $\{s_l\}_{l=1}^{25}$ from reference model, we can derive our final contrastive loss function L_{cross} in Eq. 3.

$$\mathcal{L}_{\text{cross}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[-\log\left(\frac{e^{\phi_\theta(\mathbf{V}, w_n)}}{e^{\phi_\theta(\mathbf{V}, w_n)} + \sum_{l=1}^{25} e^{\phi_\theta(\mathbf{V}, s_l)}}\right)\right]. \tag{3}$$

$$\mathcal{L}_{\text{inner}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[-\log\left(\frac{e^{\phi_\theta(\mathbf{C}, w_n)}}{e^{\phi_\theta(\mathbf{C}, w_n)} + \sum_{l=1}^{25} e^{\phi_\theta(\mathbf{C}, s_l)}}\right)\right]. \tag{4}$$

The formulation for tag-caption pairs is similar but has a little difference. For tag-caption, we enforce contrastive learning only between the tag and noun caption words, but not between the region and all caption words. This is because the caption has many unrelated tokens such as "the, a, is", which will harm our alignment process and cause redundant compute costs. Finally, we can derive our final contrastive loss function L_{inner} in Eq. 4, where $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^{P}$ is the caption words representation.

3.3. Uncertainty-Aware Selection and Re-weighting

It is natural for the raw contrastive samples to exist errors such as false-positives and false-negatives. These noisy samples would harm the discriminability and robustness of our contrastive learning process significantly. Therefore, designing a sample selection strategy to deal with the noisy contrastive samples (e.g., false positives/negatives) is desirable. On the one hand, recent work [39] focuses on selecting useful positive object-samples related to corresponding

regions but ignores crucial discriminability for object vocabulary. On the other hand, [49] adopts a image-level similarity function to remove false-negative samples with fixed threshold but the unchanged and fixed threshold will arbitrarily some useful negatives under the threshold. Though these methods show their effectiveness in dealing with noisy samples to some extent, they are still hindered by their inherent weakness. Our proposed UASR could not only select the reliable positive and negative object tags to improve the object-level features with more discriminability and robustness, but enhance the reliable con- trastive samples via modeling uncertainty from multi-modal semantics instead of using a fixed threshold.

Uncertainty-Aware Selection Specifically, we propose "L2 Normalization + dot product operation" uncertainty scores Eq(5) to filter false positives/negatives, it is equivalent to calculating cosine similarity and calculating the correlation between samples. The stronger the correlation, the high uncertain the negative sample is and more reliable for the positive samples.

$$\mathrm{u}(\mathbf{v}, \mathbf{t}) = \frac{\mathbf{v}^\top \mathbf{t}}{\|\mathbf{v}\| \, \|\mathbf{t}\|}, \tag{5}$$

Then we retreive M samples as $h = \{t_l\}_{l=1}^{M}$ for given M region features $\{\mathbf{v}_i\}_{i=1}^{M}$. Where $\arg\max$ picks the index j of the tag in the s which has the highest probability score with corresponding region feature $\{\mathbf{t}_i\}$.

$$j = \arg\max_j(\mathrm{u}(\mathbf{v}_i, \mathbf{s}_j)), \quad \hat{t}_i \leftarrow t_j, \tag{6}$$

If the score between region and tags are high, the corresponding top-M tags will be listed as confusion samples for negatives (actually positive), and these confusion samples will be removed from the original s. Similarly, we will also choose the reliable positives based on these top-M tags since samples with low-similarity score can be regarded as outliers to the dataset and they are too sparse to positively influence the shape of the embedding. Finally, we can derive Eq(7), where $\setminus$ is the set operation of removing element, $\cap$ is the intersection set operation. $\bar{s}$ and $\bar{w}$ is the final negative and positive samples.

$$\bar{s} = s \setminus h, \bar{w} = w \cap h \tag{7}$$

Uncertainty-Aware Re-weighting Moreover, we suggest using the uncertainty function 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. 8.

$$q_1(\bar{w}_n) = \exp\left(\mathrm{u}(\mathbf{v}_k, \bar{\mathbf{w}}_n)\right), q_2(\bar{w}_n) = p_n^1 \tag{8}$$

$$k = \arg\max_k(\mathrm{u}(\mathbf{v}_k, \bar{\mathbf{w}}_n)), \quad \hat{v}_i \leftarrow v_k \tag{9}$$

Where q_i is the uncertainty score computed by the most relevant region with the corresponding tag. Moreover, we further consider how certain each sample is and combine the cross-modality uncertainty score to reduce the influence of unreliable samples in Eq. 10. Where $q_i^2 = p(w_n)$ is the corresponding classification probabilities.

$$\bar{q}(\bar{w}_n) = q_1(\bar{w}_n) * q_2(\bar{w}_n) \tag{10}$$

Finally, we could re-write our function in Eq. 11, Eq. 12.

$$\mathcal{L}_{\texttt{cross}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[q(\bar{w}_n) - \log\left(\frac{e^{\phi_\theta(\mathbf{V}, \bar{w}_n)}}{e^{\phi_\theta(\mathbf{V}, \bar{w}_n)} + \sum_{l=1}^{25} e^{\phi_\theta(\mathbf{V}, \bar{s})}}\right)\right]. \tag{11}$$

$$\mathcal{L}_{\texttt{inner}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[q(\bar{w}_n) - \log\left(\frac{e^{\phi_\theta(\mathbf{C}, \bar{w}_n)}}{e^{\phi_\theta(\mathbf{C}, \bar{w}_n)} + \sum_{l=1}^{25} e^{\phi_\theta(\mathbf{C}, \bar{s})}}\right)\right]. \tag{12}$$

4. Experiments

4.1. Experimental Setup

Datasets. Our main experiments and ablation studies are based on the Nocaps [1] dataset. We use the Open Images V5 challenge training set, which has 1.7M images. Following [18], we also add the image-level tags to pretrain a reference model, 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 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 [14], which is a subset of MS COCO [23] 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.

Evaluation. We use the object detector from [48] 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). We use a maximum of 50 image regions and 15 tag tokens per image following [18] to conduct pre-training. The model is trained for 100K iterations with a batch size of 1024 and a learning rate of 4 × 10-5. In fine-tuning, we set the maximum caption length to 40 and the maximum tag length to 30. The model is trained

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 by our method are better than those by human in most cases. We note that our results on 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 [22]	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 [18]	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 [48]	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 [18, 48]	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 [39]	**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

for 30 epochs with a batch size of 512 and a learning rate of 1 × 10-4 , optimized using the cross-entropy loss and our contrastive loss. To further boost the performance, we also perform the SCST optimization (Rennie et al. 2017) 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 V-100 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 Nocaps benchmark. The other compared state-of-art results are from [39]. Note that all the methods use the Bert-base [9] for a fair comparison. We compare our method with UpDown [1, 4], OSCAR [22],VINVL [48],VIVO [18],NOC-REK [39], which holds 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 our model is optimized using SCST [32]. Our generated captions also adopt Constrained Beam Search (CBS) following [2].

In Table 1, we present our results of SPICE and CIDEr scores 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 and OCL-NOC is our final result after CIDER optimization. Noted that we won't compare with data-driven methods [17, 41, 47] which use the paired caption-images and fail to follow the rule of Nocaps. By leveraging contrastive learning in fine-tuning stage, our method has achieved significant improvement compared to all prior works. It is worth noting that our first stage result could generate a comparable result with [18, 22, 48]. After the CIDER optimization process, our method could overpass the last state-of-art method [39] over a large margin, which are 5.07, 1.2(CIDER,SPICE) better than the [39] in Open-Images 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 tags and ground-truth tags in 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 in F1-score substantially, especially for out-of-domain objects. This again verifies the effectiveness of OCL-NOC contrastive learning to describe and discriminate novel objects.

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

Results on the Held-Out COCO dataset. To further prove the generalization ability of our method, we also conduct experiments on Held-Out COCO dataset. As we can see from Table 3, our method consistently beats baseline

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

Methods	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	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**

VINVL over a large margin, the improvements are 7.7, 2., 5.3, and 10.0 with Bert-base in the F1, SPICE, METEOR, and CIDEr metrics. Additionally, our method is superior to all the other state-of-art methods. In particular, compared with recent state-of-art method [39], the recent state-of-art on the Held-Out COCO dataset, OCL-NOC achieves a 4.4 improvement in CIDEr (from 138.4 to 142.8) and 3.1 improvement in METEOR (from 32.8 to 35.9). The 3.2 improvement in the F1 score also suggests that OCL-NOC performs better in describing novel objects.

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 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 the different components. To further understand the advantages brought by different components. We conduct ablation studies step by step and progressively examine every component's effectiveness. Table 4 reports the results. By training plain framework without strategy described in section 3.1, 3.2, 3.3, we could achieve $93.0, 13.1$ in CIDEr,SPICE separately. Besides this, the CIDEr, SPICE from the plain framework could be further improved $1.6, 0.2$ by training with the raw contrastive samples (CSG), which can be attributed to taking the merit of the rich semantic information in reference mode. On top of CSG, by merging CIAM into our framework, it further obtains improvements of $95.9, 13.5$ by enhancing the modality interactions and alignment. Additionally, UASR further boosts performance to $98.1, 13.9$ by filtering the raw contrastive sample and enhancing the reliable one with uncertainty estimation, which shows the effectiveness of UASR.

Effectiveness of the external resource. By utilizing the external knowledge from the referenece model to extract multi-modal semantics, we can introduce more novel information to our contrastive learning framework. To further prove the generalization ability of our method, We change the external knowledge for contrastive samples generation and add comparison experiments of CLIP [30], NOC-REK [39] to retrieve the contrastive samples. For CLIP [30], we feed the whole image into the model and derive the ranking probabilities $p()$. For NOC-REK [39], we get $p()$ through similarity-based retrieval. The result shows that using NOC-REK to generate the contrastive samples could only slightly improve our framework, which brings $1.4, 0.1$ gains compared with the baseline. This may be caused by the limited linguistic knowledge defined in Bert, which may only cover limited novel categories and lack the multi-modality understanding of external knowledge. To get more sensible and rich contrastive samples, a natural idea is to generate contrastive samples by heavy-pretrained model CLIP [30] due to its strong zero-shot capabilities for novel categories. which could bring $4.5, 0.4$ extra boosts compared with the baseline. Finally, we empirically find that [18] behaves best for extracting object-level contrastive samples. The reason is that [18] is pretrained on corpora with rich object-level information and the cross-modality alignment ability for [18] between representations of image regions and tags.

Table 4. Ablation study on the effectiveness of different components, including CSG (Contrastive Samples Generation), CIAM (Cross-modality and Inner modality Alignment Module), URAM (Uncertainty guided Reiweighting And Filtering strategy).

CSG	CIAM	UASR	Overall	
			CIDEr	SPICE
			94.3	13.1
✓			96.4	13.3
✓	✓		99.1	13.6
✓	✓	✓	102.5	14.1

Table 5. Impact of the size of the external knowledge for the reference model. Changes in the type of reference model result in changes in performance. Higher score is better.

Reference model	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
NOC-REK	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

Effectiveness of the CIAM. We ablate each component of CIAM step by step. Table 6 shows that directly using cross-modality contrastive loss could bring $3.5, 0.5$ boosts in CIDEr, SPICE by aligning region-tag features. On top of IAM, further extending our method into CIAM yields $1.6, 0.2$ gains, which may be attributed to inner-modality alignment for caption-tag features. Finally, our method improves the performance by $5.1, 0.7$ by combining IAM and CAM. This is because CIAM could not only achieve region-tag and caption-tag alignment but also enable region-caption interaction to some extent, the result shows that the proposed CIAM is a more powerful tool for the novel object caption task.

Effectiveness of the UASR. We investigate the performance of our UASR (Uncertainty-Aware Selection And Reweighting) on the Nocaps validation in this part and add UAF_N experiment to ablate the selection process separately. Table. 7 shows that UAR could effectively enhance the reliable contrastive samples and bring $3.5, 0.5$ boosts in CIDEr, SPICE. In addition, the false positives/negatives severely harm our training process. By further incorporating UAF_N into our method, our model's performance is boosted to $97.2, 13.7$ in CIDEr, SPICE. When filtering false positives/negatives simultaneously, we could reach $98.9, 13.9$ in CIDEr, SPICE. Finally, our method improves

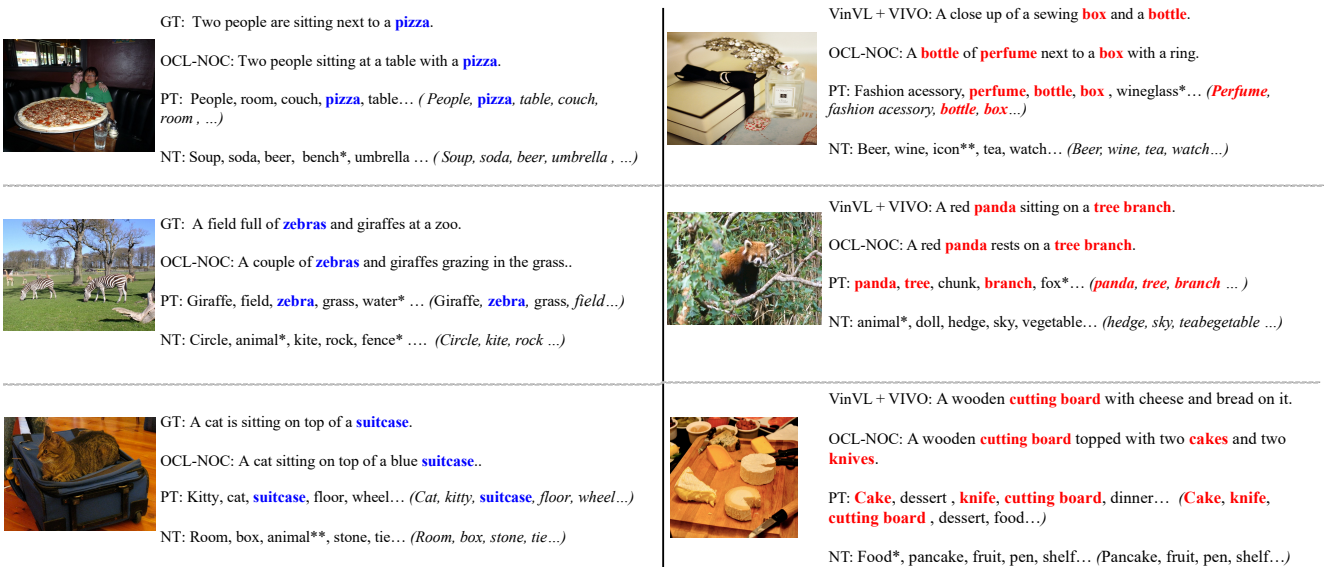

Figure 4. Examples of generated captions and contrastive samples by compared methods on Held-Out COCO (left) and Nocaps (right). We show the ground-truth captions (GT) on Held-Out COCO for reference. PT and NT denote the positive tags, positive tags after UASR and negative tags, negative tags after UASR. **Blue**/**Red** texts a indicate novel objects in Held-Out COCO/Nocaps and $*$ indicates the false positives/negatives.

Table 6. Effectiveness of the CIAM. Performance analysis over CIAM. CAM: Cross-modality Alignment Module, IAM: Inner-modality Alignment Module, CIAM: Cross-modality and Inner-modality Alignment Module

Reference model	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

Table 7. Effectiveness of the UASR. Performance analysis over UASR. UAR: Uncertainty-Aware Re-weight, UAF: Uncertainty-Aware Filtering, UAF_N: Uncertainty-Aware Filtering (only filter false-negatives), UASR: Uncertainty-Aware Selection and Re-weight

Reference model	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
UAR	105.3	14.1	97.0	13.6	90.1	12.3	97.1	13.5
UAF_N	105.2	14.2	96.1	13.7	90.4	12.6	97.2	13.7
UAF	107.8	14.3	98.0	13.8	96.6	13.0	98.9	13.9
UASR	109.7	14.7	100.86	14.2	100.7	13.2	102.5	14.1

the performance by $5.1, 0.7$ by combining UAR and UAF, which illustrates the proposed UASR is a more powerful tool for the novel object caption task.

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 contrasting the positive pairs with negative pairs. On Nocaps, VinVL+VIVO [18, 48] sometimes cannot include the novel objects in the captions (first and third examples) or generates wrong caption (second example), this shows our OCL-NOC could successfully generate correct, and coherent captions via differentiating between hard-to-distinguish object classes. We show top-5 Positive and Negative Tag results extracted from the reference model for better illustration.

5. Conclusion

In this paper, we propose a novel Object-level Contrastive Learning in Novel Object Captioning called OCL-NOC to enhance the discriminative object-level feature learning with two simple yet effective means from two core aspects. Our proposed CIAM not only enforces region-tag alignment in an explicit manner, but also further learns caption-tag semantics to connect related caption tokens and enable region-caption interaction. In addition, we propose UASR to select the reliable positives and negatives and mitigate the effect of the low-quality contrastive samples via modeling uncertainty. 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 detection tags has already been proved in recent works. However, there are few studies to further explore why the tags could perform well and this problem 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. 5, 6

[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, 6

[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. 2

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

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

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

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

[10] 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. 2

[11] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. 2

[12] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Neurips*, 2020. 3

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

[14] 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, 5

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. 2019. 3

[16] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco

Loog, editors, *Similarity-Based Pattern Recognition - Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015, Proceedings*, volume 9370 of *Lecture Notes in Computer Science*, pages 84–92. Springer, 2015. 2

[17] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 6

[18] 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, 4, 5, 6, 7, 8

[19] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard Negative Mixing for Contrastive Learning. *NeurIPS*, 2020. 3

[20] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. Collective generation of natural image descriptions. In *ACL*, 2012. 1

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

[22] 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, 4, 6

[23] 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. 5

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

[25] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Learning audio-visual representations with active contrastive coding, 2020. 2

[26] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020. 2

[27] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment, 2020. 2

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[29] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *CoRR*, abs/2008.03800, 2020. 2

[30] 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, 4, 7

[31] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017. 1

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

[33] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive Learning with Hard Negative Samples. *ICLR*, 2021. 3

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

[35] 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

[36] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2

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

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

[39] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge, 2022. 4, 6, 7

[40] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. *CoRR*, abs/1404.4661, 2014. 2

[41] 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

[42] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. *CoRR*, abs/1511.06078, 2015. 2

[43] 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

[44] 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

[45] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. 1

[46] Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011. 1

[47] 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

[48] 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, 4, 5, 6, 8

[49] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021. 3, 5