NOC-REK: Novel Object Captioning with Retrieved Vocabulary from External Knowledge

Duc Minh Vo
The University of Tokyo, Japan
vmduc@nlab.ci.i.u-tokyo.ac.jp

Hong Chen
The University of Tokyo, Japan
chen@nlab.ci.i.u-tokyo.ac.jp

Akihiro Sugimoto
National Institute of Informatics, Japan
sugimoto@nii.ac.jp

Hideki Nakayama
The University of Tokyo, Japan
nakayama@ci.i.u-tokyo.ac.jp

Abstract

*Novel object captioning aims at describing objects absent from training data, with the key ingredient being the provision of object vocabulary to the model. Although existing methods heavily rely on an object detection model, we view the detection step as vocabulary retrieval from an external knowledge in the form of embeddings for any object's definition from Wiktionary, where we use in the retrieval image region features learned from a transformers model. We propose an end-to-end **N**ovel **O**bject **C**aptioning with **R**etrieved vocabulary from **E**xternal **K**nowledge method (NOC-REK), which simultaneously learns vocabulary retrieval and caption generation, successfully describing novel objects outside of the training dataset. Furthermore, our model eliminates the requirement for model retraining by simply updating the external knowledge whenever a novel object appears. Our comprehensive experiments on held-out COCO and Nocaps datasets show that our NOC-REK is considerably effective against SOTAs.*

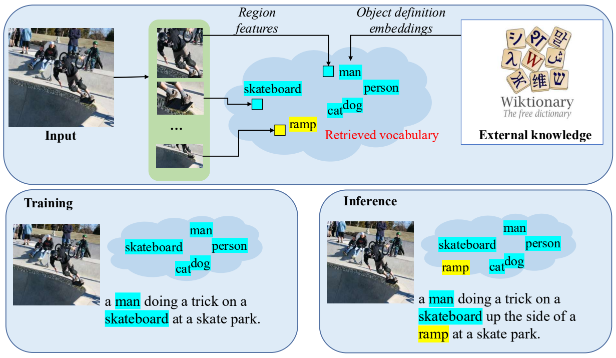

Region features
Object definition embeddings
Input
...
skateboard
man
person
cat
dog
ramp
Retrieved vocabulary
Wiktionary
The free dictionary
External knowledge
Training
a man doing a trick on a skateboard at a skate park.
Inference
a man doing a trick on a skateboard up the side of a ramp at a skate park.

Figure 1. (upper part) Our NOC-REK first retrieves (object) vocabulary by computing similarity of region features and object definition embeddings from an external knowledge. The caption is then generated using those retrieved vocabulary and region features. (below part) At training phase, the model is trained on paired image-caption data and external knowledge where both of them cover a limited number of objects (blue rectangles and texts). At inference time, our model generally describes novel objects by simply updating the external knowledge (yellow text) without any extra training.

1. Introduction

Describing *novel objects* not observable in training data expands the real-world applications of image captioning models. As we progress toward the new goal, the image captioning task shifts to novel object captioning. The conventional image captioning models [1–3], however, fail to describe novel objects because they only learn the correspondence between images and sentences provided in the image–caption pair data. To overcome this limitation, the most straightforward way is to collect more data [4, 5] so that images and captions contain the novel objects that we need; and then retrain the whole system. The newly collected data will provide the model with at least object names that appear in the given image [6]. The caption will then include novel objects. Even if such methods [4, 5] are effective to some extent, the data collection process is highly expensive and time-consuming. How to efficiently provide object vocabulary to the model is, therefore, desired.

A less expensive yet more effective approach [6–12] that relies on object detection model has been proposed, demonstrating breakthroughs in this task. Despite the fact that modern object detection models (e.g., Faster RCNN [13], zero-shot object detection [14]) can recognize a wide range of objects including novel ones, using object detection in novel object captioning models [3–12] brings a new challenge. Additional efforts are still required to improve a portion of the detection model in order to broaden the model's knowledge of novel objects. Such efforts become massive as the number of objects is unlimited in the wild. This study aims to reduce such workload whenever a new object becomes available by providing object vocabulary during inference without additional training or data.

Humans usually build our knowledge of objects by defining any object at a concrete level based on its appearance (e.g., *cat*, *dog*) or an abstract level in conjunction with other objects (e.g., *kitchen*). Therefore, we can easily match an object and its definition, answering its name, regardless of whether we have previously encountered the object. Motivated by that fact, we simplify the object detection by viewing it as vocabulary retrieval from a set of objects' definitions (hereafter referred to as external knowledge). As shown in Fig. 1, we match region features and object definition embeddings using their similarity, returning relevant vocabulary. Note that the dataset is used to train our model only to cover a limited number of objects. When a new object appears after training, we simply update the external knowledge at a much lower cost than previous methods. Our model, therefore, can generally describe any image that contains the novel object (e.g., *a man doing a trick on a skateboard up the side of a* ***ramp*** *at a skate park* in Fig. 1).

We propose **N**ovel **O**bject **C**aptioning with **R**etrieved vocabulary from **E**xternal **K**nowledge method, abbreviated as NOC-REK. Inspired by significant advances of transformers [15] in object detection [16] and vision-language models [10–12], NOC-REK makes full use of a shared-parameter transformers model to unify vocabulary retrieval and caption generation steps in an end-to-end manner. More specifically, we prepare the external knowledge by using object definitions from Wiktionary[1] and a pre-trained BERT model [17]. NOC-REK first learns region features from a set of Faster R-CNN [13] regions of interest (ROIs). Then, we perform vocabulary retrieval using object definition embeddings (in the external knowledge) and region features by computing their similarity. Finally, the caption is generated using the retrieved vocabulary and region features (Fig. 1). We train the vocabulary retrieval using Hungarian loss [16] with our modification, while at the first training stage, we use cross-entropy, and at the second training stage, we use SCST [18] with a reward for the appearance of novel objects to train caption generation. Our contributions are:

- We simplify the object detection step used in the image captioning model by viewing it as vocabulary retrieval from external knowledge.
- We propose an end-to-end NOC-REK model which retrieves vocabulary from external knowledge and generates captions using shared-parameter transformers.
- Our method provides object vocabulary during inference, effectively eliminating the necessity for either retraining model or extra image and/or caption data when a novel object appears.

2. Related work

2.1. Novel object captioning

This task aims at describing objects unseen during the training phase (called novel objects) where many methods [3–12] have been proposed. Hendricks et al. (DCC) [4] and Venugopalan et al. (NOC) [5] purposely use unpaired labeled image and sentence data to learn semantically visual concepts. Lu et al. (NBT) [3], Wu et al. (DNOC) [7], and Demirel et al. (ZSC) [8] fill the generated template sentence with objects detected by object/novel object detectors. Tanaka et al. (OVE) [6] propose a low-cost method to expand word embeddings from a few images of the novel objects. Chen et al. (ANOC) [9] combine object detector and human attention to identify novel objects. On the other hand, Li et al. (Oscar) [10], Hu et al. (VIVO) [11] and Zhang et al. (VinVL) [12] pre-train large-scale vision-language transformers models and then finetune the pre-trained model to adopt downstream tasks.

Generally, the above mentioned methods follow a two-step approach where the detection step is prior to the caption generation step. The former step leverages off-the-shelf object detectors such as Faster RCNN [13], zero-shot object detection [14] to identify objects, which requires additional training for new objects. The latter step employs either LSTM [3–5, 7, 8] or transformers [10–12] in which transformers generate better captions. The two steps are trained independently, requiring much efforts to include the novel objects into the caption [5].

Different from the aforementioned methods, we view the detection step as vocabulary retrieval, allowing our method to be trained in an end-to-end way on a transformers model. Moreover, our method does not require any re-training or additional data even when a novel object arises.

2.2. Transformers-based vision-language pre-training models

Recent vision-language methods [10–12] are BERT-like [17] models which successfully learn the vision-language cross-modal by using a concatenated-sequence of words—object tags—visual regions as its input, showing breakthroughs in many vision-language tasks such as image captioning, image-text retrieval. In fact, those models are pre-trained on a large image-text corpus that probably contains novel objects, so their performance on the novel object captioning task is doubly ambiguous. Like [10–12], our method also uses a concatenated-sequence. However, we do not pre-train the model to avoid biases.

[1] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

3. Proposed NOC-REK

3.1. Idea of NOC-REK

We simplify the object detection step, as discussed above, by viewing it as vocabulary retrieval from external knowledge. We then unify vocabulary retrieval and caption generation in an end-to-end manner. Two critical challenges arise here: (1) how to realize a knowledge-based vocabulary retrieval and (2) how to design an architecture that allows the entire model to be jointly end-to-end trained.

In order to tackle challenge (1), the straightforward solution is to compute the similarity of image features and object definition embeddings from external knowledge. The main difficulty of training is to score retrieved vocabulary with respect to ground-truth objects. This is because the order of the ground-truth objects and that of the retrieved vocabulary differs. We thus use the Hungarian loss proposed in [16] to train our model. However, unlike [16], which rejects objects that are not in training data, we introduce a simple yet effective modification to ensure that the model adopts objects that are not in ground-truth.

Transformers-based model shows impressive results in object detection [16], motivating us to build our retrieval upon transformers. At the same time, following the success of transformers-based vision-language models [10–12] in image captioning, we adopt the model proposed in [12] for our caption generation where the input is a concatenated-sequence of words—object tags—ROIs. Therefore, we propose using shared-parameter transformers for both vocabulary retrieval and caption generation, thereby answering (2) (Fig. 2). The usage of shared-parameters reduces training costs while improving the learning of the vocabulary retrieval step. This is thanks to the ability of image features to cross-attend to language information, resulting in better alignment between vision and language spaces prior to performing the retrieval step. In what follows, we describe the precise details of our method as the preceding discussions.

3.2. Knowledge-based vocabulary retrieval

We conceptualize and mathematically formulate our knowledge-based vocabulary retrieval in this section. The technical integration of vocabulary retrieval and caption generation is presented in Section 3.3.

Selection of external knowledge. We use a free dictionary Wiktionary that describes all words using definitions and descriptions to build our external knowledge. We first crawl the definition of each object name in the target dataset. Assuming that we obtain M vocabulary (i.e., object names and our defined ’no object’) with their corresponding definitions, we employ a pre-trained BERT [17] to embed the definition of each word v into a embedding with the size of 1×768: $\mathbf{d} = \mathrm{BERT}(v)$ (’no object’ is set to the vector of all-zeroes.) As a result, each word will be presented as a key-value pair $(\mathbf{d}, v)$. The external knowledge now consists of M key-value pairs.

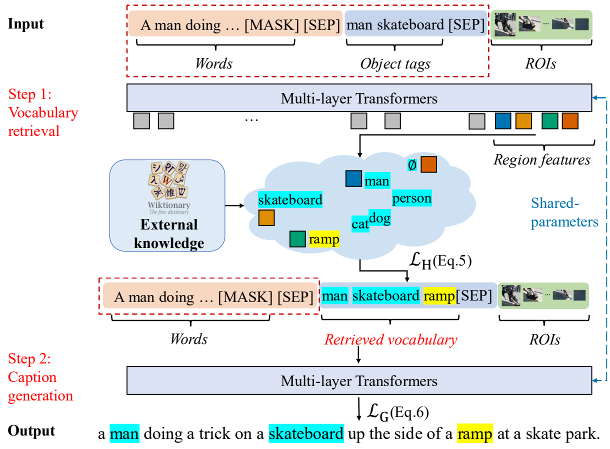


Figure 2. The overall pipeline of our proposed NOC-REK, which includes two steps: Vocabulary retrieval and Caption generation. Our model is made up of multi-layer transformers with shared parameters. The transformers is first used in the knowledge-based vocabulary retrieval step, where it return the objects in the given image. For the caption generation step, we use the same transformers as before. After training, the dashed red parts will be removed. `[MASK]` and `[SEP]` are special tokens to mask some tokens in the sequence of words and separate the functions of different inputs.

Object detection as vocabulary retrieval. We first define the similarity between the region feature and definition embedding $\mathbf{d}$ in the external knowledge:

$$\mathrm{sim}(\mathbf{r}, \mathbf{d}) = \frac{\mathbf{r}^\top \mathbf{d}}{\|\mathbf{r}\| \, \|\mathbf{d}\|}, \tag{1}$$

where $\mathbf{r}$ with the size of 1×768 is learned from NOC-REK as seen in Section 3.3.

Using Eq. (1), we can easily retrieve K vocabulary $\mathcal{V} = \{\hat{v}_i\}_{i=1}^{K}$ for K region features $\{\mathbf{r}_i\}_{i=1}^{K}$ from M pairs of key-value $(\mathbf{d}, v)$. In particular, we first obtain the index j of the vocabulary in the external knowledge which has the highest similarity score with the region feature r_i. Then, we assign v_j as the retrieved vocabulary $\hat{v}_i$. The retrieval process can be summarized as follows:

$$j = \arg\max_{j}(\mathrm{sim}(\mathbf{r}_i, \mathbf{d}_j)), \quad \hat{v}_i \leftarrow v_j, \tag{2}$$

where $j \in [0, M)$. Because an image may contain multiple identical objects, there is no requirement that each vocabulary is distinct from the others. Training the retriever is now to find the optimal bipartite match between the retrieved vocabulary and the ground truth objects. The following is the description of the retriever’s training loss function.

Let us denote $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ be a set of N ground truth objects. Assuming that $N < K$, we expand $\mathcal{Y}$ as a set of size K padded with $\varnothing$ (no object) as in [16, 19]. However, at this time, the expanded $\mathcal{Y}$ solely consists of seen and no objects, preventing the model from learning to retrieve novel objects. We thus introduce a simple yet effective modification that replaces 15% of the number of $\varnothing$ with randomly selected vocabulary from the external knowledge. As a result, our expanded $\mathcal{Y}$ composes of ground truth objects, no objects and novel vocabulary (objects) (now $|\mathcal{Y}| = |\mathcal{V}| = K$). Following [16,19], we first find a bipartite matching between $\mathcal{Y}$ and $\mathcal{V}$ by searching for a permutation of K elements $\sigma \in \mathfrak{S}_N$ with the lowest cost:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{i=1}^{K} \mathcal{C}_{\text{match}}(y_i, \hat{v}_{\sigma(i)}), \tag{3}$$

where $\mathcal{C}_{\text{match}}(y_i, \hat{v}_{\sigma(i)})$ is a pair-wise *matching cost* between y_i and a retrieved vocabulary with index $\sigma(i)$:

$$\mathcal{C}_{\text{match}}(y_i, \hat{v}_{\sigma(i)}) = -\mathbb{1}_{y_i \neq \varnothing} \, \text{sim}(\mathbf{y}_i, \hat{\mathbf{v}}_{\sigma(i)}), \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ is the same with Eq. (1), $\mathbf{y}_i = \text{BERT}(y_i)$, and $\hat{\mathbf{v}}_{\sigma(i)} = \text{BERT}(\hat{v}_{\sigma(i)})$.

Finally, we employ *Hungarian loss* to compute loss for all above pairs matched:

$$\mathcal{L}_{\text{H}}(\mathcal{Y}, \mathcal{V}) = \sum_{i=1}^{K} -\log \text{sim}(\mathbf{y}_i, \hat{\mathbf{v}}_{\hat{\sigma}(i)}), \tag{5}$$

where $\hat{\sigma}$ is the optimal matching solution in Eq. (3). Like [16], we down-weight the log term by a factor of 10 when $y_i = \varnothing$ to avoid class imbalance.

3.3. NOC-REK architecture

Fig. 2 depicts the overall NOC-REK architecture which is surprisingly simple, consisting of *only one* multi-layer transformers model in which the parameters are shared between two steps: (1) Vocabulary retrieval and (2) Caption generation. Like [10–12], we employ a pre-trained BERT model [17] to implement our model because BERT is trained on a large corpus of sentences, resulting in a better understanding of grammar and sentence structure. Following [10, 12], we feed our model a concatenated-sequence of words—object tags—ROIs. This is because, as discussed in [10, 12], the concatenated sequence allows for better alignment of vision and language. We do not pre-train our model on a large text-image corpus to eliminate the model's ability to bias to novel objects. Note that our model only receives ROIs (visual regions) as its input at the testing time.

Pre-processing. For a sequence of words $\{w_i\}_{i=1}^{L}$ (i.e., ground truth sentence), we randomly masked out 15% words of the sequence (maximum 3 words) with a special token [MASK]. For a given image, we use Faster RCNN [13] trained on the COCO dataset to extract a set of N object tags and a set of K ROIs. We remark that N tags are used as ground truth objects $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ in our vocabulary retrieval as described in Section 3.2. Each ROI $\mathbf{f}_i$, in contrast, is a vector with the size of 1×2054 which includes feature 1×2048 (noticing that this is not region feature used in retrieval step), and the region position 1×6 (the coordinates of top-left and bottom-right corners, height, and width).

Step 1: Vocabulary retrieval. This step implements the knowledge-based vocabulary retrieval described in Section 3.2. To ensure that we can retrieve the novel objects added at inference time, we enforce our model as a neural retriever [20–22], allowing the model to optimize the similarity of image features and external knowledge while training. Consequently, we only need to update the external knowledge when new objects appear; retraining the whole model is not required. To this end, we need to optimize both region feature $\mathbf{r}$ and embeddings $\mathbf{d}$ (i.e., training knowledge encoder $\text{BERT}(\cdot)$). However, because updating $\text{BERT}(\cdot)$ during training is costly and our external knowledge is sufficiently smaller than that used in NLP tasks, we instead fix the parameter of knowledge encoder while training the region feature encoder, similarly to [22].

We begin with encoding each word w_i and object tag y_i into a vector 1×768 using an embeddings layer. Simultaneously, for each ROI $\mathbf{f}_i$, we use another embeddings layer to reduce its size from 1×2054 to 1×768. The vectors are then concatenated with two special tokens [SEP] to distinguish their functions. More precisely, we have $\{\mathbf{w}_1, \ldots, \mathbf{w}_L, \texttt{[SEP]}, \mathbf{y}_1, \ldots, \mathbf{y}_N, \texttt{[SEP]}, \mathbf{f}_1, \ldots, \mathbf{f}_K\}$ as an input. After feeding the above input to the model, we obtain K region features $\{\mathbf{r}_i\}_{i=1}^{K}$. Finally, we perform the retrieval using K region features and the external knowledge, returning retrieved vocabulary $\mathcal{V}$.

Step 2: Caption generation. This step aims to generate a caption from the given image and the vocabulary retrieved in the previous step. This step, like the vocabulary retrieval step, takes as its input a concatenated-sequence of words—object tags—ROIs. However, we use our retrieved vocabulary to replace the object tags detected by Faster RCNN [13]. As a result, the input is changed to $\{\mathbf{w}_1, \ldots, \mathbf{w}_L, \texttt{[SEP]}, \hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_K, \texttt{[SEP]}, \mathbf{f}_1, \ldots, \mathbf{f}_K\}$.

Inference. Our model's input is an image that has been preprocessed to obtain ROIs. It automates the steps of vocabulary retrieval and caption generation in an end-to-end manner. To avoid significant changes in input between training and testing, we create $(L + N)$ [MASK] as pseudo words.

3.4. Loss function

We define our loss function as: $\mathcal{L} = \mathcal{L}_{\text{H}} + \mathcal{L}_{\text{G}}$. $\mathcal{L}_{\text{H}}$ is used for the vocabulary retrieval step as defined in Eq. 5 while $\mathcal{L}_{\text{G}}$ works for the caption generation step. $\mathcal{L}_{\text{G}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{G}} = \begin{cases} \mathrm{cross\_entropy}(S, S_{\mathrm{GT}}) & \text{at 1st training stage} \\ \mathrm{CIDEr}(S) + \alpha \times C & \text{at 2nd training stage} \end{cases}, \tag{6}$$

where S, S_{GT} are generated and ground truth captions. $\mathrm{cross\_entropy}(\cdot,\cdot)$ is the cross entropy loss function. Meanwhile, $\mathrm{CIDEr}(\cdot)$ plays as SCST function [18] to improve caption quality (e.g., grammar, structure) and C provides an additive reward for each retrieved vocabulary appearing in the caption. A small α balances the two terms.

4. Experiments

4.1. Implementation and training details

NOC-REK was built with PyTorch, and we used a pre-trained BERT-base model from Huggingfaces [23] for parameters initialization. Our model was trained in an end-to-end manner with two-stage optimization using AdamW, with $\mathcal{L}_{\mathrm{H}}$ used in both stages and $\mathcal{L}_{\mathrm{G}}$ changed with respect to Eq. (6). We set the learning rate to 3e-5 and the batch size to 128 for 30 epochs during the first training stage. The learning rate, batch size, and epoch at the second training stage, in contrast, are 8e-7, 6, and 25, respectively. On a PC with two GTX-3090 GPUs, our model takes 8 days to train. NOC-REK* denotes our method after the first training stage, and NOC-REK denotes the fully-trained model.

We set the length of the word sequence $L = 35$, the number of ground-truth objects $N = 20$, the number of ROIs $K = 50$, and $\alpha = 0.3$. During training, we use seen objects in held-out COCO [4] as the external knowledge ($M = 72 + 1 = 73$). During inference, novel objects in held-out COCO [4] and Nocaps [24] will be added to the knowledge ($M = 600 + 1 = 601$). We retain top-5 retrieved vocabulary with the highest similarity score at inference time (i.e., 5 objects per image). We use CBS [25] with the beam size of 5 to generate caption.

4.2. Compared methods and evaluation metrics

Dataset and compared methods. We evaluate NOC-REK on held-out COCO [4] and Nocaps [24] datasets. Held-out COCO [4] is made by dividing the original COCO [26] into known classes and 8 novel classes which include *bottle, bus, couch, microwave, pizza, racket, suitcase, zebra*. On the other hand, Nocaps [24] is the main challenging dataset for novel object captioning task, which is constructed from 513 classes out of 600 classes of OpenImage. In particular, it consists of 119 classes in COCO dataset that are *in-domain*, 394 classes not in COCO dataset are *out-domain*, and the images that includes both in-domain and out-domain are regarded as *near-domain*. We validate our method using Nocaps validation and test sets.

We compare NOC-REK with state-of-the-art methods: UpDown [1], DCC [4], NOC [5], NBT [3], DNOC [7], ZSC [8], OVE [6], ANOC [9], Oscar [10], VIVO [11], and VinVL [12] in Sections 4.3 and 4.4. We note that all compared scores are from published results, whereas for qualitative comparison, we use publicly available caption results (NOC [27]) and pre-trained models (VinVL+VIVO [28]). In Section 4.5, we go over the performance of two variants of our method: NOC-REK and NOC-REK*.

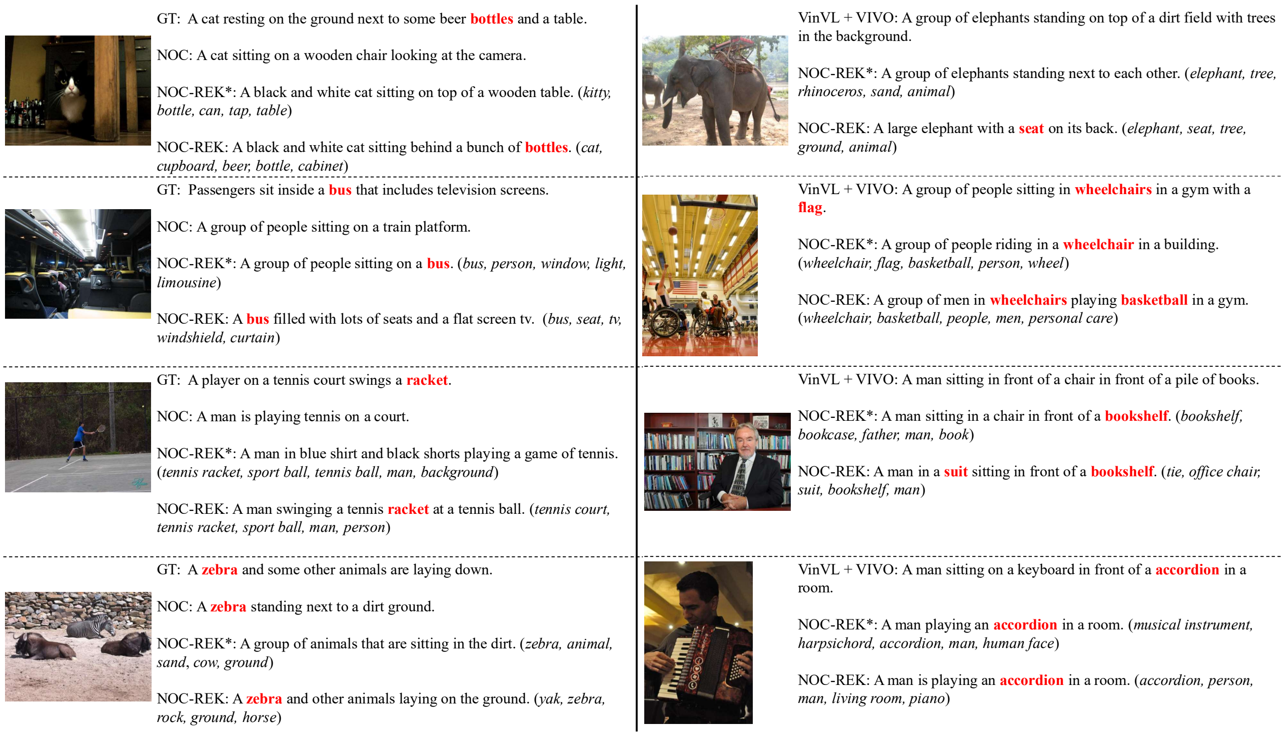
Evaluation metrics. We primarily report CIDEr [29] and SPICE [30] language scores in comparisons on both held-out COCO and Nocaps datasets, drawing on prior work. We also report object detection results on held-out COCO using F1-score and language score METEOR [31].

4.3. Results on held-out COCO dataset

Qualitative evaluation. Fig. 3 (left) shows randomly picked-up captions generated by our method and NOC [5]. We can see that our method successfully retrieves novel objects and includes some of them in the captions in a sensible fashion. Meanwhile, NOC [5] usually fails to generate a caption with a novel object (see first three examples). Furthermore, NOC [5] generates a pretty weird caption (fourth example) that is unrelated to the image's context. This is because NOC [5] uses text-corpus as its external knowledge, resulting in language biases (e.g., a *zebra* is usually described with the word *standing* rather than *laying* in the text-corpus). On the other hand, our method is free of such biases thanks to our usage of object definitions as external knowledge. We also see that our top-5 retrieved vocabulary are reasonable, reflecting objects in the given images. Thanks to end-to-end training, our model can remove less relevant vocabulary in the captions. In more detail, our captions are correct, fluent, and coherent, just like ground-truth captions, proving that our usage of a pre-trained BERT model is reasonable.

Quantitative evaluation. We first evaluate whether a novel object appears in the generated caption using the F1-score. The per-object F1-score and average F1-score of all the compared methods are shown in Table 1 (2nd – 10th columns). Notably, despite the lower cost of updating novel object information, our method yields significantly higher F1-scores than the other methods (an improvement of 12% in average F1-score). This indicates that our method successfully retrieves the novel object from the external knowledge (Fig. 3), and includes those objects in the caption.

Next, we quantitatively evaluate the quality of generated captions using SPICE, METEOR, and CIDEr scores (Table 1, 11th – 13rd columns). Because our method can include novel objects in captions, it is not surprising that we outperform the other methods by a considerable margin. Together with the F1-score, we can conclude that our method outperforms the other methods marginally. More importantly, these observations strongly support the advantage of using an end-to-end model in this task.



GT: A cat resting on the ground next to some beer **bottles** and a table.

NOC: A cat sitting on a wooden chair looking at the camera.

NOC-REK*: A black and white cat sitting on top of a wooden table. (*kitty, bottle, can, tap, table*)

NOC-REK: A black and white cat sitting behind a bunch of **bottles**. (*cat, cupboard, beer, bottle, cabinet*)

GT: Passengers sit inside a **bus** that includes television screens.

NOC: A group of people sitting on a train platform.

NOC-REK*: A group of people sitting on a **bus**. (*bus, person, window, light, limousine*)

NOC-REK: A **bus** filled with lots of seats and a flat screen tv. (*bus, seat, tv, windshield, curtain*)

GT: A player on a tennis court swings a **racket**.

NOC: A man is playing tennis on a court.

NOC-REK*: A man in blue shirt and black shorts playing a game of tennis. (*tennis racket, sport ball, tennis ball, man, background*)

NOC-REK: A man swinging a tennis **racket** at a tennis ball. (*tennis court, tennis racket, sport ball, man, person*)

GT: A **zebra** and some other animals are laying down.

NOC: A **zebra** standing next to a dirt ground.

NOC-REK*: A group of animals that are sitting in the dirt. (*zebra, animal, sand, cow, ground*)

NOC-REK: A **zebra** and other animals laying on the ground. (*yak, zebra, rock, ground, horse*)

VinVL + VIVO: A group of elephants standing on top of a dirt field with trees in the background.

NOC-REK*: A group of elephants standing next to each other. (*elephant, tree, rhinoceros, sand, animal*)

NOC-REK: A large elephant with a **seat** on its back. (*elephant, seat, tree, ground, animal*)

VinVL + VIVO: A group of people sitting in **wheelchairs** in a gym with a **flag**.

NOC-REK*: A group of people riding in a **wheelchair** in a building. (*wheelchair, flag, basketball, person, wheel*)

NOC-REK: A group of men in **wheelchairs** playing **basketball** in a gym. (*wheelchair, basketball, people, men, personal care*)

VinVL + VIVO: A man sitting in front of a chair in front of a pile of books.

NOC-REK*: A man sitting in a chair in front of a **bookshelf**. (*bookshelf, bookcase, father, man, book*)

NOC-REK: A man in a **suit** sitting in front of a **bookshelf**. (*tie, office chair, suit, bookshelf, man*)

VinVL + VIVO: A man sitting on a keyboard in front of a **accordion** in a room.

NOC-REK*: A man playing an **accordion** in a room. (*musical instrument, harpsichord, accordion, man, human face*)

NOC-REK: A man is playing an **accordion** in a room. (*accordion, person, man, living room, piano*)

Figure 3. Examples of generated captions by compared methods on held-out COCO (left) and Nocaps (right). We show the ground-truth captions (GT) on held-out COCO for reference. On held-out COCO, NOC [5] usually fails to generate captions with novel objects (first three examples) or generate caption with not related to image's context (fourth example). On Nocaps, VinVL+VIVO [11, 12] sometimes cannot include the novel objects in the captions (first and third examples) or generates weird caption (fourth example). Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.

Table 1. Quantitative comparison against other methods on held-out COCO dataset. We report F1-score for each novel class (2nd - 9th columns), average F1-score on all novel classes (10th column), and language scores (11th - 13rd columns)^{a}. Higher score is better.

Method	F1-score								Avg. F1-score	SPICE	METEOR	CIDEr
	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra				
DCC [4]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8	13.4	21.0	59.1
NOC [5]	17.8	68.8	25.6	24.7	69.3	55.3	39.9	**89.0**	48.8	–	21.4	–
NBT [3]	7.1	73.7	34.4	61.9	59.9	20.2	42.3	**88.5**	48.5	15.7	22.8	77.0
DNOC [7]	33.0	**76.9**	**54.0**	46.6	75.8	33.0	59.5	84.6	57.9	–	21.6	–
ZSC [8]	2.4	75.2	26.6	24.6	29.8	3.6	0.6	75.4	29.8	14.2	21.9	–
OVE [6]	–	–	–	–	–	–	–	–	–	16.9	22.6	80.8
ANOC [9]	–	–	–	–	–	–	–	–	64.3	18.2	25.2	94.7
NOC-REK*	**39.3**	76.2	47.4	**62.0**	**79.5**	**78.6**	**72.6**	85.8	**67.7**	**23.4**	**30.3**	**109.3**
NOC-REK	**59.4**	**79.3**	**67.8**	**73.4**	**82.1**	**81.1**	**79.9**	87.2	**76.3**	**26.9**	**32.8**	**138.4**
Δ	26.4↑	2.4↑	13.8↑	11.5↑	12.8↑	25.8↑	20.4↑	1.8↓	12.0↑	8.7↑	7.6↑	43.7↑

^{a}For all the tables in this paper, **Blue** indicates the best results among compared method (not applicable to results by human), **Red** indicates the second best results, gray background indicates results obtained by our method, and Δ indicates the improvement over state-of-the-art methods.

Table 2. Caption generation evaluation using CIDEr on held-out COCO dataset. We report the score for each class (higher score is better). Our method significantly outperforms OVE [6].

Method	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra
OVE [6]	67.2	79.7	**97.4**	**98.4**	86.5	**103.9**	56.2	**84.1**
NOC-REK*	**102.1**	**81.0**	89.2	71.9	**87.6**	60.0	**84.3**	56.2
NOC-REK	**141.1**	**118.9**	**124.3**	**108.0**	**110.0**	**89.1**	**116.9**	**99.0**
Δ	73.9↑	39.2↑	26.9↑	9.6↑	23.5↑	14.8↓	60.7↑	14.9↑

Finally, as shown in Table 2, we look further into the CIDEr score for each class. While our method achieves a higher CIDEr score for almost all classes, we notice that it underperforms on the *racket* class. This is due to a mismatch between the word forms *racket* and *racquet* in our generated captions and the ground truth captions. We believe that this is a minor issue that does not detract from our superiority over the other methods in general.

4.4. Results on Nocaps dataset

Qualitative evaluation. Fig. 3 (right) shows examples of generated captions obtained by our method and VinVL+VIVO [11, 12]. VinVL+VIVO occasionally cannot generate captions consisting of novel objects (first and third examples). In addition, VinVL+VIVO also generates a weird caption (fourth example). In contrast, our method effectively allows the appearance of novel objects in the generated captions. This advantage can be attributed to our usage of the reward for encouraging retrieved vocabulary to appear in the caption. Note that top-5 retrieved vocabulary by our method are reasonable though some are not correct because Nocaps is more challenging than held-out COCO.

Quantitative evaluation. Table 3 quantitatively compares caption generation on Nocaps validation and test sets using SPICE and CIDEr. Except for CIDEr on the out-domain of the test set, we see that our method outperforms the others by a large margin. The performance of VinVL+VIVO is comparable to ours because they use a visual vocabulary pre-trained model that covers all the objects in Nocaps. However, without VIVO, the VinVL itself cannot defeat our method. Therefore, we conclude that NOC-REK consistently works for all domains, particularly the out-domain, which contains all novel objects. Nonetheless, despite having higher scores than humans in most cases, our method performs noticeably worse than humans on the test set's out-domain. This demonstrates the task's difficulty, as well as the requirement for further improvement. Note that our method surpasses VIVO [11] and VinVL [12] on the Nocaps online leaderboard (accessed on Oct. 9, 2021).

4.5. Detailed analysis

Visualization of external knowledge. We visually investigate our collected knowledge to better understand why our method can retrieve the vocabulary added during inference time. We use t-SNE [32] to reduce the dimension of each $\mathbf{d}$ from 1×768 to 1×2, and then plot the reduced $\mathbf{d}$ into 2-D plan (Fig. 4). Note that we display the vocabulary for corresponding reduced $\mathbf{d}$ in Fig. 4. For the sake of simplicity, we show some clusters at random here, while a full visualization is provided in the supplementary. Fig. 4 shows that our external knowledge not only adequately clusters the vocabulary (black texts) but also locates new vocabulary (red and blue texts) into appropriate clusters. We optimize the similarity of image features and embeddings in external knowledge, which means that the image feature is also located in the relevant vocabulary cluster. The image feature of a novel object, on the other hand, would be similar to some seen objects to some extent [14]. When computing its similarity to the external knowledge, the novel object will intuitively fall into the same cluster as its related objects. Consequently, we can pick up on the new vocabulary properly. In fact, the model sometimes cannot retrieve a reasonable vocabulary.

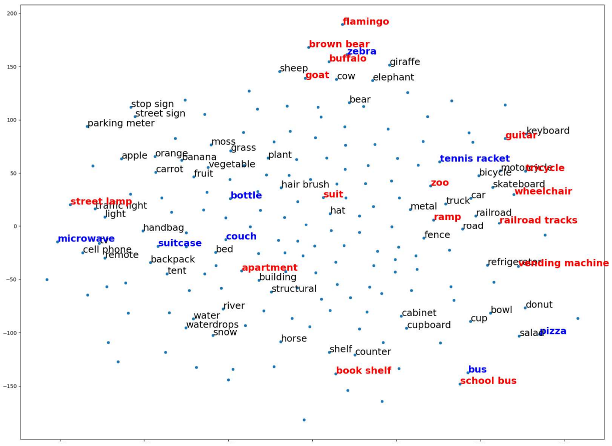


Figure 4. Visualization of external knowledge using t-SNE. We see that the related vocabulary (we use objects from the seen classes of held-out COCO dataset) fall in the same cluster (black text). When we add more objects from novel classes of held-out COCO dataset (blue text) and Nocaps dataset (red text), the novel vocabulary are located at the appropriate cluster.

In the fourth example on Nocaps (Fig. 3, right), for example, our model retrieves both *accordion* and *piano* because their appearances are too similar.

Ablation study. We evaluate the performance of using different loss term in Eq. 6 by comparing NOC-REK* and NOC-REK. As shown in Fig. 3, both NOC-REK and NOC-REK* are capable of retrieving appropriate vocabulary and generating reasonable captions. However, while NOC-REK can include more novel objects in the captions, NOC-REK* cannot. When we quantitatively evaluate caption generation, NOC-REK* performs worse than NOC-REK (see Tables 1, 2 and 3). The following is an explanation for the deterioration. We use cross-entropy loss between generated and ground truth captions in the first training stage to enforce the model to generate captions similar to those in the training dataset. Because the dataset contains no novel objects, it is difficult for NOC-REK* to describe them. In contrast, the CIDEr optimizer loss is not strictly related to ground truth captions at the second training stage, giving the model a better chance of describing novel objects. Furthermore, because we encourage more novel objects to appear in the second training stage, NOC-REK performs better than NOC-REK*. Note that both NOC-REK* and NOC-REK produce results that are at least comparable to the other methods.

We do not investigate the ablated model where the vocabulary retrieval and caption generation are trained independently because, as discussed in [5], it raises the difficulty to include novel objects in the captions. Indeed, most of our compared methods perform far worse than our method, highlighting the drawbacks of independently training.

Impact of the size of the external knowledge. We inves-

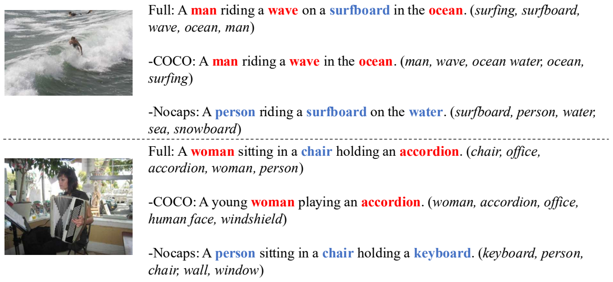
Table 3. Caption generation evaluation using SPICE and CIDEr on the Nocaps validation and test sets. We achieve the best scores for in-domain, near-domain, out-domain and Overall (excepting for CIDEr on near-domain of test set). Notably, the captions by our method are better than those by human in most 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 [1]	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
OVE [6]	79.5	11.5	74.7	11.1	78.2	10.7	76.1	11.1	–	–	–	–	–	–	–	–
ANOC [9]	86.1	12.0	80.7	11.9	73.7	10.1	80.1	11.6	85.8	12.4	79.7	11.8	68.5	10.0	78.5	11.6
Oscar [10]	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 [11]	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 [12]	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 [11, 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*	88.3	12.5	83.0	12.2	79.0	10.8	82.9	12.0	89.2	13.3	86.6	13.1	70.2	11.1	84.0	12.7
NOC-REK	**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**
Δ	1.0↑	1.1↑	4.6↑	0.7↑	13.2↑	1.1↑	6.6↑	0.9↑	2.0↑	0.5↑	0.5↑	0.2↑	0.6↓	0.1↑	0.5↑	0.3↑
Human [24]	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

[b] https://eval.ai/web/challenges/challenge-page/355/leaderboard/1011

Table 4. Impact of the size of the external knowledge on caption generation evaluation. Changes in the size of external knowledge result in changes in performance. Higher score is better.

Size of external knowledge	in-domain		near-domain		out-of-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Full (COCO + Nocaps)	104.7	14.8	100.2	14.1	100.7	13.0	100.9	14.0
–25%	90.9	12.8	85.5	12.4	83.1	11.3	85.8	12.3
–50%	87.8	12.5	83.5	12.3	79.2	10.9	83.0	12.0
–75%	86.4	12.4	83.1	12.2	79.2	10.8	83.2	12.0
–COCO	86.3	12.3	82.9	12.2	79.0	10.8	82.9	12.0
–Nocaps	88.0	12.5	83.2	12.2	77.2	10.7	82.4	11.9



Full: A **man** riding a **wave** on a **surfboard** in the **ocean**. (*surfing, surfboard, wave, ocean, man*)

-COCO: A **man** riding a **wave** in the **ocean**. (*man, wave, ocean water, ocean, surfing*)

-Nocaps: A **person** riding a **surfboard** on the **water**. (*surfboard, person, water, sea, snowboard*)

Full: A **woman** sitting in a **chair** holding an **accordion**. (*chair, office, accordion, woman, person*)

-COCO: A young **woman** playing an **accordion**. (*woman, accordion, office, human face, windshield*)

-Nocaps: A **person** sitting in a **chair** holding a **keyboard**. (*keyboard, person, chair, wall, window*)

Figure 5. Examples of generated captions using different external knowledge on Nocaps dataset: full knowledge (Full), knowledge without COCO objects (–COCO), knowledge without Nocaps objects (–Nocaps). Words in parentheses are top-5 retrieved vocabulary. **Blue** text is object in COCO, **Red** text is object in Nocaps.

tigate the performance of our model on the Nocaps validation set under various sizes of external knowledge (we cannot use the Nocaps test set because only 5 times submissions are allowed for the test set). We use various scenarios to drop vocabulary from a full external knowledge (COCO + Nocaps): we remove 25%, 50%, and 75% of vocabulary at random. We also remove any vocabulary found in COCO or Nocaps (out-domain) datasets. Table 4 demonstrates that reducing our vocabulary (25%, 50%, and 75%) leads to degrading the performance. Moreover, dropping either COCO or Nocaps vocabulary results in a trade-off between in-domain and out-domain performance. Our model can deal with out-domain but not in-domain without COCO vocabulary and vice versa without Nocaps vocabulary. These observations also confirm our model's ability to retrieve novel vocabulary from external knowledge and incorporate it into captions. This experiment also demonstrates how the size of our external knowledge affects our performance, implying that more novel objects are more effective. Fig. 5 illustrates that using different vocabulary, we obtain other captions. We remark that we do not retrain our model in this experiment.

Limitations. First, our model includes a pre-processing step to extract ROIs from a given image, which may not fully explore all potential objects in the image. To improve the capability of our method, one possible solution is to divide the image into multiple patches to which the transformers can directly attend, as discussed in [33]. Second, the quality of the external knowledge has a significant impact on our method, as seen in Table 4 and Fig. 5. As previously stated, it is reasonable to fix the knowledge embeddings as our external knowledge is sufficiently small. However, in the case of an explosion of novel vocabulary, training both image features and vocabulary embeddings is preferable. We have left detailed investigations for our future work.

5. Conclusion

We streamline the pipeline of novel object captioning by introducing the end-to-end NOC-REK model that includes vocabulary retrieval and caption generation steps. NOC-REK learns to retrieve vocabulary from external knowledge and generates captions using shared-parameters transformers. Our model does not require retraining; instead, it updates the external knowledge whenever new objects become available. We thoroughly compare our method with SOTAs on held-out COCO and Nocaps datasets, demonstrating significant superiority of NOC-REK.

Acknowledgement. This work was supported by the Institute of AI and Beyond of the University of Tokyo, Next Generation Artificial Intelligence Research Center of the University of Tokyo, and JSPS KAKENHI Grant Number JP19H04166, 22K17947. Duc Minh Vo and Hong Chen are thankful to Yi-Pei Chen's valuable comments on this work.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018. 1, 5, 8

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015. 1

[3] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *CVPR*, 2018. 1, 2, 5, 6

[4] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, S. Kate, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *CVPR*, 2016. 1, 2, 5, 6

[5] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *CVPR*, 2017. 1, 2, 5, 6, 7

[6] M. Tanaka and T. Harada, "Captioning images with novel objects via online vocabulary expansion," in *ECCV*, 2020. 1, 2, 5, 6, 8

[7] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," in *ACM MM*, 2018. 1, 2, 5, 6

[8] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, "Image captioning with unseen objects," in *BMVC*, 2019. 1, 2, 5, 6

[9] X. Chen, M. Jiang, and Q. Zhao, "Leveraging human attention in novel object captioning," in *IJCAI*, 2021. 1, 2, 5, 6, 8

[10] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020. 1, 2, 3, 4, 5, 8

[11] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu, "Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training," in *AAAI*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[12] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Making visual representations matter in vision-language models," in *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[13] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks." in *NIPS*, 2015. 1, 2, 4

[14] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, "Zero-shot object detection by hybrid region embedding," in *BMVC*, 2018. 1, 2, 7

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017. 2

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020. 2, 3, 4

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019. 2, 3, 4

[18] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017. 2, 5

[19] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *CVPR*, 2021. 4

[20] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *arXiv preprint arXiv:2002.08909*, 2020. 4

[21] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *EMNLP*, 2020. 4

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *NeurIPS*, 2020. 4

[23] https://huggingface.co/transformers/modeldoc/bert.html. 5

[24] H. Agrawal, P. Anderson, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, and S. Lee, "nocaps: novel object captioning at scale," in *ICCV*, 2019. 5, 8

[25] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided open vocabulary image captioning with constrained beam search," in *EMNLP*, 2017. 5

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014. 5

[27] http://vsubhashini.github.io/noc.html. 5

[28] https://github.com/pzzhang/VinVLl. 5

[29] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015. 5

[30] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016. 5

[31] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007. 5

[32] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, 2008. 7

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021. 8