nocaps: novel object captioning at scale

Harsh Agrawal[*1] Karan Desai[*1] Xinlei Chen[2] Rishabh Jain[1]
Dhruv Batra[1,2] Devi Parikh[1,2] Stefan Lee[1] Peter Anderson[1]
[1]Georgia Institute of Technology, [2]Facebook AI Research
[1]{hagrawal9, kdexd, rishabhjain, dbatra, parikh, steflee, peter.anderson}@gatech.edu
[2]{xinleic}@fb.com https://nocaps.org

Abstract

Image captioning models have achieved impressive results on datasets containing limited visual concepts and large amounts of paired image-caption training data. However, if these models are to ever function in the wild, a much larger variety of visual concepts must be learned, ideally from less supervision. To encourage the development of image captioning models that can learn visual concepts from alternative data sources, such as object detection datasets, we present the first large-scale benchmark for this task. Dubbed 'nocaps', for novel object captioning at scale, our benchmark consists of 166,100 human-generated captions describing 15,100 images from the Open Images validation and test sets. The associated training data consists of COCO image-caption pairs, plus Open Images image-level labels and object bounding boxes. Since Open Images contains many more classes than COCO, more than 500 object classes seen in test images have no training captions (hence, nocaps). We evaluate several existing approaches to novel object captioning on our challenging benchmark. In automatic evaluations these approaches show modest improvements over a strong baseline trained only on image-caption data. However, even when using ground-truth object detections, the results are significantly weaker than our human baseline – indicating substantial room for improvement.

1. Introduction

Image captioning, the task of generating natural language descriptions of visual content [1–6], has seen rapid progress over the past several years. This progress is largely attributed to the development and dissemination of large-scale datasets comprising image-caption pairs [7–9]. However, despite continual modeling improvements and ever-increasing benchmark performance [10–13], existing captioning models generalize poorly to images in the wild [14]. This is a natural consequence of training models using image-caption pairs that capture only a tiny fraction of the visual concepts encountered by humans in everyday life.

[*]Denotes equal contribution.

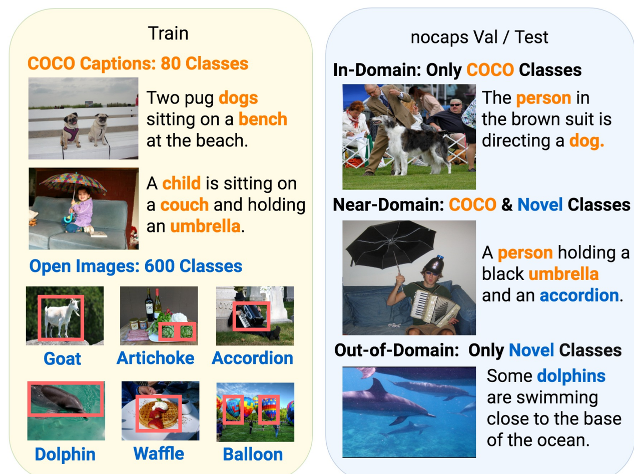


Figure 1: The nocaps benchmark for novel object captioning (at scale): Image captioning models must exploit object detection training data (bottom left) to successfully describe novel objects which are not present in the available image-caption training data (top left). This capability is crucial if image captioning models are to acquire the vast number of visual concepts required to function in the wild. The nocaps benchmark (right) rigorously evaluates performance over in-domain, near-domain and out-of-domain subsets of images containing only COCO classes, both COCO and Open Images classes, and only Open Images classes, respectively.

For example, models trained on COCO captions [9] can typically describe images containing dogs, people and umbrellas, but not accordions or dolphins. This limits the usefulness of these models in real-world applications, such as providing assistance for people with impaired vision, or for improving natural language query-based image retrieval.

To generalize better 'in the wild', we argue that captioning models should be able to learn from alternative data sources – such as object detection datasets – in order to describe objects not present in the caption corpora they are trained with. Such objects are referred to as *novel objects* and the task of describing images containing novel objects is termed *novel object captioning* [15–21]. Until now, approaches to novel object captioning have been evaluated using only 8 novel object classes held out from the COCO dataset [15]. This has left the large-scale performance of these methods open to question, particularly as each of these novel object classes was deliberately selected to be semantically similar to a cluster of in-domain classes. Therefore, given the emerging interest and practical necessity of this task, we introduce `nocaps`, the first rigorous and large-scale benchmark for novel object captioning, containing over 500 novel object classes.

In detail, the `nocaps` benchmark consists of a validation set and a test set comprised of 4,500 and 10,600 images, respectively, sourced from the Open Images object detection dataset [22] and annotated with 11 human-generated captions per image (comprising 10 reference captions for automatic evaluation plus a human baseline). Crucially, we provide no additional paired image-caption data for training. Instead, as illustrated in Figure 1, training data for the `nocaps` benchmark is image-caption pairs from the COCO 2017 [9] training set (containing 118K images reflecting 80 object classes), plus the Open Images V4 training set (containing 1.7M images annotated with bounding boxes for 600 object classes and image labels from 20K categories).

To be successful, image captioning models must utilize COCO paired image-caption data to learn to generate syntactically correct captions, while leveraging the massive Open Images detection dataset to learn many more visual concepts. As with previous work, this task setting is motivated by the observation that collecting human-annotated captions is expensive and scales poorly as object diversity grows, while on the other hand, large-scale object classification and detection datasets already exist [22, 23] and can often be scaled semi-automatically [24, 25]. To provide finer-grained insight into model performance, we report automatic evaluation metrics for the entire dataset, as well as in-domain, near-domain and out-of-domain subsets of images containing only COCO classes, both COCO and Open Images classes, and only Open Images classes, respectively.

To establish the state-of-the-art on our much more demanding benchmark, we evaluate two of the best performing existing approaches [17, 19] on the `nocaps` test set using automatic metrics. We find that performance improves only marginally over a strong baseline trained only on image-caption data [13]. Furthermore, even when leveraging ground-truth object detections, performance is significantly lower than our human baseline, suggesting substantial opportunities for future work.

In summary, we make two main contributions:

- We collect `nocaps` – the first large-scale benchmark for novel object captioning, containing 500+ novel objects.
- We undertake a detailed investigation of the performance and limitations of two state-of-the-art models for this task, which we calibrate against human performance.

We will provide an evaluation server hosting the validation and test splits, and a leaderboard to benchmark progress. We believe that improvements on this benchmark will accelerate progress towards image captioning in the wild.

2. Related Work

Novel Object Captioning A variety of approaches have been proposed to describe images containing visual concepts for which paired image-caption data does not exist. The Deep Compositional Captioner [15] and its extension, the Novel Object Captioner [16], both attempt to leverage object detection datasets and external text corpora by decomposing the captioning model into visual and textual components that can be trained with separate loss functions as well as jointly using the available image-caption data.

Several alternative approaches elect to use the output of object detectors more explicitly. Two concurrent works, Neural Baby Talk [19] and the Decoupled Novel Object Captioner [20], take inspiration from Baby Talk [26] and propose neural approaches to generate slotted caption templates, which are then filled using visual concepts identified by modern state-of-the-art object detectors. Related to this work, the LSTM-C [18] model augments a standard recurrent neural network sentence decoder with a copying mechanism which may select words corresponding to object detector predictions to appear in the output sentence. All of these models can be applied to the novel object captioning task if the object detector used is trained using detection datasets containing the novel objects.

In contrast to these works, several approaches to novel object captioning are architecture agnostic. Constrained beam search [17] is a decoding algorithm that can be used to enforce the inclusion of selected words in captions during inference, such as novel object classes predicted by an object detector. Building on this approach, partially-specified sequence supervision (PS3) [21] uses constrained beam search as a subroutine to estimate complete captions for images containing novel objects. These complete captions are then used as training targets in an iterative algorithm inspired by expectation maximization (EM) [27].

In this work, we investigate constrained beam search (CBS) [17] and Neural Baby Talk (NBT) [19] on our more challenging benchmark. We choose these models because they represent diverse approaches to this task with available codebases, and because both methods recently claimed the highest performance on the simple held-out COCO experimental procedure [15] that is currently used for evaluation.

Image Caption Datasets

In the past, two paradigms for collecting image-caption datasets have emerged: direct annotation and filtering. Direct-annotated datasets, such as Flickr 8K [7], Flickr 30K [8] and COCO Captions [9] are collected using crowd workers who are given explicit instructions to control the

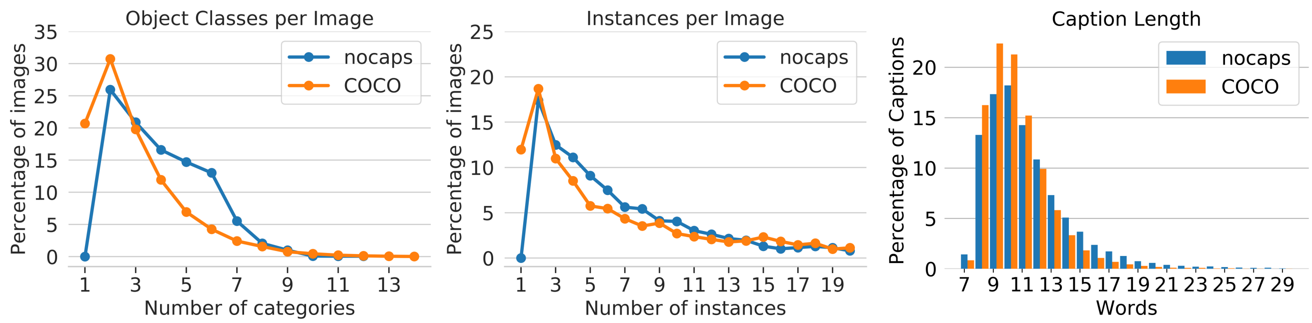


Figure 2: Compared to COCO Captions [9], on average `nocaps` images have more object classes per image (4.0 vs. 2.9), more object instances per image (8.0 vs. 7.4), and longer captions (11 words vs. 10 words). These differences reflect both the increased diversity of the underlying Open Images data [22], and our image subset selection decisions (refer Section 3.1).

quality and style of the resulting captions. To improve the reliability of automatic evaluation metrics, these datasets typically contain five or more captions per image. However, even the largest of these datasets, COCO Captions, is still based on a relatively small set of 80 object classes.

In contrast, filtered datasets, such as Im2Text [28], Pinterest40M [29] and Conceptual Captions [30], contain large numbers of image-caption pairs harvested from online resources such as Flickr, Pinterest, and the web respectively. These datasets contain many more visual concepts, but are also more likely to contain non-visual content in the description due to the automated nature of the collection pipelines. Furthermore, these datasets lack human baselines, and only include one caption per image, which reportedly decreases correlation between automatic evaluation metrics and human judgments [31,32]. Our benchmark, `nocaps`, aims to fill the gap between these datasets, by providing a high-quality benchmark with 10 reference captions per image and many more visual concepts than COCO. To the best of our knowledge, `nocaps` is the only image captioning benchmark in which humans outperform state-of-the-art models in automatic evaluation.

3. nocaps

In this section we detail the caption collection process, compare our proposed benchmark to COCO Captions [9], and introduce the evaluation protocol.

3.1. Caption Collection

The images in `nocaps` are sourced from the Open Images V4 [22] validation and test sets. Open Images is currently the largest available human-annotated object detection dataset, containing 1.9M images of complex scenes annotated with object bounding boxes for 600 classes (with an average of 8.4 object instances per image in the training set). Moreover, out of these 600 classes, more than 500 are never or exceedingly rarely mentioned in COCO captions [9] (which we select as image-caption training data), making these images an ideal basis for our challenging novel object captioning benchmark. In addition to object bounding box annotations, Open Images also contains both positive and negative image-level labels from 20K categories. In this work, we utilize only the object bounding boxes annotations, but the variety of available annotations may provide interesting directions for future work.

Image Subset Selection Since Open Images is primarily an object detection dataset, a large fraction of images contain well-framed iconic perspectives of single objects. Furthermore, the distribution of object classes is highly unbalanced, with a long-tail of object classes that appear relatively infrequently. However, for image captioning, images containing multiple objects and rare object co-occurrences are more interesting and challenging. Therefore, in order to make `nocaps` as rigorous as possible, we select subsets of images from the Open Images validation and test splits by applying the following sampling procedure.

First, since Open Images contains images for which the correct image rotation is unknown, we exclude all images for which the correct image rotation is non-zero or unknown. Next, based on the ground-truth object annotations, we exclude all images that contain instances from just a single object category, since image captioning may too closely resemble object detection for these images. Then, to capture as many visually complex images as possible, we include all images containing more than 6 unique classes. Finally, we iteratively select from the remaining images using a sampling procedure that encourages even representation both in terms of object classes and image complexity (based on the number of unique classes per image). Concretely, we divide the remaining images into 5 pools based on the number of unique classes present in the image (from 2–6 inclusive). Then, taking each pool in turn, we uniformly sample n images and among these, we select the image that when added to our benchmark results in the highest entropy over object classes. This encourages diversity, and prevents our benchmark from being overly dominated by frequently occurring object classes such as person, car or plant. In total, we select 4,500 validation images (from 41,620) and 10,600 test images (from 125,436). On average, the selected images contain 4 object classes per image and 8 instances per image (refer Figure 2).

Labels: Sombrero, Woman, Clothing

No Priming: A brown haired girl with a big straw hat.

Priming: Woman wearing a giant sombrero-type sun hat.

Labels: Red Panda, Tree

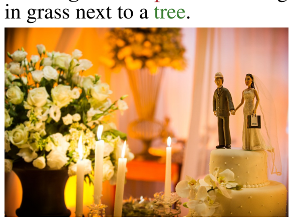
No Priming: A brown rodent climbing up a tree in the woods.

Priming: A red panda is sitting in grass next to a tree.

Labels: Gondola, Tree, Vehicle

No Priming: A man and a woman being transported in a boat by a sailor through canals

Priming: Some people enjoying a nice ride on a gondola with a tree behind them.



Labels: Woman, Man, Flower, Cake

No Priming: A wedding cake with bouquet and lighted candles in the foreground.

Priming: A vase of flowers next to a wedding cake with a bride and groom on top.

Figure 3: We conducted pilot studies to evaluate caption collection interfaces. Since Open Images contains rare and fine-grained classes (such as red panda, top right) we found that priming workers with the correct object categories resulted in more accurate and descriptive captions, on average, as illustrated by these examples.

Collecting Human Image Descriptions To enable model-generated image captions to be evaluated on nocaps, we collected 11 English captions for each selected image using a large pool of crowd-workers on Amazon Mechanical Turk (AMT). From these 11 captions, one caption per image was randomly sampled to constitute a human baseline for the task, and the other 10 captions are used as reference captions for automatic evaluations. Previous work suggests that automatic caption evaluation metrics correlate better with human judgment when more reference captions are provided [31, 32], motivating us to collect a larger number of reference captions than COCO (which contains 5 reference captions for the majority of images).

Our image caption collection interface closely resembles the interface used for collection of the COCO Captions dataset, albeit with one important difference. Since the nocaps dataset contains more rare and fine-grained classes than COCO, in initial pilot studies we found that human annotators could not always correctly identify the objects in the image. For example, as illustrated in Figure 3, a red panda was incorrectly described as a brown rodent. We therefore experimented with priming workers by displaying the list of ground-truth object classes contained in the image. To minimize the potential for this priming to reduce the language diversity of the resulting captions,

Dataset	1-grams	2-grams	3-grams	4-grams
COCO	6,913	46,664	92,946	119,582
nocaps	8,291	59,714	116,765	144,577

Table 1: Unique n-grams in equally-sized representative samples (4,500 images / 22,500 captions) from the COCO and nocaps validation sets. The increased visual variety in nocaps demands a larger vocabulary compared to COCO (1-grams), but also more diverse language compositions (2-, 3- and 4-grams).

the object classes were presented as 'keywords', and workers were explicitly instructed that it was not necessary to mention all the displayed keywords. To reduce the amount of redundant information provided, we did not display object classes which are classified in Open Images as parts (such as human hand, tire and door handle). Further pilot studies demonstrated that when workers were primed in this manner, the resulting image captions were qualitatively more accurate and descriptive (refer Figure 3). Therefore, all nocaps captions, including our human baselines, were collected using this modified COCO collection interface. Although priming has the potential to reduce language diversity, as we shown in Section 3.2, nocaps captions are nonetheless more diverse than COCO.

To help maintain the quality of the collected captions, we used only US-based workers who had completed a minimum of 5K previous tasks on AMT with at least a 95% approval rate. Additionally, we regularly spot-checked the captions written by each worker and blocked workers providing low-quality captions. Captions written by these workers were then discarded and replaced with captions written by high-quality workers. Overall, 727 qualified workers participated, writing 228 captions each on average (for a total of 166,100 captions).

3.2. Dataset Analysis

In this section, we compare our proposed nocaps benchmark to COCO Captions [9]. Most obviously, nocaps contains images spanning 600 object classes, while COCO is limited to 80 classes. As illustrated in Figure 2, consistent with this greater visual diversity, nocaps contains more object classes per image (4.0 vs 2.9), and slightly more object instances per image (8.0 vs 7.4), than COCO. Furthermore, there are no iconic images in nocaps containing just one object class, whereas 20% of the COCO dataset consists of such images. Similarly, less than 10% of all COCO images contain more than 6 object classes, while such images constitutes almost 22% of nocaps dataset.

Since nocaps images are visually more complex than COCO, on average the captions collected to describe these images tend to be slightly longer (11 words vs. 10 words) and more diverse than the captions in the COCO dataset. As illustrated in Table 1, taking representative samples over the same number of images and captions in each dataset, we show that not only do nocaps captions utilize a larger vocabulary than COCO captions (reflecting the increased number of visual concepts present), but the number of unique 2, 3 and 4-grams is also significantly higher for `nocaps`. This suggests that the captions in our benchmark contain a greater variety of unique language compositions.

3.3. Evaluation

Requirements As previously outlined, the `nocaps` benchmark requires image captioning models to utilize COCO paired image-caption data to learn to generate syntactically correct captions, while leveraging the massive Open Images detection dataset to learn many more visual concepts. Other datasets, such as external text corpora, knowledge bases, and additional object detection datasets may also be used during training or inference to tackle this challenging task. However, to ensure consistent evaluation, *the only dataset of paired image-captions that should be used is the COCO 2017 training split*, containing 118K images. Since we aim to encourage the development of more general models that can assimilate information from multiple data modalities, improving evaluation scores by simply leveraging additional paired image-caption datasets such as Flickr 30K [8] or Conceptual Captions [30] would be contrary to this aim. For similar reasons, captions from the `nocaps` validation set should not be used for training. We also note that ground-truth object detection annotations are available for Open Images validation and test splits (and hence, for the `nocaps` validation and test splits). While ground-truth object annotations may be used to establish performance upper bounds on the validation set, they should not be used for any submission to the evaluation server.

Metrics As with existing captioning benchmarks, we rely on automatic metrics to evaluate the quality of model-generated captions. We focus primarily on CIDEr [31] and SPICE [32], which have been shown to have the strongest correlation with human judgments [33], but we also report Bleu [34], Meteor [35] and ROUGE [36]. On the COCO dataset, state-of-the-art captioning models routinely outperform human baselines by a wide margin in terms of these automatic evaluation metrics. This may invite questions regarding the interpretation and meaning of further improvements. In contrast, as illustrated in Figure 5, `nocaps` includes a larger set of reference captions, capturing more of the salient content of the images. As discussed further in Section 4, due to the larger number of reference captions, the use of priming during caption collection (refer Section 3.1), and the increased difficulty of the task relative to COCO, model results on `nocaps` are significantly weaker than our human baseline. This makes improvements on automatic metrics easier to interpret in the context of human performance, and arguably more meaningful.

In addition to evaluating captioning models and the human baseline on the entire `nocaps` dataset, we also report results across three subsets of the validation and test splits.

Near-Domain | Out-of-Domain

Near-Domain:

1. A man sitting in the saddle on a camel.
2. A person is sitting on a camel with another camel behind him.
3. A man with long hair and blue jeans sitting on a camel.
4. Man sitting on a camel with a standing camel behind them.
5. Long haired man wearing sitting on blanket draped camel
6. A camel stands behind a sitting camel with a man on its back.
7. The standing camel is near a sitting one with a man on its back.
8. Someone is sitting on a camel and is in front of another camel.
9. Two camels in the dessert and a man sitting on the sitting one.
10. Two camels are featured in the sand with a man sitting on one of the seated camels.

Out-of-Domain:

1. A tank vehicle stopped at a gas station.
2. A tank and a military jeep at a gas station
3. A jeep and a tan colored tank getting gas at a gas station.
4. A tank and a truck sit at a gas station pump.
5. An Army humvee is at getting gas from the 76 gas station.
6. An army tank is parked at a gas station.
7. A land vehicle is parked in a gas station fueling.
8. A large military vehicle at the gas pump of a gas station.
9. A tanker parked outside of an old gas station
10. Multiple military vehicles getting gasoline at a civilian gas station.

Figure 4: Examples of images belonging to the near-domain and out-of-domain subsets of the `nocaps` validation set. Each image is annotated with 10 reference captions, capturing more of the salient content of the image and improving the accuracy of automatic evaluations [31, 32].

We first map COCO classes to Open Images classes to identify the novel object classes in Open Images. Subsets are then determined as follows:

1. **In-Domain** images contain only objects belonging to the 80 classes in COCO. Since these objects have been described in the paired image-caption training data, we expect the performance trends on this subset to be closer to COCO, albeit with some negative impact due to image domain shift. This subset contains 816 test images (8K captions) covering 74 COCO classes.
2. **Near-Domain** images contain both COCO object classes and novel object classes from Open Images. These images can be challenging for image captioning models trained only on COCO paired image-caption data, specially when the most salient objects in the image are novel objects. This subset contains 6,790 test images (68K captions) consisting of 80 COCO classes and 347 novel classes.
3. **Out-of-Domain** images do not contain any COCO classes, and are therefore visually very distinct from COCO images. We expect models trained only on COCO data to make ‘embarrassing errors’ [33] on this subset, reflecting the current performance of COCO trained models in the wild. There are 3,021 test images (30K captions) in this subset spanning 412 novel classes.

	COCO val 2017					nocaps val							
			Overall			In-Domain		Near-Domain		Out-of-Domain		Overall	
Method	Bleu-1	Bleu-4	Meteor	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Up-Down	**77.4**	**36.6**	27.3	**115.6**	20.4	75.4	10.6	63.2	9.9	34.7	6.8	55.4	9.0
Up-Down + VGOI feat	74.2	33.8	26.0	105.2	19.0	55.4	8.8	46.9	8.0	23.8	5.1	40.4	7.2
Up-Down + Embed	76.4	35.8	**27.4**	113.5	**20.6**	71.6	10.3	63.9	10.1	38.4	7.0	56.6	9.1
Up-Down + CBS	74.6	33.3	26.3	105.9	19.4	72.3	10.2	63.2	10.0	41.4	7.4	57.2	9.2
Up-Down + CBS + GT	-	-	-	-	-	72.7	10.6	74.0	10.8	75.2	9.3	74.2	10.3
NBT	74.3	33.1	27.1	106.5	20.3	70.9	11.1	58.7	10.0	35.0	6.8	52.4	9.1
NBT + CBS	70.8	28.2	24.8	87.7	17.8	62.5	10.3	54.5	9.7	37.7	7.1	49.9	9.0
NBT + GT	-	-	-	-	-	63.3	10.8	55.6	10.1	43.6	7.6	52.5	9.4
Human	66.3	21.7	25.2	85.4	19.8	**83.3**	**13.9**	**85.5**	**14.3**	**91.4**	**13.7**	**87.1**	**14.1**

Table 2: Single model image captioning performance on the COCO and `nocaps` validation sets. We begin with a strong baseline in the form of the Up-Down [13] image captioning model trained on COCO captions, using features from Faster R-CNN trained on Visual Genome. We then investigate the use of Faster R-CNN image features trained jointly on Visual Genome and Open Images (+VGOI feat), the addition of fixed pretrained GloVe [37] and dependency-based [38] word embeddings (+ Embed), and decoding using constrained beam search [17] based on object detections from the VGOI Faster R-CNN (+ CBS) and ground-truth object detections (+ CBS + GT), respectively. We observe modest improvements over the baseline model using word embeddings and constrained beam search decoding, particularly for Out-of-Domain `nocaps` images. In panel 2, we review the performance of Neural Baby Talk (NBT) [19], illustrating similar performance trends. Even when using ground-truth object detections, all approaches lag well behind the human baseline on `nocaps`. Note: Scores on COCO and `nocaps` should not be directly compared, see Section 3.3 for discussion. COCO human scores are reported for the test split.

Note that when comparing model performance on COCO and `nocaps`, the absolute value of the automatic evaluation scores on each dataset should not be directly compared. This is partly due to the different number of reference captions used in each dataset. Increasing the number of reference captions improves fidelity [32], but since SPICE is an F-score, the increased number of true positive propositions tends to decrease the absolute value of the scores received. Similarly, the CIDEr metric [31] uses corpus-wide statistics to calculate n-gram weightings, and these will also differ across datasets.

4. Experiments

We now describe our experiments investigating the performance of Up-Down [13] and Neural Baby Talk (NBT) [19] with and without constrained beam search (CBS) [17] in the context of our more challenging `nocaps` benchmark. We first discuss the training of object detectors / image feature representations for the task.

Object Detectors / Image Features Recent work has demonstrated substantial benefits from using feature representations derived from object detectors trained on large numbers of object and object attribute classes [13] for the task of image captioning. In the context of COCO image captioning, it is natural to use object detection features trained on Visual Genome [39] annotations, since the underlying images are sourced from COCO. However, in the context of the novel object captioning task, the image features used must also adequately represent the novel objects and not just the visual concepts in the image-caption training set. Therefore, in experiments we investigate two object detectors / image feature representations: the Faster R-CNN [40] pretrained on Visual Genome by Anderson et al. [13], and a Faster-RCNN model trained on a combination of the Visual Genome and Open Images datasets by us. We refer the former as VG features and latter as VGOI features. We will also use the VGOI model as an object detector where required in CBS and NBT.

To create the combined VGOI training dataset containing both COCO object classes and Open Images novel object classes, we begin with the 1,600 Visual Genome object classes used in previous work [13]. We then carefully map Open Images object classes to their equivalent Visual Genome classes where possible, while creating new classes as necessary. Since Visual Genome classes often categorize instances of multiple objects as separate classes (e.g. person, people), we make use of the Open Images 'IsGroupOf' annotation to ensure that Open Images annotations are mapped to the appropriate singular or plural class in Visual Genome. The resulting combined dataset contains 1,913 object classes, as well as 400 attribute classes (which are only found in Visual Genome).

To obtain VGOI features, we train a Faster-RCNN [40] model based on the ResNeXt [41] backbone architecture using the open-source Detectron framework [42]. The ResNeXt is configured with width 4, cardinality 64 and depth 101, and pretrained on ImageNet [43]. Since the Open Images training set is an order of magnitude larger than Visual Genome, when forming training minibatches we sample Visual Genome images with three times more

	nocaps test												
	In-Domain		Near-Domain		Out-of-Domain		Overall						
Method	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	Bleu-1	Bleu-4	Meteor	ROUGE_L	CIDEr	SPICE
Up-Down	70.9	10.7	61.5	10.0	33.3	6.5	73.7	18.6	22.4	50.2	54.2	9.1
Up-Down + CBS	69.1	10.6	61.3	9.9	40.1	7.2	73.8	17.1	22.4	49.8	55.8	9.2
NBT	66.0	10.6	56.6	10.0	33.1	6.7	71.9	16.6	22.4	49.8	50.6	9.1
NBT + CBS	56.6	10.1	51.4	9.7	35.5	7.0	71.3	14.3	21.6	48.6	47.3	9.0
Human	**75.8**	**14.2**	**84.8**	**14.7**	**89.1**	**14.0**	**76.6**	**19.5**	**28.2**	**52.8**	**85.3**	**14.5**

Table 3: Single model image captioning performance on the nocaps test split. We evaluate four models, including the Up-Down model [13] trained only on COCO, as well as three model variations based on constrained beam search (CBS) [17] and Neural Baby Talk (NBT) [19] that leverage the Open Images training set. Although these models differ widely in approach, performance on nocaps is remarkably similar and well below the human baseline, indicating substantial room for improvement.

likelihood than Open Images. For captioning, we extract the fc7 features to represent each region, similarly to previous work [13]. To ensure an even-handed evaluation, all of our baseline models are trained with cross-entropy loss and use the same image features where possible.

COCO Baselines and Constrained Beam Search To establish a strong baseline model trained exclusively on paired image-caption data, we select the Up-Down image captioning model [13], which is close to the state-of-the-art for a single model and has code available. In Table 2 rows 1 and 2, we report results for this model using both the original VG features (Up-Down), and alternatively using VGOI features (Up-Down + VGOI feat). Although the VGOI features have been trained on the larger combined Visual Genome and Open Images dataset, surprisingly, the original VG features trained on Visual Genome perform considerably better on both COCO and nocaps validation sets. We suspect that this may be due to the sparsity of annotations in Open Images, which has been identified as a noteworthy challenge of the dataset [44]. We therefore use VG features in all remaining experiments with the Up-Down model.

Next, we apply constrained beam search (CBS) [17] decoding to the Up-Down model. CBS is a multi-beam decoding algorithm capable of finding high-probability output sequences that satisfy certain constraints – in this case, the inclusion of object class names predicted by the VGOI model (Up-Down + CBS), or ground truth object class names from Open Images (Up-Down + CBS + GT) to establish an upper performance bound. Following the original work, we deal with out-of-vocabulary novel objects by initializing both the input and output layers of the captioning model with fixed, concatenated GloVe [37] and dependency-based [38] word embeddings. The performance of the base model with just this modification is reported in Table 2 as (Up-Down + Embed). Interestingly, just the introduction of pretrained word embeddings improves performance on the nocaps Out-of-Domain subset, which is further improved by CBS and (much more substantially) by the introduction of the ground-truth object detections. However, the performance improvements over the Up-Down baseline are modest overall. Even when using ground-truth object detections, automatic evaluation scores are significantly worse than human.

Neural Baby Talk Neural Baby Talk (NBT) [19] performs captioning in two stages, first generating a hybrid textual template with slots explicitly tied to specific image regions, and then filling slots with words by recognizing the content in the corresponding image regions. This gives NBT the capability to caption novel objects, when combined with an appropriate pretrained object detector. To evaluate NBT on nocaps, we replace the model's original 80-class COCO-based object detector with our VGOI model trained on 1,913 classes from both Open Images and Visual Genome. In addition, since the original model used less powerful CNN features, we replace the input to one of the attention layers of the language model with VG features. Similarly to the Up-Down model, we use fixed GloVe embeddings [37] in both the language model and the visual feature representation for an object region. The original NBT model performed caption template refinement by predicting fine-grained object classes and the plurality of detection labels. However, with 1,913 object classes now directly predicted by our object detector (including, in many cases, separate plural and singular classes), we drop the fine-grained classification head used in the original work.

As illustrated in Table 2, the NBT model receives lower scores on COCO than the Up-Down model (partly due to the removal of the fine-grained class classification head, which is difficult to scale to a large number of classes). However, performance on nocaps is similar to the other models. As with Up-Down, NBT can also be decoded using constrained beam search (NBT + CBS), which improves performance on the nocaps Out-of-Domain subset. Using ground-truth object detections in conjunction with NBT (NBT + GT) further increases performance, but as with CBS, the results are still significantly weaker than the human baseline.

To qualitatively assess some of the differences between the various approaches, in Figure 5 we illustrate some ex-

Method	In-Domain	Near-Domain	Out-of-Domain
Up-Down	A statue of a woman riding a bike.	A man in a red shirt holding a baseball bat.	A close up of a bird on a tree.
Up-Down+CBS	A **woman** is riding a bike in the street.	A **man** holding a baseball bat on a field.	A large yellow and black **lizard** in a rock.
Up-Down+CBS(GT)	A young **girl** is riding on the back of a bike.	A man holding a baseball bat in a **shotgun rifle**.	A long insect **caterpillar** sitting on top of a rock.
NBT	A **woman** is riding a horse with a large group of bananas.	A baseball player holding a bat on a field.	A banana that is sitting on a rock.
NBT+CBS	A **woman** is riding a horse with a large group of bananas.	A baseball bat player holding a bat in a **baseball uniform**.	A **insect** sitting on a rock next to a pile of leaves.
NBT+GT	A **woman** riding a horse with a group of people.	A man in a red shirt holding a **baseball bat**.	A small black and white caterpillar with a small black and white face.
Human	People are preforming in the open a cultural dance.	A man in a red hat is holding a shotgun in the air.	A black and yellow centipede is crawling on leaves.

Figure 5: Some challenging images from nocaps and the corresponding captions generated by existing approaches. While the models hallucinate bike or horse in the In-Domain images, they confuse shotgun with a baseball bat for Near-Domain images. Furthermore, the models fail to describe the insect in the Out-of-Domain image even though the detector had recognized the caterpillar in the image. The constraints given to the CBS are shown in **blue**. The grounded visual words associated with NBT are shown in **red**.

amples of the captions generated using various model configurations. As expected, the Up-Down model trained only on COCO fails to identify novel objects such as the shotgun / rifle and the insect / centipede. The remaining models leverage the Open Images training data, enabling them to potentially describe these novel object classes. However, the results are sporadically successful at best. To provide baselines for future work, in Table 3 we report nocaps test set results for our best performing model variations with hyperparameters tuned on the validation set.

5. Conclusion

In this work, we motivate the need for a stronger and more rigorous benchmark to assess progress on the task of novel object captioning. We introduce nocaps, a large-scale benchmark consisting of 166,100 human-generated captions describing 15,100 images containing more than 600 unique object classes (and many more visual concepts). Evaluating recent novel object captioning models [17, 19] – which have previously only been evaluated on a 'toy' dataset of 8 held-out COCO objects – on nocaps, we discover that our more challenging dataset is largely resistant to these approaches, which make minimal improvements over a strong baseline trained exclusively on paired image-caption data. We posit four possible explanations for this:

1. The Open Images object detection task is strictly more difficult than the COCO detection task, by virtue of the much larger number of object classes and the finer-grained distinctions between them. This has a direct bearing on the performance of approaches such as NBT [19] and CBS [17] (both methods improve significantly with access to ground-truth detections).
2. In order to produce humanlike descriptions, captioning models must determine when novel object detections should be mentioned. However, existing novel object captioning models [15–21] do not have explicit notions of visual saliency.
3. In the previous held-out COCO evaluation [15], the 8 novel objects were specifically selected to be semantically similar to objects described in the paired image-caption training data. In contrast, nocaps contains many novel objects such as insects, weapons and sea creatures that are very dissimilar to COCO classes, making fluent caption generation much more challenging.
4. Captioning models trained on COCO and evaluated on nocaps experience domain shift in the image domain that was not evident in the previous evaluation (for example, Open Images includes images with cropped white backgrounds, which are not present COCO).

We hope that the proposed nocaps benchmark will benefit the community by providing a rigorous evaluation for the challenging task of novel object captioning, and by encouraging researchers to consider the shortcomings of the existing approaches to this task. We strongly believe that improvements on this benchmark will accelerate progress towards image captioning in the wild, and the many tangible benefits associated with its real-world applications.

Acknowledgments

We thank Jiasen Lu for helpful discussions about Neural Baby Talk. This work was supported in part by NSF, AFRL, DARPA, Siemens, Samsung, Google, Amazon, ONR YIPs and ONR Grants N00014-16-1-{2713,2793}. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

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

[2] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015. 1

[3] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2015. 1

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015. 1

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015. 1

[6] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *CVPR*, 2015. 1

[7] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013. 1, 2

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014. 1, 2, 5

[9] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015. 1, 2, 3, 4, 11

[10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017. 1

[11] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017. 1

[12] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen, "Review networks for caption generation," in *NIPS*, 2016. 1

[13] 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, 2, 6, 7

[14] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz, "Rich Image Captioning in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016. 1

[15] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, 2016. 1, 2, 8

[16] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. J. Mooney, T. Darrell, and K. Saenko, "Captioning Images with Diverse Objects," in *CVPR*, 2017. 1, 2, 8

[17] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided open vocabulary image captioning with constrained beam search," in *EMNLP*, 2017. 1, 2, 6, 7, 8, 11

[18] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *CVPR*, 2017. 1, 2, 8

[19] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *CVPR*, 2018. 1, 2, 6, 7, 8

[20] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," *CoRR*, vol. abs/1804.03803, 2018. 1, 2, 8

[21] P. Anderson, S. Gould, and M. Johnson, "Partially-supervised image captioning," in *NIPS*, 2018. 1, 2, 8

[22] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy, "Openimages: A public dataset for large-scale multi-label and multi-class image classification.," *Dataset available from https://github.com/openimages*, 2017. 2, 3

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 2

[24] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "Extreme clicking for efficient object annotation," in *ICCV*, 2017. 2

[25] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "We don't need no bounding-boxes: Training object class detectors using only human verification," in *CVPR*, 2016. 2

[26] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *PAMI*, vol. 35, no. 12, pp. 2891–2903, 2013. 2

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, 1977. 2

[28] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NIPS*, 2011. 3

[29] J. Mao, J. Xu, K. Jing, and A. L. Yuille, "Training and evaluating multimodal word embeddings with large-scale web annotated images," in *NIPS*, 2016. 3

[30] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of ACL*,

2018. 3, 5
[31] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *CVPR*, 2015. 3, 4, 5, 6, 12, 13
[32] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *ECCV*, 2016. 3, 4, 5, 6, 12, 13
[33] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *ICCV*, 2017. 5
[34] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002. 5
[35] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): Second Workshop on Statistical Machine Translation*, 2007. 5
[36] C. Lin, "Rouge: a package for automatic evaluation of summaries," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) Workshop: Text Summarization Branches Out*, 2004. 5
[37] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *EMNLP*, 2014. 6, 7
[38] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *ACL*, 2014. 6, 7
[39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv preprint arXiv:1602.07332*, 2016. 6
[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015. 6
[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *arXiv preprint arXiv:1611.05431*, 2016. 6
[42] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron." `https://github.com/facebookresearch/detectron`, 2018. 6
[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *IJCV*, 2015. 6
[44] T. Akiba, T. Kerola, Y. Niitani, T. Ogawa, S. Sano, and S. Suzuki, "PFDet: 2nd place solution to open images challenge 2018 object detection track," *arXiv preprint arXiv:1809.00778*, 2018. 7

Appendix

This supplementary document provides additional details regarding the data collection interface and provides qualitative examples from the nocaps benchmark. It also includes additional constrained beam search implementation details and further examples of model's prediction on the three (in-domain, near-domain and out-of-domain) subsets of the dataset.

1. CBS Implementation Details

In this section, we provide further details regarding the application of constrained beam search (CBS) [17] decoding to the Up-Down model. As in the original work, when using CBS we decoded the model in question while enforcing the inclusion of words corresponding to selected object classes in the generated caption. The number of constraint words to be included was determined on the validation set. When using the tag predictions from the VGOI model (Up-Down + CBS), predictions are drawn from 1,913 combined Open Images and Visual Genome classes, as described in the main paper. In this case, we found that the highest validation set scores were gained by using the top three predictions as constraints, and then selecting the caption with the highest log-probability that satisfies *at least one* of these constraints. When using ground-truth object classes drawn from the 600 Open Images classes (Up-Down + CBS + GT), then up to three randomly select object classes were used as constraint words, and we select the caption with the highest log-probability that satisfies *at least two* of these constraints.

2. Dataset Collection Interface

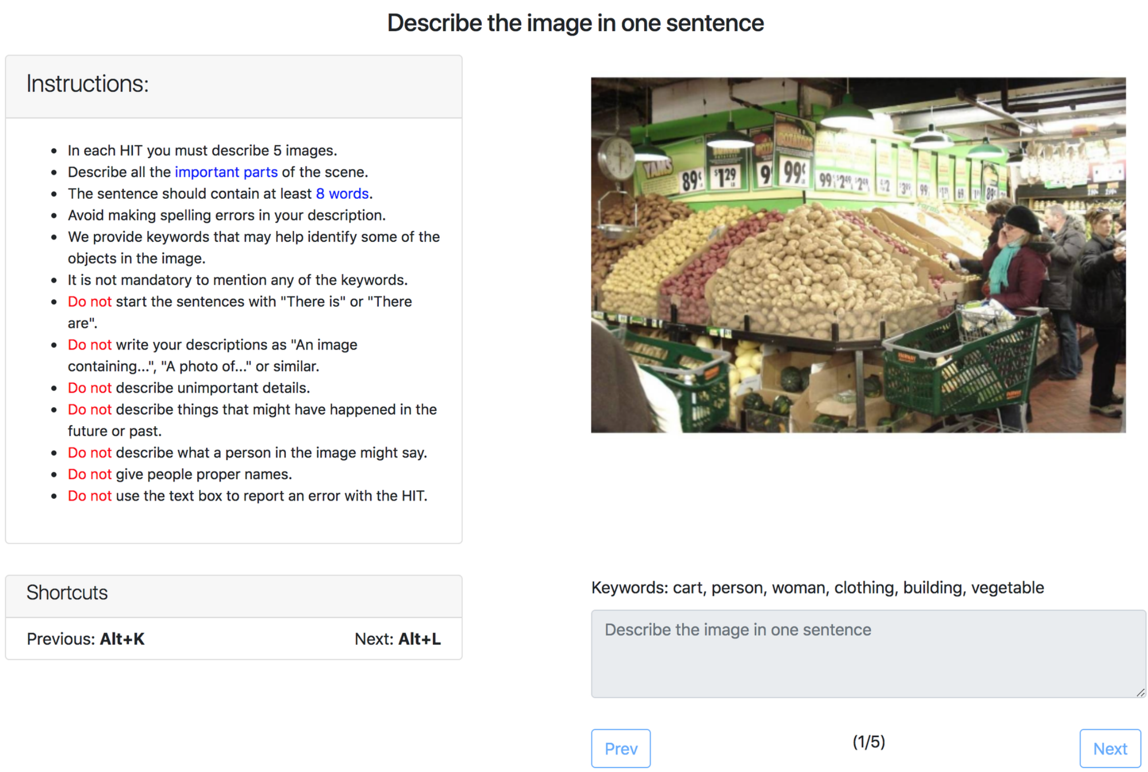


Figure 6: User Interface with priming for gathering captions. The interface shows a subset of object categories present in the image as keywords. Note that the instruction explicitly states that it is not mandatory to mention any of the displayed keywords. Other instructions are similar to the interface described in [9]

3. Examples of reference captions from nocaps

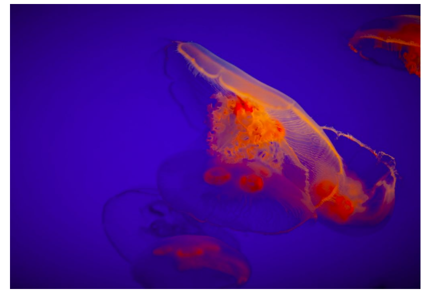
In-Domain

1. Two hardcover books are on the table
2. Two magazines are sitting on a coffee table.
3. Two books and many crafting supplies are on this table.
4. a recipe book and sewing book on a craft table
5. Two hardcover books are laying on a table.
6. A table with two different books on it.
7. Two different books on sewing and cooking/baking on a table.
8. Two magazine books are sitting on a table with arts and craft materials.
9. A couple of books are on a table.
10. The person is there looking into the book.

Near-Domain

1. Men in military uniforms playing instruments in an orchestra.
2. Military officers play brass horns next to each other.
3. Two men in camouflage clothing playing the trumpet.
4. Two men dressed in military outfits play the french horn.
5. Two people dressed in camoflauge uniforms playing musical instruments.
6. A couple people in uniforms holding tubas by their mouths.
7. Two people in uniform are playing the tuba.
8. A couple of military men playing the french horn.
9. A man in uniform plays a French horn.
10. Two men are playing the trumpet standing nearby.

Out-of-Domain



1. Some red invertebrate jellyfishes in dark blue water.
2. orange and clear jellyfish in dark blue water
3. A red jellyfish is swimming around with other red jellyfish.
4. Orange jelly fish swimming through the water.
5. Bright orange and clear jellyfish swim in open water.
6. The fish is going through the very blue water.
7. A bright orange jellyfish floating in the water.
8. Several red jellyfish swimming in bright blue water.
9. An orange jellyfish swimming with a blue background
10. A very vibrantly red jellyfish is seen swimming in the water.

1. Jockeys on horses racing around a track.
2. Several horses are running in a race thru the grass.
3. several people racing horses around a turn outside
4. Uniformed jockeys on horses racing through a grass field.
5. Se veral horse jockies are riding horses around a turn.
6. Six men and six horses are racing outside
7. A group of men wearing sunglasses and racing on a horse
8. Six horses with riders are racing, leaning over at an incredible angle.
9. Seveal people wearing goggles and helmets racing horses.
10. a row of horses and jockeys running in the same direction in a line

1. Two people in a fencing match with a woman walking by in the background.
2. Two people in masks fencing with each other.
3. Two people in white garbs are fencing while people watch.
4. Two people in full gear fencing on white mat.
5. A couple of people in white outfits are fencing.
6. Two fencers in white outfits are dueling indoors.
7. A couple of people doing a fencing competition inside.
8. Two people in white clothes fencing each other.
9. Two people in an room competing in a fencing competition.
10. Two people in all white holding swords and fencing.

1. A panda bear sitting beside a smaller panda bear.
2. The panda is large and standing over the plant.
3. Two panda are eating sticks from plants.
4. Two panda bears sitting with greenery surrounding them.
5. two panda bears in the bushes eating bamboo sticks
6. two pandas sitting in the grass eating some plants
7. two pandas are eating a green leaf from a plant
8. Two pandas are eating bamboo in a wooded area.
9. Pandas enjoy the outside and especially with a friend.
10. Two black and white panda bears eating leaf stems

Figure 7: Examples of images belonging to in-domain, near-domain and out-of-domain subsets of the nocaps validation set. Each image is annotated with 10 reference captions, capturing more of the salient content of the image and improving the accuracy of automatic evaluations [31, 32].

In-Domain

1. A dog sitting beside a man walking on the lawn
2. A small dog looking up at a person standing next to him.
3. A young dog looks up at their owner.
4. A little puppy looking up at a person.
5. A dog sitting on the grass next to a human.
6. A dog is looking up at the person who is wearing jeans.
7. The tan dog sits patiently beside the person.l
8. The white dog is sitting in the grass by a person who is standing up.
9. The tan dog happily accompanies the human on the grass.
10. A dog is on the grass is looking to a person

Near-Domain

1. A woman is sitting with a camera in front of her
2. A beautiful brown haired woman next to a camera.
3. A woman poses to take a picture in a mirror.
4. A women with brown hair holding a camera.
5. A woman behind a camera on a tripod.
6. A woman sits with her head leaned behind a camera.
7. A girl looking pity behind a camera on a tripod.
8. A woman sits and tilts her head while behind a camera.
9. A woman is sitting behind a camera with tripod.
10. A woman with a camera in front of her.

Out-of-Domain

1. Some decorations are have red lights you can see at night.
2. A couple of red lanterns floating in the air.
3. Many red Chinese lanterns are hung outside at night
4. Floating lighted lanterns on a dark night in the city.
5. Dozens of glowing paper lanterns floating off into the sky.
6. A black night sky with red, bright floating lanterns.
7. Red lanterns floating up to the dark night sky.
8. Chinese lanterns that are red are floating into the sky.
9. This town has many lit Chinese lanterns hanging between the buildings.
10. The street is filled with light from hanging lanterns.

1. People are standing on the side of a food truck
2. A food truck parked with people standing in line.
3. People standing outside and ordering from a food truck in the daytime.
4. The food truck has a line of people in front of the window.
5. A woman standing in front of a food truck.
6. A food truck outside of a small business with several people eating
7. People stand in line to get food from a food truck.
8. A large metal truck serving food to people in a parking lot.
9. Men and women speaking in front of a grey food truck that is open for business.
10. on the street woman truck clothing footwear jeans

1. A room with a hot tub and sauna.
2. A white hot tub is next to some wood.
3. A jacuzzi sitting on rocks inside of a patio.
4. A hot tub sits in the middle of the room.
5. A jacuzzi sitting near some rocks and a sauna
6. A hot tub in a room with wooden flooring.
7. A room is shown with a hot tub, decorative plants and some paintings ont he wall.
8. A room with a large hot tub and a sauna.
9. A water filled jacuzzi surrounded by smooth river rocks and a wooden deck.
10. A white and grey jacuzzi around rock building

1. Large silver tanks behind the counter at a restaurant.
2. Shiny metal containers with writing are beside each other.
3. A brewery with big, silver, metal containers and a sign.
4. A brew station inside of a restaurant.
5. Large steel breweries sit behind a chalkboard displaying different food and drink deals.
6. A cabinetry with big tin cans and a chalkboard on the top
7. A man works on machinery inside a brewery.
8. The many silver tanks are used for beverage making.
9. A menu is hanging above a craft brewery.
10. A man peers at a brewing tank while standing on a step ladder.

Figure 8: More examples of images belonging to in-domain near-domain and out-of-domain subsets of the `nocaps` validation set. Each image is annotated with 10 reference captions, capturing more of the salient content of the image and improving the accuracy of automatic evaluations [31, 32].

4. Example Model Predictions

	In-Domain	Near-Domain	Out-of-Domain
			
Method \ Labels	Boy, Person, Man, Girl	Man, Plant, Tank, Tree, Wheel	Flower, Moths and butterflies
Up-Down	a group of people standing on top of a blue floor	two men are sitting on top of a truck	a small bird sitting on top of a tree
Up-Down+CBS	a **man** in a white shirt is playing a game	a **man** and a person on a vehicle	a small bird insect is flying through the orange
Up-Down+CBS(GT)	a group of people that are standing on a court	a group of men riding on top of a **tank** truck	a small bird insect is flying through the **flower**
NBT	a crowd of **people** standing around each other	a couple of **men** are loading a truck	a close up of a bird on a tree branch
NBT+CBS	a group of people in white **mans** in a room	a man is loading a **tree** into a truck	a small **bee** is sitting on a branch
NBT+GT	a group of **people** standing around each other	a couple of **men** that are touching a truck	there is a small blue and white **flower** on the ground
Human	Two people in karate uniforms spar in front of a crowd.	Two men sitting on a tank parked in the bush.	A bee is pollinating a white flower with a yellow center.

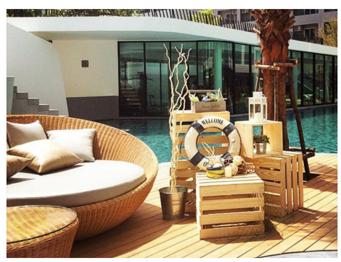
			
Method \ Labels	Person, Umbrella	Studio couch, House, Coffee table, Swimming pool, Building	Dessert, Fruit,Baked goods
Up-Down	a group of chairs and umbrellas on a beach	a couple of chairs sitting on top of a wooden table	a table with a variety of donuts on it
Up-Down+CBS	a group of people sitting on top of a beach	a table that has a **couch** and chairs	a variety of donuts sitting on a donut
Up-Down+CBS(GT)	a group of people sitting on top of a beach.	a large **swimming pool** sitting on a wooden deck.	a variety of **dessert** baked sitting on a table.
NBT	a couple of chairs and a **person** on a beach	a room with a table chairs and a table	a bunch of different types of doughnuts on a table
NBT+CBS	a beach **sky** with chairs and umbrellas on the beach	a room with a table **tree** and a table	a pile of **dessert** sitting on top of a table
NBT+GT	a sandy beach with many chairs and **umbrellas**	a room with a **couch** and a table	a bunch of different types of food on a table
Human	A couple of chairs that are sitting on the beach.	On the deck of a pool is a couch and a display of a safety ring.	Four pie shaped breads on paper with blueberries on top.

Figure 9: Some challenging images from nocaps and corresponding captions generated by existing approaches. The constraints given to the CBS are shown in **blue**. The visual words associated with NBT are shown in **red**.



Labels / Method	In-Domain	Near-Domain	Out-of-Domain
Labels	Person, Woman, Man, Clothing, Footwear	Person, Billboard	Red panda, Tree
Up-Down	a group of people are in a crowd	a large white bus on a city street	a brown bear is laying in the grass
Up-Down+CBS	a group of people watching a man on a stage	a **billboard** on the side of a building	a dog that is standing in the grass
Up-Down+CBS(GT)	a group of **people** standing in front of a crowd	a **billboard** on the side of a building	a large **red panda** walking through the grass
NBT	a group of people in a crowd with a horse	a large white bus on a city street	a large brown bear walking across a lush green field
NBT+CBS	a **man** in a crowd of people with a woman in a crowd	a large white **billboard** truck on a city street	a **tree** that is standing in the grass
NBT+GT	a group of people standing around each other	a large white **billboard** on a city street	a brown and black **red panda** is laying in the grass
Human	Two sumo wrestlers are wrestling while a crowd of men and women watch	A man is standing on the ladder and working at the billboard	The red panda trots across the forest floor.



Labels / Method			
Labels	Suit, Man ,Human face	Bicycle, Person, Land vehicle, Wheel, Wheelchair	Tank, Tree
Up-Down	a woman wearing a white hat is looking at the camera	a person in a yellow chair is riding a bicycle	an old train is parked on the tracks
Up-Down+CBS	a man in a white shirt holding a cell phone	a **person** in a yellow chair in a field	a green train sitting on top of a **tree**
Up-Down+CBS(GT)	a woman in a white **suit** holding a cell phone	a person in a **wheelchair** riding a yellow bike	a **tank** of a train sitting on a track
NBT	a woman in a white shirt is talking on a cellphone	a woman sitting in a cart with a baby in a cart	an old fashioned train engine sitting on the tracks
NBT+CBS	a woman is talking on a **man** with a smile	a woman sitting on a **person** next to a bike	a black and silver **tree** engine on a dirt road
NBT+GT	a **man** and a woman who are looking at something	a woman is sitting in a cart with a cart	a black and white photo of an old fashioned **tank**
Human	The man has a wrap on his head and a white beard.	A person sitting in a yellow chair with wheels.	A tank with an American flag on the side on display.

Figure 10: Some challenging images from `nocaps` and corresponding captions generated by existing approaches. The constraints given to the CBS are shown in **blue**. The visual words associated with NBT are shown in **red**.