nocaps: novel object captioning at scale

Harsh Agrawal[*1] Karan Desai[*1,4] Yufei Wang[2] Xinlei Chen[3] Rishabh Jain[1]
Mark Johnson[2] Dhruv Batra[1,3] Devi Parikh[1,3] Stefan Lee[1,5] Peter Anderson[1]

[1]Georgia Institute of Technology, [2]Macquarie University, [3]Facebook AI Research
[4]University of Michigan, [5]Oregon State University

[1]{hagrawal9, kdexd, rishabhjain, dbatra, parikh, steflee, peter.anderson}@gatech.edu

[2]{yufei.wang, mark.johnson}@mq.edu.au   [3]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, nearly 400 object classes seen in test images have no or very few associated training captions (hence, nocaps). We extend existing novel object captioning models to establish strong baselines for this benchmark and provide analysis to guide future work.

1. Introduction

Recent progress in image captioning, the task of generating natural language descriptions of visual content [9,10,16, 17,41,44], can be largely attributed to the publicly available large-scale datasets of image-caption pairs [5,14,48] as well as steady modeling improvements [4,24,35,46]. However, these models generalize poorly to images in the wild [37] despite impressive benchmark performance, because they are trained on datasets which cover a tiny fraction of the long-tailed distribution of visual concepts in the real world. For example, models trained on COCO Captions [5] can typically describe images containing dogs, people and um-
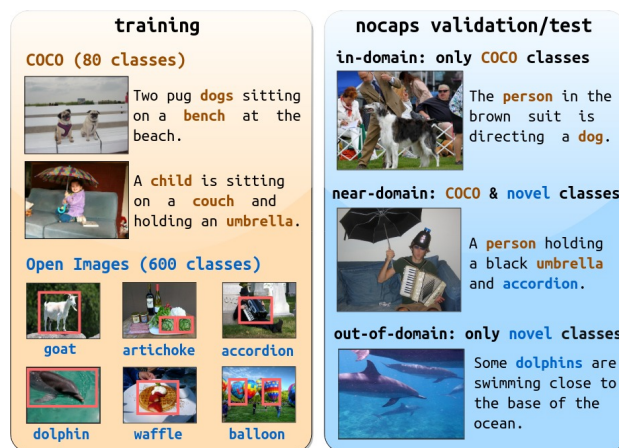

Figure 1: The nocaps task setup: Image captioning models must exploit the Open Images object detection dataset (bottom left) to successfully describe novel objects not covered by the COCO Captions dataset (top left). The nocaps benchmark (right) evaluates performance over in-domain, near-domain and out-of-domain subsets of images containing only COCO classes, both COCO and novel classes, and only novel classes, respectively.

brellas, 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 leverage alternative data sources – such as object detection datasets – in order to describe objects not present in the caption corpora on which they are trained. Such objects which have detection annotations but are not present in caption corpora are referred to as *novel objects* and the task of describing images containing novel objects is termed *novel object captioning* [2,3,13,25,40,43,47]. Until now, novel object captioning approaches have been evaluated using a proof-of-concept dataset introduced in [12]. This dataset has restric-

tive assumptions – it contains only 8 novel object classes held out from the COCO dataset [13], deliberately selected to be highly similar to existing ones (e.g. horse is seen, zebra is novel). This has left the large-scale performance of these methods open to question. Given the emerging interest and practical necessity of this task, we introduce **nocaps**, the first large-scale benchmark for novel object captioning, containing nearly 400 novel object classes.

In detail, the **nocaps** benchmark consists of a validation and test set comprised of 4,500 and 10,600 images, respectively, sourced from the Open Images object detection dataset [18] and annotated with 11 human-generated captions per image (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 Captions 2017 [5] training set (118K images containing 80 object classes), plus the Open Images V4 object detection training set (1.7M images annotated with bounding boxes for 600 object classes and image labels for 20K categories).

To be successful, image captioning models may 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. Our key scientific goal is to disentangle 'how to recognize an object' from 'how to talk about it'. After learning the name of a novel object, a human can immediately talk about its attributes and relationships. It is therefore intellectually dissatisfying that existing models, having already internalized a huge number of caption examples, can't also be taught new objects. As with previous work, this task setting is also motivated by the observation that collecting human-annotated captions is resource intensive and scales poorly as object diversity grows, while on the other hand, large-scale object classification and detection datasets already exist [8, 18] and their collection can be massively scaled, often semi-automatically [28, 29].

To establish the state-of-the-art on our challenging benchmark, we evaluate two of the best performing existing approaches [2, 25] and report their performance based on well-established evaluation metrics – CIDEr [39] and SPICE [1]. To provide finer-grained analysis, we further break performance down over three subsets – **in-domain**, **near-domain** and **out-of-domain**– corresponding to the similarity of depicted objects to COCO classes. While these models do improve over a baseline model trained only on COCO Captions, they still fall well short of human performance on this task – indicating there is still work to be done to scale to 'in-the-wild' image captioning.

In summary, we make three main contributions:

- We collect **nocaps** – the first large-scale benchmark for novel object captioning, containing ∼400 novel objects.

- We undertake a detailed investigation of the performance and limitations of two existing state-of-the-art models on this task and contrast them against human performance.
- We make improvements and suggest simple heuristics that improve the performance of constrained beam search significantly on our benchmark.

We believe that improvements on **nocaps** will accelerate progress towards image captioning in the wild. We are hosting a public evaluation server on EvalAI [45] to benchmark progress on **nocaps**. For reproducibility and to spur innovation, we have also released code to replicate our experiments at: **https://github.com/nocaps-org**.

2. Related Work

Novel Object Captioning Novel object captioning includes aspects of both transfer learning and domain adaptation [6]. Test images contain previously unseen, or 'novel' objects that are drawn from a target distribution (in this case, Open Images [18]) that differs from the source/training distribution (COCO [5]). To obtain a captioning model that performs well in the target domain, the Deep Compositional Captioner [13] and its extension, the Novel Object Captioner [40], both attempt to transfer knowledge by leveraging 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 [25] and the Decoupled Novel Object Captioner [43], take inspiration from Baby Talk [19] 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 Neural Baby Talk, the LSTM-C [47] 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.

In contrast to these works, several approaches to novel object captioning are architecture agnostic. Constrained Beam Search [2] 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) [3] 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) [7].

In this work, we investigate two different approaches: Neural Baby Talk (NBT) [25] and Constrained Beam Search (CBS) [2] on our challenging benchmark – both of which recently claimed state-of-the-art on the proof-of-concept novel object captioning dataset [13].
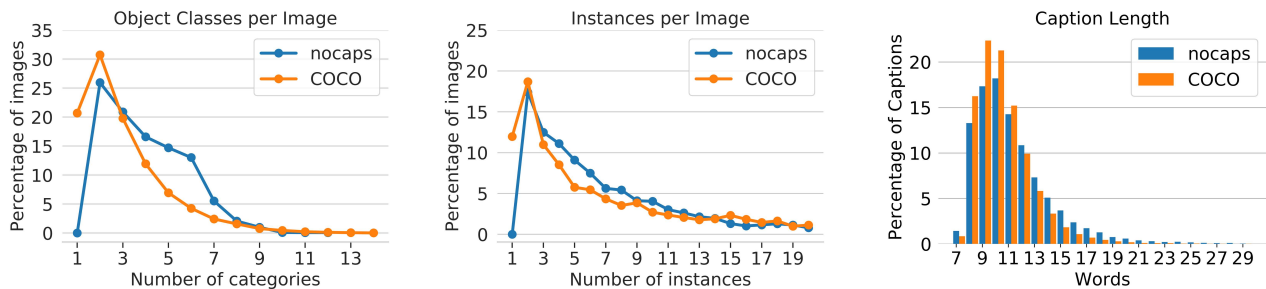
Figure 2: Compared to COCO Captions [5], on average **nocaps** images have more object classes per image (4.0 vs. 2.9), more object instances per image (8.0 vs. 7.4), and longer captions (11 words vs. 10 words). These differences reflect both the increased diversity of the underlying Open Images data [18], and our image subset selection strategy (refer Section 3.1).

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 [14], Flickr 30K [48] and COCO Captions [5] are collected using crowd workers who are given instructions to control the 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, COCO Captions, is based only on a relatively small set of 80 object classes. In contrast, filtered datasets, such as Im2Text [27], Pinterest40M [26] and Conceptual Captions [36], contain large numbers of image-caption pairs harvested from the web. These datasets contain many diverse 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 may not include enough captions per image for good correlation between automatic evaluation metrics and human judgments [1, 39].

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 **nocaps** collection process, constrast it with COCO Captions [5], and introduce the evaluation protocol and benchmark guidelines.

3.1. Caption Collection

The images in **nocaps** are sourced from the Open Images V4 [18] validation and test sets. [1] 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 the 500 classes that are not overly broad (e.g. 'clothing') or infrequent (e.g. 'paper cutter'),

nearly 400 are never or rarely mentioned in COCO Captions [5] (which we select as image-caption training data), making these images an ideal basis for our benchmark.

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, we select subsets of images from the Open Images validation and test splits by applying the following sampling procedure.

First, we exclude all images for which the correct image rotation is non-zero or unknown. Next, based on the ground-truth object detection annotations, we exclude all images that contain only instances from a single object category. Then, to capture as many visually complex images as possible, we include all images containing more than 6 unique object 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 randomly 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 prevents **nocaps** from being overly dominated by frequently occurring object classes such as person, car or plant. In total, we select 4,500 validation images (from a total of 41,620 images in Open Images validation set) and 10,600 test images (from a total of 125,436 images in Open Images test set). On average, the selected images contain 4.0 object classes and 8.0 object instances each (see Figure 2).

Collecting Image Captions from Humans To evaluate model-generated image captions, we collected 11 English captions for each image from a large pool of crowd-workers on Amazon Mechanical Turk (AMT). Out of 11 captions, we randomly sample one caption per image to establish hu-

---

[1]The images used in **nocaps** come from the Open Images V4 dataset and are provided under their original license (CC BY 2.0)

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: 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.

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.

man performance on **nocaps** and use the remaining 10 captions as reference captions for automatic evaluation. Prior work suggests that automatic caption evaluation metrics correlate better with human judgment when more reference captions are provided [1, 39], motivating us to collect more reference captions than COCO (only 5 per image).

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 present in the image. To minimize the potential for this priming to reduce the language diversity of the resulting captions, 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 clutter, we did not display object classes which are classified in Open Images as parts, e.g. human hand, tire, door handle. Pilot studies comparing captions collected with and without priming demonstrated that primed workers produced more qualitative accurate and descriptive captions (see Figure 3). Therefore, all **nocaps** captions, including our human baselines, were collected using this priming-modified COCO collection interface.

To help maintain the quality of the collected captions, we used only US-based workers who had completed at least 5K previous tasks on AMT with more than 95% approval rate. We also 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 workers participated, writing 228 captions each on average for a grand total of 166,100 captions of **nocaps**.

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 (4,500 images / 22,500 captions) uniformly randomly selected subset 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).

3.2. Dataset Analysis

In this section, we compare our **nocaps** benchmark to COCO Captions [5] in terms of both image content and caption diversity. Based on ground-truth object detection annotations, **nocaps** contains images spanning 600 object classes, while COCO contains only 80. 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) as shown in Figure 2. Further, **nocaps** contains no iconic images containing just one object class, whereas 20% of the COCO dataset consists of such images. Similarly, less than 10% of COCO images contain more than 6 object classes, while such images constitute almost 22% of **nocaps**.

Although priming the workers with object classes as keywords during data collection has the potential to reduce language diversity, **nocaps** captions are nonetheless more diverse than COCO. 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 uniformly random 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. The number of unique 2, 3 and 4-grams is also significantly higher for **nocaps**– suggesting a greater variety of unique language compositions as well.

Additionally, we compare visual and linguistic similarity between COCO, **in-domain** and **out-of-domain** in Figure 4. We observe that **in-domain** classes shows high visual similarity to equivalent COCO classes (e.g. cat, book) while many **out-of-domain** classes are visually and linguistically different from **in-domain** classes (e.g. jellyfish, beetle, cello). **out-of-domain** also covers many visually and linguistically similar concepts to COCO but rarely described in COCO (e.g. tiger, lemon)

3.3. Evaluation

The aim of **nocaps** is to benchmark progress towards models that can describe images containing visually novel concepts in the wild by leveraging other data sources. To facilitate evaluation and avoid exposing the novel object captions, we host an evaluation server for **nocaps** on
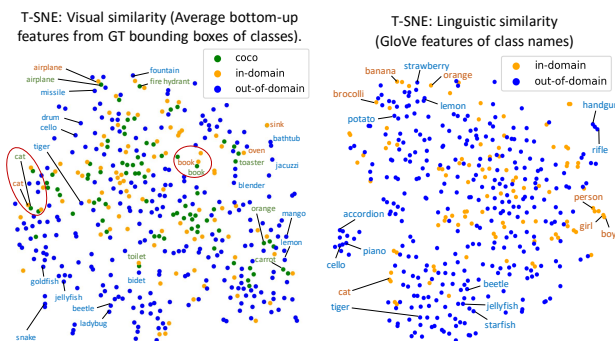
Figure 4: T-SNE [38] plots comparing visual (left) and linguistic (right) similarity in COCO, **in-domain** and **out-of-domain** classes. We observe that: (a) **in-domain** shows high visual similarity to COCO (e.g. cat, book (left)). (b) Many **out-of-domain** classes are visually and linguistically different from **in-domain** classes (e.g. jellyfish, beetle, cello). (c) **out-of-domain** also covers many visually and linguistically similar concepts to COCO, which are not well-covered in COCO (e.g. tiger, lemon).

EvalAI [45] – as such, we put forth these guidelines for using **nocaps**:

– **Do not use additional paired image-caption data collected from humans.** Improving evaluation scores by leveraging additional human-generated paired image-caption data is antithetical to this benchmark – *the only paired image-caption dataset that should be used is the COCO Captions 2017 training split*. However, external text corpora, knowledge bases, and object detection datasets may be used during training or inference.

– **Do not leverage ground truth object annotations.** We note that ground-truth object detection annotations are available for Open Images validation and test splits (and hence, for **nocaps**). While ground-truth annotations may be used to establish performance upper bounds on the validation set, they should never be used in a submission to the evaluation server unless this is clearly disclosed.

We anticipate that researchers may wish to investigate the limits of performance on **nocaps** without any restraints on the training datasets. We therefore maintain a separate leaderboard for this purpose "**nocaps (XD)**"[2] leaderboard.

Metrics As with existing captioning benchmarks, we rely on automatic metrics to evaluate the quality of model-generated captions. We focus primarily on CIDEr [39] and SPICE [1], which have been shown to have the strongest correlation with human judgments [23] and have been used in prior novel object captioning work [3,12,25], but we also report Bleu [30], Meteor [20] and ROUGE [22]. These metrics test whether models mention novel objects accurately [41] as well as describe them fluently [20].It is worth noting that the absolute scale of these metrics is not comparable across datasets due to the differing number of reference captions and corpus-wide statistics.

near-domain	out-of-domain
1. A **man** sitting in the saddle on a **camel**.	1. A **tank** vehicle stopped at a gas station.
2. A **person** is sitting on a **camel** with another **camel** behind him.	2. A **tank** and a military jeep at a gas station
3. A **man** with long hair and blue jeans sitting on a **camel**.	3. A jeep and a tan colored **tank** getting gas at a gas station.
4. **Man** sitting on a **camel** with a standing **camel** behind them.	4. A **tank** and a **truck** sit at a gas station pump.
5. Long haired **man** wearing sitting on blanket draped **camel**	5. An Army humvee is at getting gas from the 76 gas station.
6. A **camel** stands behind a sitting **camel** with a **man** on its back.	6. An army **tank** is parked at a gas station.
7. The standing **camel** is near a sitting one with a **man** on its back.	7. A **land vehicle** is parked in a gas station fueling.
8. Someone is sitting on a **camel** and is in front of another **camel**.	8. A large military vehicle at the gas pump of a gas station.
9. Two **camels** in the dessert and a **man** sitting on the sitting one.	9. A tanker parked outside of an old gas station
10. Two **camels** are featured in the sand with a **man** sitting on one of the seated **camels**.	10. Multiple military vehicles getting gasoline at a civilian gas station.

Figure 5: Examples of **near-domain** and **out-of-domain** images from the **nocaps** validation set. The image on the left belongs to the **near-domain** subset (COCO and Open Images categories), while the image on the right belongs to **out-of-domain** subset (only Open Images categories).

Evaluation Subsets We further break down performance on **nocaps** over three subsets of the validation and test splits corresponding to varied 'nearness' to COCO.

To determine these subsets, we manually map the 80 COCO classes to Open Images classes. We then select an additional 39 Open Images classes that are not COCO classes, but are nonetheless mentioned more than 1,000 times in the COCO captions training set (e.g. 'table', 'plate' and 'tree'). We classify these 119 classes as in-domain relative to COCO. There are 87 Open Images classes that are not present in **nocaps**[3]. The remaining 394 classes are out-of-domain. Image subsets are then determined as follows:

– **in-domain** images contain only objects belonging to in-domain classes. Since these objects have been described in the paired image-caption training data, we expect caption models trained only on COCO to perform reasonably well on this subset, albeit with some negative impact due to image domain shift. This subset contains 1,311 test images (13K captions).

– **near-domain** images contain both in-domain and out-of-domain object classes. These images are more challeng-

ing for COCO trained models, especially when the most salient objects in the image are novel. This is the largest subset containing 7,406 test images (74K captions).

– **out-of-domain** images do not contain any in-domain classes, and are visually very distinct from COCO images. We expect this subset to be the most challenging and models trained only on COCO data are likely to make 'embarrassing errors' [23] on this subset, reflecting the current performance of COCO trained models in the wild. This subset contains 1,883 test images (19K captions).

4. Experiments

To provide an initial measure of the state-of-the-art on **nocaps**, we extend and present results for two contemporary approaches to novel object captioning – Neural Baby Talk (NBT) [25] and Constrained Beam Search (CBS) [2] inference method which we apply both to NBT and to the popular UpDown captioner [4]. We briefly recap these approaches for completeness but encourage readers to seek the original works for further details.

Bottom-Up Top-Down Captioner (UpDown) [4] reasons over visual features extracted using object detectors trained on a large numbers of object and attribute classes and produces near state-of-the-art for single model captioning performance on COCO. For visual features, we use the publicly available Faster R-CNN [34] detector trained on Visual Genome by [4] to establish a strong baseline trained exclusively on paired image-caption data.

Neural Baby Talk (NBT) [25] first generates a hybrid textual template with slots explicitly tied to specific image regions, and then fill these slots with words associated with visual concepts identified by an object detector. This gives NBT the capability to caption novel objects when combined with an appropriate pretrained object detector. To adapt NBT to the **nocaps** setting, we incorporate the Open Images detector and train the language model using Visual Genome image features. We use fixed GloVe embeddings [31] in the visual feature representation for an object region for better contextualization of words corresponding to novel objects.

Open Images Object Detection. Both CBS and NBT make use of object detections; we use the same pretrained Faster R-CNN model trained on Open Images for both. Specifically, we use a model[4] from the Tensorflow model zoo [15] which achieves a detection mean average precision at 0.5 IoU (mAP@0.5) of 54%.

Constrained Beam Search (CBS) [2] CBS is an inference-time procedure that can force language models to include specific words referred to as constraints – achieving this by casting the decoding problem as a finite state machine with transitions corresponding to constraint satisfaction. We apply CBS to both the baseline UpDown model and NBT based on detected objects. Following [2], we use a Finite State Machine (FSM) with 24 states to incorporate up to

three selected objects as constraints, including two and three word phrases. After decoding, we select the highest log-probability caption that satisfies at least two constraints.

Constraint Filtering Although the original work [2] selected constraints from detections randomly, in preliminary experiments in the **nocaps** setting we find that a simple heuristic significantly improves the performance of CBS. To generate caption constraints from object detections, we refine the raw object detection labels by removing 39 Open Images classes that are 'parts' (e.g. human eyes) or rarely mentioned (e.g. mammal). Specifically, we resolve overlapping detections (IoU $\geqslant 0.85$) by removing the higher-order of the two objects (e.g. , a 'dog' would suppress a 'mammal') based on the Open Images class hierarchy (keeping both if equal). Finally, we take the top-3 objects based on detection confidence as constraints.

Language Embeddings To handle novel vocabulary, CBS requires word embeddings or a language model to estimate the likelihood of word transitions. We extend the original model – which incorporated GloVe [31] and dependency embeddings [21] – to incorporate the recently proposed ELMo [32] model, which increased performance in our preliminary experiments. As captions are decoded left-to-right, we can only use the forward representation of ELMo as input encodings rather than the full bidirectional model as in [11, 42]. We also initialize the softmax layer of our caption decoder with that of ELMo and fix it during training to improve the model's generalization to unseen or rare words.

Training and Implementation Details. We train all models on the COCO training set and tune parameters on the **nocaps** validation set. All models are trained with cross-entropy loss, i.e. we do not use RL fine-tuning to optimize for evaluation metrics [35].

5. Results and Analysis

We report results on the **nocaps** test set in Table 2. While our best approach (UpDown + ELMo + CBS, which is explained further below) outperforms the COCO-trained UpDown baseline captioner significantly ($\sim$19 CIDEr), it still under-performs humans by a large margin ($\sim$12 CIDEr). As expected the most sizable gap occurs for **out-of-domain** instances ($\sim$25 CIDEr). This shows that while existing novel object captioning techniques do improve over standard models, captioning in-the-wild still presents a considerable open challenge.

In the remainder of this section, we discuss detailed results on the **nocaps** and COCO validation sets (Table 3) to help guide future work. Overall, the evidence suggests that further progress can be made through stronger object detectors and stronger language models, but open questions remain – such as the best way to combine these elements, and the extent to which that solution should involve learning vs. inference techniques like CBS. We align these discussions in the context of a series of specific questions below.

---

[4]tf_faster_rcnn_inception_resnet_v2_atrous_oidv4

| | 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
UpDown	74.3	11.5	56.9	10.3	30.1	8.1	74.0	19.2	23.0	50.9	54.3	10.1
UpDown + ELMo + CBS	76.0	11.8	74.2	11.5	66.7	9.7	76.6	18.4	24.4	51.8	73.1	11.2
NBT	60.9	9.8	53.2	9.3	48.7	8.2	72.3	14.7	21.5	48.9	53.4	9.2
NBT + CBS	63.0	10.1	62.0	9.8	58.5	8.8	73.4	12.9	22.1	48.7	61.5	9.7
Human	**80.6**	**15.0**	**84.6**	**14.7**	**91.6**	**14.2**	**76.6**	**19.5**	**28.2**	**52.8**	**85.3**	**14.6**

Table 2: Single model image captioning performance on the `nocaps` test split. We evaluate four models, including the UpDown model [4] trained only on COCO, as well as three model variations based on constrained beam search (CBS) [2] and Neural Baby Talk (NBT) [25] that leverage the Open Images training set.

| | 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
(1) UpDown	**77.0**	**37.2**	**27.8**	**116.2**	**21.0**	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1
(2) UpDown + CBS	73.3	32.4	25.8	97.7	18.7	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
(3) UpDown + ELMo + CBS	72.4	31.5	25.7	95.4	18.2	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
(4) UpDown + ELMo + CBS + GT	-	-	-	-	-	84.2	12.6	82.1	11.9	86.7	10.6	83.3	11.8
(5) NBT	72.7	29.4	23.8	88.3	16.5	62.7	10.1	51.9	9.2	54.0	8.6	53.9	9.2
(6) NBT + CBS	70.2	28.2	25.1	80.2	15.8	62.3	10.3	61.2	9.9	63.7	9.1	61.9	9.8
(7) NBT + CBS + GT	-	-	-	-	-	68.9	10.7	68.6	10.3	76.9	9.8	70.3	10.3
(8) Human	66.3	21.7	25.2	85.4	19.8	**84.4**	**14.3**	**85.0**	**14.3**	**95.7**	**14.0**	**87.1**	**14.2**

Table 3: Single model image captioning performance on the COCO and `nocaps` validation sets. We begin with a strong baseline in the form of the UpDown [4] trained on COCO captions. We then investigate decoding using Constrained Beam Search [2] based on object detections from the Open Images detector (+ CBS), as well as the impact of incorporating a pretrained language model (+ ELMo) and ground-truth object detections (+ GT), respectively. In panel 2, we review the performance of Neural Baby Talk (NBT) [25], 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). COCO human scores refer to the test split.

– **Do models optimized for `nocaps` maintain their performance on COCO?** We find significant gains in `nocaps` performance correspond to large losses on COCO (rows **2-3 vs 1** – dropping ∼20 CIDEr and ∼3 SPICE). Given the similarity of the collection methodology, we do not expect to see significant differences in linguistic structure between COCO and `nocaps`. However, recent work has observed significant performance degradation when transferring models across datasets even when the new target dataset is an exact recreation of the old dataset [33]. Limiting this degradation in the captioning setting is a potential focus for future work.

– **How important is constraint filtering?** Applying CBS greatly improves performance for both UpDown and NBT (particularly on the `out-of-domain` captions), but success depends heavily on the quality of the constraints. Without our 39-class blacklist and overlap filtering, we find overall `nocaps` validation performance falls ∼8 CIDEr and ∼3 SPICE for our UpDown + ELMo + CBS

model – with most of the losses coming from the blacklisted classes. It seems likely that more sophisticated constraint selection techniques that consider image context could improve performance further.

– **Do better language models help in CBS?** To handle novel vocabulary, CBS requires representations for the novel words. We compare using ELMo encoding (row **3**) as described in Section 4 with the setting in which word embeddings are only learned during COCO training (row **2**). Note that in this setting the embedding for any word not found in COCO is randomly initialized. Surprisingly, the trained embeddings perform on par with the ELMo embeddings for the `in-domain` and `near-domain` subsets, although the model with ELMo performs much better on the `out-of-domain` subset. It appears that even relatively rare occurrences of `nocaps` object names in COCO are sufficient to learn useful linguistic models, but not visual grounding as shown by the COCO-only model's poor scores (row **1**).

	in-domain	**near-domain**	**out-of-domain**

Method			
UpDown	A beach with chairs and umbrellas on it.	A man in a red shirt holding a baseball bat.	A bird on the ocean in the ocean.
+ ELMo	A beach with chairs and umbrellas on it.	A man in a red shirt holding a baseball bat.	A bird that is floating on the water.
+ ELMo + CBS	A beach with chairs and **umbrellas** and **kites**.	A man in a red **hat** holding a baseball **rifle**.	A **dolphin** swimming in the ocean on a sunny day.
+ ELMo + CBS + GT	A beach with chairs and **umbrellas** on it.	A man in a red **hat** holding a baseball **rifle**.	A **whale dolphin** swimming in the ocean on the ocean.
NBT	A beach with a bunch of lawn chairs and **umbrellas**.	A baseball **player** holding a baseball bat in the field.	A **dolphin** sitting in the water.
+ CBS	A beach with a bunch of **umbrellas** on a beach.	A baseball **player** holding a baseball **rifle** in the field.	A **marine mammal** sitting on a **dolphin** in the ocean.
+ CBS + GT	A beach with many **umbrellas** on a beach.	A baseball **player** holding a baseball **rifle** in the field.	A black **dolphin** swimming in the ocean on a sunny day.
Human	A couple of chairs that are sitting on a beach.	A man in a red hat is holding a shotgun in the air.	A dolphin fin is up in the water..

Figure 6: Some challenging images from **nocaps** and the corresponding captions generated by our baseline models. The constraints given to the CBS are shown in **blue**, and the grounded visual words associated with NBT are shown in **red**. Models perform reasonably well on **in-domain** images but confuse objects in **near-domain** and **out-of-domain** images with visually similar **in-domain** objects, such as rifle (with baseball bat) and fin (with bird). On the difficult **out-of-domain** images, the models generate captions with repetitions, such as "in the ocean on the ocean", and produce incoherent captions, such as "marine animal" and "dolphin" referring to the same entity in the image.

– **Do better object detectors help?** To evaluate reliance on object detections, we supply ground truth detections sorted by decreasing area to our full models (rows **4** and **7**). These ground truth detections undergo the same constraint filtering as predicted ones. Comparing to prediction-reliant models (rows **3** and **6**), we see large gains on all splits (rows **4 vs 3** – ∼9 CIDEr and ∼0.6 SPICE gain for UpDown). As detectors improve, we expect to see commensurate gains on **nocaps** benchmark.

To qualitatively assess some of the differences between the various approaches, in Figure 6 we illustrate some examples of the captions generated using various model configurations. As expected, all our baseline models are able to generate accurate captions for **in-domain** images. For **near-domain** and **out-of-domain**, our UpDown model trained only on COCO fails to identify novel objects such as rifle and dolphin, and confuses them with known objects such as baseball bat or bird. The remaining models leverage the Open Images training data, enabling them to potentially describe these novel object classes. While they do produce more reasonable descriptions, there remains much room for improvement in both grounding and grammar.

6. 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 500 unique object classes and many more visual concepts. Compared to the existing proof-of-concept dataset for novel object captioning [12], our benchmark contains a fifty-fold increase in the number of novel object classes that are rare or absent in training captions (394 vs 8). Further, we collected twice the number of evaluation captions per image to improve the fidelity of automatic evaluation metrics.

We extend two recent approaches for novel object captioning to provide strong baselines for the **nocaps** benchmark. While our final models improve significantly over a direct transfer from COCO, they still perform well below the human baseline – indicating there is significant room for improvement on this task. We provide further analysis to help guide future efforts, showing that it helps to leverage large language corpora via pretrained word embeddings and language models, that better object detectors help (and can be a source of further improvements), and that simple heuristics for determining which object detections to mention in a caption have a significant impact.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016. 2, 3, 4, 5

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017. 1, 2, 6, 7

[3] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. In *NIPS*, 2018. 1, 2, 5

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 6, 7

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. arXiv:1504.00325, 2015. 1, 2, 3, 4

[6] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *Advances in Computer Vision and Pattern Recognition*, 2017. 2

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1977. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[9] Jeffrey Donahue, Lisa A. Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1

[10] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 1

[11] Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369. Association for Computational Linguistics, 2018. 6

[12] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. In *CVPR*, 2016. 1, 5, 8

[13] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond J. Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 1, 2

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 1, 3

[15] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wo-

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[17] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, 2015. 1

[18] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin 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

[19] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Babytalk: Understanding and generating simple image descriptions. *PAMI*, 35(12):2891–2903, 2013. 2

[20] Alon Lavie and Abhaya 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

[21] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL*, 2014. 6

[22] Chin-yew 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

[23] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *ICCV*, 2017. 5, 6

[24] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 1

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 1, 2, 5, 6, 7

[26] Junhua Mao, Jiajing Xu, Kevin Jing, and Alan L Yuille. Training and evaluating multimodal word embeddings with large-scale web annotated images. In *NIPS*, 2016. 3

[27] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 3

[28] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*, 2016. 2

[29] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017. 2

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP*, 2014. 6

[32] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer.

Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018. 6

[33] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv:1806.00451*, 2018. 7

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 6

[35] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 1, 6

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 3

[37] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich Image Captioning in the Wild. In *CVPR Workshops*, 2016. 1

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 5

[39] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 3, 4, 5

[40] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Captioning Images with Diverse Objects. In *CVPR*, 2017. 1, 2

[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 5

[42] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019. 6

[43] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. *CoRR*, abs/1804.03803, 2018. 1, 2

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1

[45] Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. Evalai: Towards better evaluation systems for ai agents. arXiv:1902.03570, 2019. 2, 5

[46] Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. Review networks for caption generation. In *NIPS*, 2016. 1

[47] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017. 1, 2

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 3