nocaps: novel object captioning at scale

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

[1]Georgia Institute of Technology, [2]Facebook AI Research, [3]Macquarie University
[1]{hagrawal9, kdexd, rishabhjain, dbatra, parikh, steflee, peter.anderson}@gatech.edu
[2]{xinleic}@fb.com [3]{yufei.wang, mark.johnson}@mq.edu.au 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 on this task.*

1. Introduction

Image captioning, the task of generating natural language descriptions of visual content [11, 12, 18, 19, 42, 45], 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 [6, 16, 48]. However, despite continual modeling improvements and ever-increasing benchmark performance [4, 26, 37, 46], existing captioning models generalize poorly to images in the wild [39]. 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. For example, models trained on COCO captions [6] 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.

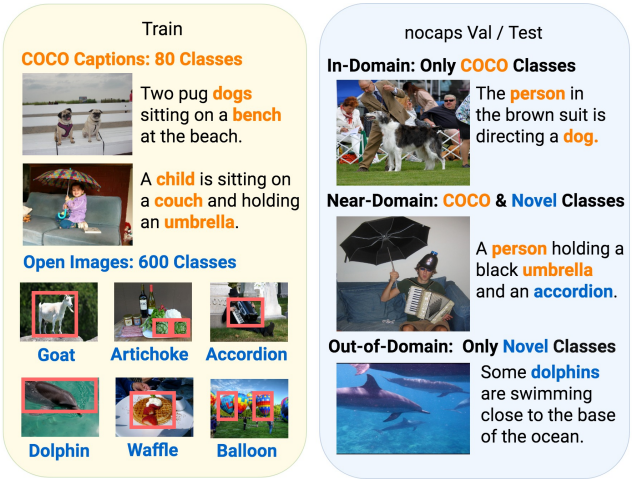


Figure 1: The **nocaps** benchmark for **n**ovel **o**bject **cap**tioning at **s**cale: Image captioning models must exploit object detection training data (bottom left) to successfully describe novel objects which are not present in the available image-caption training data (top left). This capability 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) 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.

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 re-

*First two authors contributed equally, listed in alphabetical order.

ferred to as *novel objects* and the task of describing images containing novel objects is termed *novel object captioning* [2, 3, 15, 27, 41, 44, 47]. Until now, approaches to novel object captioning have been evaluated using a proof-of-concept dataset introduced in [14] with restrictive assumptions. It contained only 8 novel object classes held out from the COCO dataset [15], all 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, particularly as these novel classes were deliberately selected to be semantically similar to clusters of seen classes. Therefore, given the emerging interest and practical necessity of this task, we introduce **nocaps**, the first large-scale benchmark for novel object captioning, containing over 500 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 [20] 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 [6] training set (118K images containing 80 object classes), plus the Open Images V4 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 expensive and scales poorly as object diversity grows, while on the other hand, large-scale object classification and detection datasets already exist [10, 20] and their collection can be massively scaled, often semi-automatically [30, 31].

To establish the state-of-the-art on our challenging benchmark, we evaluate two of the best performing existing approaches [2, 27] and report their performance based on well-established evaluation metrics – CIDEr [40] 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 will host a public evaluation server and leaderboard to benchmark progress. For reproducibility and to spur innovation, we will release code to replicate our experiments. We believe that improvements on **nocaps** will accelerate progress towards image captioning in the wild.

2. Related Work

Novel Object Captioning Novel object captioning includes aspects of both transfer learning and domain adaptation [8]. Test images contain previously unseen, or 'novel' objects that are drawn from a target distribution (in this case, Open Images [20]) that differs from the source/training distribution (COCO [6]). To obtain a captioning model that performs well in the target domain, we wish to transfer knowledge from a large dataset of target-domain object detection annotations, and possibly other datasets such as external text corpora. A variety of approaches have been proposed for this specific setting. The Deep Compositional Captioner [15] and its extension, the Novel Object Captioner [41], 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 [27] and the Decoupled Novel Object Captioner [44], take inspiration from Baby Talk [21] 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. 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

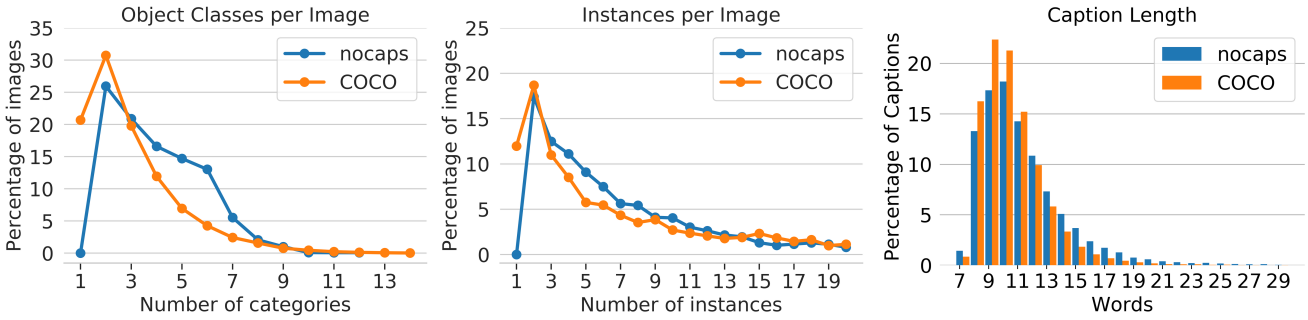


Figure 2: Compared to COCO Captions [6], on average **nocaps** images have more object classes per image (4.0 vs. 2.9), more object instances per image (8.0 vs. 7.4), and longer captions (11 words vs. 10 words). These differences reflect both the increased diversity of the underlying Open Images data [20], and our image subset selection strategy (refer Section 3.1).

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) [9].

In this work, we investigate Neural Baby Talk (NBT) [27] and Constrained Beam Search (CBS) [2] in our challenging benchmark. We choose these approaches because they represent diverse methods for this task, and because both recently claimed state-of-the-art on a simple proof-of-concept novel object captioning dataset [15].

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 [16], Flickr 30K [48] and COCO Captions [6] 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 [29], Pinterest40M [28] and Conceptual Captions [38], 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 only include one caption per image, which reportedly decreases correlation between automatic evaluation metrics and human judgments [1, 40].

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 [6], and introduce the evaluation protocol and benchmark guidelines.

3.1. Caption Collection

The images in **nocaps** are sourced from the Open Images V4 [20] 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 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 [6] (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

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.

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 Human Image Captions 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 oracle 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 [1,40], motivating us to collect a larger number of reference captions than COCO (only 5 per image).

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).

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, 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, etc. Early 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 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 grand total of 166,100 captions about novel objects.

3.2. Dataset Analysis

In this section, we compare our **nocaps** benchmark to COCO Captions [6] 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.

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 will host an evaluation server for **nocaps**– as such, we put forth these guidelines for using **nocaps**:

– **Do not use additional paired image-caption data.** Improving evaluation scores by leveraging additional paired data is antithetical to this benchmark – *the only paired image-caption dataset that should be used is the COCO 2017 training split*. However, other datasets such as external text corpora, knowledge bases, and additional 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 object 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.

Metrics As with existing captioning benchmarks, we rely on automatic metrics to evaluate the quality of model-generated captions. We focus primarily on CIDEr [40] and SPICE [1], which have been shown to have the strongest correlation with human judgments [25] and have been used in prior novel object captioning work [3, 14, 27], but we also report Bleu [32], Meteor [22] and ROUGE [24]. These metrics test whether models mention novel objects accurately [42] as well as describe them fluently [22]. 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

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 with 10 reference captions. The image on the left belongs to the **near-domain** subset which contains both **COCO** and **Open Images** categories, while the image on the right belongs to **out-of-domain** subset which contains 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**[1]. 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-

[1]These classes are not included either because they are not present in the underlying Open Images val and test splits, or because they got filtered out by our image subset selection strategy favoring more complex images.

domain object classes. These images are more challenging 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 therefore 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' [25] on this subset, reflecting the current performance of COCO trained models in the wild. There are 1,883 test images (19K captions) in this subset.

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) [27] and Constrained Beam Search (CBS) [2] inference method which we apply both to NBT and to the popular Up-Down 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 (Up-Down) [4] To establish a strong baseline model trained exclusively on paired image-caption data, we select the Bottom-Up Top-Down (Up-Down) image captioning model [4], which is near state-of-the-art for single model captioning performance on COCO and has code available. In [4], the authors find significant improvements by reasoning over visual representations extracted using object detectors trained on a large numbers of object and attribute classes. Following this, we use the publicly-available features extracted using Faster R-CNN [36] trained on Visual Genome by [4] for all models.

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[2] from the Tensorflow model zoo [17] which achieves a detection mean average precision at 0.5 IoU (mAP@0.5) of 54%.

Neural Baby Talk (NBT) [27] Neural Baby Talk (NBT) [27] performs captioning in two stages, first generating a hybrid textual template with slots explicitly tied to specific image regions, and then filling these 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 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 [33] in both the language model and the visual feature representation for an object region for better contextualization of words corresponding to novel objects.

[2] `tf_faster_rcnn_inception_resnet_v2_atrous_oidv2`

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 Up-Down 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 [33] and dependency embeddings [23] – to incorporate the recently proposed ELMo [34] 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 [13, 43]. 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 [37]. We will release all code to reproduce these experiments upon acceptance.

5. Results and Analysis

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

	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	73.7	11.6	57.2	10.3	30.4	8.1	74.1	18.9	22.9	50.7	54.5	10.1
Up-Down + 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	62.8	10.3	51.9	9.4	48.9	8.4	71.8	14.2	21.8	48.0	54.3	9.4
NBT + CBS	61.9	10.4	57.3	9.6	61.8	8.6	69.6	12.4	21.6	46.7	59.9	9.5
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 Up-Down model [4] trained only on COCO, as well as three model variations based on constrained beam search (CBS) [2] and Neural Baby Talk (NBT) [27] 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) Up-Down	**77.0**	**37.2**	**27.8**	**116.2**	**21.0**	77.6	11.6	58.4	10.4	32.3	8.3	55.8	10.2
(2) Up-Down + 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) Up-Down + 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) Up-Down + ELMo + CBS + GT	-	-	-	-	-	84.2	12.6	82.1	11.9	86.7	10.6	83.3	11.8
(5) NBT	72.2	31.5	25.3	94.1	18.0	62.6	10.0	52.7	9.4	51.8	8.6	54.0	9.3
(6) NBT + CBS	70.2	28.2	25.1	92.8	18.1	62.1	10.1	58.3	9.4	62.4	8.9	60.2	9.5
(7) NBT + CBS + GT	-	-	-	-	-	62.4	10.1	59.7	9.5	64.9	9.1	62.3	9.6
(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 Up-Down [4] image captioning model 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 using ground-truth object detections (+ GT), respectively. In panel 2, we review the performance of Neural Baby Talk (NBT) [27], 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 refer to the test split.

dard 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.

– **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). This is less pronounced for NBT (row **6 vs 5**), although the impact of applying CBS to this model was less overall. 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 [35]. 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 Up-Down 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 Up-Down + 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.

Method	in-domain 	near-domain 	out-of-domain 
Up-Down	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.
Up-Down + 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.
Up-Down + 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.
Up-Down + 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.
NBT + 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.
NBT + 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 5: 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**. While models perform reasonably well on **in-domain** images, they 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). Furthermore, 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 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**).

– **Do better object detectors help?** To evaluate reliance on object detections, we supply ground truth detections to our full models (rows **4** and **7**). Note that ground truth detections undergo the same constraint filtering as predicted ones, except they are sorted by area rather than confidence (which is ill-defined). Comparing to prediction-reliant models (rows **3** and **6**), we see large gains on all splits for our Up-Down based model (∼9 CIDEr and ∼0.6 SPICE), but lesser gains for NBT. As detectors improve, we expect to see commensurate gains on **nocaps** captioning performance.

To qualitatively assess some of the differences between the various approaches, in Figure 5 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 Up-Down 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 [14], 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.

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] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016. 2, 3, 4, 5, 12, 13

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017. 2, 3, 6, 7, 16, 19

[3] P. Anderson, S. Gould, and M. Johnson. Partially-supervised image captioning. In *NIPS*, 2018. 2, 3, 5

[4] 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, 6, 7, 16

[5] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 19

[6] 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*, abs/1504.00325, 2015. 1, 2, 3, 4

[7] X. Chen, T.-Y. L. Hao Fang, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 11

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

[9] 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. 3

[10] 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*, pages 248–255, 2009. 2

[11] 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

[12] 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

[13] L. He, K. Lee, O. Levy, and L. 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

[14] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. In *CVPR*, 2016. 2, 5, 8

[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)*, pages 1–10, 2016. 2, 3

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

[17] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 6

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

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

[20] 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. Open-images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 2, 3, 16, 17

[21] 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*, 35(12):2891–2903, 2013. 2

[22] 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

[23] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *ACL*, 2014. 6, 17

[24] 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

[25] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *ICCV*, 2017. 5, 6

[26] 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

[27] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *CVPR*, 2018. 2, 3, 5, 6, 7, 16

[28] 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

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

[30] 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

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

[32] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[33] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP*, 2014. 6, 17

[34] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. 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, 19

[35] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018. 7

[36] 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, 16

[37] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 1, 6

[38] 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

[39] 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

[40] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 3, 4, 5, 12, 13

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

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 5

[43] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019. 6

[44] Y. Wu, L. Zhu, L. Jiang, and Y. Yang. Decoupled novel object captioner. *CoRR*, abs/1804.03803, 2018. 2

[45] 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

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

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

[48] 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*, 2:67–78, 2014. 1, 3

Appendix

In the first section we provide additional details in relation to the **nocaps** benchmark, including the data collection interface and further qualitative examples and analysis. In the second section we provide implementation details for our baseline models and further examples of predicted captions on the three (**in-domain**, **near-domain** and **out-of-domain**) subsets of the **nocaps** validation set.

1. Additional Details about the nocaps Benchmark

1.1. Collection Interface

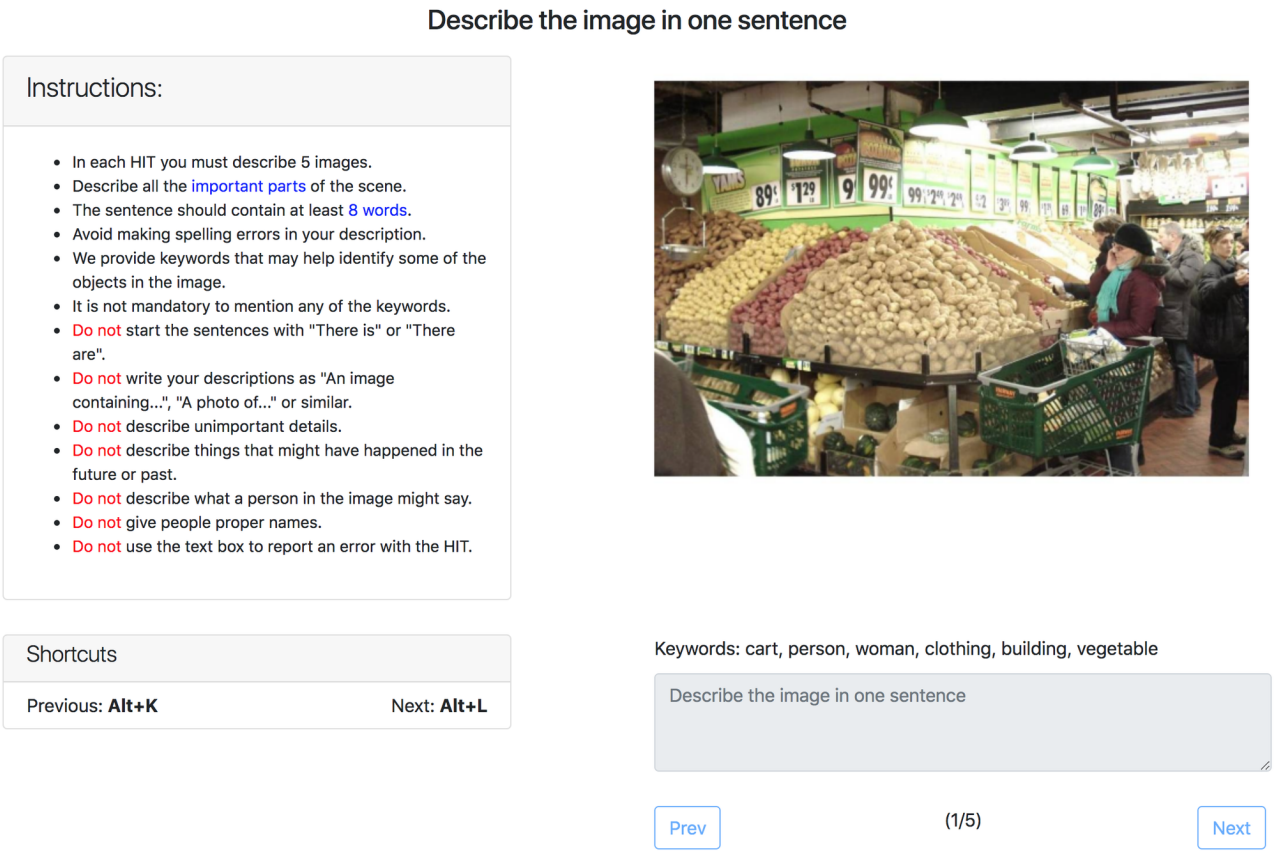


Figure 6: Amazon Mechanical Turk (AMT) user interface with priming for gathering captions. The interface shows a subset of object categories present in the image as keywords. Note that the instruction explicitly states that it is not mandatory to mention any of the displayed keywords. Other instructions are similar to the interface described in [7]

1.2. Example 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**.

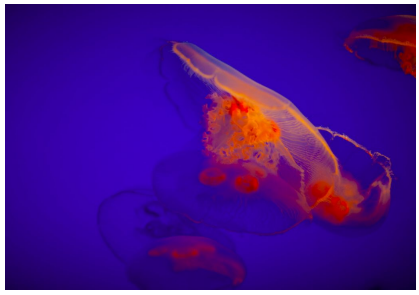
1. Jockeys on **horses** racing around a track.
2. Several **horses** are running in a race thru the grass.
3. several **people** racing **horses** around a turn outside
4. Uniformed jockeys on **horses** racing through a grass field.
5. Several **horse** jockies are riding **horses** around a turn.
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

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.

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.

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 **jellyfish** 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. 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 the **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 [1, 40]. Categories in **orange** are **in-domain** object classes while categories in **blue** are **out-of-domain** classes. Note that not all captions mention the ground-truth object classes consistent with the instructions provided on the data collection interface.

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. woman wearing jeans in front of the truck

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 the **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 [1, 40]. Categories in **orange** are **in-domain** object classes while categories in **blue** are **out-of-domain** classes. Note that not all captions mention the ground-truth object classes consistent with the instructions provided on the data collection interface.

1.3. Evaluation Subsets

As outlined in Section 3.3 of the main paper, to determine the **in-domain**, **near-domain** and **out-of-domain** subsets of **nocaps**, we first classify Open Images classes as either **in-domain** or **out-of-domain** with respect to COCO. To identify the **in-domain** Open Images classes, 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'), and we classify all 119 of these classes as **in-domain**. The remaining classes are considered to be **out-of-domain**.

To put this in perspective, in Figure 9 we plot the number of mentions of both the **in-domain** classes (in orange) and the **out-of-domain** classes (in blue) in the COCO Captions training set using a log scale. As intended, the **in-domain** object classes occur much more frequently in COCO Captions compared to **out-of-domain** object classes. However, it is worth noting that the **out-of-domain** are not necessarily absent from COCO Captions, but they are relatively infrequent which makes these concepts hard to learn from COCO.

Open Images classes ignored during image subset selection: We also note that 87 Open Images classes were not considered during the image subset selection procedure to create **nocaps**, for one of the following reasons:

- **Parts:** In our image subset selection strategy (refer Section 3.1 of the main paper), we ignored 'part' categories such as 'vehicle registration plate', 'wheel', 'human-eye', which always occur with parent categories such car, person;
- **Super-categories:** Our image subset selection strategy also ignored super-categories such as 'sports equipment', 'home appliance', 'auto part' which are often too broad and subsumes both COCO and Open Images categories;
- **Solo categories:** Certain categories such as 'chime' and 'stapler' did not appear in images alongside any other classes, and so were filtered out by our image subset selection strategy; and
- **Rare categories:** Some rare categories such as 'armadillo', 'pencil sharpener' and 'pizza cutter' do not actually occur in the underlying Open Images val and test splits.

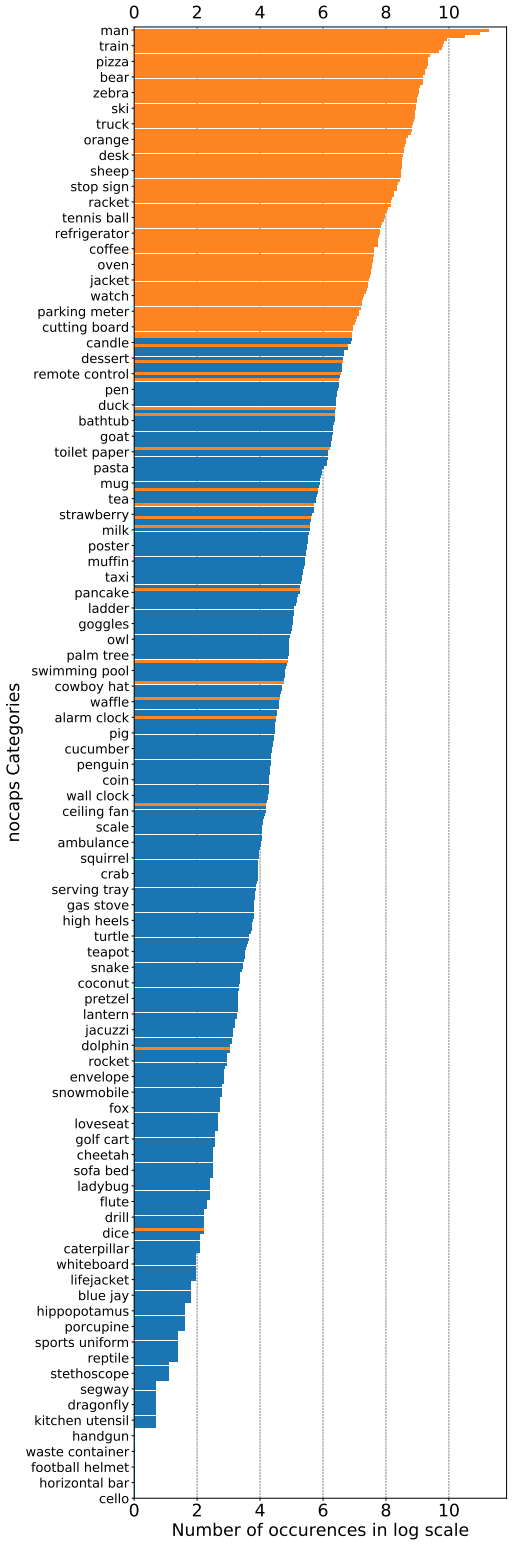


Figure 9: Histogram of mentions in the COCO Captions training set for various Open Images object classes. In **nocaps**, classes in orange are considered to be **in-domain** while classes in blue are classified as **out-of-domain**. *Zoom in for details.*

1.4. Linguistic Similarity to COCO

Overall, our collection methodology closely follows COCO. However, we do introduce keyword priming to the collection interface (refer Figure 6) which has the potential to introduce some linguistic differences between **nocaps** and COCO. To quantitatively assess linguistic differences between the two datasets, we review the performance of COCO-trained models on the **nocaps** validation set while controlling for visual similarity to COCO. As a proxy for visual similarity to COCO, we use the average cosine distance in FC7 CNN feature space between each **nocaps** image and the 10 closest COCO images.

As illustrated in Table 4, the baseline Up-Down model (trained using COCO) exceeds human performance on the decile of **nocaps** images which are most similar to COCO images (decile=1, avg. cosine distance=0.15), consistent with the trends seen in the COCO dataset. This suggests that the linguistic structure of COCO and **nocaps** captions is extremely similar. As the **nocaps** images become visually more distinct from COCO images, the performance of Up-Down drops consistently. This suggests that no linguistic variations have been introduced between COCO and **nocaps** due to priming and the degradation in the performance is due to visual differences. Similar trends are observed for our best model (Up-Down + ELMo + CBS) although the performance degradation with increasing visual dissimilarity to COCO is much less.

	nocaps test CIDEr scores										
Decile	1	2	3	4	5	6	7	8	9	10	Overall
Avg Cosine Dist from COCO	0.15	0.18	0.20	0.21	0.23	0.24	0.25	0.27	0.30	0.35	
Up-Down	**82.6**	72.6	63.9	61.1	55.9	55.0	50.7	48.5	39.2	28.7	54.5
Up-Down + ELMo + CBS	**81.8**	77.3	75.4	72.8	77.1	78.2	72.3	71.7	70.6	65.1	73.1
Human	77.8	**78.0**	**82.4**	**84.0**	**86.2**	**88.8**	**89.4**	**91.2**	**97.3**	**95.6**	**85.3**

Table 4: CIDEr scores on **nocaps** test deciles split by visual similarity to COCO (using CNN features). Our models exceed human performance on the decile of **nocaps** images that are most visually similar to COCO. This suggests that after controlling for visual variations the linguistic structure of COCO and **nocaps** captions is highly similar.

2. Additional Implementation Details for Baseline Models

2.1. Neural Baby Talk (NBT)

In this section, we describe our modifications to the original authors' implementation of Neural Baby Talk (NBT) [27] to enable the model to produce captions for images containing novel objects present in **nocaps**.

Grounding Regions for Visual Words

Given an image, NBT leverages an object detector to obtain a set of candidate image region proposals, and further produces a caption template, with slots explicitly tied to specific image regions. In order to accurately caption **nocaps** images, the object detector providing candidate region proposals must be able to detect the object classes present in **nocaps** (and broadly, Open Images). Hence, we use a Faster-RCNN [36] model pre-trained using Open Images V4 [20] (referred as **OI detector** henceforth), to obtain candidate region proposals as described in Section 4 of the main paper. This model can detect 601 object classes of Open Images, which includes the novel object classes of **nocaps**. In contrast, the authors' implementation uses a Faster-RCNN trained using COCO.

For every image in COCO train 2017 split, we extract image region proposals after the second stage of detection, with an IoU threshold of 0.5 to avoid highly overlapping region proposals, and a class detection confidence threshold of 0.5 to reduce false positive detections. This results in number of region proposals per image varies up to a maximum of 18.

Bottom-Up Visual Features

The language model in NBT (Refer Figure 4 in [27]) has two separate attention layers, and takes visual features as input in three different manners:

- The first attention layer learns an attention distribution over region features, extracted using ResNet-101 + RoI Align layer.
- The second attention layer learns an attention distribution over spatial CNN features from the last convolutional layer of ResNet-101 (7 x 7 grid, 2048 channels).
- The word embedding input is concatenated with FC7 features from ResNet-101 at every time-step.

All the three listed visual features are extracted using ResNet-101, with the first being specific to visual words, while the second and third provide the holistic context of the image. We replace the ResNet-101 feature extractor with the publicly available Faster-RCNN model pre-trained using Visual Genome (referred as **VG detector** henceforth), same as [4]. Given a set of candidate region proposals obtained from **OI detector**, we extract 2048-dimensional bottom-up features using the **VG detector** and use them as input to first attention layer (and also for input to the Pointer Network). For input to the second attention layer, we extract top-36 bottom-up features (class agnostic) using the **VG detector**. Similarly, we perform mean-pooling of these 36 features for input to the language model at every time-step.

Fine-grained Class Mapping

NBT fills the slots in each caption template using words corresponding to the object classes detected in the corresponding image regions. However, object classes are coarse labels (e.g. 'cake'), whereas captions typically refer entities in a fine-grained fashion (e.g. 'cheesecake', 'cupcake', 'coffeecake' etc.). To account for these linguistic variations, NBT predicts a fine-grained class for each object class using a separate MLP classifier. To determine the output vocabulary for this fine-grained classifier we extend the fine-grained class mapping used for COCO (Refer Table 5 in [27]), adding Open Images object classes. Several fine-grained classes in original mapping are already present in Open Images (e.g. 'man', 'woman' – fine-grained classes of 'person'), we drop them as fine-grained classes from original mapping and retain them as Open Images object classes.

Visual Word Prediction Criterion

In order to ensure correctness in visual grounding, the authors' implementation uses three criteria to decide whether a particular region proposal should be tied with a "slot" in the caption template. At any time during decoding, when the Pointer Network attends to a visual feature (instead of the visual sentinel), the corresponding region proposal is tied with the "slot" if:

- The class prediction threshold of this region proposal is higher than 0.5.
- The IoU of this region proposal with at least one of the ground truth bounding boxes is greater than 0.5.
- The predicted class is same as the object class of ground truth bounding box having highest IoU with this region proposal.

We drop the third criterion, as the **OI detector** can predict several fine-grained classes in context of COCO, such as 'man' and 'woman' (while the ground truth object class would be 'person'). Keeping the third criterion intact in **nocaps** setting would suppress such region proposals, and result in lesser visual grounding, which is not desirable for NBT. Relaxation of this criterion might introduce false positives from detection in the caption but prevents reduction in visual grounding.

We use the same optimization hyper-parameters as the authors' implementation. We encourage the reader to refer the authors' implementation for further details. We will release code for our modifications.

2.2. Constrained Beam Search (CBS)

Determining Constraints

When using constrained beam search (CBS) [2], we decoded the model in question while forcing the generated caption to include words corresponding to object classes detected in the image. For object detection, we use the same Faster-RCNN [36] model pre-trained using

Open Images V4 [20] (**`OI detector`**) that is used in conjunction with NBT. However, not all detected object classes are used as constraints. We perform constraint filtering by removing the 39 object classes listed in Table 5 from the constraint set, as these classes are either object parts, or classes that we consider to be either too rare or too broad. We also suppress highly overlapping objects as described in Section 4 of the main paper.

Parts	Too Rare or Too Broad
Human Eye	Clothing
Human Head	Footwear
Human Face	Fashion Accessory
Human Mouth	Sports Equipment
Human Ear	Hiking Equipment
Human Nose	Mammal
Human Hair	Personal Care
Human Hand	Bathroom Accessory
Human Foot	Plumbing Fixture
Human Arm	Tree
Human Leg	Building
Human Beard	Plant
Human Body	Land Vehicle
Vehicle Registration Plate	Person
Wheel	Man
Seat Belt	Woman
Tire	Boy
Bicycle Wheel	Girl
Auto Part	
Door Handle	
Skull	

Table 5: Remove Class List for object filtering

To quantify the impact of this simple constraint filtering heuristic, in Table 6 we report the results of the following ablation studies:

- Using all the object classes for constraints (w/o class),
- Using overlapping objects for constraints (w/o overlap), and
- Using no filtering heuristic at all (w/o both).

Note that in all cases we rank objects based on confident score for detected objects and pick the top-3 as the constraints. We report results for three models, the baseline model (Up-Down), the baseline model using Glove [33] and dependency-based [23] word embeddings (Up-Down + GD) and our ELMo-based model (Up-Down + ELMo +CBS). Table 6 shows that removing the above 39 classes significantly improves the performance of constrained beam search and removing overlapping objects can also slightly improve the performance. This conclusion is consistent across the three models.

Finite State Machine

Constrained Beam Search implements constraints in the decoding process using a Finite State Machine (FSM). In all experiments we use a 24 state FSM. We use 8 states for standard three single word constraints D_1, D_2 and D_3. As shown in Figure 12, the outputs of this FSM are the captions that mention at least two constraints out of three. Each D_i (i = 1,2,3) represents a set of alternative constraint words (e.g., bike, bikes). D_i can also be multi-word expressions. Our FSM dynamically support two-word or three-word phrases in D_i by extending additional one states (see Figure 10) or two states (see Figure 11) for two-word or three-word phrases respectively. Since D_1, D_2 and D_3 are all used 4 times in the base eight-state FSM, we need to allocate 4 states for a single two-word expression and 8 states for a single three-word expression.

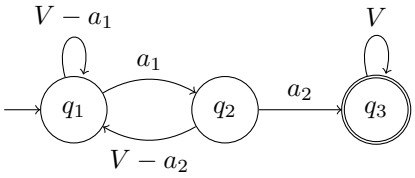


Figure 10: FSM for a two-word phrase $\{a_1, a_2\}$ constraint

	In-Domain		Near-Domain		Out-of-Domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Up-Down + CBS w/o both	73.4	11.2	68.0	10.9	65.2	9.8	68.2	10.7
Up-Down + CBS w/o class	72.8	11.2	68.6	10.9	65.5	9.7	68.6	10.8
Up-Down + CBS w/o overlap	80.6	12.0	73.5	11.3	66.4	9.8	73.1	11.1
Up-Down + CBS	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
Up-Down + GD + CBS w/o both	72.8	11.2	68.4	10.8	66.3	9.8	68.6	10.7
Up-Down + GD + CBS w/o class	72.3	11.2	68.6	10.9	66.9	9.7	68.8	10.7
Up-Down + GD + CBS w/o overlap	77.0	12.0	73.5	11.4	67.2	9.7	72.8	11.1
Up-Down + GD + CBS	77.0	12.0	73.6	11.4	69.5	9.7	73.2	11.1
Up-Down + ELMo + CBS w/o both	73.3	11.5	68.6	10.9	70.0	10.8	69.6	10.8
Up-Down + ELMo + CBS w/o class	73.5	11.5	69.2	11.0	69.9	9.9	70.0	10.9
Up-Down + ELMo + CBS w/o overlap	79.8	12.3	73.7	11.4	72.0	9.9	74.2	11.2
Up-Down + ELMo + CBS	**79.3**	**12.4**	**73.8**	**11.4**	**71.7**	**9.9**	**74.3**	**11.2**
Human	**83.3**	**13.9**	**85.5**	**14.3**	**91.4**	**13.7**	**87.1**	**14.1**

Table 6: We investigate the effect of different object filtering strategies in Constrained Beam Search and report the model performance in **nocaps** val. We find that using both strategies with the ELMo model performs best.

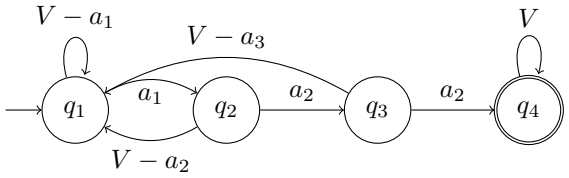


Figure 11: FSM for a three-word phrase $\{a_1, a_2, a_3\}$ constraint

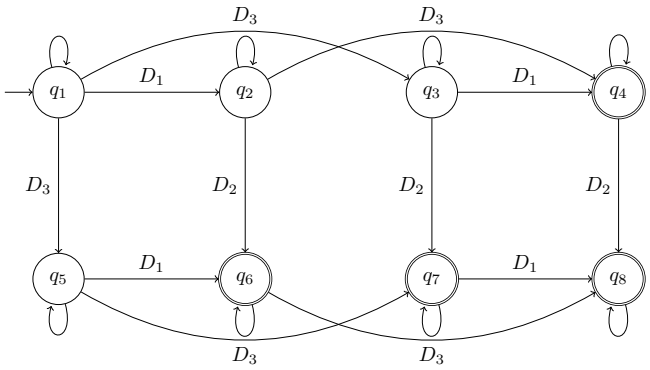


Figure 12: FSM for D_1, D_2, D_3 constraint

Integrating Up-Down Model with ELMo

When using ELMo [34], we use a dynamic representation of w_c, $\bar{h}_t^1$ and $\bar{h}_t^2$ as the input word embedding w_{ELMo}^t for our caption model. w_c is the character embedding of input words and $\bar{h}_t^i$ $(i = 1, 2)$ is the hidden output of i^{th} LSTM layer of ELMo. We combine them via:

$$w_{ELMo}^t = \gamma_0 \cdot w_c + \gamma_1 \cdot \bar{h}_t^1 + \gamma_2 \cdot \bar{h}_t^2 \tag{1}$$

where γ_i (i=0, 1, 2) are three trainable scalars. When using w_{ELMo}^t as the external word representation of other models, we fixed all the parameters of ELMo but γ_i (i=0, 1, 2).

In addition, to handle unseen objects in training data, following [2], we initialize the softmax layer matrix (W_p, b_p) using word embedding and keep this layer fixed during training. This allow our caption model to produce similar logits score for the words that share similar vectors and values in W_p and b_p. We have:

$$W_p = W_{ELMo} \tag{2}$$

$$b_p = b_{ELMo} \tag{3}$$

where W_{ELMo} and b_{ELMo} is the softmax layer in original ELMo language model. To align the different dimension in softmax layer and LSTM hidden state, we add an additional fully connected layer with a non-linearity function $tanh$. We have:

$$v_t = tanh(W_t h_t^2 + b_t) \tag{4}$$

$$P(y_t|y_{1:t-1}, I) = softmax(W_p v_t + b_p) \tag{5}$$

where $W_t \in \mathbb{R}^{H \times E}$, $b_t \in \mathbb{R}^E$, H is LSTM hidden dimension, E is the word embedding dimension, $W_p \in \mathbb{R}^{E \times D}$, $b_p \in \mathbb{R}^D$ and D is the vocabulary size.

Other details of using ELMo

In our experiment, we use the full tensorflow checkpoint trained on 1 Billion Word Language Model Benchmark[3] from official ELMo tensorflow implementation project[4].

When selecting vocabularies for our model, we first extract all words from COCO captions and open image object labels. We then extend the open image object labels to both singular and plural word forms. Finally, we remove all the words that are not in ELMo output vocabularies. This allow us to use ELMo LM prediction for each decoding step.

Our Up-Down + ELMo model is optimized by SGD [5]. We conduct hyper-parameter tuning the model and choose the model based on its performance on **nocaps** val. Table 7 shows the chosen hyper-parameters for the Up-Down Model in the paper.

Parameter	Value	Parameter	Value
Batch Size	150	Attention Size	768
LSTM Hidden Size	1200	Word Dropout	0.2
Image Feature	2048	ELMo Embedding	512
Learning Rate	0.015	Momentum	0.9
Clip Gradients	12.5	Weight Decay	0.001

Table 7: Hyper-parameters for Up-Down Model

[3] http://www.statmt.org/lm-benchmark/

[4] https://github.com/allenai/bilm-tf/

2.3. Example Model Predictions

Method	in-domain 	near-domain 	out-of-domain 
Up-Down	A man in a white shirt is playing baseball.	A couple of men standing on top of a truck.	A group of vases sitting on top of a table.
Up-Down + ELMo	A group of people standing around a blue table.	A couple of men standing next to a truck.	Two vases sitting next to each other on a table.
Up-Down + ELMo + CBS	A group of people standing near a blue table.	A couple of men standing on top of a **tank**.	A **teapot** sitting on top of a table next to a **vase**.
Up-Down + ELMo + CBS + GT	A group of people standing around a blue table.	A couple of men standing on top of a **tank**.	A couple of **kettle jugs** sitting next to each other.
NBT	A group of **men** standing in a field.	A **man** standing on the back of a **tank**.	A couple of **kettles** are sitting on a table.
NBT + CBS	A couple of **men** standing on a tennis court.	A **man** standing on top of a **tank** with a truck.	A close up of a **kettle** on a table.
NBT + CBS + GT	A group of **men** are standing in a field.	A man standing on top of a **tank plant**.	Two **kettles** and **teapot jugs** are sitting on a table.
Human	Two people in karate uniforms spar in front of a crowd.	Two men sitting on a tank parked in the bush.	Ceramic jugs are on display in a glass case.

Method	in-domain 	near-domain 	out-of-domain 
Up-Down	A woman riding a bike with a statue on her head.	A couple of chairs sitting in front of a building.	A bird sitting on the ground in the grass.
Up-Down + ELMo	There is a woman that is riding a bike.	A room that has a lot of furniture in it.	A dog laying on the ground next to a stuffed animal.
Up-Down + ELMo + CBS	There is a woman that is riding a bike.	Two **pillows** and a table in the **house**.	A dog laying on the ground next to a **tortoise**.
Up-Down + ELMo + CBS + GT	There is a woman that is riding a bike.	Two **couches** and a table in a **house**.	A dog laying on the ground next to a **tortoise**.
NBT	A man is riding a **clothing** on a bike.	A table with a **couch** and a table.	A **tortoise** is laying on top of the ground.
NBT + CBS	A woman is riding a **clothing** in the street.	A couple of **pillows** on a wooden table in a **couch**.	A **tortoise** that is sitting on the ground.
NBT + CBS + GT	A man is riding a **clothing** on a **person**.	A **house** and a **studio couch** of **couches** in a room.	A **tortoise** is laying on the ground in the grass.
Human	People are performing in an open cultural dance.	On the deck of a pool is a couch and a display of a safety ring.	Three tortoises crawl on soil and wood chips in an enclosure.

Figure 13: 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**.

Method	in-domain	near-domain	out-of-domain
			
Up-Down	A group of people are playing a game.	A large white sign on a city street.	A bear laying in the grass near a tree.
Up-Down + ELMo	A woman in a pink dress is holding a child.	A large white bus parked on the side of a road.	A bear that is laying down in the grass.
Up-Down + ELMo + CBS	A woman in a pink dress is holding a child.	A **billboard** that has a **street light** on it.	A **red panda** is walking through the grass.
Up-Down + ELMo + CBS + GT	A woman in a pink dress is holding a child.	A large white bus parked next to a **billboard**.	A **red panda** is walking through the grass.
NBT	A group of **man** are standing in a field.	A **billboard** sign on the side of a building.	A brown **red panda** is laying on the grass.
NBT + CBS	A group of **man** are playing a baseball game.	A picture of **billboard** sign on the **street light**.	A **tree** and a brown **red panda** in a field.
NBT + CBS + GT	A group of **man** are standing on a field.	A **billboard** sign on the side of a building.	A brown **red panda** lying on top of a field.
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.

Method	in-domain	near-domain	out-of-domain
			
Up-Down	A woman wearing a white shirt and a white shirt.	A person riding a yellow bike in the field.	A close up of a cat looking at a bird.
Up-Down + ELMo	A woman wearing a white shirt and white shirt.	A woman sitting on a yellow bike.	A close up of a bird with its mouth open.
Up-Down + ELMo + CBS	A woman wearing a white **suit** and a white shirt.	A person sitting on a bicycle with a **wheelchair**.	A **sea lion** standing with its mouth open.
Up-Down + ELMo + CBS + GT	A woman wearing a white **suit** and a white shirt.	A person sitting on a **bicycle** with a **wheelchair**.	A **sea lion** standing next to a **harbor seal**.
NBT	A woman wearing a white shirt is wearing a hat.	A man sitting on a **wheelchair** with a **bike**.	A close up of a **sea lion** and **harbor seal** with its head.
NBT + CBS	A **suit** of woman wearing a white suit.	A man sitting on a **wheelchair** and a **bike**.	A close up of a **harbor seal** of a **sea lion**.
NBT + CBS + GT	A **suit** of woman wearing a white shirt.	A **bicycle** sitting on a **wheelchair** with a **bike**.	A close up of a **harbor seal** of a **sea lion**.
Human	The man has a wrap on his head and a white beard.	A person sitting in a yellow chair with wheels.	A brown and gray sea lion looking at the photographer.

Figure 14: 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**.