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 [9,10,16,17,40,43], can be largely attributed to the publicly available large-scale datasets of image-caption pairs [5,14,47] as well as steady modeling improvements [4, 24, 35, 45]. 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 umbrellas, but not accordions or dolphins. This limits the usefulness of these models in real-world applications, such as providing assistance for people with impaired vision, or for improving natural language query-based image retrieval.

To generalize better ‘in the wild’, we argue that captioning models should be able to 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 [2,3,13,25,39,42,46]. Until now, novel object captioning approaches have been evaluated using a proof-of-concept dataset introduced in [12]. This dataset has restric-
In summary, we make three main contributions:

- 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 [44] 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 [39], 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 [42], 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 [46] 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 [47] 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, 38].

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, contrast 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 human performance on nocaps and use the remaining 10 cap-
for a grand total of 166,100 captions of workers participated, writing 228 captions each on average written by high-quality workers. Overall, 727 qualified workers were then discarded and replaced with captions providing low-quality captions. Captions written by these captions written by each worker and blocked workers pro-

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 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 [44] – 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 separate a leaderboard for this purpose "nocaps (XD)" 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 [38] 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 [40] 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.

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 nocaps2. 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-

2XD stands for “extra data”

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.
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 number 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 model⁴ 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 ≥ 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, 41]. 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 (~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 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.

⁴tf_faster_rcnn_inception_resnet_v2_atrous_oidv4
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Bleu-1</th>
<th>Bleu-4</th>
<th>Meteor</th>
<th>ROUGE_L</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>UpDown</td>
<td>74.3</td>
<td>11.5</td>
<td>56.9</td>
<td>10.3</td>
<td>30.1</td>
<td>8.1</td>
<td>74.0</td>
<td>19.2</td>
<td>23.0</td>
<td>50.9</td>
<td>54.3</td>
<td>10.1</td>
</tr>
<tr>
<td>UpDown + ELMo + CBS</td>
<td>76.0</td>
<td>11.8</td>
<td>74.2</td>
<td>11.5</td>
<td>66.7</td>
<td>9.7</td>
<td>76.6</td>
<td>18.4</td>
<td>24.4</td>
<td>51.8</td>
<td>73.1</td>
<td>11.2</td>
</tr>
<tr>
<td>NBT</td>
<td>60.9</td>
<td>9.8</td>
<td>53.2</td>
<td>9.3</td>
<td>48.7</td>
<td>8.2</td>
<td>72.3</td>
<td>14.7</td>
<td>21.5</td>
<td>48.9</td>
<td>53.4</td>
<td>9.2</td>
</tr>
<tr>
<td>NBT + CBS</td>
<td>63.0</td>
<td>10.1</td>
<td>62.0</td>
<td>9.8</td>
<td>58.5</td>
<td>8.8</td>
<td>73.4</td>
<td>12.9</td>
<td>22.1</td>
<td>48.7</td>
<td>61.5</td>
<td>9.7</td>
</tr>
<tr>
<td>Human</td>
<td>80.6</td>
<td>15.0</td>
<td>84.6</td>
<td>14.7</td>
<td>91.6</td>
<td>14.2</td>
<td>76.6</td>
<td>19.5</td>
<td>28.2</td>
<td>52.8</td>
<td>85.3</td>
<td>14.6</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bleu-1</th>
<th>Bleu-4</th>
<th>Overall</th>
<th>Meteor</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) UpDown</td>
<td>77.0</td>
<td>37.2</td>
<td>27.8</td>
<td>116.2</td>
<td>21.0</td>
<td></td>
</tr>
<tr>
<td>(2) UpDown + CBS</td>
<td>73.3</td>
<td>32.4</td>
<td>25.8</td>
<td>97.7</td>
<td>18.7</td>
<td></td>
</tr>
<tr>
<td>(3) UpDown + ELMo + CBS</td>
<td>72.8</td>
<td>31.5</td>
<td>25.7</td>
<td>95.4</td>
<td>18.2</td>
<td></td>
</tr>
<tr>
<td>(4) UpDown + ELMo + CBS + GT</td>
<td>72.7</td>
<td>29.4</td>
<td>23.8</td>
<td>88.3</td>
<td>16.5</td>
<td></td>
</tr>
<tr>
<td>(5) NBT</td>
<td>70.2</td>
<td>28.2</td>
<td>25.1</td>
<td>80.2</td>
<td>15.8</td>
<td></td>
</tr>
<tr>
<td>(6) NBT + CBS</td>
<td>68.9</td>
<td>10.7</td>
<td>68.6</td>
<td>10.3</td>
<td>76.9</td>
<td>9.8</td>
</tr>
<tr>
<td>(7) NBT + CBS + GT</td>
<td>66.3</td>
<td>21.7</td>
<td>25.2</td>
<td>85.4</td>
<td>19.8</td>
<td></td>
</tr>
<tr>
<td>(8) Human</td>
<td>66.3</td>
<td>21.7</td>
<td>25.2</td>
<td>85.4</td>
<td>19.8</td>
<td></td>
</tr>
</tbody>
</table>

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).

Do better object detectors help? To evaluate reliance on object detections, we supply ground truth detections
500 unique object classes and many more visual concepts. Captions describing 15,100 images containing more than scale benchmark consisting of 166,100 human-generated nocaps novel object captioning. We introduce more rigorous benchmark to assess progress on the task of

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.

Acknowledgements: We thank Jiasen Lu for helpful discussions. The Georgia Tech effort was supported in part by NSF, AFRL, DARPA, ONR YIPs, ARO PECASE. 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


[28] 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


