Captioning Images with Diverse Objects
Supplementary Material

Subhashini Venugopalan[†]
Raymond Mooney[†]
[†]UT Austin
{vsub,mooney}@cs.utexas.edu

Lisa Anne Hendricks[*]
Trevor Darrell[*]
[*]UC Berkeley
{lisa_anne, rohrbach, trevor}@eecs.berkeley.edu

Marcus Rohrbach[*]
Kate Saenko[‡]
[‡]Boston Univ.
saenko@bu.edu

Supplement

This supplement presents further qualitative results of our Novel Object Captioner (NOC) model on Imagenet images (in Sec. A), details pertaining to the quantitative results on COCO held-out objects (in Sec. B), as well as the interface used by Mechanical Turk workers comparing NOC with prior work (in Sec. C).

A. ImageNet Qualitative Examples

We present additional examples of the NOC model's descriptions on Imagenet images. We first present some examples where the model is able to generate descriptions of an object in different contexts. Then we present several examples to demonstrate the diversity of objects that NOC can describe. We then present examples where the model generates erroneous descriptions and categorize these errors.

A.1. Context

Fig. 7 shows images of eight objects, each in two different settings from ImageNet. Images show objects in different backgrounds (Snowbird on a tree branch and on a rock, Hyena on a dirt path and near a building); actions (Caribou sitting vs lying down); and being acted upon differently (Flounder resting and a person holding the fish, and Lychees in a bowl vs being held by a person). NOC is able to capture the context information correctly while describing the novel objects (eartherware, caribou, warship, snowbird, flounder, lychee, verandah, and hyena).

A.2. Object Diversity

Fig. 8 and Fig. 9 present descriptions generated by NOC on a variety of object categories such as birds, animals, vegetable/fruits, food items, household objects, kitchen utensils, items of clothing, musical instruments, indoor and outdoor scenes among others. While almost all novel words (nouns in Imagenet) correspond to objects, NOC learns to use some of them more appropriately as adjectives ('chiffon' dress in Fig. 8, 'brownstone' building and 'tweed' jacket in Fig. 9 as well as 'woollen' yarn in Fig. 4 (main paper).

Comparison with prior work. Additionally, for comparison with the DCC model from [1], Fig. 9 presents images of objects that both models can describe, and captions generated by both DCC and NOC.

A.3. Categorizing Errors

Fig. 10 presents some of the errors that our model makes when captioning Imagenet images. While NOC improves upon existing methods to describe a variety of object categories, it still makes a lot of errors. The most common error is when it simply fails to recognize the object in the image (e.g. image with 'python') or describes it with a more generic hyponym word (e.g. describing a bird species such as 'wren' or 'warbler' in Fig. 10 as just 'bird'). For objects that the model is able to recognize, the most common errors are when the model tends to repeat words or phrases (e.g. descriptions of images with 'balaclava', 'mousse' and 'cashew'), or just hallucinate other objects in the context that may not be present in the image (e.g. images with 'butte', 'caldera', 'lama', 'timber'). Sometimes, the model does get confused between images of other similar looking objects (e.g. it confuses 'levee' with 'train'). Apart from these the model does make mistakes when identifying gender of people (e.g. 'gymnast'), or just fails to create a coherent correct description even when it identifies the object and the context (e.g. images of 'sunglass' and 'cougar').

Relevant but Minor Errors. Fig. 11 presents more examples where NOC generates very relevant descriptions but makes some minor errors with respect to counting (e.g. images of 'vulture' and 'aardvark'), age (e.g. refers to boy wearing 'snorkel' as 'man'), confusing the main object cate-

Metric	Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg.
F1	DCC	4.63	29.79	**45.87**	**28.09**	64.59	52.24	13.16	79.88	39.78
	NOC (ours)	**17.78**	**68.79**	25.55	24.72	**69.33**	**55.31**	**39.86**	**89.02**	**48.79**
METEOR	DCC	18.1	**21.6**	**23.1**	22.1	22.2	20.3	18.3	22.3	21.00
	NOC (ours)	**21.2**	20.4	21.4	21.5	21.8	**24.6**	18.0	21.8	**21.32**

Table 7. MSCOCO Captioning: F1 and METEOR scores (in %) of NOC (our model) and DCC [1] on the held-out objects not seen jointly during image-caption training, along with the average scores of the generated captions across images containing these objects.

Model	F1 (%)	METEOR (%)
DCC with word2vec	39.78	21.00
DCC with GloVe	38.04	20.26
NOC (ours, uses GloVe)	**48.79**	**21.32**

Table 8. DCC and NOC both using GloVe on MSCOCO dataset.

gory (e.g. 'macaque' with 'bear' and person as 'teddy bear') or makes minor word repetitions, and grammatical errors.

B. MSCOCO Quantitative Results

We present detailed quantitative results comparing DCC and NOC on the 8 held-out objects.

B.1. F1 and METEOR

While Table. 1 (main paper) presents the F1 scores comparing the DCC model [1] and our NOC model for each of the eight held-out objects in the test split, Table. 7 supplements this by also providing the individual meteor scores for the sentences generated by the two models on these eight objects. In case of NOC, we sampled sentences (25) and picked one with lowest log probability. Using, beam search with a beam width of 1 produces sentences with METEOR score 20.69 and F1 of 50.51. In Tables 2 and 3 (main paper), all lines except the last line corresponding to NOC use beam search with a beam-width of 1.

B.2. Word-embedding for DCC and NOC

One aspect of difference between NOC and DCC is that NOC uses GloVe embeddings in it's language model whereas DCC uses word2vec embeddings to select similar objects for transfer. In order to make a fair comparison of DCC with NOC, it is also important to consider the setting where both models use the same word-embedding. We modify the transfer approach in DCC and replace word2vec with GloVe embeddings. From Table. 8 we note that the difference in DCC is not significant. Thus, the embeddings themselves do not play as significant a role as the joint training approach.

B.3. Joint Training with Auxiliary Objectives

When performing joint training and considering the overall optimization objective as the sum of the image-specific loss, the text-specific loss and image-caption loss, we can define the objective more generally as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{CM}} + \alpha \mathcal{L}_{\mathcal{IM}} + \beta \mathcal{L}_{\mathcal{LM}} \tag{1}$$

where α and β are hyper-parameters which determine the weighting between different losses. In our experiments setting $\alpha = 1$ and $\beta = 1$ provided the best performance on the validation set. Other values of $(\alpha, \beta) \in \{(1, 2), (2, 1)\}$ resulted in lower F1 and METEOR scores.

C. Mechanical Turk Interface

Fig. 12 presents the interface used by mechanical turk workers when comparing sentences generated by our model and previous work. The workers are provided with the image, the novel object category (word as well as meaning) that is present in the image, and two sentences (one each from our model and previous work). The sentence generated by the NOC model is randomly chosen to be either Sentence 1 or Sentence 2 for each image (with the other sentence corresponding to the caption generated by previous work [1]). Three workers look at each image and the corresponding descriptions. The workers are asked to judge the captions based on how well it incorporates the novel object category in the description, and which sentence describes the image better.

D. Future directions

One interesting future direction would be to create a model that can learn on new image-caption data after it has already been trained. This would be akin to [2], where after an initial NOC model has already been trained we might want to add more objects to the vocabulary, and train it on few image-caption pairs. The key novelty would be to improve the captioning model by re-training only on the new data instead of training on all the data from scratch.

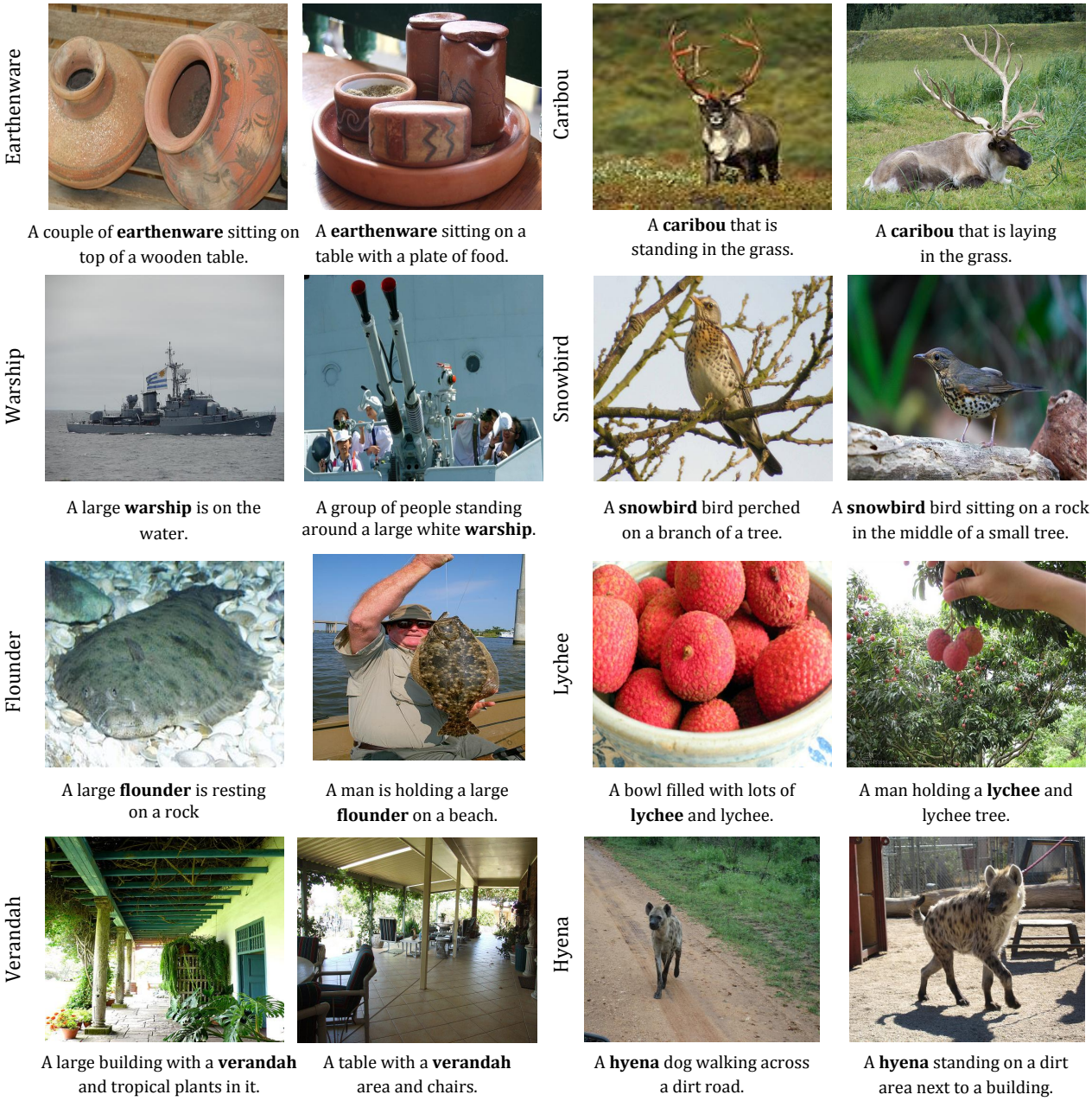


Figure 7. Examples showing descriptions generated by NOC for ImageNet images of eight objects, each in two different contexts. NOC is often able to generate descriptions incorporating both the novel object name as well as the background context correctly.

Figure 8. Examples of sentences generated by our NOC model on ImageNet images of objects belonging to a diverse variety of categories including food, instruments, outdoor scenes, household equipment, and vehicles. The novel objects are in **bold**. The last row highlights common errors where the model tends to repeat itself or hallucinate objects not present in the image.

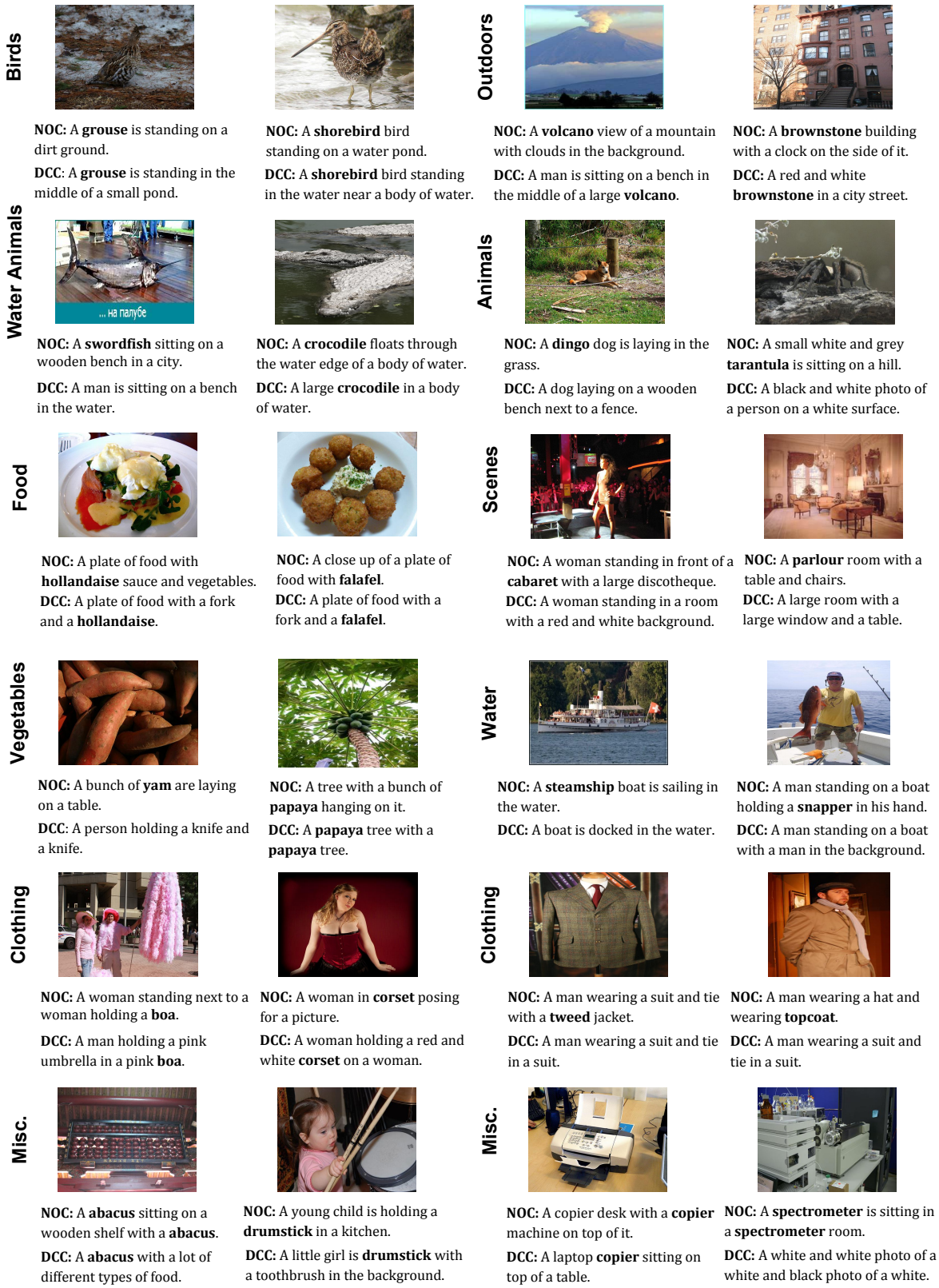
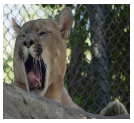


Figure 9. Examples comparing sentences generated by DCC [1] and our NOC model on ImageNet images of object categories that *both* models can describe including food, animals, vegetables/fruits, indoor and outdoor scenes, and clothing. The novel objects are in **bold**.

Superhighway (n04358491) Error: Synonym
NOC: A road with a traffic light and a red **superhighway**.

Sunglass (n04355933) Error: Grammar
NOC: A **sunglass** mirror reflection of a mirror in a mirror.

Caddie (n09886403) Error: Semantics
NOC: A man holding a **caddie** in his hand.



Cougar (n02125311) Error: Description
NOC: A **cougar** with a **cougar** in its mouth.

Warbler (n01563128) Error: Synonym
NOC: A bird sitting on a tree with a **warbler** on it

Cashew (n12759273) Error: Repetition
NOC: A tree branch with **cashew** tree branch.

Levee (n03658635) Error: Hallucination
NOC: A train going down the tracks near a **levee**.

Javelin (n03594148) Error: Description
NOC:A **javelin** player is getting ready to hit a ball.

Caldera (n09231117) Error: Hallucination
NOC: A man is surfing on a **caldera** in the mountains.

Python (n01743605) Error: Recognition
NOC: A tree branch with a tree in the background.

Butte (n09230202) Error: Hallucination
NOC: A **butte** is sitting on a rock near a body of water.

Gymnast (n10153594) Error: Gender, Hallucination
NOC: A man **gymnast** in a blue shirt doing a trick on a skateboard.

Balaclava (n02776825) Error: Repetition
NOC: A **balaclava** black and white photo of a man in a **balaclava.**

Lama (n10243664) Error: Hallucination
NOC: A man **lama** holding a cell phone while standing in the background.

Timber (n04436329) Error: Hallucination
NOC: A man in a **timber** factory with a dog on his back.

Boatman (n09861946) Error: Incomplete
NOC: A **boatman** paddling on a lake with a rowing.

Chemist (n10421470) Error: Semantics
NOC: A man in a **chemist** kitchen preparing food.

Spectacles (n04272054) Error: Hallucination
NOC: A **spectacles** glasses is on a white surface.

Wren (n01584225) Error: Recognition
NOC:A bird sitting on a tree branch with leaves in the background.

Mousse (n07611991) Error: Repetition
NOC: A **mousse** with a red strawberry mousse sits on a table.

Figure 10. Examples of images where the model makes errors when generating descriptions. The novel object is in **bold** and the errors are underlined. NOC often tends to repeat words in its description, or hallucinate objects not present in the image. The model sometime misidentifies gender, misrepresents the semantics of the novel object, or just makes grammatical errors when composing the sentence.

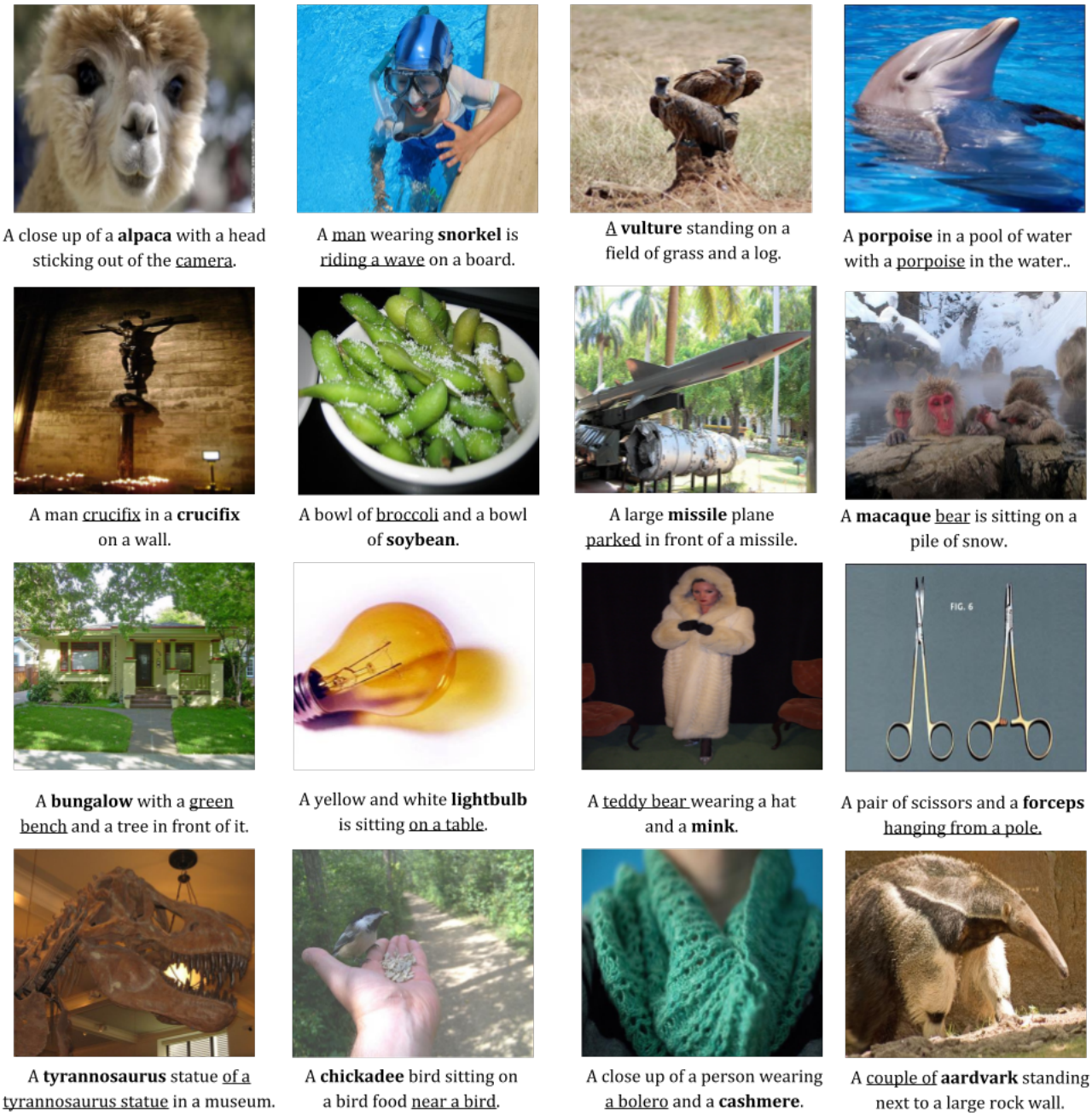


Figure 11. Some examples where NOC makes minor errors when describing the image. The novel object is in **bold** and the word or segment corresponding to the error is underlined. Counting, repetitions, confusing object categories (e.g. 'macaque', 'bear'), grammatical errors, and hallucinating objects that are absent are some common errors that the model makes. However, the generated description is still meaningful and relevant.

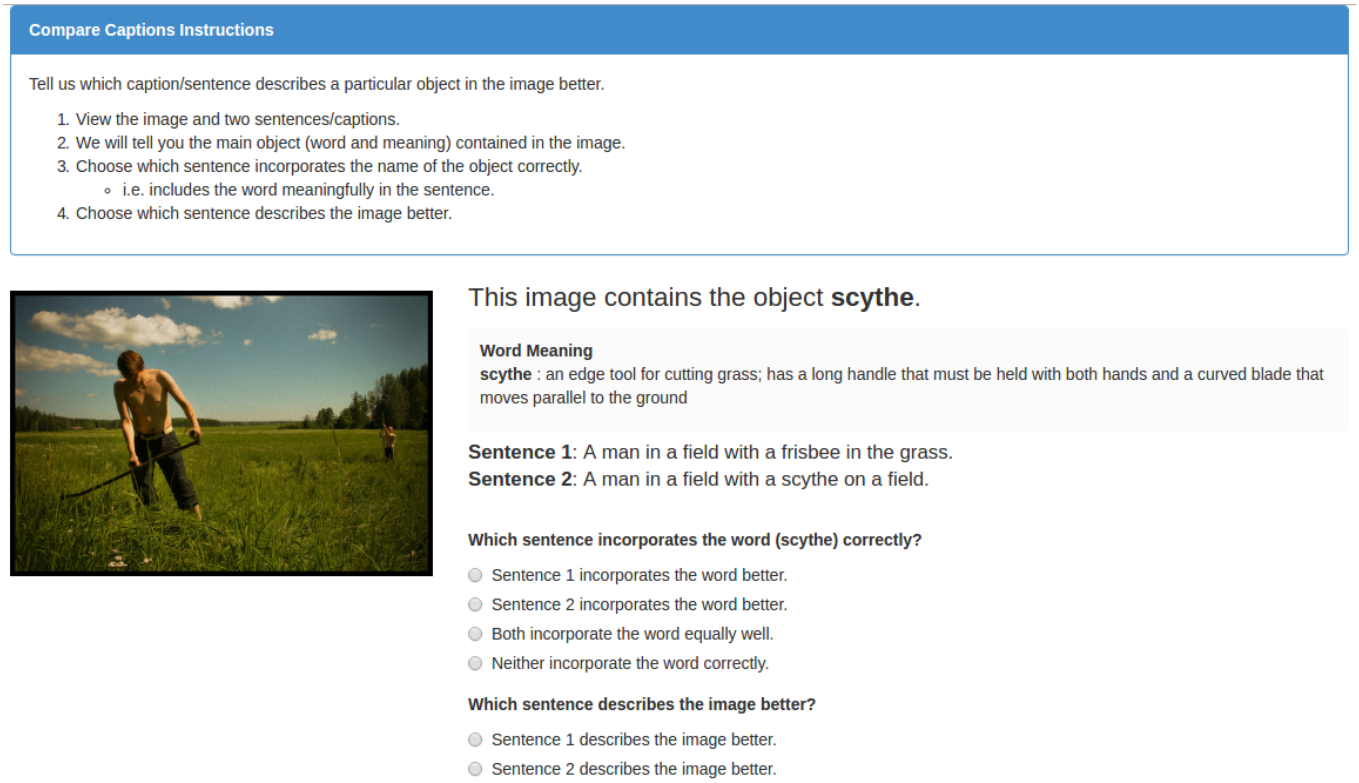


Figure 12. Interface used by mechanical turk workers when comparing captions/sentences generated by our NOC model with previous work (DCC [1]). The workers are asked to compare on both Word Incorporation i.e. how well each model incorporates the novel object in the sentence, as well as Image Description i.e. which caption describes the image better.

References

[1] 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. 1, 2, 5, 8

[2] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, 2015. 2