

Captioning Images with Diverse Objects

Subhashini Venugopalan*

UT Austin
vsub@cs.utexas.edu

Lisa Anne Hendricks*

UC Berkeley
lisa_anne@berkeley.edu

Marcus Rohrbach

UC Berkeley, ICSI
rohrbach@berkeley.edu

Raymond Mooney

UT Austin
mooney@cs.utexas.edu

Trevor Darrell

UC Berkeley, ICSI
trevor@eecs.berkeley.edu

Kate Saenko

UMass Lowell
saenko@cs.uml.edu

Abstract

We propose Novel Object Captioner (NOC), a deep visual semantic captioning model that can describe a large number of object categories not present in existing image-caption datasets. Recent captioning models are limited in their ability to scale and describe concepts outside of paired image-text corpora. Our model takes advantage of external sources - labeled images from object recognition datasets, and semantic knowledge extracted from unannotated text - and combines them to generate descriptions about novel objects. We propose minimizing a joint objective which can learn from diverse data sources and leverage distributional semantic embeddings, enabling the model to generalize and describe novel objects outside of image-caption datasets. We demonstrate that our model exploits semantic information to generate captions for hundreds of object categories in the ImageNet object recognition dataset that are not observed in image-caption training data, as well as many categories that are observed very rarely.

1 Introduction

Modern visual classifiers (He et al., 2016; Simonyan and Zisserman, 2014) can recognize thousands of object categories, some of which are basic or entry-level (e.g. television), and others that are fine-grained and task specific (e.g. dial-phone, cell-phone). However, recent state-of-the-art visual captioning systems (Vinyals et

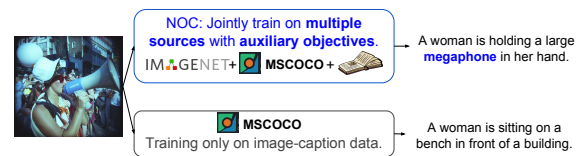


Figure 1: We propose a model that learns simultaneously from multiple data sources with auxiliary objectives to describe a variety of objects.

al., 2015; Fang et al., 2015; Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Kiros et al., 2015; Mao et al., 2015a) that learn directly from image and description pairs fail in their ability to generalize and describe this vast set of recognizable objects in context. While such systems could be scaled by building larger image/video description datasets, obtaining such captioned data would be expensive and laborious. Furthermore, visual description is challenging because models must not only correctly identify visual concepts contained in an image, but must also compose these concepts into a coherent sentence.

Recent deep captioning models rely entirely on paired images and captions (to model language and vision) and are hence unable to describe objects which do not explicitly appear in image-caption datasets. Recent work from Hendricks et al. (2016) shows that, to incorporate the vast knowledge of current visual recognition models without explicit paired captions, models must learn from external sources and learn to compose sentences about visual concepts which are infrequent or non-existent in image-description corpora as in Fig. 1. In this spirit, we propose NOC, an architecture that integrates knowledge

*Equal contribution.

from external visual recognition datasets as well as semantic information from independent unannotated text corpora to generate captions for a variety of objects.

Specifically, we introduce auxiliary objectives which allows our network to learn a description model simultaneously with a deep language model and visual recognition system. Unlike previous work, our auxiliary objectives allow our model to learn relevant information from multiple data sources in an end-to-end fashion. Another notable aspect of our model is that it implicitly leverages distributional embeddings trained on external text corpora enabling it to describe unseen object categories based on the descriptions of the training images. Additionally, unlike previous captioning approaches, our model combines textual and visual confidences at the semantic (word) level, fully exploiting the visual classifier’s ability to recognize a wide range of objects for the novel object captioning task. We demonstrate that our model’s auxiliary training objectives along with its ability to leverage visual and semantic information, enable it to generate descriptions for hundreds of objects not present in description datasets. Compared to previous work (Hendricks et al., 2016) for the task, our model is end-to-end trainable and provides improved performance. Additionally, our model directly extends to generate descriptions for objects which appear infrequently/rarely in image description datasets.

2 Related Work

Visual Description has seen many different approaches over the years (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2014), and more recently deep models have gained popularity because of their high performance and their potential for end-to-end training. Deep visual description frameworks first encode an image into a fixed length feature vector and then generate a description by either conditioning text generation on image features (Vinyals et al., 2015; Donahue et al., 2015; Karpathy and Fei-Fei, 2015) or embedding image features and previously generated words into a multimodal space (Mao et al.,

2015a; Kiros et al., 2014; Kiros et al., 2015) before predicting the next word. Though most models represent images with an intermediate representation from a convolutional neural network (CNN) (such as fc7 activations from an object recognition CNN), other models represent images as a vector of confidences over a fixed number of visual concepts (Hendricks et al., 2016; Fang et al., 2015). Also, in almost all cases, the parameters of the visual pipeline are initialized with weights trained on the ImageNet classification task. On the caption generation side, recurrent networks (RNNs) are a popular choice to model language, but log bilinear models (Kiros et al., 2014) and maximum entropy language models (Fang et al., 2015) have also been explored. Our model is similar to the CNN-RNN frameworks in Mao et al. (2015a) and Hendricks et al. (2016).

Novel object captioning Recently Mao et al. (2015b) proposed an approach, that extends a model’s capability to describe a small set of novel concepts (e.g. *quidditch*, *samisen*) from paired training data while retaining its ability to describe older concepts learned previously. On the other hand, Hendricks et al. (2016) introduce a model that can describe many objects already existing in English corpora and object recognition datasets (ImageNet) but not in the caption corpora (e.g. *pheasant*, *otter*). Our focus is on the latter case. Hendricks et al. (2016) integrate information from external text and visual sources, and explicitly transfer parameters from objects seen in image-caption data to unseen ImageNet objects to generate descriptions for the new objects. While this works well for many ImageNet classes, their parameter transfer is explicit and cannot be trained end-to-end. Furthermore, their model cannot caption objects for which few paired training examples already exist. Our proposed framework integrates information from external sources implicitly, making it end-to-end trainable, as well as extends directly to caption ImageNet objects with few or no descriptions.

Multi-modal and Zero-Shot Learning Another closely related line of research takes advantage of distributional semantics to learn a joint embedding space using visual and textual infor-

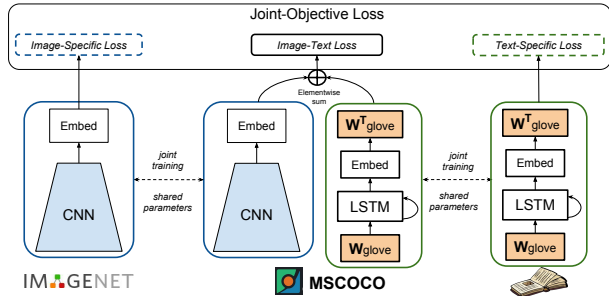


Figure 2: Our model consists of a visual recognition network (left), an LSTM-based language model (right), and an image caption network (center). The networks share parameters, and are trained simultaneously on different sources with different objectives enabling the model to describe a variety of objects.

mation for zero-shot labeling of novel object categories (Frome et al., 2013; Norouzi et al., 2013), as well as retrieval of images with text (Socher et al., 2014; Lazaridou et al., 2014). Visual description itself can be cast as a multimodal learning problem in which caption words w_0, \dots, w_{n-1} and an image are projected into a joint embedding space before the next word in a caption, w_n , is generated (Mao et al., 2015a; Kiros et al., 2015). Although our approach uses distributional word embeddings, our model differs in the sense that it can be trained with unpaired text and visual data but still combine the semantic information at a later stage during caption generation.

3 Novel Object Captioner (NOC)

Our NOC model is illustrated in Fig. 2. It not only learns from paired image-caption data, but also leverages semantic information from unannotated text and integrates it with a visual recognition model. By introducing auxiliary loss functions (objectives) and jointly training different components on multiple data sources, we obtain a visual description model which simultaneously learns an independent object recognition model, as well as a language model.

We start by first training a simple LSTM-based language model (LM) (Sundermeyer et al., 2010) for sentence generation. Our LM incorporates dense representations for words from distributional embeddings (GloVe, Pennington et al. (2014)) pre-trained on external text corpora.

Simultaneously, we also train a state-of-the-art visual recognition network to provide confidences over words in the vocabulary given an image. To generate descriptions conditioned on image content, we combine the predictions of our language and visual recognition networks by summing textual and visual confidences at the word level.

In this way, we decompose our model into discrete textual and visual pipelines which can be trained exclusively using unpaired text and unpaired image data (networks on left and right of Fig. 2). When training our description model, we introduce auxiliary image-specific (\mathcal{L}_{IM}), and text-specific (\mathcal{L}_{LM}) objectives along with the paired image-caption (\mathcal{L}_{CM}) loss function. These loss functions when trained jointly influence our model to not only produce reasonable image descriptions, but also predict visual concepts as well as generate cohesive text (language modeling). We first discuss the auxiliary objectives and the joint training, and then discuss how we leverage embeddings trained with external text to compose descriptions about novel objects.

3.1 Auxiliary Training Objectives

Our motivation for introducing auxiliary objectives is to learn how to describe images without losing the ability to recognize more objects. Typically, image-captioning models incorporate a visual classifier pre-trained on a source domain (e.g. ImageNet dataset) and then tune it to the target domain (the image-caption dataset). However, important information from the source dataset can be suppressed if similar information is not present when fine-tuning, leading the network to forget (over-write weights) for objects not present in the target domain. This is problematic in our scenario in which the model relies on the source datasets to learn a large variety of visual concepts not present in the target dataset. However, with pre-training as well as the complementary auxiliary objectives the model maintains its ability to recognize a wider variety of objects and is encouraged to describe objects which are not present in the target dataset at test time. For the ease of exposition, we abstract away the details of the language and the visual models and first describe the joint training objectives of the

complete model, i.e. the text-specific loss, the image-specific loss, and the image-caption loss. We will then describe the language and the visual models.

Image-specific Loss Our visual recognition model (Fig. 2a, left) is a neural network parametrized by θ_I and is trained on object recognition datasets. Unlike typical visual recognition models that are trained with a single label on a classification task, for the task of image captioning an image model that has high confidence over multiple visual concepts occurring in an image simultaneously would be preferable. Hence, we choose to train our model using multiple labels (more in Sec. 4.2) with a multi-label loss. If l denotes a label and z_l denotes the binary groundtruth value for the label, then the optimization objective for the visual model is given by the cross-entropy loss (\mathcal{L}_{IM}):

$$\mathcal{L}_{IM}(I; \theta_I) = - \sum_l z_l \log(S_l(f_{IM}(I; \theta_I))) + (1 - z_l)(1 - \log(S_l(f_{IM}(I; \theta_I)))) \quad (1)$$

where $S_i(x)$ is the output of a softmax function over index i and input x , and f_{IM} , is the activation of the final layer of the visual recognition network.

Text-specific Loss Our language model (Fig. 2, right) is based on LSTM Recurrent Neural Networks. We denote the parameters of this network by θ_L , and the activation of the final layer of this network by f_{LM} . The language model is trained to predict the next word w_t in a given sequence of words w_0, \dots, w_{t-1} . This is optimized using the softmax loss \mathcal{L}_{LM} which is equivalent to the maximum-likelihood:

$$\mathcal{L}_{LM}(w_0, \dots, w_{t-1}; \theta_L) = - \sum_t \log(S_{w_t}(f_{LM}(w_0, \dots, w_{t-1}; \theta_L))) \quad (2)$$

Image-caption Loss The goal of the image captioning model (Fig. 2, center) is to generate a sentence conditioned on an image (I). NOC predicts the next word in a sequence, w_t , conditioned on previously generated words (w_0, \dots, w_{t-1}) and

an image (I), by summing activations from the deep language model, which operates over previous words, and the deep image model, which operates over an image. We denote these final (summed) activations by f_{CM} . Then, the probability of predicting the next word is given by,

$$P(w_t | w_0, \dots, w_{t-1}, I) = S_{w_t}(f_{CM}(w_0, \dots, w_{t-1}, I; \theta)) = S_{w_t}(f_{LM}(w_0, \dots, w_{t-1}; \theta_L) + f_{IM}(I; \theta_I))$$

Given pairs of images and descriptions, the caption model optimizes the parameters of the underlying language model (θ_L) and image model (θ_I) by minimizing the caption model loss \mathcal{L}_{CM} :

$$\mathcal{L}_{CM}(w_0, \dots, w_{t-1}, I; \theta_L, \theta_I) = - \sum_t \log(S_{w_t}(f_{CM}(w_0, \dots, w_{t-1}, I; \theta_L, \theta_I))) \quad (3)$$

Joint Training with Auxiliary Losses

While many previous approaches have been successful on image captioning by pre-training the image and language models and tuning the caption model alone (Eqn. 3), this is insufficient to generate descriptions for objects outside of the image-caption dataset since the model tends to “forget” (over-write weights) for objects only seen in external data sources. To remedy this, we propose to train the image model, language model, and caption model simultaneously on different data sources. The NOC model’s final objective simultaneously minimizes the three individual complementary objectives:

$$\mathcal{L} = \mathcal{L}_{CM} + \alpha \mathcal{L}_{IM} + \beta \mathcal{L}_{LM} \quad (4)$$

where α and β are hyper-parameters which determine the weighting between different losses (we use $\alpha = 1$ and $\beta = 1$). By sharing the weights of the caption model’s network with the image network and the language network (as depicted in Fig. 2 (a)), the model can be trained simultaneously on independent image-only data, unannotated text data, as well as paired image-caption data. Consequently, co-optimizing different objectives aids the model in recognizing categories outside of the paired image-sentence data.

3.2 Language Model with Semantic Embeddings

Our language model consists of the following components: a continuous lower dimensional embedding space for words (W_{glove}), a single recurrent (LSTM) hidden layer, and two linear transformation layers where the second layer (W_{glove}^T) maps the vectors to the size of the vocabulary. Finally a softmax activation function is used on the output layer to produce a normalized probability distribution. The cross-entropy loss which is equivalent to the maximum-likelihood is used as the training objective.

In addition to our joint objective (Eqn.4), we also employ semantic embeddings in our language model to help generate sentences when describing novel objects. Specifically, the initial input embedding space (W_{glove}) is used to represent the input (one-hot) words into semantically meaningful dense fixed-length vectors. While the final transformation layer (W_{glove}^T) reverses the mapping of a dense vector back to the full vocabulary with the help of a softmax activation function. These distributional embeddings (Mikolov et al., 2013; Pennington et al., 2014) share the property that words that are semantically similar have similar vector representations. The intuitive reason for using these embeddings in the input and output transformation layers is to help the language model treat words unseen in the image-text corpus to (semantically) similar words that have previously been seen so as to encourage compositional sentence generation i.e. encourage it to use new/rare word in a sentence description based on the visual confidence.

3.3 Visual Classifier

The other main component of our model is the visual classifier. We employ a recent convolutional neural network (VGG-16 (Simonyan and Zisserman, 2014)) as the visual recognition network for the image classifier. We modify the final layers of the network to incorporate the multi-label loss (Eqn. 1) to predict visual confidence over multiple labels in the vocabulary. The rest of the classification network remains unchanged.

4 Datasets

In this section we describe the image description dataset as well as the external text and image datasets used in our experiments.

4.1 External Text Corpus (WebCorpus)

We extract sentences from Gigaword, the British National Corpus (BNC), UkWaC, and Wikipedia. Stanford CoreNLP 3.4.2 (Manning et al., 2014) was used to extract tokenizations. This dataset was used to train the LSTM language model. For the dense word representation in the network, we use GloVe (Pennington et al., 2014) pre-trained on 6B tokens of external corpora including Gigaword and Wikipedia. To create our LM vocabulary we identified the 80,000 most frequent tokens from the combined external corpora. We refine this vocabulary further to a set of 72,700 words that also had GloVe embeddings.

4.2 Image Caption data

To empirically evaluate the ability of NOC to describe new objects we use the training and test set from Hendricks et al. (2016). The dataset consists of a subset of MSCOCO (Lin et al., 2014) which excludes all sentences that describe one of the following eight objects: bottle, bus, couch, microwave, pizza, racket, suitcase, or zebra. We then evaluate how well NOC can generate descriptions about these eight held-out words.

4.3 Image data

We also qualitatively evaluate sentences generated by NOC on approximately 700 different ImageNet (Russakovsky et al., 2014) objects which are not present in the MSCOCO dataset. We choose this set by identifying objects that are present in both ImageNet and our language corpus (vocabulary), but not present in MSCOCO. Chosen words span a variety of categories including fine-grained categories (e.g., “bloodhound” and “chrysanthemum”), adjectives (e.g., “chiffon”), and entry level words (e.g., “toad”).

Rare objects In order to understand if including additional data enables NOC to better generate sentences about rare words, we select 52 objects which are in ImageNet as well as the

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg.
DCC	4.63	29.79	45.87	28.09	64.59	52.24	13.16	79.88	39.78
NOC (ours)	19.02	69.34	33.25	26.46	69.16	62.45	34.65	89.75	50.51

Table 1: COCO Compositional Captioning: F1 scores of NOC (our model) and DCC (Hendricks et al., 2016) on held-out objects not seen jointly during image-caption training. Our approach uses auxiliary objectives and distributional vector representation and, unlike previous methods, is end-to-end trainable.

MSCOCO vocabulary. Selected words occur with varying frequency in the MSCOCO training set, with about 52 mentions on average (median 27) across all training sentences. For example, words such as bonsai only appear 5 times, while some others such as pumpkin appears 58 times, and scarf appears 144 times.

5 Experiments

5.1 Empirical Evaluation on MSCOCO

We empirically evaluate the ability of our proposed model to describe novel objects by following the experimental setup of (Hendricks et al., 2016). We optimize each loss in our model with the following datasets: the caption model, which jointly learns the parameters θ_L and θ_I , is trained only on the subset of MSCOCO without the 8 objects (see section 4.2), the image model, which updates parameters θ_I , is optimized using labeled images, and the language model which updates parameters θ_L , is trained using the corresponding descriptions. We first present evaluations for the in-domain setting in which the image classifier is trained with all COCO training images and the language model is trained with all sentences in the training set before comparing training from different data sources. We use the F1 metric, which determines if the model mentions the object name in the description of the images containing the object, and the METEOR metric (Denkowski and Lavie, 2014) to evaluate description quality.

COCO heldout objects. Results of the models on the 8 held-out COCO objects are presented in Table 1. This compares the F1 score achieved by NOC to the previously existing method, DCC (Hendricks et al., 2016), and demonstrates that NOC is able to integrate new vocabulary words into description better for every

object in the held-out dataset except “couch” and “microwave”, outperforming DCC considerably.

Ablations. Table 2 compares how different aspects of training impact the overall performance. *Vision Contribution (tuned)*: The model that does not incorporate Glove or LM pre-training has poor performance (METEOR 15.78, F1 14.41); this ablation shows the contribution of the vision model alone in recognizing and describing the held out objects. *Glove contribution*: The model trained without the auxiliary objective, performs better with F1 of 25.38 and METEOR of 19.80; this improvement comes largely from the GloVe embeddings which help in captioning novel object classes. *Vision Contribution (No-Tuning)*: It’s interesting to note that when we fix the pre-trained classifier’s weights (trained on all objects), before training on the image-caption COCO subset, the F1 increases substantially to 39.70 suggesting that the visual model recognizes many objects but can “forget” objects learned by the classifier when fine-tuned on the image-caption data (without the 8 objects). *Auxiliary Objective*: Finally, incorporating the auxiliary objectives F1 improves remarkably to 47.02 and incorporating all aspects does best (F1 50.51, METEOR 20.69) significantly outperforming DCC.

Training data source. To study the effect of different data sources, we also evaluate our model in an out-of-domain setting where classifiers for held out objects are trained with images from ImageNet and the language model is trained on text mined from external corpora. Table 3 reports average scores across the eight held-out objects. We compare our NOC model to results from Hendricks et al. (2016) (DCC), as well as a competitive image captioning model - LRCN (Donahue et al., 2015) trained only on the held-out MSCOCO dataset. In the out-of-domain setting (line 2), for

Contributing factor	Glove	LM pretrain	Tuned Visual Classifier	Auxiliary Objective	METEOR	F1
Tuned Vision	-	-	✓	✓	15.78	14.41
LM & Embedding	✓	✓	✓	-	19.80	25.38
LM & Pre-trained Vision	✓	✓	Fixed	-	18.91	39.70
Auxiliary Objective	✓	-	✓	✓	19.69	47.02
All	✓	✓	✓	✓	20.69	50.51

Table 2: Ablations comparing the contributions of the Glove embedding, LM pre-training, and auxiliary objectives, of the NOC model. Our auxiliary objective along with Glove have the largest impact in captioning novel objects.

	Image	Text	Model	METEOR	F1
1		Baseline (no transfer)	LRCN	19.33	0
			DCC	19.90	0
2	Image Net	Web Corpus	DCC	20.66	34.94
			NOC	17.55	36.50
3	COCO	Web Corpus	NOC	19.15	41.74
4	COCO	COCO	DCC	21.00	39.78
			NOC	20.69	50.51

Table 3: Comparison with different training data sources on 8 held-out COCO objects. Having in-domain data helps both the DCC (Hendricks et al., 2016) and our NOC model caption novel objects.

the chosen set of 8 objects, NOC performs slightly better on F1 and a bit lower on METEOR compared to DCC. However, we note that in DCC one needs to explicitly identify a set of “seen” object classes from which weights are transferred to the new unseen objects, whereas our model is directly able to caption novel objects without any post-processing. This transfer mechanism not only makes DCC unwieldy to train (it’s not end-to-end) but also leads to peculiar descriptions. E.g., to describe “racket”, DCC transfers parameters from the word “tennis”. Consequently, given a picture of a man holding a tennis racket, DCC produces the caption “A man is playing racket on the court.” whereas our NOC model produces the caption “A man is swinging a racket at a ball” (Fig. 3).

With COCO image training (line 3), F1 scores of our NOC model improves considerably even with the Web Corpus LM training. Finally in the in-domain setting (line 4) NOC outperforms

DCC on F1 by over 10 points while maintaining a similar METEOR score. This suggests that NOC is able to associate the objects with captions better with in-domain training, and the auxiliary objectives and embedding help the model to generalize and describe novel objects.

5.2 Scaling to ImageNet

Describing Novel words To demonstrate the scalability of our proposed architecture, we describe objects in ImageNet for which no paired image-sentence data exists. Table 4 compares models on 638 novel object categories using the following metrics: (i) Describing novel objects (%) refers to the percentage of the selected ImageNet objects mentioned in descriptions, i.e. for each novel word (e.g., “otter”) the model should incorporate the word (“otter”) into at least one description about an ImageNet image of the object (otter). While DCC is able to recognize and describe 56.85% (363) of the selected ImageNet objects in descriptions, NOC recognizes several more objects and is capable of describing 91.27% (582 of 638) ImageNet objects. (ii) Accuracy refers to the percentage of images from each category where the model is able to correctly identify and describe the category. We report the average accuracy across all categories. DCC incorporates a new word correctly 11.08% of the time, in comparison, NOC improves this appreciably to 24.74%. (iii) F1 score is computed based on precision and recall of mentioning the object in the description. Again, NOC outperforms with average F1 33.76% to DCC’s 14.47%. In short, NOC is both able to describe more categories, and correctly integrates new vocabulary word into descriptions more frequently.



Figure 3: Descriptions produced by NOC on a variety of objects, including “caddie”, “saucepan”, and “flounder”. (Right) NOC makes errors and (top right) fails to describe the new object (“sitar”). More in supplement.

Model	Desc. Novel (%)	Acc (%)	F1 (%)
DCC	56.85	11.08	14.47
NOC	91.27	24.74	33.76

Table 4: ImageNet: Comparing our model against DCC (Hendricks et al., 2016) on % of novel classes described, average accuracy of mentioning the class in the description, and mean F1 scores for object mentions.

Fig. 3 (column 3) shows examples where NOC describes a large variety of objects. Fig. 3 (right) outlines some errors. Failing to describe a new object is one common error for NOC. E.g. Fig. 3 (top right), NOC incorrectly describes a man holding a “sitar” as a man holding a “baseball bat”. Other common errors include generating non-grammatical or nonsensical phrases (example with “aardvark”) and repeating a specific object (“A barracuda ... with a barracuda”).

Describing Rare Words In addition to describing novel words, NOC can also help describe words which are mentioned rarely in paired corpora. We demonstrate this by choosing 52 words which are rarely described in paired text corpora including “teapot” (30 annotations), “swan” (60 annotations), and “whisk” (11 an-

notations). When tested on ImageNet images containing these concepts, a model trained only with MSCOCO paired data incorporates rare words into sentences 2.93% of the time with an average F1 score of 4.58%. In contrast, when integrating outside data using our NOC model can incorporate rare words into descriptions 35.15% of the time with an average F1 score of 47.58%.

6 Conclusion

We present an end-to-end trainable architecture that incorporates auxiliary training objectives and distributional semantics to generate descriptions for object classes not seen in paired image-caption data. Our model outperforms previous captioning frameworks in recognizing novel objects while generating descriptions of comparable quality. We demonstrate our model’s captioning capabilities on a held-out set of MSCOCO objects as well as several hundred ImageNet objects. We also present an analysis of the contributions from different network modules, training objectives, and data sources. Additionally, our model directly extends to generate captions for ImageNet objects mentioned rarely in the image-caption corpora.

Birds			Outdoors		
	A small pheasant is standing in a field.	A osprey flying over a large grassy area.		A large glacier with a mountain in the background.	A group of people are sitting in a baobab .
Water Animals			Misc		
	A humpback is flying over a large body of water.	A man is standing on a beach holding a snapper .		A table with a cauldron in the dark.	A woman is posing for a picture with a chiffon dress.
Food			Kitchen		
	A close up of a plate of food with a scone .	A dumpling sitting on top of a wooden table		A saucepan and a pot of food on a stove top.	A large colander with a piece of food on it.
Instruments			Vehicles		
	A man holding a banjo in a park.	A large chime hanging on a metal pole		A snowplow truck driving down a snowy road.	A group of people standing around a large white warship .
Land Animals			Household		
	An okapi is in the grass with a okapi .	A small brown and white jackal is standing in a field.		A large metal candelabra next to a wall.	A black and white photo of a corkscrew and a corkscrew .
Errors					
	A chainsaw is sitting on a chainsaw near a chainsaw .	A man is sitting on a bike in front of a waggon .		A volcano view of a volcano in the sun.	A trampoline with a trampoline in the middle of it.

Figure 4: Examples of sentences generated by our NOC model on ImageNet images belonging to different categories.

Acknowledgements

Lisa Anne Hendricks is supported by the NDSEG Fellowship. Trevor Darrell was supported in part by DARPA; AFRL; DoD MURI award N000141110688; NSF awards IIS-1212798, IIS-1427425, and IIS-1536003, and the Berkeley Artificial Intelligence Research (BAIR) Lab. Raymond Mooney and Kate Saenko are supported

in part by DARPA under AFRL grant FA8750-13-2-0026 and a Google Grant. Mooney is also supported by ONR ATL Grant N00014-11-1-010.

References

[Denkowski and Lavie2014] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target lan-

- guage. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.
- [Donahue et al.2015] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.
- [Fang et al.2015] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *CVPR*.
- [Frome et al.2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129.
- [He et al.2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- [Hendricks et al.2016] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*.
- [Karpathy and Fei-Fei2015] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
- [Kiros et al.2014] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603.
- [Kiros et al.2015] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*.
- [Kuznetsova et al.2014] Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, UNC Chapel Hill, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. In *TACL*.
- [Lazaridou et al.2014] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL*.
- [Lin et al.2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.
- [Manning et al.2014] C. Manning, M Surdeanu, J Bauer, J Finkel, S J Bethard, and Dx McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- [Mao et al.2015a] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015a. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*.
- [Mao et al.2015b] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L. Yuille. 2015b. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*.
- [Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
- [Mitchell et al.2012] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *EACL*.
- [Norouzi et al.2013] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.
- [Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.
- [Russakovsky et al.2014] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ILSVRC.
- [Simonyan and Zisserman2014] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.
- [Socher et al.2014] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL*.
- [Sundermeyer et al.2010] M. Sundermeyer, R. Schluter, and H. Ney. 2010. Lstm neural networks for language modeling. In *INTERSPEECH*.

- [Vinyals et al.2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.
- [Yang et al.2011] Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *EMNLP*.