Captioning Images with Diverse Objects

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

Abstract

Recent captioning models are limited in their ability to scale and describe concepts unseen in paired image-text corpora. We propose the 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. Our model takes advantage of external sources – labeled images from object recognition datasets, and semantic knowledge extracted from unannotated text. We propose minimizing a joint objective which can learn from these 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 MSCOCO image-caption training data, as well as many categories that are observed very rarely. Both automatic evaluations and human judgements show that our model considerably outperforms prior work in being able to describe many more categories of objects.

1. Introduction

Modern visual classifiers [6, 22] 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 [2, 3, 8, 10, 15, 26] that learn directly from images and descriptions, rely solely on paired image-caption data for supervision and 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 have to not only correctly identify visual concepts contained in an image, but must also compose these concepts into a coherent sentence.

Figure 1. We propose a model that learns simultaneously from multiple data sources with auxiliary objectives to describe a variety of objects unseen in paired image-caption data.

Recent work [7] shows that, to incorporate the vast knowledge of current visual recognition networks without explicit paired caption training data, caption models can learn from external sources and learn to compose sentences about visual concepts which are infrequent or non-existent in image-description corpora. However, the pioneering DCC model from [7] is unwieldy in the sense that the model requires explicit transfer ("copying") of learned parameters from previously seen categories to novel categories. This not only prevents it from describing rare categories and limits the model's ability to cover a wider variety of objects but also makes it unable to be trained end-to-end. We instead propose the Novel Object Captioner (NOC), a network that can be trained end-to-end using a joint training strategy to integrate knowledge from external visual recognition datasets as well as semantic information from independent unannotated text corpora to generate captions for a diverse range of rare and novel objects (as in Fig. 1).

Specifically, we introduce auxiliary objectives which allow our network to learn a captioning model on image-caption pairs simultaneously with a deep language model and visual recognition system on unannotated text and labeled images. Unlike previous work, the auxiliary objectives allow the NOC model to learn relevant information from multiple data sources simultaneously in an end-to-end fashion. Furthermore, NOC implicitly leverages pre-trained distributional word embeddings enabling it to describe unseen and rare object categories. The main contributions of our work are 1) an end-to-end model to describe objects not present in paired image-caption data, 2) auxiliary/joint

1063-6919/17 $31.00 © 2017 IEEE
DOI 10.1109/CVPR.2017.130

training of the visual and language models on multiple data sources, and 3) incorporating pre-trained semantic embeddings for the task. We demonstrate the effectiveness of our model by performing extensive experiments on objects held out from MSCOCO [13] as well as hundreds of objects from ImageNet [21] unseen in caption datasets. Our model substantially outperforms previous work [7] on both automated as well as human evaluations.

2. Related Work

Visual Description. This area has seen many different approaches over the years [27, 11, 18], and more recently deep models have gained popularity for both their performance and 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 [2, 8, 26] or embedding image features and previously generated words into a multimodal space [9, 10, 15] before predicting the next word. Though most models represent images with an intermediate representation from a convolutional neural network (such as fc_7 activations from a CNN), other models represent images as a vector of confidences over a fixed number of visual concepts [3, 7]. In almost all cases, the parameters of the visual pipeline are initialized with weights trained on the ImageNet classification task. For caption generation, recurrent networks (RNNs) are a popular choice to model language, but log bilinear models [9] and maximum entropy language models [3] have also been explored. Our model is similar to the CNN-RNN frameworks in [7, 15] but neither of these models can be trained end-to-end to describe objects unseen in image-caption pairs.

Novel object captioning. [16] proposed an approach that extends a model's capability to describe a small set of novel concepts (e.g. *quidditch, samisen*) from a few paired training examples while retaining its ability to describe previously learned concepts. On the other hand, [7] 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. [7] integrate information from external text and visual sources, and explicitly transfer ('copy') parameters from objects seen in image-caption data to unseen ImageNet objects to caption these novel categories. While this works well for many ImageNet classes it still limits coverage across diverse categories 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 distributional semantic embeddings implicitly, obviating the need for any explicit transfer and making it end-to-end trainable. It also extends directly to caption ImageNet objects with few or no descriptions.

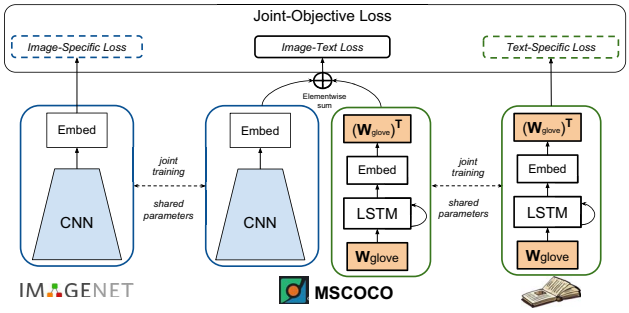

Figure 2. Our NOC image caption network. During training, the visual recognition network (left), the LSTM-based language model (right), and the caption model (center) are trained simultaneously on different sources with different objectives but with shared parameters, thus enabling novel object captioning.

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 information for zero-shot labeling of novel object categories [4, 19], as well as retrieval of images with text [12, 23]. Visual description itself can be cast as a multimodal learning problem in which caption words $w_0, ..., w_{n-1}$ and an image are projected into a joint embedding space before the next word in a caption, w_n, is generated [10, 15]. 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. This is similar in spirit to works in natural language processing that use monolingual data to improve machine translation [5].

3. Novel Object Captioner (NOC)

Our NOC model is illustrated in Fig. 2. It consists of a language model that leverages distributional semantic embeddings trained on unannotated text and integrates it with a visual recognition model. We introduce auxiliary loss functions (objectives) and jointly train different components on multiple data sources, to create a visual description model which simultaneously learns an independent object recognition model, as well as a language model.

We start by first training a LSTM-based language model (LM) [24] for sentence generation. Our LM incorporates dense representations for words from distributional embeddings (GloVe, [20]) 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. This decomposes 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). To generate descriptions conditioned on image content, we combine the predictions of our language and visual recognition networks by summing (element-wise) textual and visual confidences

over the vocabulary of words. During training, we introduce auxiliary image-specific ($\mathcal{L}_{\mathcal{IM}}$), and text-specific ($\mathcal{L}_{\mathcal{LM}}$) objectives along with the paired image-caption ($\mathcal{L}_{\mathcal{CM}}$) loss. 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.

3.1.1 Image-specific Loss

Our visual recognition model (Fig. 2, 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. 5.1) with a multi-label loss. If l denotes a label and z_l denotes the binary ground-truth value for the label, then the objective for the visual model is given by the cross-entropy loss ($\mathcal{L}_{IM}$):

$$\begin{aligned}\mathcal{L}_{\mathcal{IM}}(I;\theta_I) = -\sum_l \Big[& z_l \; log(S_l(f_{IM}(I;\theta_I))) \\ & + (1-z_l) \; log(1 - S_l(f_{IM}(I;\theta_I)))\Big]\end{aligned} \tag{1}$$

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

3.1.2 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 $w_0, ..., w_{t-1}$. This is optimized using the $\mathrm{softmax}$ loss $\mathcal{L}_{\mathcal{LM}}$ which is equivalent to the maximum-likelihood:

$$\begin{aligned}&\mathcal{L}_{\mathcal{LM}}(w_0, ..., w_{t-1}; \theta_L) = \\ &-\sum_t log(S_{w_t}(f_{LM}(w_0, ..., w_{t-1}; \theta_L)))\end{aligned} \tag{2}$$

3.1.3 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 $(w_0, ..., w_{t-1})$ 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, $P(w_t|w_0, ..., w_{t-1}, I)$

$$\begin{aligned}&= S_{w_t}(f_{CM}(w_0, ..., w_{t-1}, I; \theta)) \\ &= S_{w_t}(f_{LM}(w_0, ..., w_{t-1}; \theta_L) + f_{IM}(I; \theta_I))\end{aligned} \tag{3}$$

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 $\mathcal{L}_{\mathcal{CM}} : \mathcal{L}_{\mathcal{CM}}(w_0, ., w_{t-1}, I; \theta_L, \theta_I)$

$$= -\sum_t log(S_{w_t}(f_{CM}(w_0, ., w_{t-1}, I; \theta_L, \theta_I))) \tag{4}$$

3.1.4 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. 4), this is insufficent 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:

$$\mathcal{L} = \mathcal{L}_{\mathcal{CM}} + \mathcal{L}_{\mathcal{IM}} + \mathcal{L}_{\mathcal{LM}} \tag{5}$$

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.5), 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 [15, 25] of a dense vector back to the full vocabulary with the help of a $\mathrm{softmax}$ activation function. These distributional embeddings [17, 20] 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. Identical to previous work [7], we employ the VGG-16 [22] convolutional network as the visual recognition network. 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 full vocabulary. The rest of the classification network remains unchanged.

Finally, we take an elementwise-sum of the visual and language outputs, one can think of this as the language model producing a smooth probability distribution over words (based on GloVe parameter sharing) and then the image signal "selecting among these based on the visual evidence when summed with the the language model beliefs.

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 [14] 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 [20] 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 [7]. This dataset is created from MSCOCO [13] by clustering the main 80 object categories using cosine distance on word2vec (of the object label) and selecting one object from each cluster to hold out from training. The training set holds out images and sentences of 8 objects (bottle, bus, couch, microwave, pizza, racket, suitcase, zebra), which constitute about 10% of the training image and caption pairs in the MSCOCO dataset. Our model is evaluated on how well it can generate descriptions about images containing the eight held-out objects.

4.3. Image data

We also evaluate sentences generated by NOC on approximately 700 different ImageNet [21] 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", "woollen"), and entry level words (e.g., "toad"). Further, to study how well our model can describe rare objects, we pick a separate set of 52 objects which are in ImageNet but mentioned infrequently in MSCOCO (52 mentions on average, with median 27 mentions across all 400k training sentences).

5. Experiments on MSCOCO

We perform the following experiments to compare NOC's performance with previous work [7]: 1. We evaluate the model's ability to caption objects that are held out from MSCOCO during training (Sec. 5.1). 2. To study the effect of the data source on training, we report performance of NOC when the image and language networks are trained on in-domain and out-of-domain sources (Sec. 5.2). In addition to these, to understand our model better: 3. We perform ablations to study how much each component of our model (such as word embeddings, auxiliary objective, etc.) contributes to the performance (Sec. 5.3). 4. We also study if the model's performance remains consistent when holding out a different subset of objects from MSCOCO (Sec. 5.4).

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 [7]. We optimize each loss in our model with the following datasets: the caption model, which

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg. F1	Avg. METEOR
DCC	4.63	29.79	**45.87**	**28.09**	64.59	52.24	13.16	79.88	39.78	21.00
NOC (ours)	**17.78**	**68.79**	25.55	24.72	**69.33**	**55.31**	**39.86**	**89.02**	**48.79**	**21.32**

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

Racket
DCC: A man playing a **racket** on a court.
NOC (Ours): A tennis player swinging a **racket** at a ball.

Bus
DCC: A group of people on a snowy road next to trees.
NOC (Ours): **Bus** driving down a snowy road next to trees.

Bottle
DCC: A glass of wine sitting on a table with a glass of wine.
NOC (Ours): A table with a **bottle** of wine and a glass of wine.

Suitcase
DCC: A close up of a person sitting on a wooden bench.
NOC (Ours): A bunch of **suitcases** stacked on top of each other.

Figure 3. COCO Captioning: Examples comparing captions by NOC (ours) and DCC [7] on held out objects from MSCOCO.

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. When training the visual network on images from COCO, we obtain multiple labels for each image by considering all words in the associated captions as labels after removing stop words. 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. We use the METEOR metric [1] to evaluate description quality. However, METEOR only captures fluency and does not account for the mention (or lack) of specific words. Hence, we also use F1 to ascertain that the model mentions the object name in the description of the images containing the object. Thus, the metrics measure if the model can both identify the object and use it fluently in a sentence.

COCO heldout objects. Table 1 compares the F1 score achieved by NOC to the previous best method, DCC [7] on the 8 held-out COCO objects. NOC outperforms DCC (by 10% F1 on average) on all objects except "couch" and "microwave". The higher F1 and METEOR demonstrate that NOC is able to correctly recognize many more instances of the unseen objects and also integrate the words into fluent descriptions.

5.2. Training data source

To study the effect of different data sources, we also evaluate our model in an out-of-domain setting where classifiers

	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	17.56	36.50
3	COCO	Web Corpus	NOC	19.18	41.74
4	COCO	COCO	DCC	21.00	39.78
			NOC	**21.32**	**48.79**

Table 2. Comparison with different training data sources on 8 held-out COCO objects. Having in-domain data helps both the DCC [7] and our NOC model caption novel objects.

for held out objects are trained with images from ImageNet and the language model is trained on text mined from external corpora. Table 2 reports average scores across the eight held-out objects. We compare our NOC model to results from [7] (DCC), as well as a competitive image captioning model - LRCN [2] trained on the same split. In the out-of-domain setting (line 2), for the chosen set of 8 objects, NOC performs slightly better on F1 and a bit lower on METEOR compared to DCC. However, as previously mentioned, DCC needs to explicitly identify a set of "seen" object classes to transfer weights to the novel classes whereas NOC can be used for inference directly. DCC's transfer mechanism also leads to peculiar descriptions. E.g., *Racket* in Fig. 3.

With COCO image training (line 3), F1 scores of NOC improves considerably even with the Web Corpus LM training. Finally in the in-domain setting (line 4) NOC outperforms DCC on F1 by around 10 points while also improving METEOR slightly. 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.

5.3. Ablations

Table 3 compares how different aspects of training impact the overall performance. *Tuned Vision contribution* The model that does not incorporate Glove or LM pretraining 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. *LM & Glove contribution:* The model trained without the

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	✓	✓	✓	✓	**21.32**	**48.79**

Table 3. 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.

Model	bed	book	carrot	elephant	spoon	toilet	truck	umbrella	Avg. F1	Avg. METEOR
NOC (ours)	53.31	18.58	20.69	85.35	02.70	73.61	57.90	54.23	45.80	20.04

Table 4. MSCOCO Captioning: F1 scores (in %) of NOC (our model) on a different subset of the held-out objects not seen jointly during image-caption training, along with the average F1 and METEOR scores of the generated captions across images containing these objects. NOC is consistently able to caption different subsets of unseen object categories in MSCOCO.

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. *LM & Pre-trained Vision:* It's interesting to note that when we fix classifier's weights (pre-trained on all objects), before tuning the LM 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:* Incorporating the auxiliary objectives, F1 improves remarkably to 47.02. We note here that by virtue of including auxiliary objectives the visual network is tuned on all images thus retaining it's ability to classify/recognize a wide range of objects. Finally, incorporating all aspects gives NOC the best performance (F1 48.79, METEOR 21.32), significantly outperforming DCC.

5.4. Validating on a different subset of COCO

To show that our model is consistent across objects, we create a different training/test split by holding out a different set of eight objects from COCO. The objects we hold out are: bed, book, carrot, elephant, spoon, toilet, truck and umbrella. Images and sentences from these eight objects again constitute about 10% of the MSCOCO training dataset. Table 4 presents the performance of the model on this subset. We observe that the F1 and METEOR scores, although a bit lower, are consistent with numbers observed in Table 1 confirming that our model is able to generalize to different subsets of objects.

6. Experiments: Scaling to ImageNet

To demonstrate the scalability of NOC, we describe objects in ImageNet for which no paired image-sentence data exists. Our experiments are performed on two subsets of ImageNet, (i) Novel Objects: A set of 638 objects which are present in ImageNet as well as the model's vocabulary but are not mentioned in MSCOCO. (ii) Rare Objects: A set of 52 objects which are in ImageNet as well as the MSCOCO vocabulary but are mentioned infrequently in the MSCOCO captions (median of 27 mentions). For quantitative evaluation, (i) we measure the percentage of objects for which the model is able to describe at least one image of the object (using the object label), (ii) we also report accuracy and F1 scores to compare across the entire set of images and objects the model is able to describe. Furthermore, we obtain human evaluations comparing our model with previous work on whether the model is able to incorporate the object label meaningfully in the description together with how well it describes the image.

6.1. Describing Novel Objects

Table 5 compares models on 638 novel object categories (identical to [7]) 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%.

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

Table 5. ImageNet: Comparing our model against DCC [7] on % of novel classes described, average accuracy of mentioning the class in the description, and mean F1 scores for object mentions.

Woollen (n04599235)
DCC: A red and white cat sitting on top of a red **woollen**.
NOC (Ours): A red and blue **woollen** yarn sitting on a wooden table.

Newsstand (n03822656)
DCC: A bunch of people are sitting on a **newsstand**.
NOC (Ours): A extremely large **newsstand** with many different items on it.

Scythe (n04158250)
DCC: A small child is holding a small child on a skateboard.
NOC (Ours): A man is standing on a green field with a **scythe**.

Circuitry (n03034405)
DCC: A large white and black and white photo of a large building.
NOC (Ours): A bunch of different types of **circuitry** on a table.

Moussaka (n07872593)
DCC: A white plate topped with a sandwich and a **moussaka**.
NOC (Ours): A **moussaka** with cheese and vegetables on a white plate.

Warship (n04552696)
DCC: A **warship** is sitting on the water.
NOC (Ours): A large **warship** is on the water.

Caribou (n02433925)
DCC: A **caribou** is in a field with a small caribou.
NOC (Ours): A **caribou** that is standing in the grass.

Pharmacy (n03249342) [Both models incorporate the word incorrectly]
DCC: A white refrigerator freezer sitting on top of a **pharmacy**.
NOC (Ours): A kitchen with a **pharmacy** and a refrigerator.

Figure 4. ImageNet Captioning: Examples comparing captions by NOC (ours) and DCC [7] on objects from ImageNet.

Although NOC and DCC [7] use the same CNN, NOC is both able to describe more categories, and correctly integrate new words into descriptions more frequently. DCC [7] can fail either with respect to finding a suitable object that is both semantically and syntactically similar to the novel object, or with regard to their language model composing a sentence using the object name, in NOC the former never occurs (i.e. we don't need to explicitly identify similar objects), reducing the overall sources of error.

Fig. 4 and Fig. 6 (column 3) show examples where NOC describes a large variety of objects from ImageNet. Fig. 4 compares our model with DCC. Fig. 5 and Fig. 6 (right) outline some errors. Failing to describe a new object is one common error for NOC. E.g. Fig. 6 (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 "gladiator", "aardvark") and repeating a specific object ("A barracuda ... with a barracuda", "trifle cake").

6.2. Describing Rare Objects/Words

The selected rare 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,"whisk" (11 annotations), "teapot" (30 annotations), and others such as pumpkin appears 58 times, "swan" (60 annotations), and on the higher side objects like scarf appear 144 times. 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, integrating outside data, our NOC model can incorporate rare words into descriptions 35.15% of the time with an average F1 score of 47.58%. We do not compare this to DCC since DCC cannot be applied directly to caption rare objects.

Gladiator (n10131815) Error: Semantics
NOC: A man wearing a **gladiator** wearing a **gladiator** hat.

Taper (n13902793) Error: Counting
NOC: A group of three **taper** sitting on a table.

Trifle (n07613480) Error: Repetition
NOC: A **trifle** cake with **trifle** cake on top of a **trifle** cake.

Lory (n01820348) Error: Recognition
NOC: A bird sitting on a branch with a colorful bird sitting on it.

Figure 5. ImageNet Captioning: Common types of errors observed in the captions generated by the NOC model.

6.3. Human Evaluation

ImageNet images do not have accompanying captions and this makes the task much more challenging to evaluate. To compare the performance of NOC and DCC we obtain human judgements on captions generated by the models on several object categories. We select 3 images each from about 580 object categories that at least one of the two models, NOC and DCC, can describe. (Note that although both models were trained on the same ImageNet object categories, NOC is able to describe almost all of the object categories that have been described by DCC). When selecting the images, for object categories that both models can describe, we make sure to select at least two images for which both models mention the object label in the description. Each image is presented to three workers. We conducted two human studies (sample interface is in the supplement): Given the image, the ground-truth object category (and meaning), and the captions generated by the models, we evaluate on:

Word Incorporation: We ask humans to choose which sentence/caption incorporates the object label meaningfully in the description. The options provided are: (i) Sentence 1 incorporates the word better, (ii) Sentence 2 incorporates the word better, (iii) Both sen-

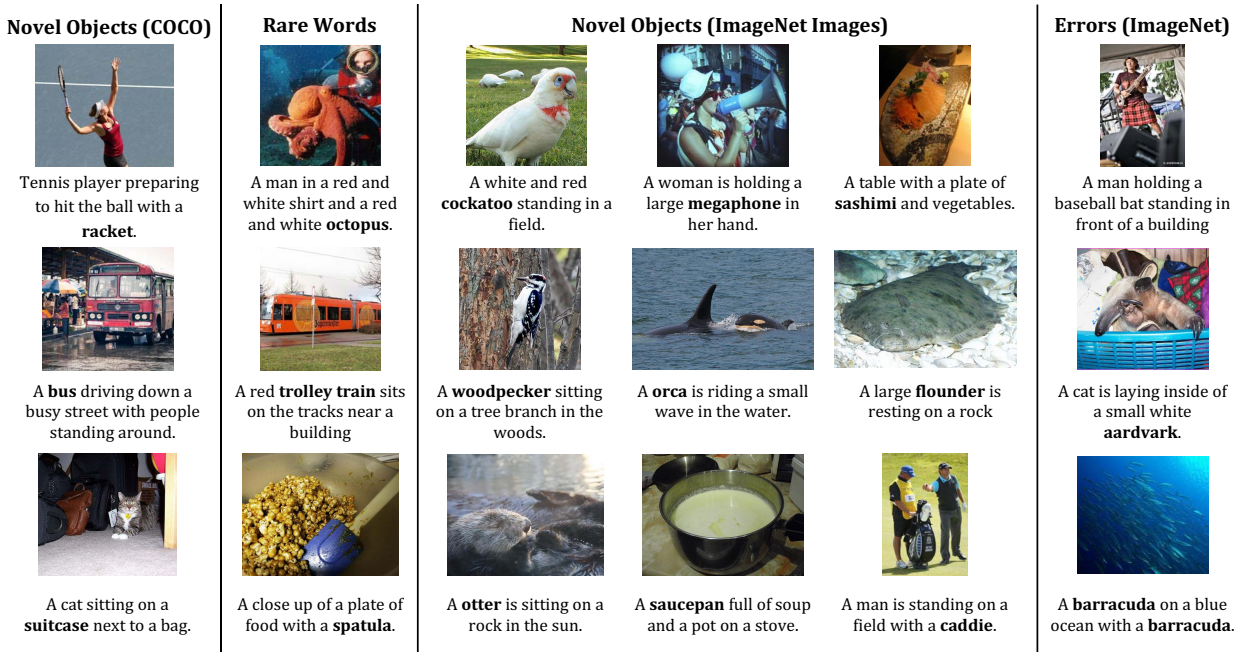

Figure 6. Descriptions produced by NOC on a variety of objects, including "caddie", "saucepan", and "flounder". (Right) NOC makes errors and (top right) fails to describe the new object ("sitar"). More categories of images and objects are in the supplement.

tences incorporate the word equally well, or (iv) Neither of them do well.

Image Description: We also ask humans to pick which of the two sentences describes the image better.

This allows us to compare both how well a model incorporates the novel object label in the sentence, as well as how appropriate the description is to the image. The results are presented in Table 6. On the subset of images corresponding to objects that both models can describe (Intersection), NOC and DCC appear evenly matched, with NOC only having a slight edge. However, looking at all object categories (Union), NOC is able to both incorporate the object label in the sentence, and describe the image better than DCC.

	Word Incorporation		Image Description	
Objects subset →	Union	Intersection	Union	Intersection
NOC is better	**43.78**	34.61	**59.84**	51.04
DCC is better	25.74	34.12	40.16	48.96
Both equally good	6.10	9.35		-
Neither is good	24.37	21.91		-

Table 6. ImageNet: Human judgements comparing our NOC model with DCC [7] on the ability to meaningfully incorporate the novel object in the description (Word Incorporation) and describe the image. 'Union' and 'Intersection' refer to the subset of objects where atleast one model, and both models are able to incorporate the object name in the description. All values in %.

7. Conclusion

We present an end-to-end trainable architecture that incorporates auxiliary training objectives and distributional semantics to generate descriptions for object classes unseen in paired image-caption data. Notably, NOC's architecture and training strategy enables the visual recognition network to retain its ability to recognize several hundred categories of objects even as it learns to generate captions on a different set of images and objects. We demonstrate our model's captioning capabilities on a held-out set of MSCOCO objects as well as several hundred ImageNet objects. Both human evaluations and quantitative assessments show that our model is able to describe many more novel objects compared to previous work. NOC has a 10% higher F1 on unseen COCO objects and 20% higher F1 on ImageNet objects compared to previous work, while also maintaining or improving descriptive quality. 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. Code is available at: https://vsubhashini.github.io/noc.html

Acknowledgements

We thank anonymous reviewers and Saurabh Gupta for helpful suggestions. Venugopalan is supported by a UT scholarship, and Hendricks was supported by a Huawei fellowship. 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 Lab. Mooney and Saenko are supported in part by DARPA under AFRL grant FA8750-13-2-0026 and a Google Grant.

References

[1] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014. 5

[2] 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, 2, 5

[3] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, 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, 2

[4] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 2

[5] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[7] 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, 4, 5, 6, 7, 8

[8] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2

[9] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603, 2014. 2

[10] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 1, 2

[11] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. In *TACL*, 2014. 2

[12] A. Lazaridou, E. Bruni, and M. Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL*, 2014. 2

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 4

[14] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. 4

[15] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015. 1, 2, 4

[16] 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

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 4

[18] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. C. Berg, T. L. Berg, and H. D. III. Midge: Generating image descriptions from computer vision detections. In *EACL*, 2012. 2

[19] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[20] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014. 2, 4

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ILSVRC, 2014. 2, 4

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 4

[23] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2014. 2

[24] M. Sundermeyer, R. Schlüter, and H. Ney. LSTM neural networks for language modeling. In *INTERSPEECH*, 2012. 2

[25] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko. Improving LSTM-based video description with linguistic knowledge mined from text. In *EMNLP*, 2016. 4

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2

[27] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011. 2