# Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data

Lisa Anne Hendricks[1] Subhashini Venugopalan[3] Marcus Rohrbach[1,2]
Raymond Mooney[3] Kate Saenko[4] Trevor Darrell[1,2]
[1] UC Berkeley, [2] ICSI, Berkeley, [3] UT Austin, [4] UMass Lowell

## Abstract

*While recent deep neural network models have achieved promising results on the image captioning task, they rely largely on the availability of corpora with paired image and sentence captions to describe objects in context. In this work, we propose the Deep Compositional Captioner (DCC) to address the task of generating descriptions of novel objects which are not present in paired image-sentence datasets. Our method achieves this by leveraging large object recognition datasets and external text corpora and by transferring knowledge between semantically similar concepts. Current deep caption models can only describe objects contained in paired image-sentence corpora, despite the fact that they are pre-trained with large object recognition datasets, namely ImageNet. In contrast, our model can compose sentences that describe novel objects and their interactions with other objects. We demonstrate our model's ability to describe novel concepts by empirically evaluating its performance on MSCOCO and show qualitative results on ImageNet images of objects for which no paired image-sentence data exist. Further, we extend our approach to generate descriptions of objects in video clips. Our results show that DCC has distinct advantages over existing image and video captioning approaches for generating descriptions of new objects in context.*

## 1. Introduction

In the past year, several deep recurrent neural network models have demonstrated promising results on the task of generating descriptions for images and videos [36, 5, 16, 15, 21]. Large corpora of paired images and descriptions, such as MSCOCO [20] and Flickr30k [11] have been an important factor contributing to the success of these methods. However, these datasets describe a relatively small variety of objects in comparison to the number of labeled objects in object recognition datasets, such as ImageNet [3]. Consequently, though modern object recognition systems have the capacity to recognize thousands of object classes, existing state-of-the-art caption models lack the ability to form compositional structures which integrate new objects with known concepts without explicit examples of image-sentence pairs. To address this limitation, we propose the Deep Compositional Captioner (DCC) which can combine visual groundings of lexical units to generate descriptions about objects which are not present in caption corpora (paired image-sentence data), but are present in object recognition datasets (unpaired image data) and text corpora (unpaired text data).

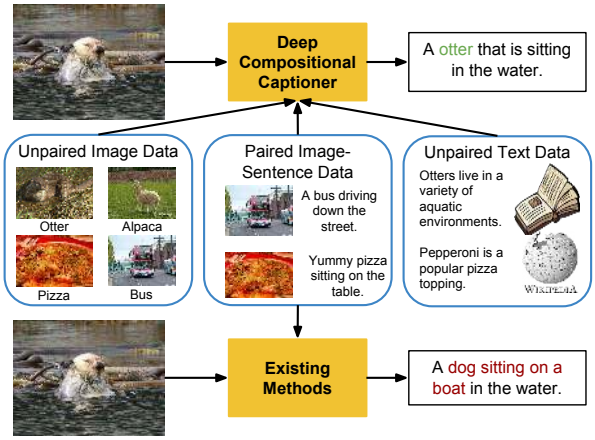


Figure 1: Existing deep caption methods are unable to generate sentences about objects unseen in caption corpora (like otter). In contrast, our model effectively incorporates information from independent image datasets and text corpora to compose descriptions about novel objects without any paired image-sentence data.

DCC builds on recent deep captioning models [36, 5, 16, 15, 21] which combine convolutional and recurrent networks for visual description. However unlike previous models which can only describe objects that are present in paired image-sentence data, DCC is compositional in the sense that it can seamlessly construct sentences about new objects by combining them with already seen linguistic expressions in paired training data. To illustrate, consider the image of the otter in Figure 1. To describe the image ac-

curately, any captioning model needs to identify the constituent visual elements such as "otter", "water" and "sitting" and combine them to generate a coherent sentence. While previous deep caption models learn to combine visual elements into a cohesive description exclusively from image and caption pairs, DCC can compose a caption to describe a new visual element such as the "otter" by understanding that "otters" are similar to "animals" and can thus be composed in the same way with other lexical expressions.

To effectively describe new objects, our model incorporates two key design elements. First, DCC consists of a separate lexical classifier and language model, which can each be trained independently on unpaired image data and unpaired text data. Additionally, the lexical classifier and language model can be combined into a deep caption model which is trained jointly on paired image-sentence data. Second, and crucial for generating compositional captions, is the multimodal layer where knowledge from known objects in paired image-sentence datasets can be transferred to new objects only seen in unpaired datasets. In this work, we leverage external text corpora to relate novel objects to concepts seen in paired data and propose two mechanisms to transfer knowledge from known objects to novel objects.

We demonstrate the ability of DCC to generate captions about new objects by empirically studying results on a training split of the MSCOCO dataset which excludes certain objects. Qualitatively, we show that our model can be used to describe a variety of objects in the Imagenet 7k dataset which are not present in caption datasets. Furthermore, we demonstrate that the efficacy of DCC is not limited to images, but can also be used to describe new objects in videos by presenting results on a collection of Youtube video clips.

## 2. Related Work

**Deep Captioning**. In the last year, a variety of models [5, 36, 15, 16, 6, 21] have achieved promising results on the image captioning task. Some [5, 36, 15] follow a CNN-RNN framework: first high-level features are extracted from a CNN trained on the image classification task, and then a recurrent model learns to predict subsequent words of a caption conditioned on image features and previously predicted words. Others [16, 21] adopt a multimodal framework in which recurrent language features and image features are embedded in a multimodal space. The multimodal embedding is then used to predict the caption word by word. Retrieval methods [4] based on comparing the k-nearest neighbors of training and test images in a deep image feature space, have also achieved competitive results on the captioning task. However, retrieval methods are limited to words and descriptions which appear in a training set of paired image-sentence data. As opposed to using high level image features extracted from a CNN, another approach [6, 37] is to train classifiers on visual concepts such as objects, attributes and scenes. A language model, such as an LSTM [37] or maximum entropy model [6], then generates a visual description conditioned on the presence of classified visual elements. Our model most closely resembles the framework suggested in [21] which uses a multimodal space to combine features from image and language, however our approach modifies this framework considerably to describe concepts that are never seen in paired image-sentence data.

**Zero-Shot Learning**. Zero-shot learning has received substantial attention in computer vision [27, 24, 19, 30, 7] since it becomes difficult to obtain sufficient labeled images as the number of object categories grows. In particular, our method draws on previous zero-shot learning work that mines object relationships from external text data [27, 30, 7]. [27] uses text corpora to determine how objects are related to each other, then classifies unknown objects based on their relationship to known objects. In [30, 7], images are mapped to semantic word vectors corresponding to their classes, and the resulting image embeddings are used to detect and distinguish between unseen and seen classes. We also exploit transfer learning via an intermediate-level semantic word vector representation, however, the above approaches focus specifically on assigning a category label, while our method generates full sentence descriptions. In [12], zero-shot object detectors are learned by transferring information about how network weights trained on the classification task differ from weights trained on the detection task. We explore a similar transfer method to transfer information from weights which are trained on image-sentence data to weights which are only trained on text data.

**Describing New Objects in Context**. Many early caption models [31, 17, 8, 18, 9] rely on first discerning visual elements from an image, such as subjects, objects, scenes, and actions, then filling in a sentence template to create a coherent visual description. These models are capable of describing objects without being provided with paired image-sentence examples containing the objects, but are restricted to generating descriptions using a fixed, predetermined template. More recently, [21] explore describing new objects with a deep caption model with only a few paired image-sentence examples during training. However, [21] do not consider how to describe objects when no paired image-sentence data is available. Our model provides a mechanism to include information from existing vision datasets as well as unpaired text data, whereas [21] relies on additional image-sentence annotations to describe novel concepts.

## 3. Deep Compositional Captioner

DCC composes novel sentences about objects unseen in paired image-sentence data. Although it is common to pretrain deep caption models on unpaired image data, unlike existing models, we are able to describe objects present

in unpaired image data but not present in paired image-sentence data. Additionally, to enhance the language structure, we train our model on independent text corpora. Further, we explore methods to transfer knowledge between semantically related words to compose descriptions of new objects. Our method consists of three stages: 1) training a deep lexical classifier and deep language model with unpaired data, then, 2) combining the lexical classifier and language model into a caption model which is trained on paired image-sentence data, and, finally, 3) transferring knowledge from words which appear in paired image-sentence data to words which do not appear in paired image-sentence data.

### 3.1. Deep Lexical Classifier

The lexical classifier (Fig 2, left) is a CNN which maps images to semantic concepts. In order to train the lexical classifier, we first mine concepts which are common in paired image-text data by extracting the part-of-speech of each word [33] and then select the most common adjectives, verbs, and nouns. We do not refine the mined concepts, which means some of the concepts, such as "use", are not strictly visual. In addition to concepts common in paired image-sentence data, the classifier is also trained on objects that we wish to describe outside of the caption datasets.

The lexical classifier is trained by fine-tuning a CNN which is pre-trained on the training split of the ILSVRC-2012 [28] dataset. When describing images, multiple visual concepts from the image influence the description. For example, the sentence "An alpaca stands in the green grass." includes the visual concepts "alpaca", "stands", "green", and "grass". In order to apply multiple labels to each image, we use a sigmoid cross-entropy loss. We denote the image feature output by the lexical classifier as $f_I$, where each index of $f_I$ corresponds to the probability that a particular concept is present in the image. Our idea of learning visual classifiers from text descriptions for captioning is similar to [26] who learn classifiers for objects, verbs, and locations and [6] who learn visual concepts using multiple instance learning.

### 3.2. Language Model

The language model (Fig 2, right) learns sentence structure using only unpaired text data and includes an embedding layer which maps a one-hot-vector word representation to a lower dimensional space, an LSTM [10], and a word prediction layer. The language model is trained to predict a word given previous words in a sentence. At each time step, the previous word is input into the embedding layer. The embedded word is input into an LSTM, which learns the recurrent structure inherent in language. The embedded word and LSTM output are concatenated to form the language features, $f_L$. $f_L$ is input to an inner product layer which outputs the next word in a generated sequence. At training

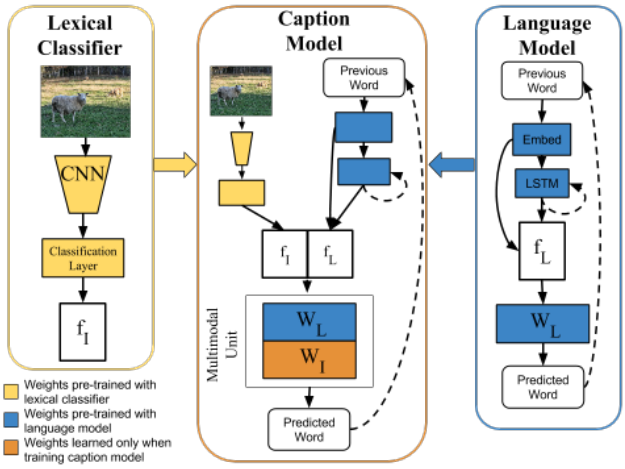


Figure 2: DCC consists of a lexical classifier, which maps pixels to semantic concepts and is trained only on unpaired image data, and a language model, which learns the structure of natural language and is trained on unpaired text data. The multimodal unit of DCC integrates the lexical classifier and language model and is trained on paired image-sentence data.

time, the ground truth word is always used as an input to the language model, but at test time we input the previous word predicted by our model. We also find that results improve by enforcing a constraint that the model cannot predict the same word twice in a row. We explore a variety of sources for unpaired text corpora as described in Section 4.1.

### 3.3. Caption Model

The caption model integrates the lexical classifier and the language model to learn a joint model for image description. As shown in Fig 2 (center) the multimodal unit in the caption model combines the image features, $f_I$ and the language features, $f_L$. The multimodal unit we use is an affine transformation of the image and language features:

$$p_w = \text{softmax}(f_I W_I + f_L W_L + b) \quad (1)$$

where $W_I$, $W_L$, and $b$ are learned weight matrices and $p_w$ is a probability distribution over the predicted word.

Intuitively, the weights in $W_I$ learn to predict a set of words which are likely to occur in a caption given the visual elements discerned by the lexical classifier. In contrast, $W_L$ learns the sequential structure of natural language by learning to predict the next word in a sequence given the previous words. By summing $f_I W_I$ and $f_L W_L$, the multimodal unit combines the visual information learned by the lexical classifier with the knowledge of language structure learned by the language model to form a coherent description of an image.

Both the language model and caption model are trained to predict a sequence of words, whereas the lexical classifier is trained to predict a fixed set of candidate visual elements

for a given image. Consequently, the weights $W_L$, which map language features to a predicted word are learned when training the language model, but the weights $W_I$ are not. Weights in $W_L$ are pretrained using unpaired text data before fine-tuning with paired image-sentence data, $W_I$ are trained purely with image-sentence data. Though we use a linear multimodal unit, our results are comparable to results achieved by other methods which include a nonlinear layer for word prediction. For example, on the MSCOCO validation set [5] achieves a METEOR score of 23.7, and DCC achieves a METEOR of 23.2.

The caption model is designed to enable easy transfer of learned weights from words which appear in the paired image-sentence data to words which do not appear in the image-sentence data. First, by using a lexical classifier to extract image features, image features have explicit semantic meaning. Consequently, it is trivial to expand the image feature to include new objects and to adjust weights in the multimodal unit which correspond to specific objects. Second, by learning language features using unpaired text data, we ensure that the model learns a good embedding for words which are not present in paired image-sentence data. Finally, by using a single-layer, linear multimodal unit, the dependence between image and language features and predicted words is straightforward to understand and easy to exploit for semantic transfer.

### 3.4. Transferring Information Between Objects

**Direct Transfer**. The first method we explore to transfer weights between objects directly transfers learned weights in $W_I$, $W_L$ and $b$ from words that appear in the paired image-sentence dataset to words which do not appear in a paired image-sentence dataset (Fig 3). Intuitively, the direct transfer model requires that a new word is described in the same way that semantically similar words are described. To illustrate, consider the new word "alpaca" which is semantically close to the known word "sheep". Let $v_a$ and $v_s$ indicate the index of the words alpaca and sheep in the vocabulary. Given image and language features, $f_I$ and $f_L$ respectively, the probability of predicting the word "sheep" is proportional to:

$$f_I W_I[:, v_s] + f_L W_L[:, v_s] + b[v_s] \quad (2)$$

In order to construct sentences with "alpaca" in the same way sentences are constructed with the word "sheep", we first directly transfer the weights $W_I[:, v_s]$, $W_L[:, v_s]$, and $b[v_s]$ (indicated in red in Fig 3) to $W_I[:, v_a]$, $W_L[:, v_a]$, and $b[v_a]$ (indicated in green in Fig 3). Additionally, we expect the prediction of the word "sheep" to be highly dependent on the likelihood that a "sheep" is present in the image. In other words, we expect $W_I[:, c_s]$ to strongly weight the output of the lexical classifier which corresponds to the word "sheep". However, $W_I[:, c_a]$ should strongly weight the lexical classifier which corresponds to the word "alpaca". To enforce this, we set $W_I[r_a, c_a] = W_I[r_s, c_s]$ where $r_a$ and $r_s$ indicate the index in the image features which correspond to the alpaca and sheep classifiers respectively. Finally, we do not expect the output of the word "alpaca" to depend on the presence of a sheep in the image and vice versa. Consequently, we set $W_I[r_s, c_a] = W_I[r_a, c_s] = 0$.

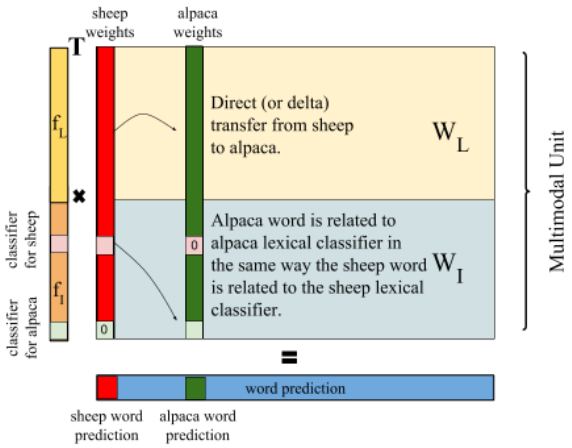


Figure 3: Method for transferring knowledge from words trained with paired image-sentence data to words trained without image-sentence data. See Section 3.4 for details.

**Delta Transfer**. Instead of directly transferring weights, we can also transfer *how weights change* when trained on paired image-text data. Again, consider transferring the word "sheep" to the word "alpaca". We determine $\Delta_L$ for a given word as:

$$\Delta_L = W_{L-caption}[:, v_s] - W_{L-language}[:, v_s] \quad (3)$$

where $W_{L-caption}$ are weights learned when training with both images and sentences and $W_{L-language}$ are weights learned when training only with language. The weights for the new word "alpaca" are updated as:

$$W_{L-caption}[:, v_a] = W_{L-language}[:, v_a] + \Delta_L \quad (4)$$

Delta transfer may be advantageous because, unlike direct transfer, it does not overwrite pretrained weights in $W_L$ during transfer. When performing delta transfer for $W_L$, we still use direct transfer for weights in $W_I$.

**Determining Concept Similarity**. Determining which words in the paired image-sentence data are semantically similar to words out of the paired image-sentence data is key for transfer. We determine semantic similarity with the word2vec [22] CBOW model which we trained on the British National Corpus (BNC), UkWaC, and Wikipedia, and estimate word similarity using cosine distance. Additionally, we restrict words that are transferred to new words to be in the lexical layer.

## 4. Experimental Framework

### 4.1. Datasets

**Image Description**. To empirically evaluate our method we create a subset of the MSCOCO [20] training set (denoted as the held-out MSCOCO training set) which excludes all image-sentence pairs which describe at least one of eight MSCOCO objects. To ensure that excluded objects are at least similar to some included ones, we cluster the 80 objects annotated in the MSCOCO segmentation challenge using the vectors from the word2vec embedding described in Section 3.4 and exclude one object from each cluster. The following words are chosen: "bottle", "bus", "couch", "microwave", "pizza", "racket", "suitcase", and "zebra". We randomly select 50% of the MSCOCO validation set for validation, and set aside the other 50% for testing. We use the validation set to determine all model hyperparameters, and present all results on the test set. We label the visual concepts in each image based on the five ground truth caption annotations provided in the MSCOCO dataset. If any of the ground truth captions mention an object, the corresponding image is considered a positive example for that object.

In addition to empirically evaluating our model, we also qualitatively examine the performance of DCC at a large scale by describing objects outside of the paired image-sentence corpora. Specifically, we select 642 objects from the full ImageNet object recognition dataset [3] which do not occur in MSCOCO and are also present in the WebCorpus text dataset (see section 4.3) vocabulary. We do no manual concept pruning; consequently some selected concepts refer to a broad variety of objects (e.g., the class "fauna" contains all animals) and other classes only contain a small number of images (e.g., there are three "discus" images). We use 75% of images from each class to train the lexical classifier, and evaluate on the rest. We stress that we do not have any descriptions for these categories.

**Video Description**. For empirical evaluation on video description, we use a collection of Youtube clips from the Microsoft Video description (MSVD) corpus [2], which contains 1,970 short annotated clips. Our basic experimental setting follows previous video description works [35, 34]. However, we hold out paired video-sentence data for some objects during training. Because there is significant variation in the number of video clips containing each object in the MSVD dataset, we hold out objects in the MSVD dataset which appear in five or fewer training videos and at least one test video and also appear in the ILSVRC2015 video object detection challenge set. [1] Our MSVD held-out set excludes paired video-sentence training data which include "zebra", "hamster", "broccoli", and "turtle".

We also qualitatively evaluate our method on the ILSVRC object detection challenge videos (initial release) which consists of 1,952 video snippets of the 30 objects from the ILSVRC2015 object detection in video. Objects which we describe in the detection challenge videos include "whale", "fox", "hamster", "lion", "zebra", and "turtle".

[1] http://image-net.org/challenges/LSVRC/2015/

### 4.2. Training the Lexical Classifier

**Image description**. We consider both MSCOCO and ImageNet as sources of labeled image data to train the lexical classifier. For all objects in paired image-sentence data, we use COCO images which are labeled with 471 visual concepts to train the lexical classifier. For the eight objects which do not appear in the paired image-sentence data, we explore training the lexical classifier using MSCOCO images (in-domain) and ImageNet images (out-of-domain). For qualitative experiments on ImageNet objects, we use Imagenet images to train the lexical classifier on new visual concepts. The lexical classifier is trained by fine-tuning a deep convolutional model (VGG-16 layer [29]) trained on the ILSVRC-2012 [28] object recognition training subset of the ImageNet dataset.

**Video Description**. Unlike images, videos consist of a sequence of frames which need to be mapped to a set of semantic concepts by the lexical classifier. To build a lexical classifier for videos, we mean-pool $fc_7$ features across all frames in a video clip before classification. We use both MSVD and ImageNet videos to train the lexical classifier. We use the VGG-16 layer model to extract $fc_7$ layer features from video frames.

### 4.3. Training the Language Model

**Image Description**. We consider three different sources for unpaired text data to train the language model: (1) **MSCOCO** consists of all captions from the MSCOCO train set (2) **Text from Image Description Corpora (CaptionTxt)** consists of text data from other paired image and video description datasets: Flickr1M [13], Flickr30k [11], Pascal-1k [25] and ImageCLEF-2012 [32] and sentence descriptions of Youtube clips from the MSVD training corpus. This corpus *does not* include sentences from MSCOCO. (3) **External text (WebCorpus)** consists of 60 million sentences from the British National Corpus (BNC), UkWaC, and Wikipedia.

**Video Description**. We consider two sources of text to train the video description language model. The first is the WebCorpus text described above. We also consider a slight variant on the CaptionTxt described above which includes descriptions from MSCOCO, Flickr-30k [11], Pascal-1k [25] and the MSVD sentence descriptions.

### 4.4. Training the Caption Model

After training the lexical classifier and language model, the weights in the multimodal layer of the caption model are trained with paired image-sentence data. For the direct

| | No Transfer | ΔT | DT |
|---|---|---|---|
| F1 | 0 | 34.89 | **39.78** |
| BLEU-1 | 62.99 | 64.00 | **64.40** |
| METEOR | 19.9 | 20.86 | **21.00** |

Table 1: We compare the the delta transfer (ΔT) and direct transfer (DT) DCC models to a model with no transfer. We measure our models ability to insert new words into a generated sentence with the F1-score. We also report Bleu-1 and METEOR, which indicates overall sentence quality. DCC successfully incorporates new words and improves sentence quality. (Values in %)

transfer method, we simply train the weights in the multimodal unit ($W_I$ and $W_L$) while freezing all other weights. For the delta transfer method, if weights in $W_L$, which are pretrained when training the language model, diverge too much from their original values, transfer does not work well. Consequently, we first hold weights in $W_L$ constant, training only $W_I$, before jointly training $W_L$ and $W_I$. The caption model is trained the same way for both image and video description.

### 4.5. Metrics

To evaluate our transfer methods, we must choose a metric that indicates whether or not a generated sentence includes a new object. Common caption metrics such as BLEU [23] and METEOR [1] measure overall sentence meaning and fluency. However, for many objects, it is possible to achieve good BLEU and METEOR scores without mentioning the new object (e.g., consider sentences describing the boy playing tennis in Figure 4). To definitively report our model's ability to integrate new vocabulary, we also report the F1-score. The F1-score considers "false positives" (when a word appears in a sentence it should not appear in), "false negatives" (when a word does not appear in a sentence it should appear in), and "true positives" (when a word appears in a sentence it should appear in). We consider generated sentences "positive" if they contain at least one mention of a held out word and ground truth sentences "positive" if a word is mentioned in any ground truth annotation that describes an image.

We train our models using *Caffe*[14]. [2]

## 5. Results

### 5.1. Image Description

As shown in Figure 4, DCC is capable of integrating new vocabulary into image descriptions in a cohesive manner.

**Direct Transfer Versus Delta Transfer**. Table 1 compares the average F1-score across the eight held-out training classes. As shown by the F1-scores reported in Table 1, both the delta transfer and direct transfer methods are capable of integrating new words into their vocabulary. We also report the BLEU-1 score, which measures the overlap between generated words and words in reference sentences. By measuring the METEOR score, we ensure that our model maintains sentence fluency when inserting new objects. DCC consistently increases METEOR scores indicating that overall sentence quality improves with DCC. The direct transfer method improves F1-scores, BLEU, and METEOR scores by a larger amount than the delta transfer method and is thus used for the remainder of our experiments.


Figure 4: Image Description: Comparison of captions generated by a model without transfer, DCC with in-domain training (MSCOCO), with out-of-domain training (ImageNet and WebCorpus), and a model trained with paired image-sentence supervision for all MSCOCO objects. DCC is capable of integrating new words and generates sentences similar to those generated when paired image-sentences for all objects are present during training.

Importantly, BLEU and METEOR scores do not decrease for objects which are present in the held-out training data set. When trained with all image-sentence training examples, our model achieves an average BLEU-1 of 69.36 and METEOR of 23.98 on held-out classes.

To illustrate which words our model works best on, we report the F1-score for individual objects in Table 2. We compare to a model which is trained with image-sentence pairs for the eight held-out objects. For all objects, DCC is able to compose sentences which include the object.

**Analysis of Transfer Words**. In general, determining word similarity with a word2vec embedding works well. Words such as "zebra"/"giraffe" and "mi-

[2] Code can be found at `http://www.eecs.berkeley.edu/~lisa_anne/dcc_project_page.html`.

| | bottle | bus | couch | microwave | pizza | racket | suitcase | zebra | average |
|---|---|---|---|---|---|---|---|---|---|
| Pair Supervision | 23.20 | 72.07 | 50.60 | 39.48 | 77.07 | 38.52 | 46.50 | 91.02 | 54.81 |
| DT | 4.63 | 29.79 | 45.87 | 28.09 | 64.59 | 52.24 | 13.16 | 79.88 | 39.78 |

Table 2: Image Description: Comparison of F1 scores for direct transfer DCC model (DT) and a model trained with image-sentence training examples for all objects. (Values in %)

| Lexical classifier | Language model | B-1 | METEOR | F1 |
|---|---|---|---|---|
| MSCOCO | MSCOCO | 64.40 | 21.00 | 39.78 |
| Imagenet | MSCOCO | 64.00 | 20.71 | 33.60 |
| Imagenet | CaptionTxt | 64.79 | 20.66 | 35.53 |
| Imagenet | WebCorpus | 64.85 | 20.66 | 34.94 |

Table 3: Image Description: We compare the effect of pre-training the lexical classifier and language model with different unpaired image and text data sets. As expected, we see the best result when using in domain MSCOCO data to train the lexical classifier and language model, though training with out of domain corpora is comparable. (Values in %)

crowave"/"refrigerator" are close in embedding space and are also used in similar ways in natural language, suggesting they will work well for transfer. Some transfer pairs ("racket"/"tennis" and "bus"/"stop") are used together frequently but play different structural roles in sentences. Consequently, the word "racket" is frequently used like the word "tennis" leading to grammatical errors. However, similar errors do not occur when transferring "stop" to "bus".

**Pre-Training with Out-of-Domain Data**. In the above experiments the lexical classifier and language model are pre-trained using MSCOCO images and text. In a real world scenario, it is unlikely that available unpaired image and text data will be from the same domain as paired image-sentence data. However, it is essential that the model learns good image and language features. Naturally, if the lexical classifier is unable to recognize certain objects, DCC will not be able to describe the objects. Perhaps more subtly, if the language model is not trained with unpaired text which includes an object, it will not learn a proper embedding for the new word and will not produce cohesive descriptions about new objects.

Table 3 demonstrates the impact of using outside image and text corpora to train the lexical classifier and language model. Our model performs best when provided with in-domain image and text for all training stages, but performance is comparable when using ImageNet images to train the lexical classifier and CaptionTxt or WebCorpus text data to train the language model.

## 5.2. Describing ImageNet Objects

We qualitatively assess our model by describing a variety of ImageNet objects which are not included in the MSCOCO data set (Fig 5). DCC accurately describes 335 new words including entry-level words like "toad" as well as fine-grained categories like "baobab". Though most Imagenet words we transfer are nouns, we are able to successfully transfer some adjectives such as "chiffon". DCC achieves more than simple noun replacement. For example, the sentence "A large giraffe standing in a tree" changes significantly to "A large baobab in a field with trees in the background" after transfer. Importantly, our model is able to compose sentences by placing objects in the correct context. For example, comparing Fig 5 (top) to the image in Fig 1, the object "otter" is correctly described as either "sitting in the water" or "standing on top of a lush green field" depending on visual context.


Figure 5: Image Description: DCC is able to describe Imagenet objects (bolded) which are not mentioned in any of the paired image-sentence data, and therefore cannot be described by existing deep caption models. X → Y indicates that the known word X is transferred to the new word Y.

Figure 6 examines a few common error types:

**New Object Not Mentioned**. (Figure 6, top) For some images, DCC produces relevant sentences, but fails to mention the new object.

**Grammatically Incorrect**. (Figure 6, second row) Some

Error: New object (lifejacket) not mentioned
DCC: A group of people sitting on a bench.

Error: Grammatically Incorrect
DCC: A woman is playing **gymnastics** on a **gymnastics** court.

Error: Object Hallucination
DCC: A woman in a **snorkel** is holding a surfboard.

Error: Irrelevant description
DCC: A dog is sitting on a white bench.

Figure 6: Image Description: We highlight four common error types generated by the DCC. See Section 5.2 for details.

sentences incorporate new words, but are grammatically incorrect. For example, though DCC describes sentences with the word "gymnastics", the resulting sentences are frequently grammatically incorrect (e.g., "A woman playing gymnastics on a gymnastics court"). This is likely because the word "tennis" is transferred to "gymnastics". Though both of these words are sports, one does not "play" gymnastics and gymnastics is not performed on a "court."

**Object Hallucination**. (Figure 6, third row) DCC frequently hallucinates objects which commonly occur in a specific visual context. For example, in a beach image, the model commonly includes the word "surfboard".

**Irrelevant Description**. (Figure 6, bottom) Some captions do not mention any salient objects correctly. Such errors can be caused by poor image recognition or because the language model is unable to construct a reasonable sentence from constituent visual elements.

More examples are in our supplemental material.

### 5.3. Video description

We believe DCC can be especially beneficial in domains, such as video description, where the amount of paired training data is small. Table 4 presents empirical results of direct transfer DCC on videos in the MSVD corpus (Section 4.1). We report the average F1 score on all held-out classes, and METEOR scores on the complete test dataset. As seen by the F1 score, transferring weights allows us to describe new objects in video. Additionally, the METEOR score improves with transfer demonstrating that DCC improves overall sentence quality. Similar to the trend seen for image captioning, training on in-domain text corpora achieves slightly better performance than training on external text. When adding ImageNet videos, both F1 and METEOR increase suggesting that including outside image data is beneficial. Including ImageNet videos to learn better lexical classifiers especially improves the F1 score, which increases from 6.0 to 22.2. Figure 7 presents qualitative results of our best model on snippets with the held out objects in MSVD corpus and the ILSVRC validation set.

| Model (Video) | METEOR | F1 |
|---|---|---|
| Baseline (No Transfer) | 28.8 | 0.0 |
| + DT | 28.9 | 6.0 |
| + ILSVRC Videos (No Transfer) | 29.0 | 0.0 |
| + ILSVRC Videos + DT | 29.1 | 22.2 |

Table 4: Video Description: METEOR scores across the test dataset and average F1 scores for the four held-out categories (All values in %) using direct transfer (DT). The DCC models were trained on videos with 4 objects removed and the language model was trained on in-domain sentences.


Figure 7: Video Description: Captions generated by DCC on videos of novel objects unseen in paired training data.

## 6. Conclusion

We present the Deep Compositional Captioner (DCC) which can be used to describe new objects which are not present in current caption copora. Our quantitative and qualitative results demonstrate our model's ability to integrate new vocabulary into generated image and video descriptions by effectively using existing vision datasets and unpaired text data. By integrating data from disparate sources and transferring knowledge between semantically related concepts, DCC improves upon current deep caption models by providing rich descriptions which are not limited by the availability of paired image-sentence corpora.

## Acknowledgements

Lisa Anne Hendricks is supported by the NDSEG Fellowship. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD). 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 Vision and Learning Center. Raymond Mooney and Kate Saenko were supported in part by DARPA under AFRL grant FA8750-13-2-0026 and a Google Grant. Mooney was also supported by ONR ATL Grant N00014-11-1-010.

## References


[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[3] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[4] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. *ACL*, 2015.

[5] 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.

[6] 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.

[7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[8] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *ICCV*, 2013.

[9] A. Guptal, P. Srinivasan, J. Shi, and L. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] P. Hodosh, A. Young, M. Lai, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014.

[12] J. Hoffman, S. Guadarrama, E. Tzeng, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *NIPS*, 2014.

[13] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

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

[16] R. Kiros, R. Salakhuditnov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015.

[17] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, 2013.

[18] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. Babytalk: Understanding and generating simple image descriptions. *TPAMI*, 2013.

[19] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 2014.

[20] 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.

[21] 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.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[24] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[25] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

[26] A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. *GCPR*, 2015.

[27] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps Where - and Why? Semantic Relatedness for Knowledge Transfer. In *CVPR*, 2010.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2014.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[30] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*. 2013.

[31] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014.

[32] B. Thomee and A. Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 12, 2012.

[33] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*, 2003.

[34] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. *ICCV*, 2015.

[35] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL*, 2015.

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

[37] Q. Wu, C. Shen, A. v. d. Hengel, L. Liu, and A. Dick. Image captioning with an intermediate attributes layer. *arXiv preprint arXiv:1506.01144*, 2015.