Obj2Text: Generating Visually Descriptive Language from Object Layouts

Xuwang Yin **Vicente Ordonez**
Department of Computer Science, University of Virginia, Charlottesville, VA.
[xy4cm, vicente]@virginia.edu

Abstract

Generating captions for images is a task that has recently received considerable attention. In this work we focus on caption generation for abstract scenes, or object layouts where the only information provided is a set of objects and their locations. We propose OBJ2TEXT, a sequence-to-sequence model that encodes a set of objects and their locations as an input sequence using an LSTM network, and decodes this representation using an LSTM language model. We show that our model, despite encoding object layouts as a sequence, can represent spatial relationships between objects, and generate descriptions that are globally coherent and semantically relevant. We test our approach in a task of object-layout captioning by using only object annotations as inputs. We additionally show that our model, combined with a state-of-the-art object detector, improves an image captioning model from 0.863 to 0.950 (CIDEr score) in the test benchmark of the standard MS-COCO Captioning task.

1 Introduction

Natural Language generation (NLG) is a long standing goal in natural language processing. There have already been several successes in applications such as financial reporting (Kukich, 1983; Smadja and McKeown, 1990), or weather forecasts (Konstas and Lapata, 2012; Wen et al., 2015), however it is still a challenging task for less structured and open domains. Given recent progress in training robust visual recognition models using convolutional neural networks, the task of generating natural language descriptions for arbitrary images has received considerable attention (Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Mao et al., 2015). In general, generating visually descriptive language can be useful for various tasks such as human-machine communication, accessibility, image retrieval, and search. However this task is still challenging and it depends on developing both a robust visual recognition model, and a reliable language generation model. In this paper, we instead tackle a task of describing object layouts where the categories for the objects in an input scene and their corresponding locations are known. Object layouts are commonly used for story-boarding, sketching, and computer graphics applications. Additionally, using our object layout captioning model on the outputs of an object detector we are also able to improve image caption-

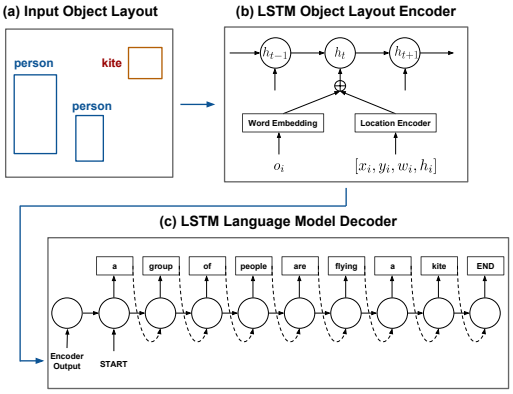


Figure 1: Overview of our proposed model for generating visually descriptive language from object layouts. The input (a) is an object layout that consists of object categories and their corresponding bounding boxes, the encoder (b) uses a two-stream recurrent neural network to encode the input object layout, and the decoder (c) uses a standard LSTM recurrent neural network to generate text.

ing models. Object layouts contain rich semantic information, however they also abstract away several other visual cues such as color, texture, and appearance, thus introducing a different set of challenges than those found in traditional image captioning.

We propose OBJ2TEXT, a sequence-to-sequence model that encodes object layouts using an LSTM network (Hochreiter and Schmidhuber, 1997), and decodes natural language descriptions using an LSTM-based neural language model[1]. Natural language generation systems usually consist of two steps: content planning, and surface realization. The first step decides on the content to be included in the generated text, and the second step connects the concepts using structural language properties. In our proposed model, OBJ2TEXT, content planning is performed by the encoder, and surface realization is performed by the decoder. Our model is trained in the standard MS-COCO dataset (Lin et al., 2014), which includes both object annotations for the task of object detection, and textual descriptions for the task of image captioning. While most previous research has been devoted to any one of these two tasks, our paper presents, to our knowledge, the first approach for learning mappings between object annotations and textual descriptions. Using several lesioned versions of the proposed model we explored the effect of object counts and locations in the quality and accuracy of the generated natural language descriptions.

Generating visually descriptive language requires beyond syntax, and semantics; an understanding of the physical word. We also take inspiration from recent work by Schmaltz et al. (2016) where the goal was to reconstruct a sentence from a bag-of-words (BOW) representation using a simple surface-level language model based on an encoder-decoder sequence-to-sequence architecture. In contrast to this previous approach, our model is grounded on visual data, and its corresponding spatial information, so it goes beyond word re-ordering. Also relevant to our work is Yao et al. (2016a) which previously explored the task of oracle image captioning by providing a language generation model with a list of manually defined visual concepts known to be present in the image. In addition, our model is able to leverage both quantity and spatial information as additional cues associated with each object/concept, thus allowing it to learn about verbosity, and spatial relations in a supervised fashion.

In summary, our contributions are as follows:

- We demonstrate that despite encoding object layouts as a sequence using an LSTM, our model can still effectively capture spatial information for the captioning task. We perform ablation studies to measure the individual impact of object counts, and locations.
- We show that a model relying only on object annotations as opposed to pixel data, performs competitively in image captioning despite the ambiguity of the setup for this task.
- We show that more accurate and comprehensive descriptions can be generated on the image captioning task by combining our OBJ2TEXT model using the outputs of a state-of-the-art object detector with a standard image captioning approach.

2 Task

We evaluate OBJ2TEXT in the task of object layout captioning, and image captioning. In the first task, the input is an object layout that takes the form of a set of object categories and bounding box pairs $\langle \boldsymbol{o}, \boldsymbol{l} \rangle = \{\langle o_i, l_i \rangle\}$, and the output is natural language. This task resembles the second task of image captioning except that the input is an object layout instead of a standard raster image represented as a pixel array. We experiment in the MS-COCO dataset for both tasks. For the first task, object layouts are derived from ground-truth bounding box annotations, and in the second task object layouts are obtained using the outputs of an object detector over the input image.

3 Related Work

Our work is related to previous works that used clipart scenes for visually-grounded tasks including sentence interpretation (Zitnick and Parikh, 2013; Zitnick et al., 2013), and predicting object dynamics (Fouhey and Zitnick, 2014). The cited advantage of abstract scene representations such as the ones provided by the clipart scenes dataset proposed in (Zitnick and Parikh, 2013) is their ability to separate the complexity of pattern recognition from semantic visual representation. Abstract scene representations also maintain

[1] We build on neuraltalk2 and make our Torch code, and an interactive demo of our model available in the following url: http://vision.cs.virginia.edu/obj2text

common-sense knowledge about the world. The works of Vedantam et al. (2015b); Eysenbach et al. (2016) proposed methods to learn common-sense knowledge from clipart scenes, while the method of Yatskar et al. (2016), similar to our work, leverages object annotations for natural images. Understanding abstract scenes has demonstrated to be a useful capability for both language and vision tasks and our work is another step in this direction.

Our work is also related to other language generation tasks such as image and video captioning (Farhadi et al., 2010; Ordonez et al., 2011; Mason and Charniak, 2014; Ordonez et al., 2015; Xu et al., 2015; Donahue et al., 2015; Mao et al., 2015; Fang et al., 2015). This problem is interesting because it combines two challenging but perhaps complementary tasks: visual recognition, and generating coherent language. Fueled by recent advances in training deep neural networks (Krizhevsky et al., 2012) and the availability of large annotated datasets with images and captions such as the MS-COCO dataset (Lin et al., 2014), recent methods on this task perform end-to-end learning from pixels to text. Most recent approaches use a variation of an encoder-decoder model where a convolutional neural network (CNN) extracts visual features from the input image (encoder), and passes its outputs to a recurrent neural network (RNN) that generates a caption as a sequence of words (decoder) (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). However, the MS-COCO dataset, containing object annotations, is also a popular benchmark in computer vision for the task of object detection, where the objective is to go from pixels to a collection of object locations. In this paper, we instead frame our problem as going from a collection of object categories and locations (object layouts) to image captions. This requires proposing a novel encoding approach to encode these object layouts instead of pixels, and allows for analyzing the image captioning task from a different perspective. Several other recent works use a similar sequence-to-sequence approach to generate text from source code input (Iyer et al., 2016), or to translate text from one language to another (Bahdanau et al., 2015).

There have also been a few previous works explicitly analyzing the role of spatial and geometric relations between objects for vision and language related tasks. The work of Elliott and Keller (2013) manually defined a dictionary of object-object relations based on geometric cues. The work of Ramisa et al. (2015) is focused on predicting preposition given two entities and their locations in an image. Previous works of Plummer et al. (2015) and Rohrbach et al. (2016) showed that switching from classification-based CNN network to detection-based Fast RCNN network improves performance for phrase localization. The work of Hu et al. (2016) showed that encoding image regions with spatial information is crucial for natural language object retrieval as the task explicitly asks for locations of target objects. Unlike these previous efforts, our model is trained end-to-end for the language generation task, and takes as input a holistic view of the scene layout, potentially learning higher order relations between objects.

4 Model

In this section we describe our base OBJ2TEXT model for encoding object layouts to produce text (section 4.1), as well as two further variations to use our model to generate captions for real images: OBJ2TEXT-YOLO which uses the YOLO object detector (Redmon and Farhadi, 2017) to generate layouts of object locations from real images (section 4.2), and OBJ2TEXT-YOLO + CNN-RNN which further combines the previous model with an encoder-decoder image captioning which uses a convolutional neural network to encode the image (section 4.3).

4.1 OBJ2TEXT

OBJ2TEXT is a sequence-to-sequence model that encodes an input object layout as a sequence, and decodes a textual description by predicting the next word at each time step. Given a training data set comprising N observations $\left\{\langle \boldsymbol{o}^{(n)}, \boldsymbol{l}^{(n)} \rangle\right\}$, where $\langle \boldsymbol{o}^{(n)}, \boldsymbol{l}^{(n)} \rangle$ is a pair of sequences of input category and location vectors, together with a corresponding set of target captions $\left\{\boldsymbol{s}^{(n)}\right\}$, the encoder and decoder are trained jointly by minimizing a loss function over the training set using stochastic gradient descent:

$$W^* = \arg\min_W \sum_{n=1}^{N} \mathcal{L}(\langle \boldsymbol{o}^{(n)}, \boldsymbol{l}^{(n)} \rangle, \boldsymbol{s}^{(n)}), \quad (1)$$

in which $W = \binom{W_1}{W_2}$ is the group of encoder parameters W_1 and decoder parameters W_2. The loss

function is a negative log likelihood function of the generated description given the encoded object layout

$$\mathcal{L}(\langle \boldsymbol{o}^{(n)}, \boldsymbol{l}^{(n)} \rangle, \boldsymbol{s}^{(n)}) = -\log p(\boldsymbol{s}^n | h_L^n, W_2), \quad (2)$$

where h_L^n is computed using the LSTM-based encoder (eqs. 3, and 4) from the object layout inputs $\langle \boldsymbol{o}^{(n)}, \boldsymbol{l}^{(n)} \rangle$, and $p(\boldsymbol{s}^n | h_L^n, W_2)$ is computed using the LSTM-based decoder (eqs. 5, 6 and 7).

At inference time we encode an input layout $\langle \boldsymbol{o}, \boldsymbol{l} \rangle$ into its representation h_L, and sample a sentence word by word based on $p(s_t | h_L, \boldsymbol{s}_{<t})$ as computed by the decoder in time-step t. Finding the optimal sentence $\boldsymbol{s}^* = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s} | h_L)$ requires the evaluation of an exponential number of sentences as in each time-step we have K number of choices for a word vocabulary of size K. As a common practice for an approximate solution, we follow (Vinyals et al., 2015) and use beam search to limit the choices for words at each time-step by only using the ones with the highest probabilities.

Encoder: The encoder at each time-step t takes as input a pair $\langle o_t, l_t \rangle$, where o_t is the object category encoded as a one-hot vector of size V, and $l_t = [B_t^x, B_t^y, B_t^w, B_t^h]$ is the location configuration vector that contains left-most position, top-most position, and the width and height of the bounding box corresponding to object o_t, all normalized in the range [0,1] with respect to input image dimensions. o_t and l_t are mapped to vectors with the same size k and added to form the input x_t to one time-step of the LSTM-based encoder as follows:

$$x_t = W_o o_t + (W_l l_t + b_l), \quad x_t \in \mathbb{R}^k, \quad (3)$$

in which $W_o \in \mathbb{R}^{k \times V}$ is a categorical embedding matrix (the word encoder), and $W_l \in \mathbb{R}^{k \times 4}$ and bias $b_l \in \mathbb{R}^k$ are parameters of a linear transformation unit (the object location encoder).

Setting initial value of cell state vector $c_0^e = 0$ and hidden state vector $h_0^e = 0$, the LSTM-based encoder takes the sequence of input $(x_1, ..., x_{T_1})$ and generates a sequence of hidden state vectors $(h_1^e, ..., h_{T_1}^e)$ using the following step function (we omit cell state variables and internal transition gates for simplicity as we use a standard LSTM cell definition):

$$h_t^e = \text{LSTM}(h_{t-1}^e, x_t; W_1). \quad (4)$$

We use the last hidden state vector $h_L = h_{T_1}^e$ as the encoded representation of the input layout $\langle \boldsymbol{o}_t, \boldsymbol{l}_t \rangle$ to generate the corresponding description $\boldsymbol{s}$.

Decoder: The decoder takes the encoded layout h_L as input and generates a sequence of multinomial distributions over a vocabulary of words using an LSTM neural language model. The joint probability distribution of generated sentence $\boldsymbol{s} = (s_1, ..., s_{T_2})$ is factorized into products of conditional probabilities:

$$p(\boldsymbol{s} | h_L) = \prod_{t=1}^{T_2} p(s_t | h_L, \boldsymbol{s}_{<t}), \quad (5)$$

where each factor is computed using a softmax function over the hidden states of the decoder LSTM as follows:

$$p(s_t | h_L, \boldsymbol{s}_{<t}) = \text{softmax}(W_h h_{t-1}^d + b_h), \quad (6)$$

$$h_t^d = \text{LSTM}(h_{t-1}^d, W_s s_t; W_2), \quad (7)$$

where W_s is the categorical embedding matrix for the one-hot encoded caption sequence of symbols. By setting $h_{-1}^d = 0$ and $c_{-1}^d = 0$ for the initial hidden state and cell state, the layout representation is encoded into the decoder network at the 0 time step as a regular input:

$$h_0^d = \text{LSTM}(h_{-1}^d, h_L; W_2). \quad (8)$$

We use beam search to sample from the LSTM as is routinely performed in previous literature in order to generate text.

4.2 OBJ2TEXT-YOLO

For the task of image captioning we propose OBJ2TEXT-YOLO. This model takes an image as input, extracts an object layout (object categories and locations) with a state-of-the-art object detection model YOLO (Redmon and Farhadi, 2017), and uses OBJ2TEXT as described in section 4.1 to generate a natural language description of the input layout and hence, the input image. The model is trained using the standard back-propagation algorithm, but the error is not back-propagated to the object detection module.

4.3 OBJ2TEXT-YOLO + CNN-RNN

For the image captioning task we experiment with a combined model (see Figure 2) where we take an image as input, and then use two separate computation branches to extract visual feature information and object layout information. These two streams of information are then passed to an LSTM neural language model to generate a description. Visual features are extracted using the

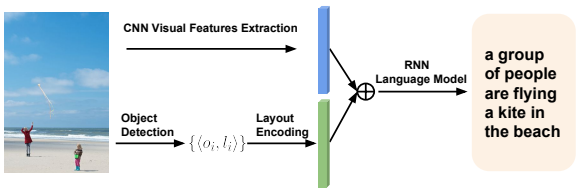


Figure 2: Image Captioning by joint learning of visual features and object layout encoding.

VGG-16 (Simonyan and Zisserman, 2015) convolutional neural network pre-trained on the ImageNet classification task (Russakovsky et al., 2015). Object layouts are extracted using the YOLO object detection system and its output object locations are encoded using our proposed OBJ2TEXT encoder. These two streams of information are encoded into vectors of the same size and their sum is input to the language model to generate a textual description. The model is trained using the standard back-propagation algorithm where the error is back-propagated to both branches but not the object detection module. The weights of the image CNN model are fine-tuned only after the layout encoding branch is well trained but no significant overall performance improvements were observed.

5 Experimental Setup

We evaluate the proposed models on the MS-COCO (Lin et al., 2014) dataset which is a popular image captioning benchmark that also contains object extent annotations. In the object layout captioning task the model uses the ground-truth object extents as input object layouts, while in the image captioning task the model takes raw images as input. The qualities of generated descriptions are evaluated using both human evaluations and automatic metrics. We train and validate our models based on the commonly adopted split regime (113,287 training images, 5000 validation and 5000 test images) used in (Karpathy et al., 2016), and also test our model in the MS-COCO official test benchmark.

We implement our models based on the open source image captioning system Neuraltalk2 (Karpathy et al., 2016). Other configurations including data preprocessing and training hyper-parameters also follow Neuraltalk2. We trained our models using a GTX1080 GPU with 8GB of memory for 400k iterations using a batch size of 16 and an Adam optimizer with alpha of 0.8, beta of 0.999 and epsilon of 1e-08. Descriptions of the CNN-RNN approach are generated using the publicly available code and model checkpoint provided by Neuraltalk2 (Karpathy et al., 2016). Captions for online test set evaluations are generated using beam search of size 2, but score histories on split validation set are based on captions generated without beam search (i.e. max sampling at each time-step).

Ablation on Object Locations and Counts:. We setup an experiment where we remove the input locations from the OBJ2TEXT encoder to study the effects on the generated captions, and confirm whether the model is actually using spatial information during surface realization. In this restricted version of our model the LSTM encoder at each time step only takes the object category embedding vector as input. The OBJ2TEXT model additionally encodes different instances of the same object category in different time steps, potentially encoding in some of its hidden states information about how many objects of a particular class are in the image. For example, in the object annotation presented in the input in Figure 1, there are two instances of "person". We perform an additional experiment where our model does not have access neither to object locations, nor the number of object instances by providing only a set of object categories. Note that in this set of experiments the object layouts are given as inputs, thus we assume full access to ground-truth object annotations, even in the test split. In the experimental results section we use the "-GT" postfix to indicate that input object layouts are obtained from ground-truth object annotations provided by the MS-COCO dataset.

Image Captioning Experiment: In this experiment we assess whether the image captioning model OBJ2TEXT-YOLO that only relies on object categories and locations could give comparable performance with a CNN-RNN model based on Neuraltalk2 (Karpathy et al., 2016) that has full access to visual image features. We also explore how much does a combined OBJ2TEXT-YOLO + CNN-RNN model could improve over a CNN-RNN model by fusing object counts and location information that is not explicitly encoded in a traditional CNN-RNN approach.

Human Evaluation Protocol. We use a two-alternative forced-choice evaluation (2AFC) ap-

proach to compare two methods that generate captions. For this, we setup a task on Amazon Mechanical Turk where users are presented with an image and two alternative captions, and they have to choose the caption that best describes the image. Users are not prompted to use any single criteria but rather a holistic assessment of the captions, including their semantics, syntax, and the degree to which they describe the image content. In our experiment we randomly sample 500 captions generated by various models for MS COCO online test set images, and use three users per image to obtain annotations. Note that three users choosing randomly between two options have a chance of 25% to select the same caption for a given image. In our experiments comparing method A vs method B, we report the percentage of times A was picked over B (Choice-all), the percentage of times all users selected the same method, either A or B, (Agreement), and the percentage of times A was picked over B only for these cases where all users agreed (Choice-agreement).

6 Results

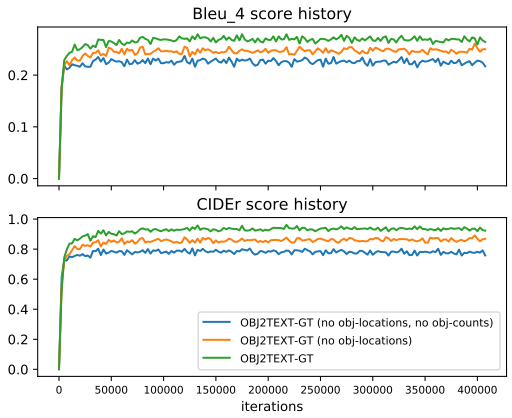
Impact of Object Locations and Counts: Figure 3a shows the CIDEr (Vedantam et al., 2015a), and BLEU-4 (Papineni et al., 2002) score history on our validation set during 400k iterations of training of OBJ2TEXT, as well as a version of our model that does not use object locations, and a version of our model that does not use neither object locations nor object counts. These results show that our model is effectively using both object locations and counts to generate better captions, and absence of any one of these two cues affects performance. Table 1 confirms these results on the test split after a full round of training.

Furthermore, human evaluation results in the first row of Table 2 show that the OBJ2TEXT model with access to object locations is preferred by users, especially in cases where all evaluators agreed on their choice (62% over the baseline that does not have access to locations). In Figure 4 we additionally present qualitative examples showing predictions side-by-side between OBJ2TEXT-GT and OBJ2TEXT-GT (no obj-locations). These results indicate that 1) perhaps not surprisingly, object counts is useful for generating better quality descriptions, and 2) object location information when properly encoded, is an important cue for generating more accurate descriptions. We additionally implemented a nearest neighbor baseline by representing the objects in the input layout using an orderless bag-of-words representation of object counts and the CIDEr score on the test split was only 0.387.

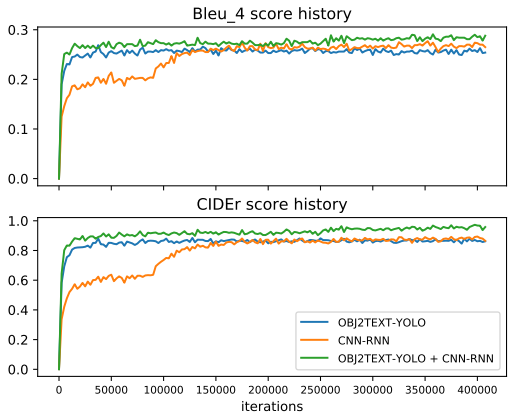
On top of OBJ2TEXT we additionally experimented with the global attention model proposed in (Luong et al., 2015) so that a weighted combination of the encoder hidden states are forwarded to the decoding neural language model, however we did not notice any overall gains in terms of accuracy from this formulation. We observed that this model provided gains only for larger input sequences where it is more likely that the LSTM network forgets its past history (Bahdanau et al., 2015). However in MS-COCO the average number of objects in each image is rather modest, so the last hidden state can capture well the overall nuances of the visual input.

Object Layout Encoding for Image Captioning: Figure 3b shows the CIDEr, and BLEU-4 score history on the validation set during 400k iterations of training of OBJ2TEXT-YOLO, CNN-RNN, and their combination. These results show that OBJ2TEXT-YOLO performs surprisingly close to CNN-RNN, and the model resulting from combining the two, clearly outperforms each method alone. Table 3 shows MS-COCO evaluation results on the test set using their online benchmark service, and confirms results obtained in the validation split, where CNN-RNN seems to have only a slight edge over OBJ2TEXT-YOLO which lacks access to pixel data after the object detection stage. Human evaluation results in Table 2 rows 2, and 3, further confirm these findings. These results show that meaningful descriptions could be generated solely based on object categories and locations information, even without access to color and texture input.

The combined model performs better than the two models, improving the CIDEr score of the basic CNN-RNN model from 0.863 to 0.950, and human evaluation results show that the combined model is preferred over the basic CNN-RNN model for 65.3% of the images for which all evaluators were in agreement about the selected method. These results show that explicitly encoded object counts and location information, which is often overlooked in traditional image captioning approaches, could boost the performance of existing models. Intuitively, object lay-



(a) Score histories of lesioned versions of the proposed model for the task of object layout captioning.



(b) Score histories of image captioning models. Performance boosts of CNN-RNN and combined model around iteration 100K and 250K are due to fine-tuning of the image CNN model.

Figure 3: Score histories of various models on the MS COCO split validation set.

Method	Bleu_4	CIDEr	METEOR	ROUGE-L
OBJ2TEXT-GT (no obj-locations, counts)	0.21	0.759	0.215	0.464
OBJ2TEXT-GT (no obj-locations)	0.233	0.837	0.222	0.482
OBJ2TEXT-GT	**0.253**	**0.922**	**0.238**	**0.507**

Table 1: Performance of lesioned versions of the proposed model on the MS COCO split test set.

out and visual features are complementary: neural network models for visual feature extraction are trained on a classification task where object-level information such as number of instances and locations are ignored in the objective. Object layouts on the other hand, contain categories and their bounding-boxes but don't have access to rich image features such as image background, object attributes and objects with categories not present in the object detection vocabulary.

Figure 5 provides a three-way comparison of captions generated by the three image captioning models, with preferred captions by human evaluators annotated in bold text. Analysis on actual outputs gives us insights into the benefits of combing object layout information and visual features obtained using a CNN. Our OBJ2TEXT-YOLO model makes many mistakes because of lack of image context information since it only has access to object layout, while CNN-RNN makes many mistakes because the visual recognition model is imperfect at predicting the correct content. The combined model is usually able to generate more accurate and comprehensive descriptions.

In this work we only explored encoding spatial information with object labels, but object labels could be readily augmented with rich semantic features that are more detailed descriptions of objects or image regions. For example, the work of You et al. (2016) and Yao et al. (2016b) showed that visual features trained with semantic concepts (text entities mentioned in captions) instead of object labels is useful for image captioning, although they didn't consider encoding semantic concepts with spatial information. In case of object annotations the MS-COCO dataset only provides object labels and bounding-boxes, but there are other datasets such as Flick30K Entities (Plummer et al., 2015), and the Visual Genome dataset (Krishna et al., 2017) that provide richer region-to-phrase correspondence annotations. In addition, the fusion of object counts and spatial information with CNN visual features could in principle benefit other vision and language tasks such as visual question answering. We leave these possible extensions as future work.

7 Conclusion

We introduced OBJ2TEXT, a sequence-to-sequence model to generate visual descriptions for object layouts where only categories and locations are specified. Our proposed model

Alternatives	Choice-all	Choice-agreement	Agreement
OBJ2TEXT-GT vs. OBJ2TEXT-GT (no obj-locations)	54.1%	62.1%	40.6%
OBJ2TEXT-YOLO vs. CNN+RNN	45.6%	40.6%	54.7%
OBJ2TEXT-YOLO + CNN-RNN vs. CNN-RNN	58.1%	65.3%	49.5%
OBJ2TEXT-GT vs. HUMAN	23.6%	9.9%	58.8%

Table 2: Human evaluation results using two-alternative forced choice evaluation. Choice-all is percentage the first alternative was chosen. Choice-agreement is percentage the first alternative was chosen only when all annotators agreed. Agreement is percentage where all annotators agreed (random is 25%).

MS COCO Test Set Performance	CIDEr	ROUGE-L	METEOR	B-4	B-3	B-2	B-1
5-Refs							
OBJ2TEXT-YOLO	0.830	0.497	0.228	0.262	0.361	0.500	0.681
CNN-RNN	0.857	0.514	0.237	0.283	0.387	0.529	0.705
OBJ2TEXT-YOLO + CNN-RNN	**0.932**	**0.528**	**0.250**	**0.300**	**0.404**	**0.546**	**0.719**
40-Refs							
OBJ2TEXT-YOLO	0.853	0.636	0.305	0.508	0.624	0.746	0.858
CNN-RNN	0.863	0.654	0.318	0.540	0.656	0.775	0.877
OBJ2TEXT-YOLO + CNN-RNN	**0.950**	**0.671**	**0.334**	**0.569**	**0.686**	**0.802**	**0.896**

Table 3: The 5-Refs and 40-Refs performances of OBJ2TEXT-YOLO, CNN-RNN and the combined approach on the MS COCO online test set. The 5-Refs performance is measured using 5 ground-truth reference captions, while 40-Refs performance is measured using 40 ground-truth reference captions.

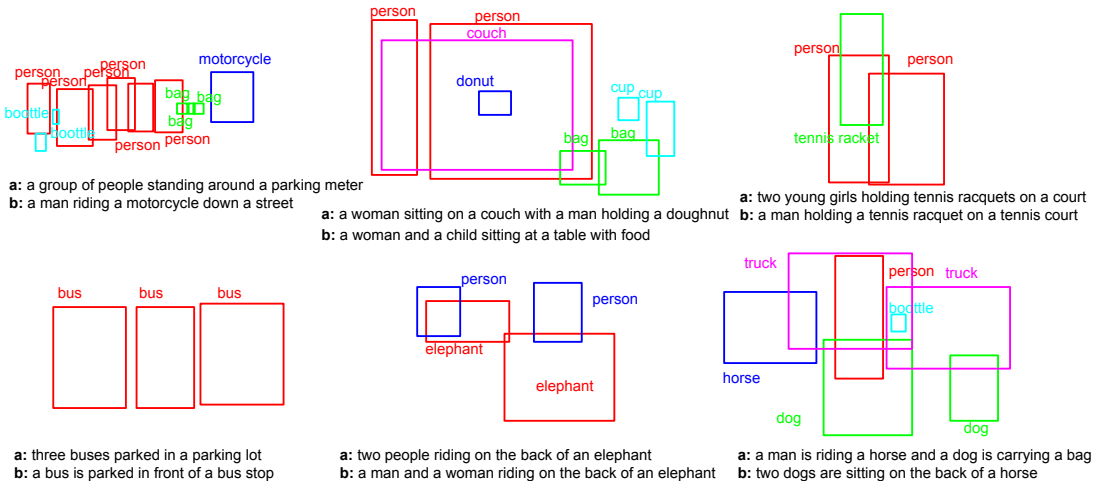
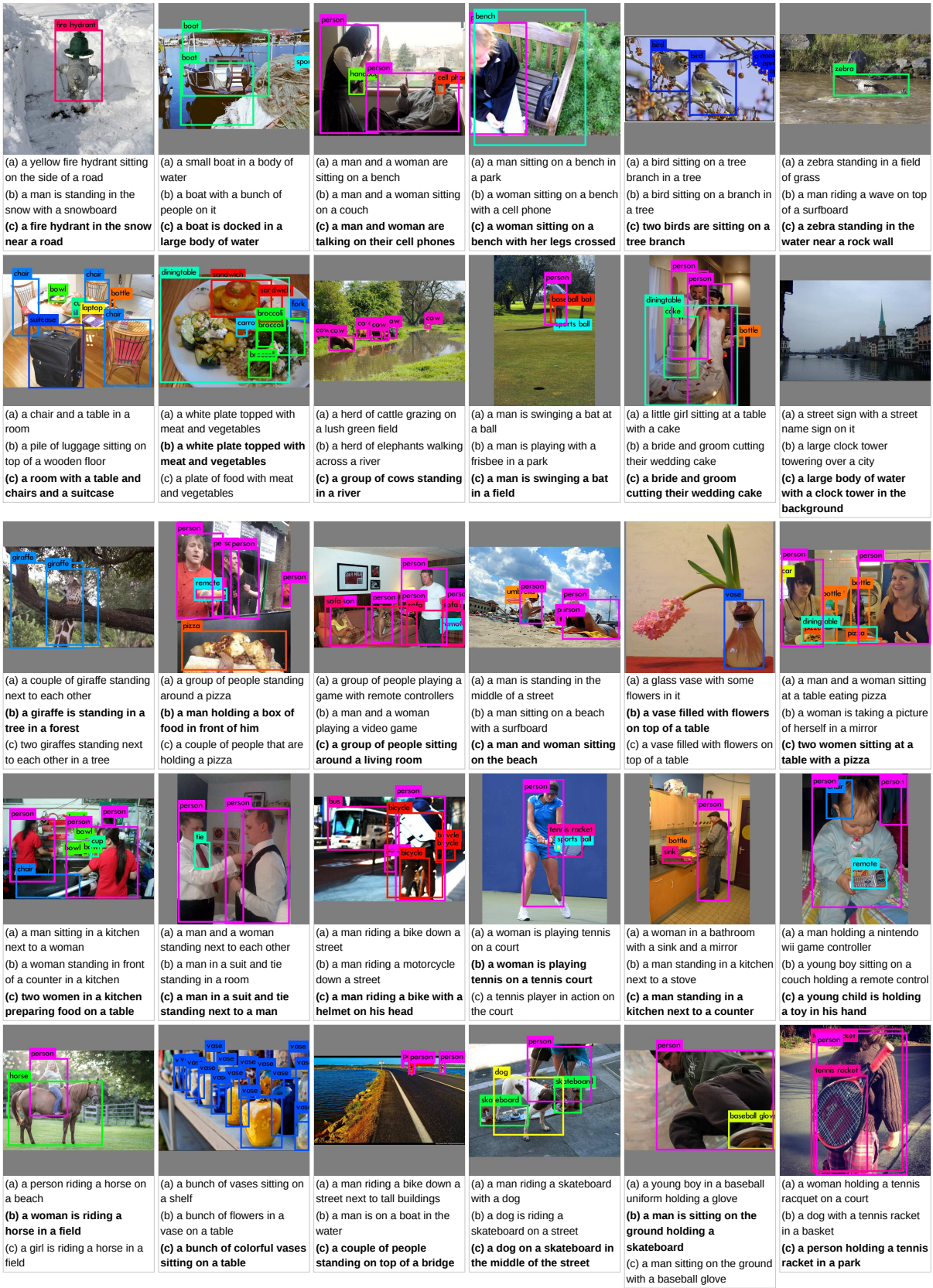


Figure 4: Qualitative examples comparing generated captions of (**a**) OBJ2TEXT-GT, and (**b**) OBJ2TEXT-GT (no obj-locations).

shows that an orderless visual input representation of concepts is not enough to produce good descriptions, but object extents, locations, and object counts, all contribute to generate more accurate image descriptions. Crucially we show that our encoding mechanism is able to capture useful spatial information using an LSTM network to produce image descriptions, even when the input is provided as a sequence rather than as an explicit 2D representation of objects. Additionally, using our proposed OBJ2TEXT model in combination with an existing image captioning model and a robust object detector we showed improved results in the task of image captioning.

Acknowledgments

This work was supported in part by an NVIDIA Hardware Grant. We are also thankful for the feedback from Mark Yatskar and anonymous reviewers of this paper.



(a) a yellow fire hydrant sitting on the side of a road
(b) a man is standing in the snow with a snowboard
(c) a fire hydrant in the snow near a road

(a) a small boat in a body of water
(b) a boat with a bunch of people on it
(c) a boat is docked in a large body of water

(a) a man and a woman are sitting on a bench
(b) a man and a woman sitting on a couch
(c) a man and woman are talking on their cell phones

(a) a man sitting on a bench in a park
(b) a woman sitting on a bench with a cell phone
(c) a woman sitting on a bench with her legs crossed

(a) a bird sitting on a tree branch in a tree
(b) a bird sitting on a branch in a tree
(c) two birds are sitting on a tree branch

(a) a zebra standing in a field of grass
(b) a man riding a wave on top of a surfboard
(c) a zebra standing in the water near a rock wall

(a) a chair and a table in a room
(b) a pile of luggage sitting on top of a wooden floor
(c) a room with a table and chairs and a suitcase

(a) a white plate topped with meat and vegetables
(b) a white plate topped with meat and vegetables
(c) a plate of food with meat and vegetables

(a) a herd of cattle grazing on a lush green field
(b) a herd of elephants walking across a river
(c) a group of cows standing in a river

(a) a man is swinging a bat at a ball
(b) a man is playing with a frisbee in a park
(c) a man is swinging a bat in a field

(a) a little girl sitting at a table with a cake
(b) a bride and groom cutting their wedding cake
(c) a bride and groom cutting their wedding cake

(a) a street sign with a street name sign on it
(b) a large clock tower towering over a city
(c) a large body of water with a clock tower in the background

(a) a couple of giraffe standing next to each other
(b) a giraffe is standing in a tree in a forest
(c) two giraffes standing next to each other in a tree

(a) a group of people standing around a pizza
(b) a man holding a box of food in front of him
(c) a couple of people that are holding a pizza

(a) a group of people playing a game with remote controllers
(b) a man and a woman playing a video game
(c) a group of people sitting around a living room

(a) a man is standing in the middle of a street
(b) a man sitting on a beach with a surfboard
(c) a man and woman sitting on the beach

(a) a glass vase with some flowers in it
(b) a vase filled with flowers on top of a table
(c) a vase filled with flowers on top of a table

(a) a man and a woman sitting at a table eating pizza
(b) a woman is taking a picture of herself in a mirror
(c) two women sitting at a table with a pizza

(a) a man sitting in a kitchen next to a woman
(b) a woman standing in front of a counter in a kitchen
(c) two women in a kitchen preparing food on a table

(a) a man and a woman standing next to each other
(b) a man in a suit and tie standing in a room
(c) a man in a suit and tie standing next to a man

(a) a man riding a bike down a street
(b) a man riding a motorcycle down a street
(c) a man riding a bike with a helmet on his head

(a) a woman is playing tennis on a court
(b) a woman is playing tennis on a tennis court
(c) a tennis player in action on the court

(a) a woman in a bathroom with a sink and a mirror
(b) a man standing in a kitchen next to a stove
(c) a man standing in a kitchen next to a counter

(a) a man holding a nintendo wii game controller
(b) a young boy sitting on a couch holding a remote control
(c) a young child is holding a toy in his hand

(a) a person riding a horse on a beach
(b) a woman is riding a horse in a field
(c) a girl is riding a horse in a field

(a) a bunch of vases sitting on a shelf
(b) a bunch of flowers in a vase on a table
(c) a bunch of colorful vases sitting on a table

(a) a man riding a bike down a street next to tall buildings
(b) a man is on a boat in the water
(c) a couple of people standing on top of a bridge

(a) a man riding a skateboard with a dog
(b) a dog is riding a skateboard on a street
(c) a dog on a skateboard in the middle of the street

(a) a young boy in a baseball uniform holding a glove
(b) a man is sitting on the ground holding a skateboard
(c) a man sitting on the ground with a baseball glove

(a) a woman holding a tennis racquet on a court
(b) a dog with a tennis racket in a basket
(c) a person holding a tennis racket in a park

Figure 5: Qualitative examples comparing the generated captions of (a) OBJ2TEXT-YOLO, (b) CNN-RNN and (c) OBJ2TEXT-YOLO + CNN-RNN. Images are selected from the 500 human evaluation images and annotated with YOLO object detection results. Captions preferred by human evaluators with agreement are highlighted in bold text.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Jeffrey 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 *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2625–2634.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *EMNLP*, volume 13, pages 1292–1302.

Benjamin Eysenbach, Carl Vondrick, and Antonio Torralba. 2016. Who is mistaken? *arXiv preprint arXiv:1612.01175*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

David F Fouhey and C Lawrence Zitnick. 2014. Predicting object dynamics in scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2026.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2016. Summarizing source code using a neural attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2073–2083, Berlin, Germany. Association for Computational Linguistics.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Andrej Karpathy et al. 2016. Neuraltalk2. https://github.com/karpathy/neuraltalk2/.

Ioannis Konstas and Mirella Lapata. 2012. Unsupervised concept-to-text generation with hypergraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 752–761.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*, pages 1097–1105.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 145–150. Association for Computational Linguistics.

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*, pages 740–755. Springer.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

Rebecca Mason and Eugene Charniak. 2014. Nonparametric method for data-driven image captioning. In *ACL (2)*, pages 592–598.

Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, III Daume, Hal, Alexander C. Berg, Yejin Choi, and Tamara L. Berg. 2015. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, pages 1–14.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of*

the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Arnau Ramisa, JK Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220. Association for Computational Linguistics.

Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Computer Vision and Pattern Recognition (CVPR)*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Allen Schmaltz, Alexander M. Rush, and Stuart Shieber. 2016. Word ordering without syntax. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2319–2324, Austin, Texas.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Frank A Smadja and Kathleen R McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Annual meeting of the Association for Computational Linguistics (ACL)*, pages 252–259.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015a. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. 2015b. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2542–2550.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1711–1721, Lisbon, Portugal.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Li Yao, Nicolas Ballas, Kyunghyun Cho, John R. Smith, and Yoshua Bengio. 2016a. Oracle performance for visual captioning. In *British Machine Vision Conference (BMVC)*.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2016b. Boosting image captioning with attributes. *arXiv preprint arXiv:1611.01646*.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3016.

C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1681–1688.