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-
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. 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 world. 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 \( \langle o, l \rangle = \{\langle o_i, l_i \rangle\} \), 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

---

1We 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 \{ $\langle o^{(n)}, l^{(n)} \rangle$ \}, where $\langle o^{(n)}, l^{(n)} \rangle$ is a pair of sequences of input category and location vectors, together with a corresponding set of target captions \{ $s^{(n)}$ \}, the encoder and decoder are trained jointly by minimizing a loss function over the training set using stochastic gradient descent:

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

(1)

in which $W = (W_1, W_2)$ 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
\[
\mathcal{L}(\langle o^{(n)}, l^{(n)} \rangle, s^{(n)}) = -\log p(s^n | h_L^n, W_2),
\]
where \(h_L^n\) is computed using the LSTM-based encoder (eqs. 3, and 4) from the object layout inputs \(\langle o^{(n)}, l^{(n)} \rangle\), and \(p(s^n | h_L^n, W_2)\) is computed using the LSTM-based decoder (eqs. 5, 6 and 7).

At inference time we encode an input layout \(\langle o, l \rangle\) into its representation \(h_L\), and sample a sentence word by word based on \(p(s_t|h_L, s_{<t})\) as computed by the decoder in time-step \(t\). Finding the optimal sentence \(s^* = \text{arg max}_s p(s|h_L)\) 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 \(l_t = [B^b_t, B^p_t, B^v_t, B^h_t]\) 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:
\[
x_t = W_o o_t + (W_l l_t + b_l), \quad x_t \in \mathbb{R}^k,
\]
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 \(c^0 = 0\) and hidden state vector \(h^0 = 0\), the LSTM-based encoder takes the sequence of input \((x_1, \ldots, x_{T_1})\) and generates a sequence of hidden state vectors \((h^e_1, \ldots, h^e_{T_1})\) using the following step function (we omit cell state variables and internal transition gates for simplicity as we use a standard LSTM cell definition):
\[
h^e_t = \text{LSTM}(h^e_{t-1}, x_t; W_l).
\]

We use the last hidden state vector \(h^e_{T_1}\) as the encoded representation of the input layout \(\langle o_t, l_t \rangle\) to generate the corresponding description \(s\).

**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 \(s = (s_1, \ldots, s_{T_2})\) is factorized into products of conditional probabilities:
\[
p(s|h_L) = \prod_{t=1}^{T_2} p(s_t|h_L, s_{<t}),
\]
where each factor is computed using a softmax function over the hidden states of the decoder LSTM as follows:
\[
p(s_t|h_L, s_{<t}) = \text{softmax}(W_s h^d_{t-1} + b_h),
\]
\[
h^d_t = \text{LSTM}(h^d_{t-1}, W_s s_t; W_2),
\]
where \(W_s\) is the categorical embedding matrix for the one-hot encoded caption sequence of symbols. By setting \(h^d_{-1} = 0\) and \(c^d_{-1} = 0\) 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:
\[
h^d_0 = \text{LSTM}(h^d_{-1}, h_L; W_2).
\]

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
A group of people are flying a kite in the beach.

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-
approach 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-
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 combining 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
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%).

<table>
<thead>
<tr>
<th>Alternatives</th>
<th>Choice-all</th>
<th>Choice-agreement</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ2TEXT-GT vs. OBJ2TEXT-GT (no obj-locations)</td>
<td>54.1%</td>
<td>62.1%</td>
<td>40.6%</td>
</tr>
<tr>
<td>OBJ2TEXT-YOLO vs. CNN+RNN</td>
<td>45.6%</td>
<td>40.6%</td>
<td>54.7%</td>
</tr>
<tr>
<td>OBJ2TEXT-YOLO + CNN-RNN vs. CNN-RNN</td>
<td>58.1%</td>
<td>65.3%</td>
<td>49.5%</td>
</tr>
<tr>
<td>OBJ2TEXT-GT vs. HUMAN</td>
<td>23.6%</td>
<td>9.9%</td>
<td>58.8%</td>
</tr>
</tbody>
</table>

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.

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

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 a street in the middle of a road | (b) a dog with a tennis racket in a basket | (c) a young child is holding a tennis racket on a court |
| a small boat in a body of water | a man holding a box of food in front of him | a woman holding a tennis racket on top of a table |
| a beagle on a skateboard on a street | a group of people standing around a living room | a woman is playing with a Wii game controller |
| (c) a group of people playing a game with remote controllers | (b) a man and a woman sitting next to each other | a man in a kitchen near a counter |
| (b) a man and a woman playing a video game | (c) a couple of people that are holding a pizza | a man holding a Nintendo Wii game controller |
| (c) a group of people sitting around a living room | (a) a man is standing in the middle of a street | (b) a young boy sitting on a couch holding a remote control |
| (c) a group of people sitting on a beach with a surfboard | (b) a man is riding a wave on top of a surfboard | (c) a young child is holding a boy in his hand |
| (c) a group of people playing tennis on a tennis court | (b) a woman is sitting on a bench with her legs crossed | (b) a woman sitting on a bench in a kitchen with a clock tower in the background |
| (c) a man and woman sitting around a pizza | (b) a man is sitting on the ground holding a blanket | (c) a zebra standing in a field of grass |
| (b) a man holding a box of food in front of him | (c) a couple of people that are holding a pizza | (b) a man riding on the side of a road |
| (c) a couple of people that are holding a pizza | (c) a group of people sitting next to each other next to a trash can | (c) a zebra standing on the water near a rock wall |

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


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


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


