HAL
open science

Areas of Attention for Image Captioning

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, Jakob Verbeek

To cite this version:

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, Jakob Verbeek. Areas of Attention for Image Captioning. International Conference on Computer Vision (ICCV), Oct 2017, Venice, Italy. hal-01428963v1

HAL Id: hal-01428963

https://inria.hal.science/hal-01428963v1

Submitted on 6 Jan 2017 (v1), last revised 25 Aug 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Areas of Attention for Image Captioning

Marco Pedersoli Thomas Lucas Cordelia Schmid Jakob Verbeek
Inria, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France
firstname.lastame@inria.fr

Abstract

We propose "Areas of Attention", a novel attention-based model for automatic image caption generation. Our approach models the interplay between the state of the RNN, image region descriptors and word embedding vectors by three pairwise interactions. It allows association of caption words with local visual appearances rather than with descriptors of the entire scene. This enables better generalization to complex scenes not seen during training. Our model is agnostic to the type of attention areas, and we instantiate it using regions based on CNN activation grids, object proposals, and spatial transformer networks. Our results show that all components of our model contribute to obtain state-of-the-art performance on the MSCOCO dataset. In addition, our results indicate that attention areas are correctly associated to meaningful latent semantic structure in the generated captions.

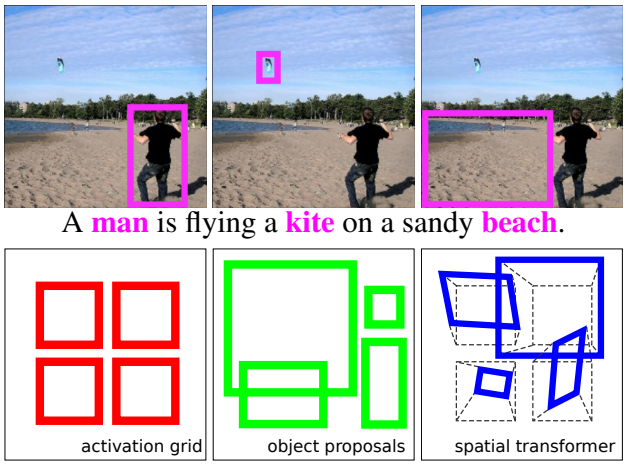


Figure 1. At each time-step, our model predicts the next caption word and the corresponding region based on the RNN state (top). We explore three different region types for our model (bottom).

1. Introduction

Image captioning, *i.e.* automatically generating natural language image descriptions, is useful for the visually impaired, and for natural language based image search. It is significantly more challenging than classic tasks such as object recognition and image classification for two reasons. First, the structured output space of well formed natural language sentences is significantly more challenging to predict over than just a set of class labels. Second, this complex output space allows to express a finer interpretation of the visual scene, and therefore also requires a more detailed visual analysis of the scene to do well at this task. Figure 1 gives an example of a typical image description that not only refers to objects in the scene, but also the scene type or location, object properties, and their interactions.

Neural encoder-decoder based approaches, similar to those used in recent machine translation models [33], have been found very effective, see *e.g.* [21, 25, 36]. These methods use a convolutional neural network (CNN) to encode the input image into a compact representation. A recurrent neural network (RNN) is used to decode this representation word-by-word to a natural language description of the image. While very effective, these models are limited in that the image analysis is (i) static, *i.e.* does not change over time as the description is produced, (ii) not spatially localized, *i.e.* describes the scene as a whole instead of local aspects relevant to parts of the description. Attention mechanisms can address these limitations by dynamically focusing on different parts of the input as the output sequence is generated. Such mechanisms are effective for a variety of sequential prediction tasks, including machine translation [1], speech recognition [5], image synthesis [13], and image captioning [37]. For some tasks the definition of parts of the input to attend to are clear and limited in number: for example the individual words in the source sentence for machine translation. For other tasks with complex inputs, such as image captioning, the notion of parts is less clear.

In this paper we present a novel attention-based image captioning model. The core of our approach is to model the interplay between the RNN state, image region descriptors, and word embedding vectors by means of three pair-

wise interactions. Our model is agnostic to the type areas of attention, and we explore the effectiveness of three types, illustrated in Figure 1. First, following [37], we use the positions in the activation grid of a CNN layer. Second, we use a set of fixed, but image specific, edge-box object proposals [42]. Third, we integrate a localization sub-network based on spatial transformers [16] in our model, that regresses the attention areas from the image content. We experimentally evaluate the relative contributions of different components of our model, together with the three types of attention regions. The results show that all components of our model contribute to improve the quality of the generated captions, which are state-of-the-art on the MSCOCO dataset. In addition, our results indicate that the attention areas are correctly associated to meaningful latent semantic structure in the generated captions.

2. Related work

Image captioning with encoder-decoder models has recently been extensively studied, see *e.g*. [2, 10, 19, 21, 25, 28, 36, 37, 38]. In its basic form a CNN processes the input image to encode it into a vectorial image representation, which is used as the initial input for an RNN. The RNN decodes its input by sequentially predicting the next word in the caption given the previous ones, without the need to restrict the temporal dependence to a fixed order as in approaches based on n-grams. The CNN image representation can be entered into the RNN in different manners. While some authors [19, 36] use it only to compute the initial state of the RNN, others enter it in each RNN iteration [10, 25].

Xu *et al*. [37] were the first to propose an attention-based approach for image captioning, in which the RNN state update includes the visual representation of an image region. Which image region is attended to is determined based on the previous state of the RNN. They propose a "soft" variant in which a convex combination of different region descriptors is used, and a "hard" variant in which a single region is selected. The latter is found to perform slightly better, but is more complex to train due to a non-differentiable sampling operator in the state update. They use the positions in the activation grid of a convolutional CNN layer as the loci of attention. Each position is described with the corresponding activation column across the layer's channels.

Several works build upon this seminal work. You *et al*. [41] learn a set of attribute detectors, similar to [11], for each word of their vocabulary. These detectors are applied to an image, and the strongest object detections are used as regions for an attention mechanism similar to that of [37]. In their work the detectors are learned prior and independently from the language model. Jin *et al*. [17] adapted the attention model of Xu *et al*. by using selective search object proposals [34] as the regions of attention. They resize the regions to a fixed size and use the VGG16 [32] penultimate layer to characterize them. Yang *et al*. [38] improved the attention based encoder-decoder model by adding a reviewer module that improves the representation passed to the decoder. They show improved results for various tasks, including image captioning. Yoa *et al*. [39] use a temporal version of the same mechanism to adaptively aggregate visual representations across video frames per word for video captioning. Yeung *et al*. [40] use a similar temporal attention model for temporal action localization.

Visual grounding of natural language expressions is a related problem [19, 30], which can be seen as an extension of weakly supervised object localization [3, 7, 31]. The goal is to localize objects referred to by natural language descriptions, while only using image-level supervision. Since the goal in visual grounding and weakly supervised localization is precise localization, methods typically rely on object proposal regions which are specifically designed to align well with object boundaries [34, 42]. Instead of localizing a given textual description, our approach uses image-level supervision to infer a latent correspondence between the sequence of caption words and a sequence of image regions that are attended to when generating the caption words.

Object proposal methods were designed to focus computation of object detectors on a selective set of image regions likely to contain objects. Recent state-of-the-art detectors, however, integrate the object proposal generation and recognition into a single network. This is computationally more efficient and leads to more accurate results [24, 29] Spatial transformer networks [16] allow similar ideas to be used in cases without location supervision. A localization sub-network computes spatial transformation parameters (*e.g*. position, scale, aspect ratio) that determine which region of the previous CNN layer is sub-sampled and passed to the next CNN layer. Johnson *et al*. [18] used spatial transformers for the task of localized image captioning, which predicts semantically relevant image regions together with their descriptions. In each region, they generate descriptions with a basic non-attentive image captioning model similar to the one used by Vinyals *et al*. [36]. They train their model from a set of bounding-boxes with corresponding captions per image. In our work we model image-wide captions, do not used bounding-box annotations, and use image regions for our attention mechanism.

Compared to previous attention models [17, 37, 38, 41], our attention mechanism, consisting of a single interaction layer, is less complex yet improves performance. Our approach generalizes weakly supervised localization methods and RNN language models. It includes a region-word interaction found in weakly supervised localization, as well as a word-state interaction found in RNN language models. In addition, our model includes a region-state interaction which forms a dynamic appearance-based salience mechanism. Our model naturally handles different types of atten-

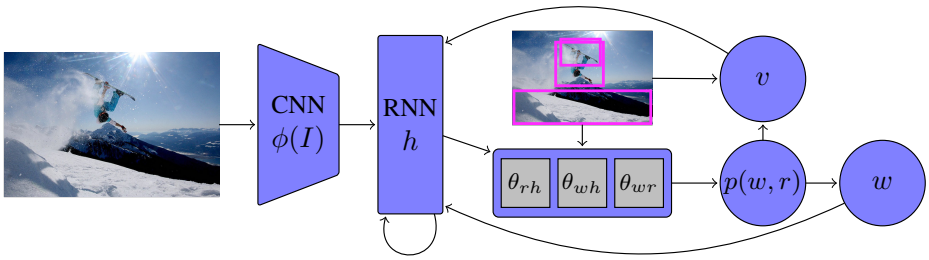


Figure 2. In our attention-based model the conditional joint distribution $p(w, r|h)$ over words and regions given the current state h is used to generate a word and to pool region descriptors in a convex combination. Both are then fed back to the state at the next time-step.

tion regions (fixed grid, object proposals, and spatial transformers). To the best of our knowledge, we are the first to systematically compare different region types for attention-based image captioning in a single model.

3. Attention in encoder-decoder captioning

In Section 3.1 we describe a baseline encoder-decoder model. We extend this baseline in Section 3.2 with our attention mechanism in a way that abstracts away from the underlying region types. In Section 3.3 we show how regions based on CNN activation grids, object proposals, and spatial transformers can be integrated in our model.

3.1. Baseline CNN-RNN encoder-decoder model

Our baseline encoder-decoder model uses a CNN to encode an image I into a vectorial representation $\phi(I) \in \mathbb{R}^{d_I}$, which is extracted from a fully connected layer of the CNN. The image encoding $\phi(I)$ is used to initialize the state of an RNN language model. Let h_t denote the RNN state vector at time t, then $h_0 = \theta_{hi}\phi(I)$, where $\theta_{hi} \in \mathbb{R}^{d_h \times d_I}$ linearly maps $\phi(I)$ to the RNN state space of dimension d_h.

The distribution over w_t, the word at time t, is given by a logistic regression model over the RNN state vector,

$$p(w_t|h_t) \propto \exp\left(w_t^\top W \theta_{wh} h_t\right), \tag{1}$$

where $w_t \in \{0,1\}^{n_w}$ is a 1-hot coding over the captioning vocabulary of n_w words, W is a matrix which contains word embedding vectors as rows, and θ_{wh} maps the word embedding space to the RNN state space. For sake of clarity, we omit the dependence on I in Eq. (1) and below.

We use an RNN based on gated recurrent units (GRU) [6], which are simpler than LSTM units [15], while we found them to be at least as effective in preliminary experiments. Abstracting away from the GRU internal gating mechanism (see supplementary material), the state update function is given by a non-linear deterministic function

$$h_{t+1} = g(h_t, W^\top w_t). \tag{2}$$

The feedback of w_t in the state update makes that w_{t+1} recursively depends on both $\phi(I)$ and the entire sequence of words, $w_{1:t} = (w_1, \ldots, w_t)$, generated so far.

During training we minimize the sum of losses induced by pairs of images I_m with corresponding captions $w_{1:l_m}$,

$$L(I_m, w_{1:l_m}, \theta) = -\sum_{t=1}^{l_m} \ln p(w_t|h_t, \theta), \tag{3}$$

where θ collectively denotes all parameters of the CNN and RNN component. This amounts to approximate maximum likelihood estimation, due to local minima in the loss.

Once the model is trained, captions for a new image can be generated by sequentially sampling $w_t \sim p(w_t|h_t)$, and updating the state $h_{t+1} = g(h_t, w_t)$. Since determining the maximum likelihood sequence is intractable, we resort to beam search if a single high-scoring caption is required.

3.2. Attention for prediction and feedback

In the baseline model the image is used only to initialize the RNN, assuming that the memory of the recurrent net is sufficient to retain the relevant information. We now extend the baseline model with a mechanism to attend to different image regions as the caption is generation word-by-word. Inspired by weakly supervised object localization methods, we score region-word pairs and aggregate these scores by marginalization to obtain a predictive distribution over the next word in the caption. The advantage is that this model allows words to be associated with image region appearances instead of global image representations, which leads to better generalization to recognize familiar scene elements in novel compositions. Importantly, we maintain the word-state interaction in Eq. (1) of the baseline model, to ensure temporal coherence in the generated word sequence by recursive conditioning on all previous words. Finally, a region-state interaction term allows the model to highlight and suppress image regions based on their appearance and the state, implementing a dynamic salience mechanism. See Figure 2 for a schematic illustration of our model.

We define a joint distribution, $p(w_t, r_t|h_t)$, over words w_t and image regions r_t at time t given the RNN state h_t. The marginal distribution over words, $p(w_t|h_t)$, is used to predict the next word at every time-step, while the marginal distribution over regions, $p(r_t|h_t)$, is used to provide a visual feedback to the RNN state update. Let $r_t \in \{0,1\}^{n_r}$

denote a 1-hot coding of the index of the region attended to among n_r regions at time t. We write the state-conditional joint distribution on words and regions as

$$p(w_t, r_t|h_t) \propto \exp s(w_t, r_t, h_t), \quad (4)$$

$$s(w_t, r_t, h_t) = w_t^\top W\theta_{wh}h_t + w_t^\top W\theta_{wr}R^\top r_t + r_t^\top R\theta_{rh}h_t + w_t^\top W\theta_w + r_t^\top R\theta_r, \quad (5)$$

where R contains the region descriptors in its rows. The score function $s(w_t, r_t, h_t)$ is composed of three bi-linear pairwise interactions. The first scores state-word combinations as in the baseline model. The second scores the compatibility between words and region appearances. The third scores region appearances given the current state, and acts as a dynamic salience term. The last two unary terms implement linear bias terms for words and regions respectively.

Given then RNN state, the next word in the image caption is predicted using the marginal word distribution, $p(w_t|h_t) = \sum_{r_t} p(w_t, r_t|h_t)$, which replaces Eq. (1) of the baseline model. The baseline model is recovered for $R = 0$.

In addition to using the image regions to extend the state-conditional word prediction model, we also use them to extend the feedback connections of the RNN state update. We use a mechanism related to the soft attention model of Xu *et al*. [37]. We compute a convex combination of region descriptors which will enter into the state-update. In contrast to Xu *et al*., we derive the region weights from the joint distribution defined above. In particular, we use the marginal distribution over regions at time t, $p(r_t|h_t) = \sum_{w_t} p(w_t, r_t|h_t)$, to pool the region descriptors as

$$v_t = \sum_{r_t} p(r_t|h_t) r_t^\top R = p_{rh}^\top R, \quad (6)$$

where $p_{rh} \in \mathbb{R}^{n_r}$ stacks all region probabilities at time t. This visual representation is concatenated to the generated word in the feedback signal of the state update, *i.e*. we replace the update of Eq. (2) of the baseline model with

$$h_{t+1} = g(h_t, [w_t^\top W \;\; v_t^\top]^\top). \quad (7)$$

In Section 4, we experimentally assess the importance of the different pairwise interactions, and the use of the attention mechanism in the state update.

3.3. Areas of attention

Our attention mechanism is agnostic to the definition of the attention regions. In this section we describe how to integrate three types of regions in our model.

Activation grid. For the most basic notion of image regions we follow the approach of Xu *et al*. [37]. In this case the regions of attention correspond to the $z = x \times y$ positions in the activation grid of a CNN layer $\gamma(I)$ with c channels. The region descriptors in the rows of $R \in \mathbb{R}^{z \times c}$ are given by the activations corresponding to each one of the z locationsof the activation grid. In this case, the receptive fields for the regions all have a fixed shape and size, independent of the image content.

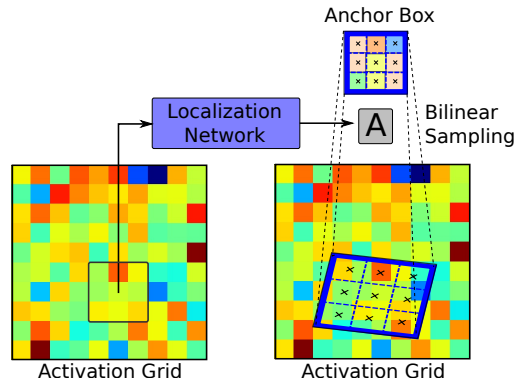


Figure 3. The localization network regresses affine transformations for all feature map positions, which are applied to the anchor boxes that are used to locally re-sample the feature map.

Object proposals. To obtain a set of attention regions that adapt to the image content, we consider the use of object detection proposals. We expect such regions to be more effective since they tend to focus on scene elements such as (groups of) objects, and their parts. In particular we use edge-boxes [42], and max-pool the activations in a CNN layer $\gamma(I)$ over each object proposal to obtain a set of fixed-size region descriptors. To ensure a high-enough resolution of the CNN layer which allows to pool activations for small proposals, we use a separate CNN model which processes the input image at a higher resolution than the one used for the global image representation $\phi(I)$. This is similar to [12, 14], but we pool to a single cell instead of using a spatial pyramid. In this case the number of proposals is not limited by the number of positions in the activation tensor of the CNN layer that is accessed for the region descriptors.

Spatial transformers. Our third type of attention regions is based on recent object detectors and localized image captioning methods with integrated the object proposal mechanisms [18, 24, 29]. In contrast to the latter methods, which rely on bounding-box annotations to learn the region proposal network, we only use image-wide captions for training. Therefore, we need a mechanism that allows back-propagation of the gradient of the captioning loss w.r.t. the region coordinates and the features extracted using them. To this end we use a bi-linear sampling approach as in [16, 18]. In contrast to the max-pooling we use for proposals, it enables differentiation w.r.t. the region coordinates.

Our approach is illustrated in Figure 3. Given an activation map $\gamma(I)$, we use a localization network that consists of two convolutional layers to locally regress an affine transformation $A \in \mathbb{R}^{2\times 3}$ for each location of the feature map. With each location of the activation map $\gamma(I)$ we associate

an "anchor box", which is centered at that position and covers 3×3 activations. The affine transformations, computed at each location, are applied to the coordinates of the anchor boxes. Locally a 3×3 patch is bi-linearly interpolated from $\gamma(I)$ over the area of the transformed anchor box. A 3×3 filter is then applied to the locally extracted patches to compute the region descriptor, which has the same number of dimensions as the activation tensor $\gamma(I)$ has channels. If the local transformations leave the anchor boxes unchanged, then this reduces to the activation grid approach.

As we have no bounding-box annotations, training the spatial transformer can get stuck at poor local minima. To alleviate this issue, we initialize the network with a model that was trained using activation grids. We initialize the transformation layers to produce affine transformations that scale the anchor boxes to twice their original size, to move away from the local optimum of the activation grid model.

4. Experimental evaluation

In this section we first present the experimental setup and implementation details in Section 4.1. The presentation of our experimental results follows in Section 4.2

4.1. Experimental setup and implementation details

Dataset and evaluation metrics. For our experimental evalutation we use the MSCOCO dataset [23]. It consists of 80K training images and 40K development images. Each image comes with five descriptive captions, see Figure 5 for example images. To generate captions we use beam-search with beam size seven, which we found in preliminary experiments to give a good trade-off in search speed and performance. We report standard metrics for the dataset: BLUE1, BLUE4 [27], METEOR [9] and CIDEr [35]. Similar to previous work, we use the first 5K development images to measure performance. We use the next 5K development images to validate the training hyper-parameters, learning rate and early stopping iteration, using CIDEr. We do not use the remaining validation images, unless specified otherwise.

CNN image encoder. We use the penultimate layer of the VGG16 architecture [32] to extract the global image representation $\phi(I)$, which is used to initialize the RNN state. The CNN processes the input image at a resolution of 224×224 pixels. The dimension of the visual region descriptors is given by the number of channels in the corresponding CNN layer of the VGG16 network, *i.e.* $d_r = 512$. The "activation grid" regions are taken from the last convolutional layer. For the "spatial transformer" regions, we use the penultimate convolutional layer to regress the transformations, which are then applied to convolve a locally transformed version of the same layer.

For the "object proposal" regions we max-pool features from the last convolutional layer. Similarly to [29], we rescale the image so that the smaller image dimension is 300 pixels while keeping the original aspect-ratio. When fine-tuning we do not share the parameters of the two CNNs.

Captioning vocabulary. We use all 6,325 unique words in the training captions that appear at least 10 times. Words that appear less frequently are replaced by a special OUT-OF-VOCABULARY token, and the end of the caption is marked with a special STOP token. The word embedding vectors of dimension $d_w = 512$ collected in the matrix W are learned along with the RNN parameters.

Training. We use RNNs with a single layer of $d_h = 512$ GRU units. We use dropout for the fully connected layers at the interface of the CNN and the RNN component of our model, but not for the RNN state because it deteriorates performance. Although they are likely to improve results, we did not experiment with RNN-specific regularization techniques such as zone-out [22].

We found it useful to train our models in two stages. In the first stage, we fix the weights of the CNN component to values that were obtained by training on the ImageNet 2010 large-scale classification challenge dataset [8]. We train all remaining model parameters using the Adam stochastic gradient descend algorithm [20], and learning rate set to 10^{-3}. In the second stage, we continue training to fine-tune all network parameters, including the ones of the CNN, using a learning rate of 10^{-5}.

To speed-up training, we sub-sample the 14×14 convolutional layers to 7×7 when using the activation grid and the spatial transformer regions. Similarly, when training with object proposals, each time we process an image we use 50 randomly selected regions.

Region visualization. To visualize the attention regions, we show the areas from which the convolutional features are pooled. For the spatial transformers, we show the transformed anchor boxes. For the activation grid regions, we show the back-projection of a 3×3 activation block, which allows for direct comparison with the spatial transformers. Note that in all cases the underlying receptive fields are significantly larger than the depicted areas. For object proposals we show the edge-boxes.

4.2. Experimental results

In this section we evaluate our model with respect to the baseline, and assess the relative importance of different components of our model. We also evaluate the effectiveness of the different types of attention regions, and the effect of jointly fine-tuning the CNN and RNN components. Finally, we compare our results to the state of the art.

Attention and visual feedback. We start with evaluating our baseline captioning system, and our attention model using activation grid regions. In Table 1 we progressively add components of our model to the baseline system. The baseline RNN uses only word-state interaction terms to predict the next word given the RNN state, this leads to a

Method	B1	B4	Meteor	CIDEr
Baseline: θ_{wh}	66.3	26.4	22.2	78.9
Ours: θ_{wh}, θ_{wr}	68.0	28.0	22.9	83.6
Ours: $\theta_{wh}, \theta_{wr}, \theta_{rh}$	68.2	28.4	23.3	85.5
Ours: conditional feedback	68.3	28.7	**23.7**	86.8
Ours: full model	**69.1**	**28.8**	**23.7**	**87.4**

Table 1. Evaluation of the baseline and our attention model using activation grid regions, including variants with certain components omitted, and word-conditional instead of marginal feedback.

CIDEr score of 78.9. Adding the word-region interaction term (second row), and predicting words from the marginal distribution over words, leads to an improvement of 4.7 points in the CIDEr score to 83.6. This demonstrates the significance of localized visual input to the RNN, which allows it to associate caption terms to local appearances rather than to global scene descriptors. Adding the third pairwise interaction term between regions and the RNN state (third row) brings another improvement of 1.9 points to 85.5 CIDEr. This shows that the RNN is also able to implement a dynamic salience mechanism that favors certain regions over others at a given time-step by scoring the compatibility between the RNN state and the region appearance. Finally we add the visual feedback mechanism to our model (87.4, last row), which drives the CIDEr score further up by 1.9 points. We also experimented with a word-conditional version of the visual feedback mechanism (86.8, last but first row), which uses $p(r_t|w_t, h_t)$ instead of $p(r_t|h_t)$ to compute the visual feedback. Although this also improves the CIDEr score, as compared to not using visual feedback, it is less effective than using the marginal distribution weights. The visualizations in Figure 5 suggest that the reason for this is that the marginal distribution already tends to focus on a single semantically meaning full area. The generated word therefore does not seem to be required to collapse the marginal distribution over regions to a single mode.

Differences between areas of attention. In our next set of experiments we compare the effectiveness of different attention regions in our model. In Figure 4 we consider the performance of the three regions types as a function of the number of regions that are used when running the trained model on test images. For activation grids and spatial transformers the number of regions are regularly sampled from the original 14×14 resolution using increasing strides. For instance, using as stride of 2 generates $7 \times 7 = 49$ regions. For object proposals we test a larger range, from 1 up to 2,000 regions, sorted by "objectness" score. For all three region types, performance quickly increases with the number of regions, and then plateaus off. The spatial transformer regions consistently improve over the activation grid ones, demonstrating the effectiveness of the region transformation sub-network. Using less than ten regions the object proposals are best, but spatial transformer regions give best overall results. In the remaining experiments, we report performance with the optimal number of regions per method: 1,000 for proposals, and 196 for grids and transformers.

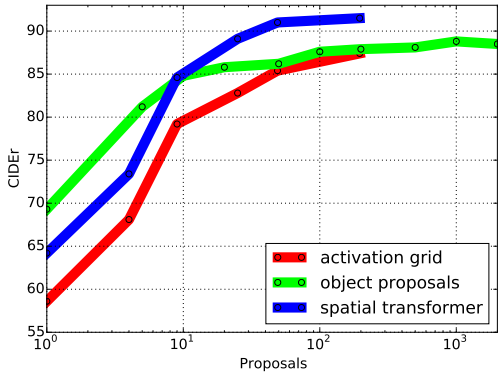


Figure 4. Image captioning performance as a function of the number of regions used. Note the log-scale on the horizontal axis.

Method	B1	B4	Meteor	CIDEr
	RNN training only			
Baseline	66.3	26.4	22.2	78.9
Activation grid	69.1	28.8	23.6	87.4
Object proposals	69.4	28.9	23.7	89.0
Spatial transformers	**70.2**	**30.2**	**24.2**	**91.1**
	CNN-RNN fine-tuning			
Baseline	68.6	28.7	23.5	87.1
Activation grid	70.4	30.3	**24.5**	92.6
Object proposals	**71.0**	30.1	**24.5**	93.7
Spatial transformers	70.8	**30.7**	**24.5**	**93.8**

Table 2. Captioning performance of the baseline and our model using different attention regions, with and without fine tuning.

Joint CNN-RNN fine-tuning. We now consider the effect of jointly fine-tuning the CNN and RNN components. In Table 2 we report the performance with and without fine-tuning for each region type, as well as the baseline performance for reference. All models are significantly improved by the fine-tuning. Although the baseline improves the most in absolute terms, its performance remains substantially behind that of our attention models. The two types of image-dependent attention regions improve over fixed activation grids, but the differences between them are reduced after fine-tuning. Our spatial transformer approach leads to comparable results as using state-of-the-art edge-box object proposals, that were designed to align with object boundaries.

Our approach offers several advantages over the object proposals however. First, it uses only 196 instead of 1,000 regions per image, which makes the RNN more efficient to evaluate. Second, in the case of object proposals two

Figure 5. Visualization of the focus of our attention model during sequential word generation for the three different region types: activation grids, object proposals, and spatial transformers. The attention areas are drawn with line widths directly proportional to weights $p(r_t|h_t)$.

separate CNNs are fine-tuned to compute the image-wide and region descriptors, the latter being significantly more costly to evaluate since it is run at a higher image resolution. Third, from a modeling perspective the spatial transformer approach is more satisfying, since it is fully end-to-end trainable, and does not rely on an external image processing pipeline. Finally, our spatial transformer approach can easily be interfaced with an object localization network, to strengthen the region regression component.

Visualizing areas of attention. In Figure 5 we provide a qualitative comparison of the attentive focus using different regions in our model, a larger selection can be found in the supplementary material. We show the generated captions together with the attention weights over the image re-

gions at each point in the sentence. The images displayed for the object proposals differ slightly from the others, since the high-resolution network used in that case uses a different cropping and scaling scheme.

Object proposals accurately capture small objects in some cases, *e.g*. the kite, but in other cases regions for background elements are missing, *e.g*. for the field and the sky. The spatial transformers tend to focus quite well on relational terms. For example, *"standing"* focuses on the area around the legs of the elephants in the first image, and *"low"* on the area between the airplane and the ground in the last image. For the spatial transformers in particular, the focus of attention tends to be stable across meaningful sub-sequences, such as noun phrases (*e.g*. *"A couple of elephants"*) and prepositional phrases (*e.g*. *"on a beach."*). This suggests that our model succeeds in implicitly recovering the latent sentence structure to some extent.

Comparison to the state of the art. We compare our results obtained using the spatial transformer regions to the state of the art in Table 3. We refer to our method as "Areas of Attention", or AoA for short. Our results compare favorably with the state of the art, in particular our CIDEr score of 93.8 improves the state-of-the-art results of Bengio *et al*. [2] by 1.7 point. Xu *et al*. [37] report the best Blue1 score (71.8), higher than ours by one point. For the other measures their performance is worse, in particular for B4 they report 25.0 where we obtain 30.7. The result closest to ours by Bengio *et al*. [2] are obtained using a "scheduled sampling" training algorithm. Using standard "teacher-forced" training, as we use in our work, they report a CIDEr score of 89.5 for the same model. We expect our model to also benefit from this improved learning algorithm.

While ensembling models can significantly improve results, we did not include such results in Table 3 for sake of comparability. Bengio *et al*. [2] report 98.7 CIDEr using an ensemble of 10 models, an improvement of 6.6 points over their result (92.1) using a single model. You *et al*. [41] report CIDEr only using an ensemble of 5 models on the test set, they obtain 94.3. We will report ensemble results for our model in the final version of our paper.

We also trained our model using 30K additional validation images on top of the 80K training images. We use the same 5K validation images and 5K images for reporting as in the other experiments. In addition we use a random horizontal flip of the images during training. For testing we generate a sentence for the original image and for its horizontal flip, and then pick the sentence with maximum likelihood among the two alternatives. Both data augmentation methods significantly improve performance, in total by 1.8 points to 95.6 CIDEr. This suggests that even more training data might bring further improvements.

For reference we report test set results, computed using the evaluation server [4], in the supplementary material.

Method	B1	B4	Meteor	CIDEr
Vinyals *et al*. [36]	-	27.7	23.7	85.5
Xu *et al*. [37], soft	70.9	24.3	23.9	-
Xu *et al*. [37], hard	**71.8**	25.0	23.0	-
Yang *et al*. [38]	-	29.0	23.7	88.6
Jin *et al*. [17]	69.7	28.2	23.5	83.8
Donahue *et al*. [10]	71.1	30.0	24.2	89.6
Ranzato *et al*. [28]	-	29.2	-	-
Bengio *et al*. [2]	-	30.6	24.3	92.1
Areas of Attention (ours)	70.8	**30.7**	**24.5**	**93.8**
AoA, data augmentation	72.1	31.1	25.0	95.6

Table 3. Comparison of our results with the state of the art.

5. Conclusion

We have presented a novel attention-based model for image captioning. Our model builds upon recent encoder-decoder image captioning models. It is based on a score function that consist of three pairwise interactions between the RNN state, image regions, and caption words. We evaluated our model with three different region types, ranging from a simple regular activation grid, to object proposals, and an integrated spatial transformer network that learns to regress the attention regions from the image content. Extensive experimental results show the importance of all model components, and the importance of image-adaptive attention regions. Our model has excellent performance, and sets a new state-of-the-art among non-ensembled methods.

In ongoing work we pursue several directions to further improve our model. First, in our current model, the spatial transformer regions are regressed once per image, they are only scored differently per generated caption word. Making the region shapes themselves dependent on RNN state might improve our model, by dynamically shaping the attention regions to what has been written so far in the caption. Second, using multiple instead of a single anchor box at each location in the spatial transformer model has been observed to improve object detection performance, and may carry over to our case. Third, in the presented work we train our models with the cross-entropy loss, using teacher forced training where the RNN updates are based on the ground-truth words instead of the last generated word. Alternative training schemes such as scheduled sampling [2], and more appropriate loss functions [28], are likely to benefit our model. Finally, we plan to extend our model to handle words not seen in the training captions. For such words we can learn embedding vectors by using an unsupervised word embedding method like word2vec [26], while fixing the embedding learned by the caption model for other words.

We will release an open source Theano-Lasagne based implementation of our model to reproduce our experiments.

Acknowledgments. We would like to thank NVIDIA

for the donation of GPUs used in this research. This work has been supported in part by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.

[3] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325, 2015.

[5] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *NIPS*, 2015.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS Deep Learning Workshop*, 2014.

[7] R. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, 2014.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] M. Denkowski and A. Lavie. Meteor universal: Language spec5ic translation evaluation 6r any target language. In *In Proceedings 6 the Ninth Workshop on Statistical Machine Translation*, 2014.

[10] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[11] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

[12] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[13] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

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

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015.

[17] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272, 2015.

[18] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

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

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[21] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, 2014.

[22] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. Ke, A. Goyal, Y. Bengio, H. Larochelle, A. Courville, and C. Pal. Zoneout: Regularizing RNNs by randomly preserving hidden activations. arXiv:1606.01305.

[23] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[25] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). *ICLR*, 2015.

[26] T. Mikolov, W.-T. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *NAACL-HLT*, 2013.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[28] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[30] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

[31] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

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

[33] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[34] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[35] R. Vedantam, C. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

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

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[38] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. Cohen. Encode, review, and decode: Reviewer module for caption generation. In *NIPS*, 2016.
[39] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.
[40] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos.
[41] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.
[42] C. Zitnick and P. Dollár. Edge boxes: locating object proposals from edges. In *ECCV*, 2014.

A. Supplementary material

In this appendix, we report results obtained on the MSCOCO test set in Section A.1. We describe the gated recurrent unit (GRU) used in our work with more detail in Section A.2. We provide additional visualizations of our attention model in Section A.3.

A.1. Results on MSCOCO test set

We submit our AoA full model on the MSCOCO test server and obtain a CIDEr score of 92.4. The model uses spatial transformer regions, and is learned from both training and validation data (110K images) using also flipped images. Ensembling more models and using more training data can further improve results.

A.2. Details on gated recurrent units (GRUs)

To alleviate the problem of vanishing or exploding gradients encountered in deep and recurrent neural networks (RNNs), gated units have been proposed. The most two well known ones are LSTMs [15] and GRUs [6]. Such units use a gating mechanism to control the flow of information depending on the input, enabling better learning of long-term dependencies. In our work we used GRUs, based on better results in initial experiments with the baseline model. We provide here a brief description of the GRU mechanism, see [6] for more details.

We use $h_t \in \mathbb{R}^{d_h}$ to denote the RNN state, and x_t as the input to the RNN at time t. To compute the state evolution, a gated recurrent unit (GRU) makes use of two gates: a "forget gate" $z_t \in [0,1]^{d_h}$ and a "read gate" $r_t \in [0,1]^{d_h}$. Both gates are computed as a sigmoid of a linear function of the input and previous state:

$$z_t = \sigma\left(\omega_{zx} x_t + \omega_{zh} h_{t-1}\right), \tag{8}$$

$$r_t = \sigma\left(\omega_{rx} x_t + \omega_{rh} h_{t-1}\right). \tag{9}$$

The read gate r_t is used, together with the previous hidden state h_{t-1} and the input x_t, to compute a "tentative" state $\tilde{h}_t$:

$$\tilde{h}_t = \tanh\left(\omega_{hr}(r_t \odot h_{t-1}) + \omega_{hx} x_t\right), \tag{10}$$

where $\odot$ denotes the element-wise product of two vectors.

Finally, the forget gate controls to what extent the previous state h_{t-1} is maintained, or replaced by the tentative state $\tilde{h}_t$:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \tag{11}$$

These updates together constitute the state update function $h_{t+1} = g(h_t, w_t)$ used in the main paper.

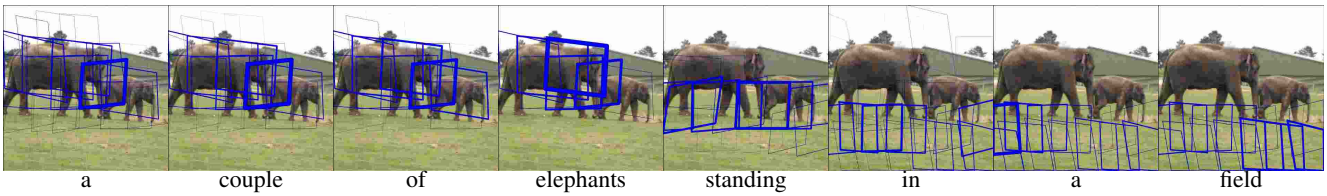
In the baseline model, presented in Section 3.1 of the main paper, the input x_t used to compute h_{t+1} is the embedding of the previously generated word $W^\top w_t$. In our attention model the input to compute h_{t+1} is a concatenation of the embedding of the previously generated word and the visual feedback vector v_t as defined in Equation 6 of the main paper.

A.3. Additional visualizations

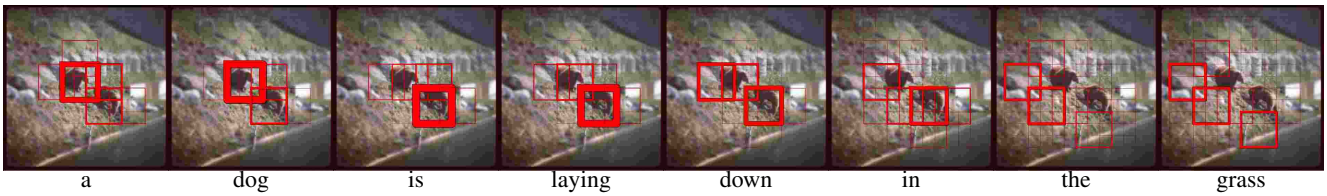
We provide visualizations similar to those of Figure 5 of the main paper. We visualize the focus of our attention model during sequential word generation for the three different region types: activation grids, object proposals, and spatial transformers. The attention areas are drawn with line widths directly proportional to weights $p(r_t|h_t)$. The images displayed for the object proposals differ slightly from the others, since the high-resolution network used in that case uses a different cropping and scaling scheme. To visualize the attention regions, we show the areas from which the convolutional features are pooled. For the spatial transformers, we show the transformed anchor boxes. For the activation grid regions, we show the back-projection of a 3×3 activation block, which allows for direct comparison with the spatial transformers. For object proposals we show the edge-boxes. Note that in all cases the underlying receptive fields are significantly larger than the depicted areas.

a bird is standing in the water at night
a bird is standing on a rock in the water
a bird perched on top of a bird feeder
a group of children sitting around a table with food
a group of people sitting around a table eating food
a little girl sitting at a table with a plate of food
a large jetliner sitting on top of an airport tarmac
a large jetliner sitting on top of an airport tarmac
a large jetliner sitting on top of an airport runway
a man riding a wave on top of a surfboard

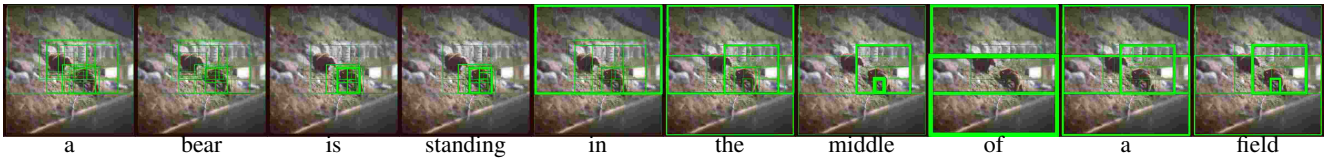
a person riding a wave on top of a surfboard
a man riding a wave on top of a surfboard
a woman sitting on a train talking on a cell phone
a woman sitting on a bed with a bag of luggage
a woman sitting on a bed talking on a cell phone
a man holding a frisbee in his hand
a man holding a frisbee in his hand
a man holding a frisbee in his hand
a couple of elephants standing next to each other
a large elephant standing in a field of grass



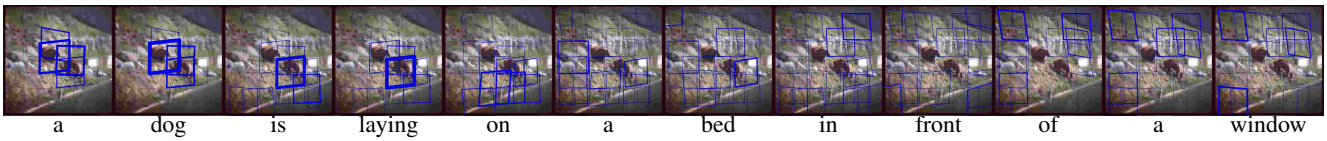
a couple of elephants standing in a field



a dog is laying down in the grass

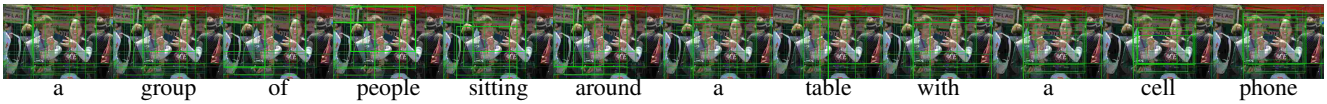


a bear is standing in the middle of a field

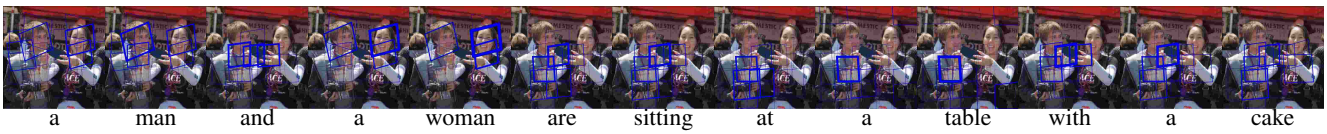


a dog is laying on a bed in front of a window

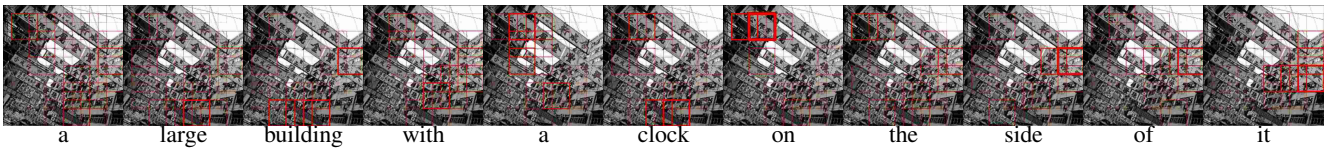
a man and woman are holding a smart phone



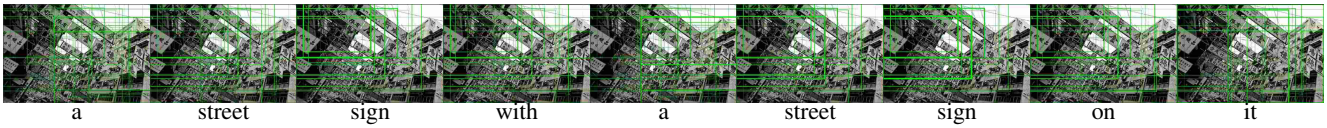
a group of people sitting around a table with a cell phone



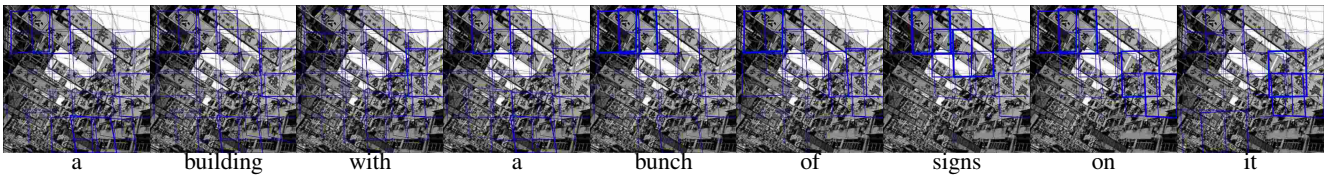
a man and a woman are sitting at a table with a cake



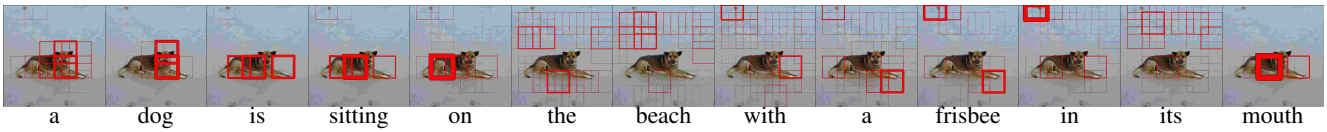
a large building with a clock on the side of it



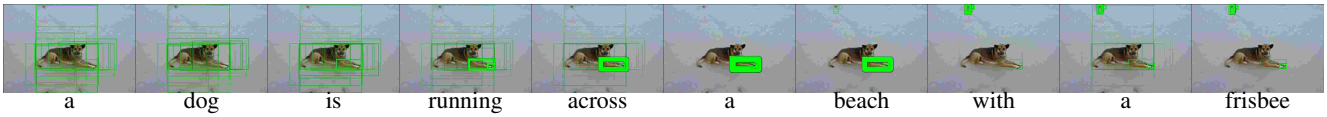
a street sign with a street sign on it



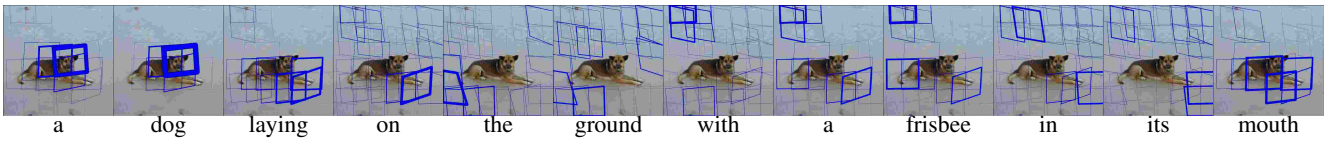
a building with a bunch of signs on it



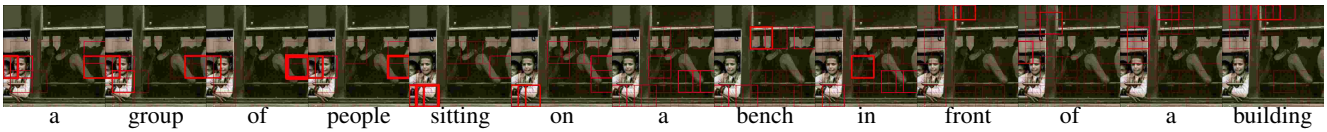
a dog is sitting on the beach with a frisbee in its mouth



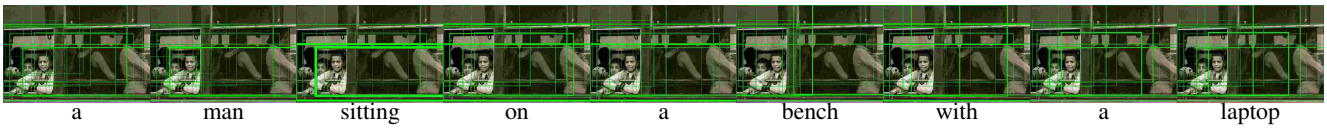
a dog is running across a beach with a frisbee



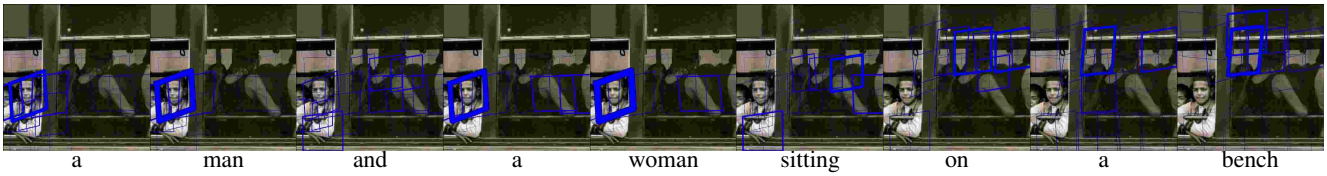
a dog laying on the ground with a frisbee in its mouth



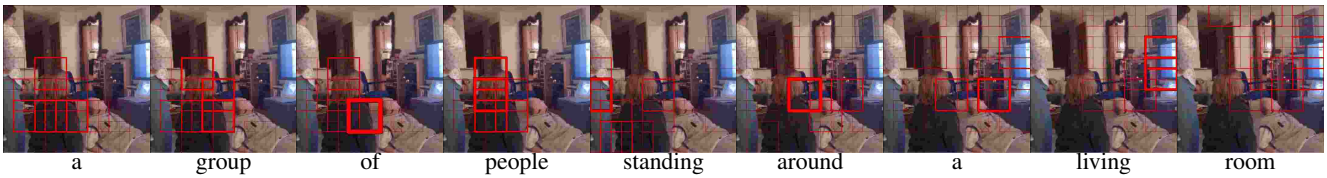
a group of people sitting on a bench in front of a building



a man sitting on a bench with a laptop

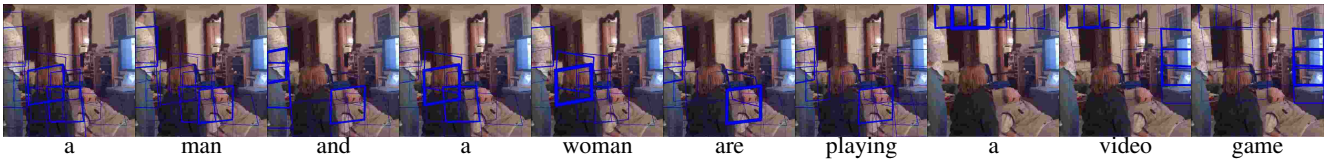


a man and a woman sitting on a bench

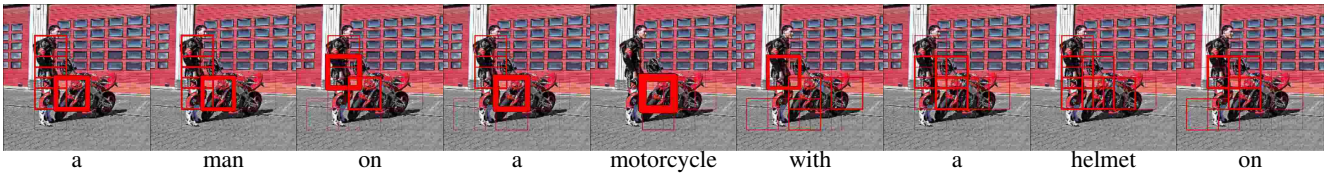


a group of people standing around a living room

a man and a woman are playing a video game

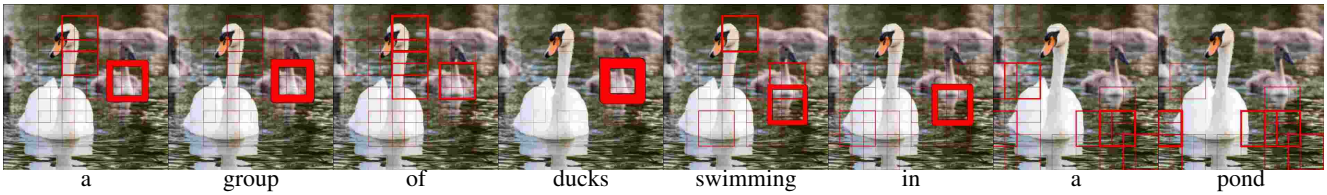


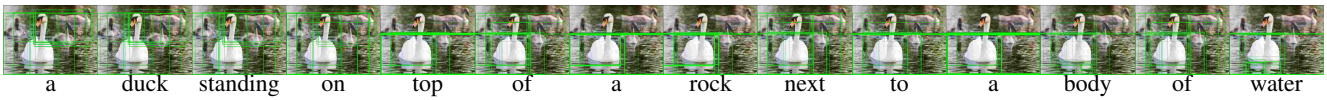
a man and a woman are playing a video game

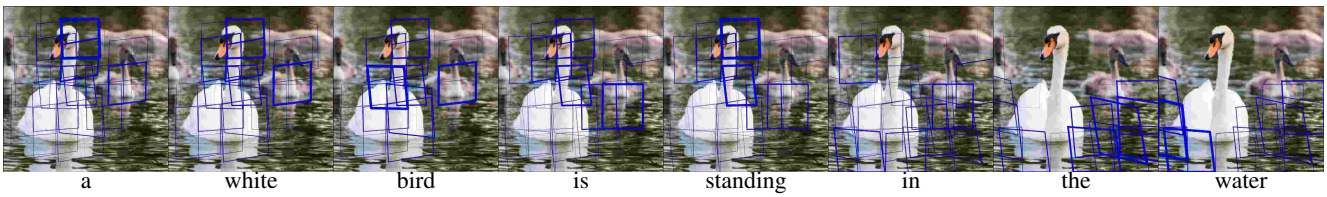


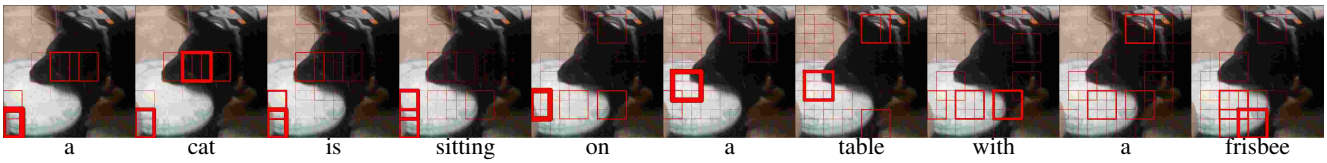
a man on a motorcycle with a helmet on

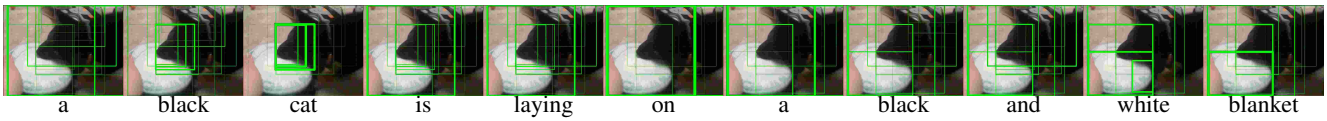
a man riding a motorcycle down a street
a man is riding a motorcycle down a street
a man riding a snowboard down a snow covered slope
a man riding a snowboard down a snow covered slope
a man riding skis down a snow covered slope
a man sitting at a table with a laptop
a man sitting at a table with a laptop
a man sitting at a table with a laptop

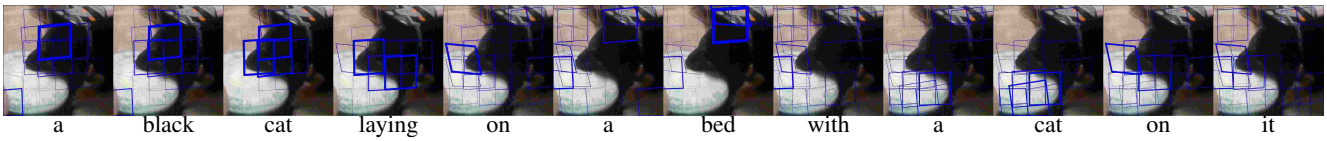

a group of ducks swimming in a pond

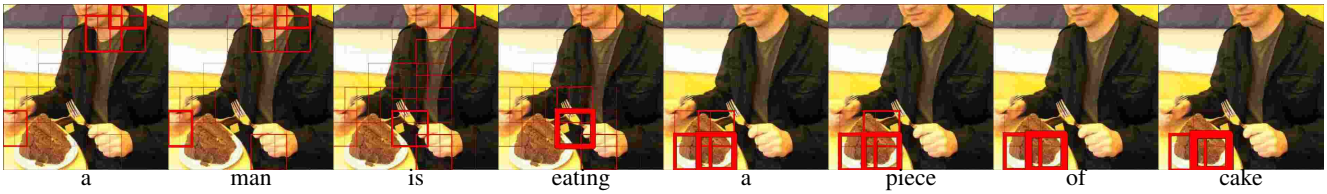

a duck standing on top of a rock next to a body of water

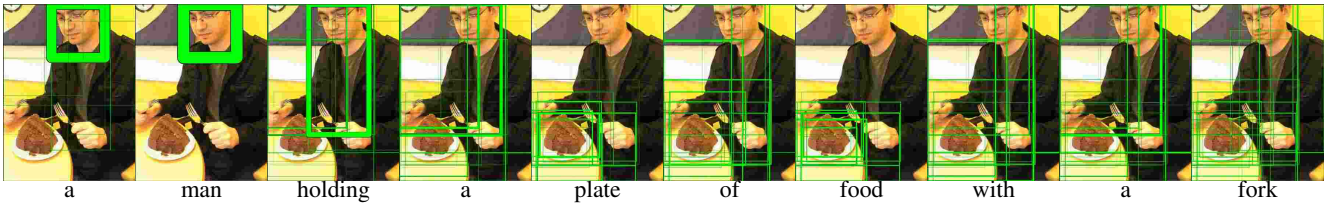

a white bird is standing in the water

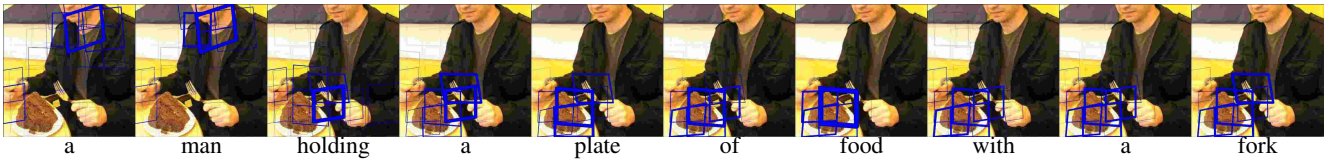

a cat is sitting on a table with a frisbee

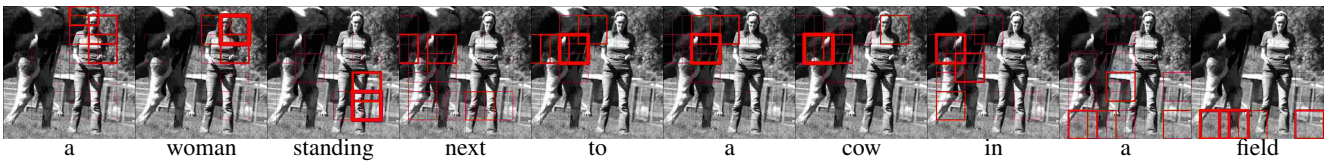

a black cat is laying on a black and white blanket


a black cat laying on a bed with a cat on it

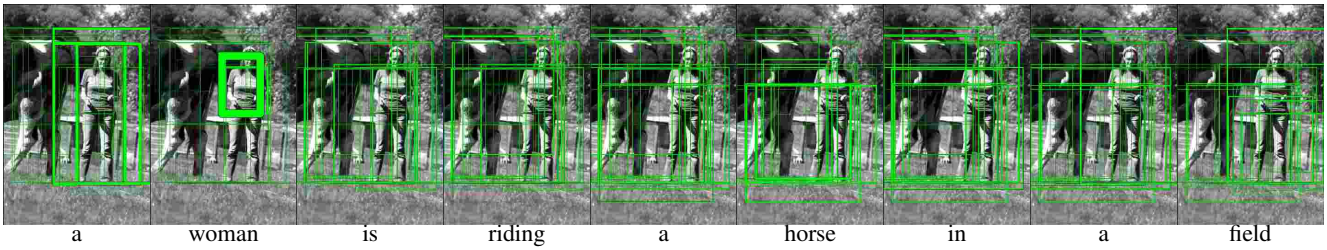

a man is eating a piece of cake


a man holding a plate of food with a fork

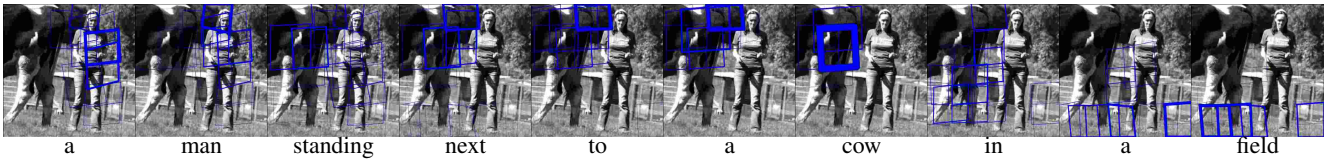

a man holding a plate of food with a fork



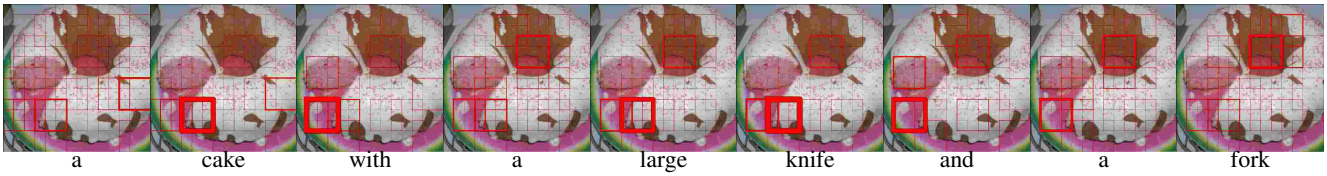
a woman standing next to a cow in a field



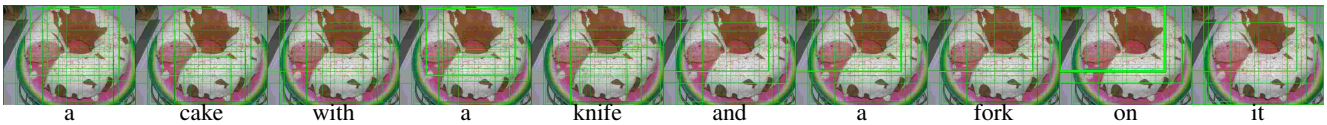
a woman is riding a horse in a field



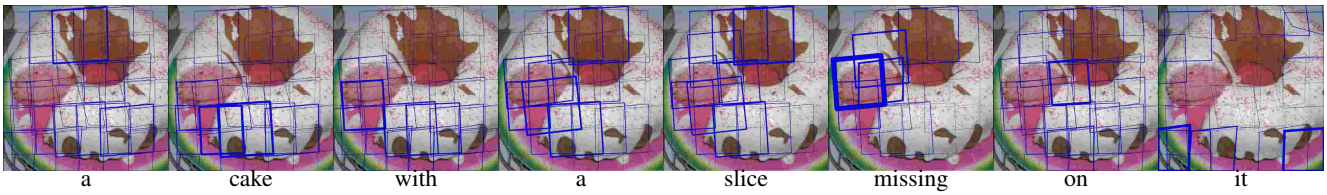
a man standing next to a cow in a field



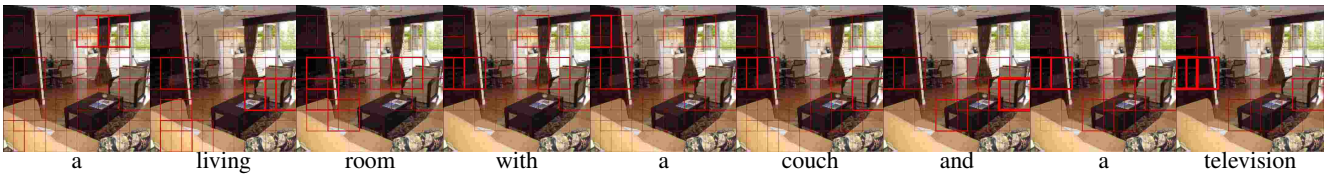
a cake with a large knife and a fork



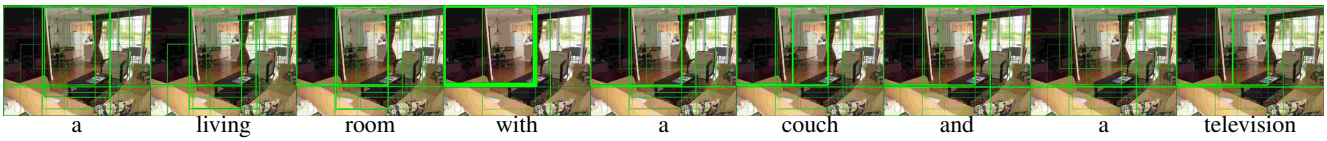
a cake with a knife and a fork on it



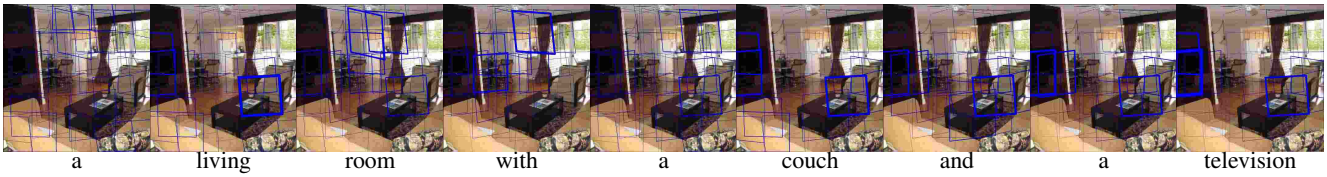
a cake with a slice missing on it



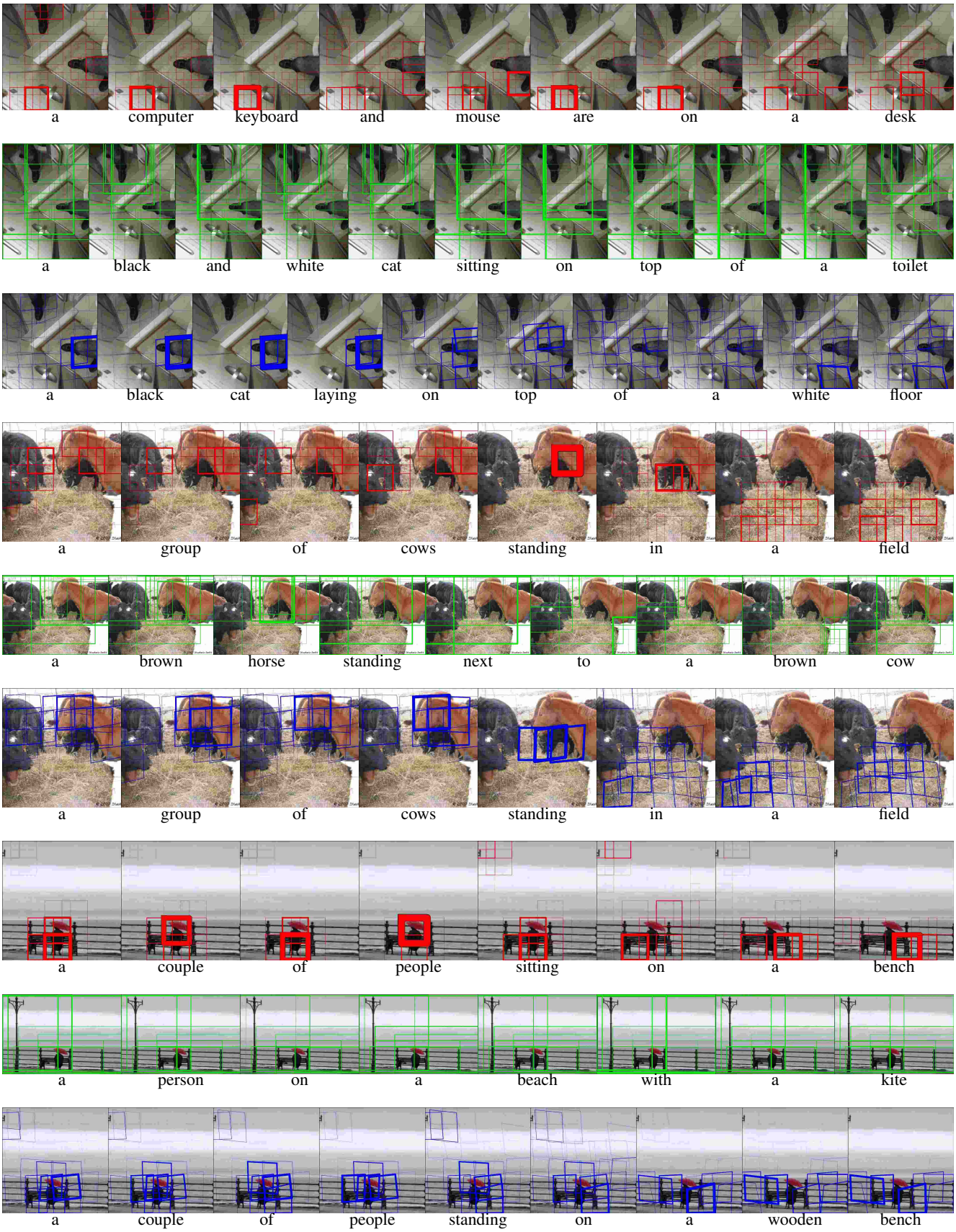
a living room with a couch and a television

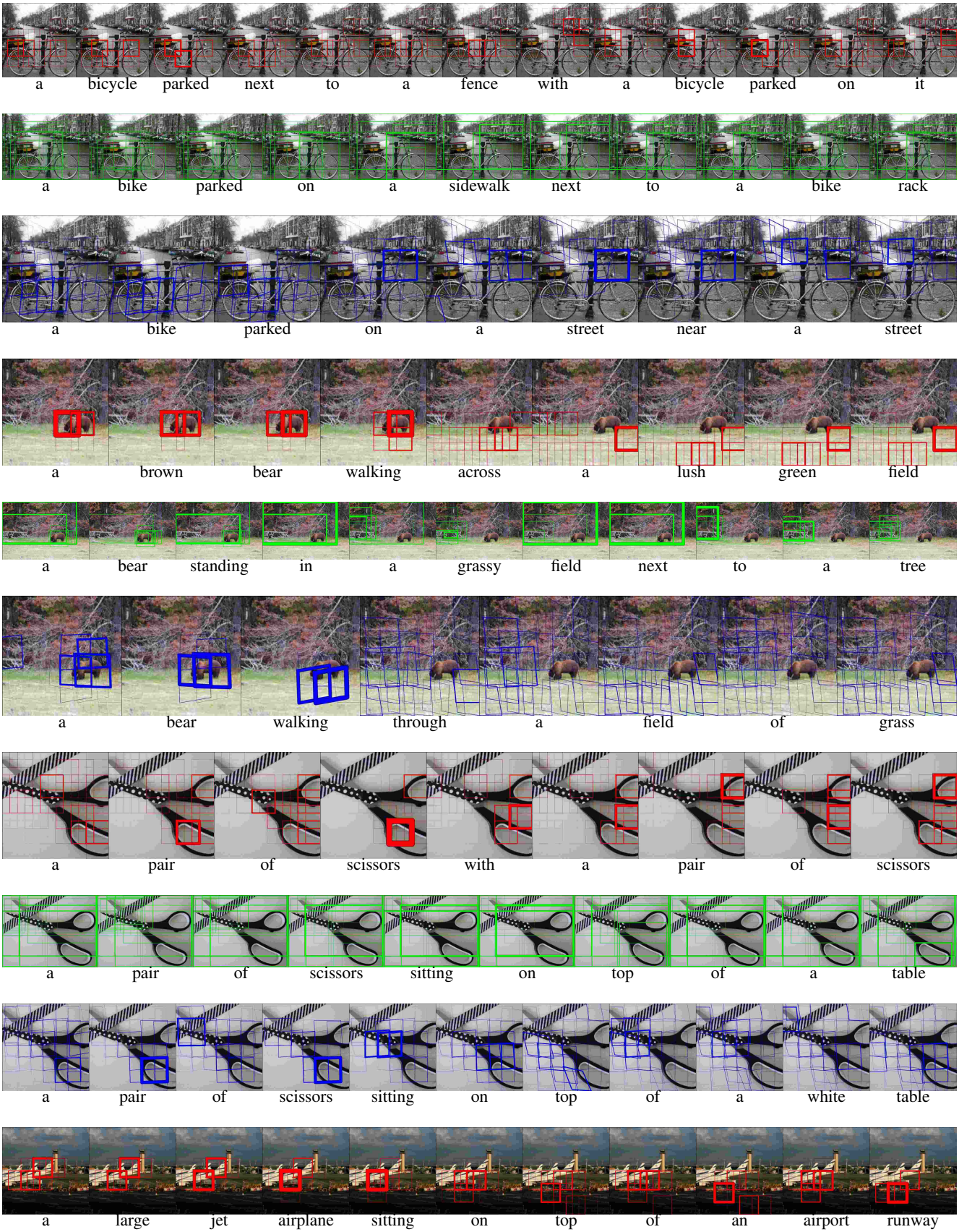


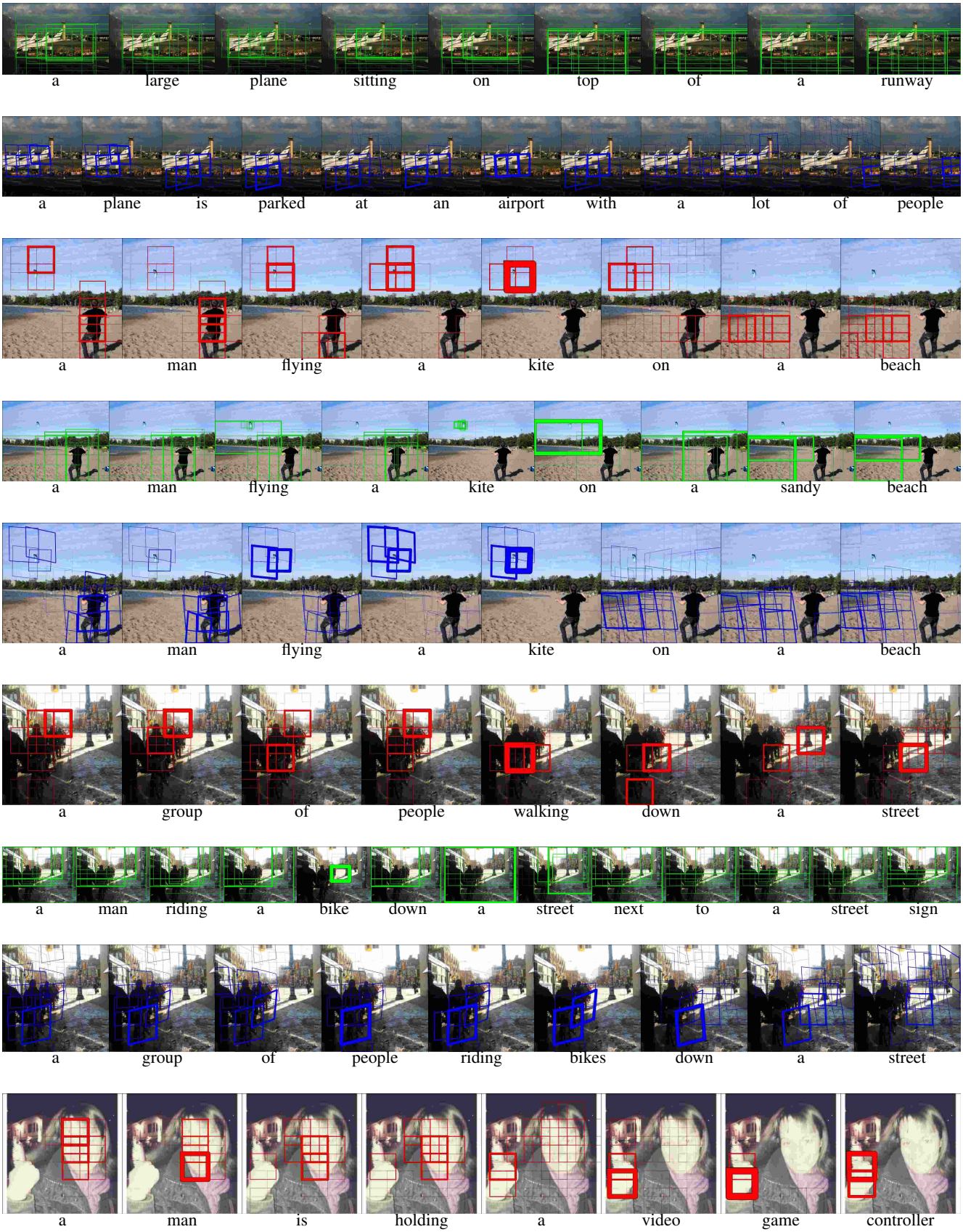
a living room with a couch and a television



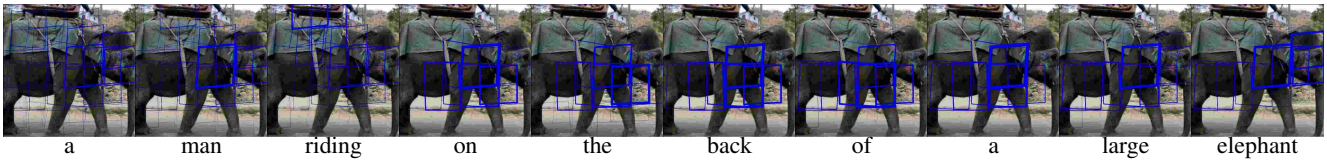
a living room with a couch and a television


a computer keyboard and mouse are on a desk
a black and white cat sitting on top of a toilet
a black cat laying on top of a white floor
a group of cows standing in a field
a brown horse standing next to a brown cow
a group of cows standing in a field
a couple of people sitting on a bench
a person on a beach with a kite
a couple of people standing on a wooden bench

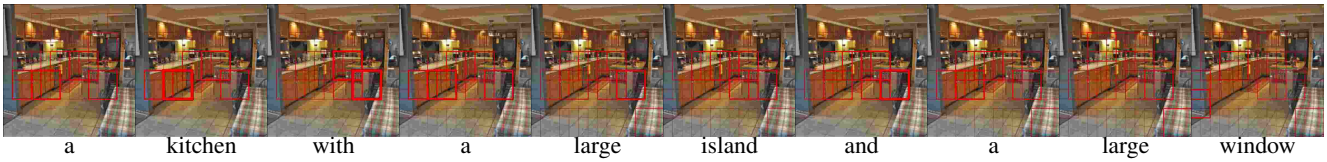

a bicycle parked next to a fence with a bicycle parked on it
a bike parked on a sidewalk next to a bike rack
a bike parked on a street near a street
a brown bear walking across a lush green field
a bear standing in a grassy field next to a tree
a bear walking through a field of grass
a pair of scissors with a pair of scissors
a pair of scissors sitting on top of a table
a pair of scissors sitting on top of a white table
a large jet airplane sitting on top of an airport runway


a large plane sitting on top of a runway
a plane is parked at an airport with a lot of people
a man flying a kite on a beach
a man flying a kite on a sandy beach
a man flying a kite on a beach
a group of people walking down a street
a man riding a bike down a street next to a street sign
a group of people riding bikes down a street
a man is holding a video game controller

a young girl is looking at a cell phone
a man in a santa hat is holding a toothbrush
a man standing on a beach holding a white umbrella
a man standing in the snow with a snowboard
a man standing in the snow with a surfboard
a man in a red shirt is standing in front of a wall
a man and a woman are standing in a kitchen
a man in a red shirt and a tie is holding a glass of wine
two elephants are standing next to each other
a man riding a elephant with a man on his back

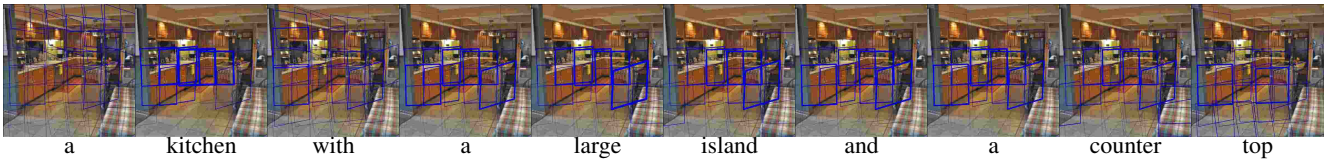


a man riding on the back of a large elephant



a kitchen with a large island and a large window

a kitchen with a large island and a large window



a kitchen with a large island and a counter top