Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy Li Fei-Fei
Department of Computer Science, Stanford University
{karpathy,feifeili}@cs.stanford.edu

Abstract

We present a model that generates free-form natural language descriptions of image regions. Our model leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between text and visual data. Our approach is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. We demonstrate the effectiveness of our alignment model with ranking experiments on Flickr8K, Flickr30K and COCO datasets, where we substantially improve on the state of the art. We then show that the sentences created by our generative model outperform retrieval baselines on the three aforementioned datasets and a new dataset of region-level annotations.

1. Introduction

A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene [8]. However, this remarkable ability has proven to be an elusive task for our visual recognition models. The majority of previous work in visual recognition has focused on labeling images with a fixed set of visual categories, and great progress has been achieved in these endeavors [36, 6]. However, while closed vocabularies of visual concepts constitute a convenient modeling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose.

Some pioneering approaches that address the challenge of generating image descriptions have been developed [22, 7]. However, these models often rely on hard-coded visual concepts and sentence templates, which imposes limits on their variety. Moreover, the focus of these works has been on reducing complex visual scenes into a single sentence, which we consider as an unnecessary restriction.

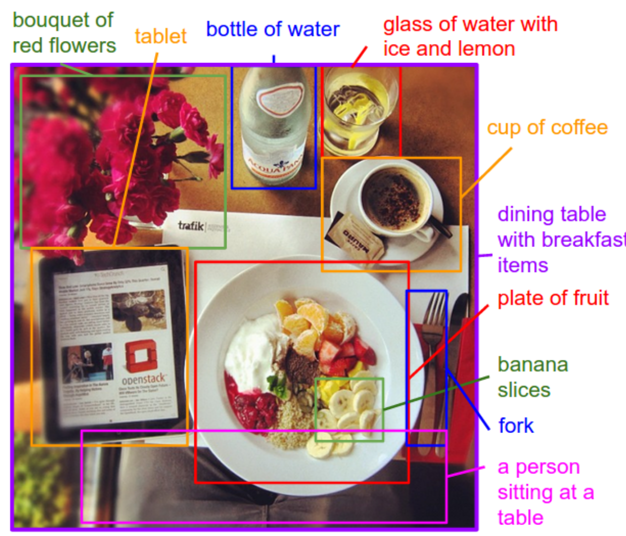

Figure 1. Our model generates free-form natural language descriptions of image regions.

In this work, we strive to take a step towards the goal of generating dense, free-form descriptions of images (Figure 1). The primary challenge towards this goal is in the design of a model that is rich enough to reason simultaneously about contents of images and their representation in the domain of natural language. Additionally, the model should be free of assumptions about specific hard-coded templates, rules or categories and instead rely primarily on training data. The second, practical challenge is that datasets of image captions are available in large quantities on the internet [14, 46, 29], but these descriptions multiplex mentions of several entities whose locations in the images are unknown.

Our core insight is that we can leverage these large image-sentence datasets by treating the sentences as weak labels, in which contiguous segments of words correspond to some particular, but unknown location in the image. Our approach is to infer these alignments and use them to learn a generative model of descriptions. Concretely, our contributions are twofold:

- We develop a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe. Our model associates the two modalities through a common, multimodal embedding space and a structured objective. We validate the effectiveness of this approach on image-sentence retrieval experiments in which we surpass the state-of-the-art.

- We introduce a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text. Our experiments show that the generated sentences significantly outperform retrieval-based baselines, and produce sensible qualitative predictions. We then train the model on the inferred correspondences and evaluate its performance on a new dataset of region-level annotations.

We make our code, data and annotations publicly available.

2. Related Work

Dense image annotations. Our work shares the high-level goal of densely annotating the contents of images with many works before us. Barnard et al. [1] and Socher et al. [38] studied the multimodal correspondence between words and images to annotate segments of images. Several works [26, 12, 9] studied the problem of holistic scene understanding in which the scene type, objects and their spatial support in the image is inferred. However, the focus of these works is on correctly labeling scenes, objects and regions with a fixed set of categories, while our focus is on richer and higher-level descriptions of regions.

Generating textual descriptions. Multiple works have explored the goal of annotating images with textual descriptions on the scene level. A number of approaches pose the task as a retrieval problem, where the most compatible annotation in the training set is transferred to a test image [14, 39, 7, 34, 17], or where training annotations are broken up and stitched together [23, 27, 24]. However, these methods rely on a large amount of training data to capture the variety in possible outputs, and are often expensive at test time due to their non-parametric nature. Several approaches have been explored for generating image captions based on fixed templates that are filled based on the content of the image [13, 22, 7, 43, 44, 4]. This approach still imposes limits on the variety of outputs, but the advantage is that the final results are more likely to be syntactically correct. Instead of using a fixed template, some approaches that use a generative grammar have also been developed [33, 45]. More closely related to our approach is the work of Srivastava et al. [40] who use a Deep Boltzmann Machine to learn a joint distribution over a images and tags. However, they do not generate extended phrases. More recently, Kiros et al. [19] developed a log-bilinear model that can generate full sentence descriptions. However, their model uses a fixed window context, while our Recurrent Neural Network model can condition the probability distribution over the next word in the sentence on all previously generated words.

Grounding natural language in images. A number of approaches have been developed for grounding textual data in the visual domain. Kong et al. [20] develop a Markov Random Field that infers correspondences from parts of sentences to objects to improve visual scene parsing in RGBD images. Matuszek et al. [30] learn joint language and perception model for grounded attribute learning in a robotic setting. Zitnick et al. [48] reason about sentences and their grounding in cartoon scenes. Lin et al. [28] retrieve videos from a sentence description using an intermediate graph representation. The basic form of our model is inspired by Frome et al. [10] who associate words and images through a semantic embedding. More closely related is the work of Karpathy et al. [18], who decompose images and sentences into fragments and infer their inter-modal alignment using a ranking objective. In contrast to their model which is based on grounding dependency tree relations, our model aligns contiguous segments of sentences which are more meaningful, interpretable, and not fixed in length.

Neural networks in visual and language domains. Multiple approaches have been developed for representing images and words in higher-level representations. On the image side, Convolutional Neural Networks (CNNs) [25, 21] have recently emerged as a powerful class of models for image classification and object detection [36]. On the sentence side, our work takes advantage of pretrained word vectors [32, 15, 2] to obtain low-dimensional representations of words. Finally, Recurrent Neural Networks have been previously used in language modeling [31, 41], but we additionally condition these models on images.

3. Our Model

Overview. The ultimate goal of our model is to generate descriptions of image regions. During training, the input to our model is a set of images and their corresponding sentence descriptions (Figure 2). We first present a model that aligns segments of sentences to the visual regions that they describe through a multimodal embedding. We then treat these correspondences as training data for our multimodal Recurrent Neural Network model which learns to generate the descriptions.

3.1. Learning to align visual and language data

Our alignment model assumes an input dataset of images and their sentence descriptions. The key challenge to inferring the association between visual and textual data is that sentences written by people make multiple references to some particular, but unknown locations in the image. For example, in Figure 2, the words *"Tabby cat is leaning"* refer

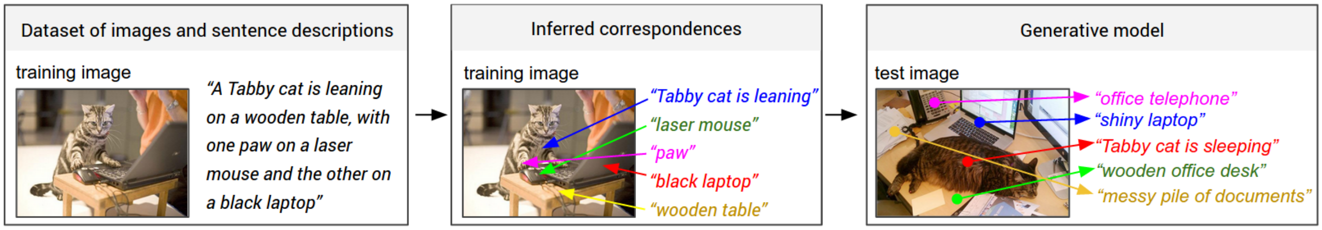

Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle) and then learns to generate novel descriptions (right).

to the cat, the words *"wooden table"* refer to the table, etc. We would like to infer these latent correspondences, with the goal of later learning to generate these snippets from image regions. We build on the basic approach of Karpathy et al. [18], who learn to ground dependency tree relations in sentences to image regions as part of a ranking objective. Our contribution is in the use of bidirectional recurrent neural network to compute word representations in the sentence, dispensing of the need to compute dependency trees and allowing unbounded interactions of words and their context in the sentence. We also substantially simplify their objective and show that both modifications improve ranking performance.

We first describe neural networks that map words and image regions into a common, multimodal embedding. Then we introduce our novel objective, which learns the embedding representations so that semantically similar concepts across the two modalities occupy nearby regions of the space.

3.1.1 Representing images

Following prior work [22, 18], we observe that sentence descriptions make frequent references to objects and their attributes. Thus, we follow the method of Girshick et al. [11] to detect objects in every image with a Region Convolutional Neural Network (RCNN). The CNN is pre-trained on ImageNet [3] and finetuned on the 200 classes of the ImageNet Detection Challenge [36]. To establish fair comparisons to Karpathy et al. [18], we use the top 19 detected locations and the whole image and compute the representations based on the pixels I_b inside each bounding box as follows:

$$v = W_m[CNN_{\theta_c}(I_b)] + b_m, \tag{1}$$

where $CNN(I_b)$ transforms the pixels inside bounding box I_b into 4096-dimensional activations of the fully connected layer immediately before the classifier. The CNN parameters θ_c contain approximately 60 million parameters and the architecture closely follows the network of Krizhevsky et al [21]. The matrix W_m has dimensions $h \times 4096$, where h is the size of the multimodal embedding space (h ranges from 1000-1600 in our experiments). Every image is thus represented as a set of h-dimensional vectors $\{v_i \mid i = 1 \dots 20\}$.

3.1.2 Representing sentences

To establish the inter-modal relationships, we would like to represent the words in the sentence in the same h-dimensional embedding space that the image regions occupy. The simplest approach might be to project every individual word directly into this embedding. However, this approach does not consider any ordering and word context information in the sentence. An extension to this idea is to use word bigrams, or dependency tree relations as previously proposed [18]. However, this still imposes an arbitrary maximum size of the context window and requires the use of Dependency Tree Parsers that might be trained on unrelated text corpora.

To address these concerns, we propose to use a bidirectional recurrent neural network (BRNN) [37] to compute the word representations. In our setting, the BRNN takes a sequence of N words (encoded in a 1-of-k representation) and transforms each one into an h-dimensional vector. However, the representation of each word is enriched by a variably-sized context around that word. Using the index $t = 1 \dots N$ to denote the position of a word in a sentence, the precise form of the BRNN we use is as follows:

$$x_t = W_w \mathbb{I}_t \tag{2}$$

$$e_t = f(W_e x_t + b_e) \tag{3}$$

$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f) \tag{4}$$

$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b) \tag{5}$$

$$s_t = f(W_d(h_t^f + h_t^b) + b_d). \tag{6}$$

Here, $\mathbb{I}_t$ is an indicator column vector that is all zeros except for a single one at the index of the t-th word in a word vocabulary. The weights W_w specify a word embedding matrix that we initialize with 300-dimensional word2vec [32] weights and keep fixed in our experiments due to overfitting concerns. Note that the BRNN consists of two independent streams of processing, one moving left to right (h_t^f) and the other right to left (h_t^b) (see Figure 3 for diagram). The final h-dimensional representation s_t for the t-th word is a function of both the word at that location and also its surrounding context in the sentence. Technically, every s_t is a function of all words in the entire sentence, but our empir-

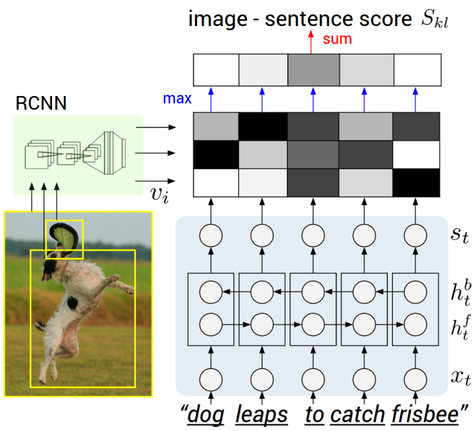

Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

ical finding is that the final word representations (s_t) align most strongly to the visual concept of the word at that location ($\mathbb{I}_t$). Our hypothesis is that the strength of influence diminishes with each step of processing since s_t is a more direct function of $\mathbb{I}_t$ than of the other words in the sentence.

We learn the parameters W_e, W_f, W_b, W_d and the respective biases b_e, b_f, b_b, b_d. A typical size of the hidden representation in our experiments ranges between 300-600 dimensions. We set the activation function f to the rectified linear unit (ReLU), which computes $f : x \mapsto max(0, x)$.

3.1.3 Alignment objective

We have described the transformations that map every image and sentence into a set of vectors in a common h-dimensional space. Since our labels are at the level of entire images and sentences, our strategy is to formulate an image-sentence score as a function of the individual scores that measure how well a word aligns to a region of an image. Intuitively, a sentence-image pair should have a high matching score if its words have a confident support in the image. In Karpathy et al. [18], they interpreted the dot product $v_i^T s_t$ between an image fragment i and a sentence fragment t as a measure of similarity and used these to define the score between image k and sentence l as:

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} max(0, v_i^T s_t). \tag{7}$$

Here, g_k is the set of image fragments in image k and g_l is the set of sentence fragments in sentence l. The indices k, l range over the images and sentences in the training set. Together with their additional Multiple Instance Learning objective, this score carries the interpretation that a sentence fragment aligns to a subset of the image regions whenever the dot product is positive. We found that the following reformulation simplifies the model and alleviates the need for additional objectives and their hyperparameters:

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t. \tag{8}$$

Here, every word s_t aligns to the single best image region. As we show in the experiments, this simplified model also leads to improvements in the final ranking performance. Assuming that $k = l$ denotes a corresponding image and sentence pair, the final max-margin, structured loss remains:

$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big]. \tag{9}$$

This objective encourages aligned image-sentences pairs to have a higher score than misaligned pairs, by a margin.

3.1.4 Decoding text segment alignments to images

Consider an image from the training set and its corresponding sentence. We can interpret the quantity $v_i^T s_t$ as the unnormalized log probability of the $t-$th word describing any of the bounding boxes in the image. However, since we are ultimately interested in generating snippets of text instead of single words, we would like to align extended, contiguous sequences of words to a single bounding box. Note that the naïve solution that assigns each word independently to the highest-scoring region is insufficient because it leads to words getting scattered inconsistently to different regions.

To address this issue, we treat the true alignments as latent variables in a Markov Random Field (MRF) where the binary interactions between neighboring words encourage an alignment to the same region. Concretely, given a sentence with N words and an image with M bounding boxes, we introduce the latent alignment variables $a_j \in \{1..M\}$ for $j = 1 \ldots N$ and formulate an MRF in a chain structure along the sentence as follows:

$$E(\mathbf{a}) = \sum_{j=1 \ldots N} \psi_j^U(a_j) + \sum_{j=1 \ldots N-1} \psi_j^B(a_j, a_{j+1}) \tag{10}$$

$$\psi_j^U(a_j = t) = v_i^T s_t \tag{11}$$

$$\psi_j^B(a_j, a_{j+1}) = \beta \mathbb{1}[a_j = a_{j+1}]. \tag{12}$$

Here, β is a hyperparameter that controls the affinity towards longer word phrases. This parameter allows us to interpolate between single-word alignments ($\beta = 0$) and

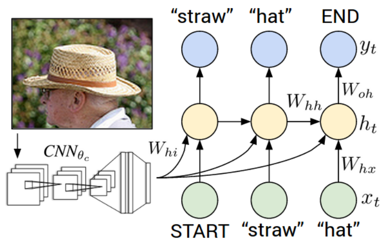

Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes an image, a word, the context from previous time steps and defines a distribution over the next word. START and END are special tokens.

aligning the entire sentence to a single, maximally scoring region when β is large. We minimize the energy to find the best alignments **a** using dynamic programming. The output of this process is a set of image regions annotated with segments of text. We now describe an approach for generating novel phrases based on these correspondences.

3.2. Multimodal Recurrent Neural Network for generating descriptions

In this section we assume an input set of images and their textual descriptions. These could be full images and their sentence descriptions, or regions and text snippets as discussed in previous sections. The key challenge is in the design of a model that can predict a variable-sized sequence of outputs. In previously developed language models based on Recurrent Neural Networks (RNNs) [31, 41, 5], this is achieved by defining a probability distribution of the next word in a sequence, given the current word and context from previous time steps. We explore a simple but effective extension that additionally conditions the generative process on the content of an input image. More formally, the RNN takes the image pixels I and a sequence of input vectors $(x_1, \dots, x_T)$. It then computes a sequence of hidden states $(h_1, \dots, h_t)$ and a sequence of outputs $(y_1, \dots, y_t)$ by iterating the following recurrence relation for $t = 1$ to T:

$$b_v = W_{hi}[CNN_{\theta_c}(I)] \tag{13}$$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + b_v) \tag{14}$$

$$y_t = softmax(W_{oh}h_t + b_o). \tag{15}$$

In the equations above, $W_{hi}, W_{hx}, W_{hh}, W_{oh}$ and b_h, b_o are a set of learnable weights and biases. The output vector y_t has the size of the word dictionary and one additional dimension for a special END token that terminates the generative process. Note that we provide the image context vector b_v to the RNN at every iteration so that it does not have to remember the image content while generating words.

RNN training. The RNN is trained to combine a word (x_t), the previous context (h_{t-1}) and the image information (b_v) to predict the next word (y_t). Concretely, the training proceeds as follows (refer to Figure 4): We set $h_0 = \vec{0}$, x_1 to a special START vector, and the desired label y_1 as the first word in the sequence. In particular, we use the word embedding for "the" as the START vector x_1. Analogously, we set x_2 to the word vector of the first word and expect the network to predict the second word, etc. Finally, on the last step when x_T represents the last word, the target label is set to a special END token. The cost function is to maximize the log probability assigned to the target labels.

RNN at test time. The RNN predicts a sentence as follows: We compute the representation of the image b_v, set $h_0 = 0$, x_1 to the embedding of the word "the", and compute the distribution over the first word y_1. We sample from the distribution (or pick the argmax), set its embedding vector as x_2, and repeat this process until the END token is generated.

3.3. Optimization

We use Stochastic Gradient Descent with mini-batches of 100 image-sentence pairs and momentum of 0.9 to optimize the alignment model. We cross-validate the learning rate and the weight decay. We also use dropout regularization in all layers except in the recurrent layers [47]. The generative RNN is more difficult to optimize, party due to the word frequency disparity between rare words, and very common words (such as the END token). We achieved the best results using RMSprop [42], which is an adaptive step size method that scales the gradient of each weight by a running average of its gradient magnitudes.

4. Experiments

Datasets. We use the Flickr8K [14], Flickr30K [46] and COCO [29] datasets in our experiments. These datasets contain 8,000, 31,000 and 123,000 images respectively and each is annotated with 5 sentences using Amazon Mechanical Turk. For Flickr8K and Flickr30K, we use 1,000 images for validation, 1,000 for testing and the rest for training (consistent with [14, 18]). For COCO we use 5,000 images for both validation and testing.

Data Preprocessing. We convert all sentences to lowercase, discard non-alphanumeric characters, and filter out the articles "an", "a", and "the" for efficiency. Our word vocabulary contains 20,000 words.

4.1. Image-Sentence Alignment Evaluation

We first investigate the quality of the inferred text and image alignments. As a proxy for this evaluation we perform ranking experiments where we consider a withheld set of images and sentences and then retrieve items in one modality given a query from the other. We use the image-sentence score S_{kl} (Section 3.1.3) to evaluate a compatibility score between all pairs of test images and sentences. We then report the median rank of the closest ground truth result in the

Model	Image Annotation R@1	R@5	R@10	Med r	Image Search R@1	R@5	R@10	Med r
Flickr8K								
DeViSE (Frome et al. [10])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [39])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
DeFrag (Karpathy et al. [18])	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Our implementation of DeFrag [18]	13.8	35.8	48.2	10.4	9.5	28.2	40.3	15.6
Our model: DepTree edges	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
Our model: BRNN	**16.5**	**40.6**	**54.2**	**7.6**	**11.8**	**32.1**	**44.7**	**12.4**
Flickr30K								
DeViSE (Frome et al. [10])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [39])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
DeFrag (Karpathy et al. [18])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [18]	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
Our model: DepTree edges	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
Our model: BRNN	**22.2**	**48.2**	**61.4**	**4.8**	**15.2**	**37.7**	**50.5**	**9.2**
COCO								
Our model: 1K test images	29.4	62.0	75.9	2.5	20.9	52.8	69.2	4.0
Our model: 5K test images	11.8	32.5	45.4	12.2	8.9	24.9	36.3	19.5

Table 1. Image-Sentence ranking experiment results. **R@K** is Recall@K (high is good). **Med** r is the median rank (low is good). In the results for our models, we take the top 5 validation set models, evaluate each independently on the test set and then report the average performance. The standard deviations on the recall values range from approximately 0.5 to 1.0.

list and Recall @K, which measures the fraction of times a correct item was found among the top K results. The results of these experiments can be found in Table 1, and example retrievals in Figure 5. We now highlight some of the takeaways.

Our full model outperforms previous work. We compare our full model ("Our model: BRNN") to the following baselines: DeViSE [10] is a model that learns a score between words and images. As the simplest extension to the setting of multiple image regions and multiple words, Karpathy et al. [18] averaged the word and image region representations to obtain a single vector for each modality. Socher et al. [39] is trained with a similar objective, but instead of averaging the word representations, they merge word vectors into a single sentence vector with a Recursive Neural Network. DeFrag are the results reported by Karpathy et al. [18]. Since we use different word vectors, dropout for regularization and different cross-validation ranges (including larger embedding sizes), we re-implemented their cost function for a fair comparison ("Our implementation of DeFrag"). In all of these cases, our full model ("Our model: BRNN") provides consistent improvements.

Our simpler cost function improves performance. We now try to understand the sources of these improvements. First, we removed the BRNN and used dependency tree relations exactly as described in Karpathy et al. [18] ("Our model: DepTree edges"). The only difference between this model and "Our reimplementation of DeFrag" is the new, simpler cost function introduced in Section 3.1.3. We see that our formulation shows consistent improvements.

BRNN outperforms dependency tree relations. Furthermore, when we replace the dependency tree relations with the BRNN, we observe additional performance improvements. Since the dependency relations were shown to work better than single words and bigrams [18], this suggests that the BRNN is taking advantage of contexts longer than two words. Furthermore, our method does not rely on extracting a Dependency Tree and instead uses the raw words directly.

COCO results for future comparisons. The COCO dataset has only recently been released, and we are not aware of other published ranking results. Therefore, we report results on a subset of 1,000 images and the full set of 5,000 test images for future comparisons.

Qualitative. As can be seen from example groundings in Figure 5, the model discovers interpretable visual-semantic correspondences, even for small or relatively rare objects such as *"seagulls"* and *"accordion"*. These details would be missed by models that only reason about full images.

4.2. Evaluation of Generated Descriptions

We have demonstrated that our alignment model produces state of the art ranking results and qualitative experiments suggest that the model effectively infers the alignment between words and image regions. Our task is now to synthesize these sentence snippets given new image regions. We evaluate these predictions with the BLEU [35] score, which despite multiple problems [14, 22] is still considered to be the standard metric of evaluation in this setting. The BLEU score evaluates a *candidate* sentence by measuring the fraction of n-grams that appear in a set of *references*.

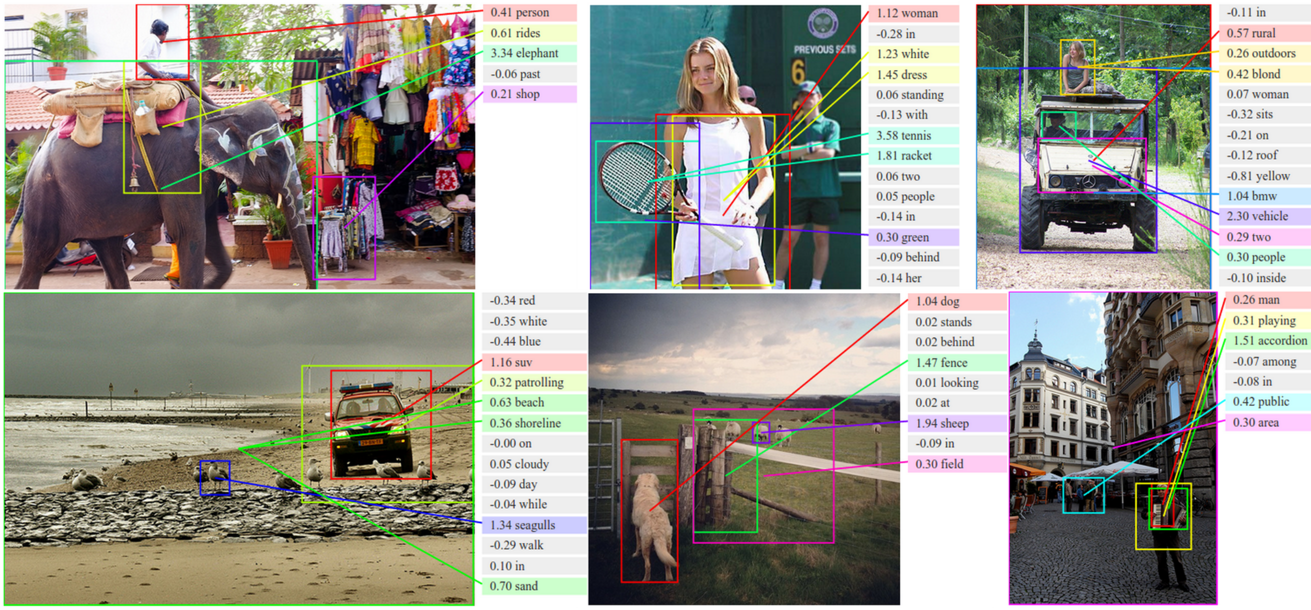

Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

	Flickr8K			**Flickr30K**			**COCO**		
Method of generating text	B-1	B-2	B-3	B-1	B-2	B-3	B-1	B-2	B-3
Human agreement	0.59	0.35	0.16	0.64	0.36	0.16	0.57	0.31	0.13
Ranking: Nearest Neighbor	0.29	0.11	0.03	0.27	0.08	0.02	0.32	0.11	0.03
Generating: RNN	**0.42**	**0.19**	**0.06**	**0.45**	**0.20**	**0.06**	**0.50**	**0.25**	**0.12**

Table 2. BLEU score evaluation of full image predictions on 1,000 images. **B-n** is BLEU score that uses up to n-grams (high is good).

Our multimodal RNN outperforms retrieval baseline. We first verify that our multimodal RNN is rich enough to support sentence generation for full images. In this experiment, we trained the RNN to generate sentences on full images from Flickr8K, Flickr30K, and COCO datasets. Then at test time, we use the first four out of five sentences as references and the fifth one to evaluate human agreement. We also compare to a ranking baseline which uses the best model from the previous section (Section 4.1) to annotate each test image with the highest-scoring sentence from the training set. The quantitative results of this experiment are in Table 2. Note that the RNN model confidently outperforms the retrieval method. This result is especially interesting in COCO dataset, since its training set consists of more than 600,000 sentences that cover a large variety of descriptions. Additionally, compared to the retrieval baseline which compares each image to all sentences in the training set, the RNN takes a fraction of a second to evaluate.

We show example fullframe predictions in Figure 6. Our generative model (shown in blue) produces sensible descriptions, even in the last two images that we consider to be failure cases. Additionally, we verified that none of these sentences appear in the training set. This suggests that the model is not simply memorizing the training data. However, there are 20 occurrences of “man in black shirt” and 60 occurrences of “is paying guitar”, which the model may have composed to describe the first image.

Region-level evaluation. Finally, we evaluate our region RNN which was trained on the inferred, intermodal correspondences. To support this evaluation, we collected a new dataset of region-level annotations. Concretely, we asked 8 people to label a subset of COCO test images with region-level text descriptions. The labeling interface consisted of a single test image, and the ability to draw a bounding box and annotate it with text. We provided minimal constraints and instructions, except to “describe the content of each box” and we encouraged the annotators to describe a large variety of objects, actions, stuff, and high-level concepts. The final dataset consists of 1469 annotations in 237 images. There are on average 6.2 annotations per image, and each one is on average 4.13 words long.

We compare three models on this dataset: The region RNN model, a fullframe RNN model that was trained on full images and sentences, and a ranking baseline. To predict descriptions with the ranking baseline, we take the number of words in the shortest reference annotation and search the training set sentences for the highest scoring segment of text

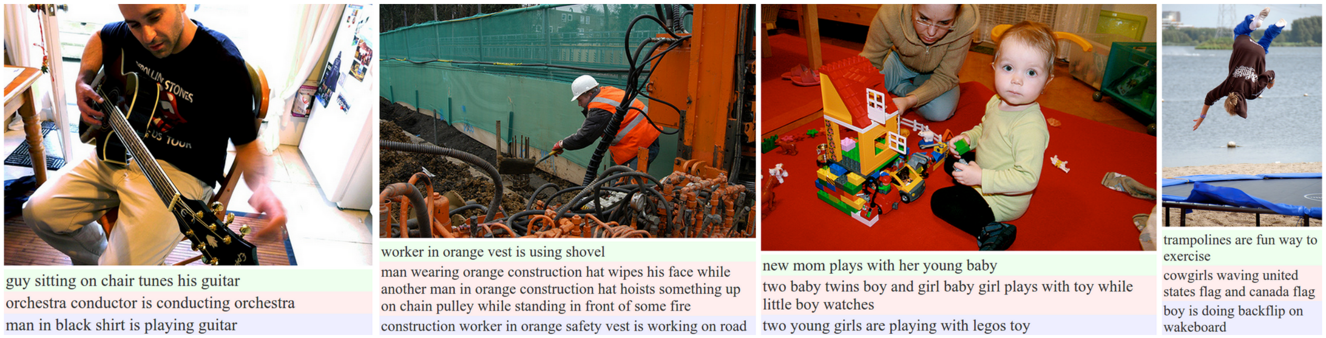

Figure 6. Example fullframe predictions. Green: human annotation. Red: Most compatible sentence in the training set (i.e. ranking baseline). Blue: Generated sentence using the fullframe multimodal RNN. We provide more examples in the supplementary material.

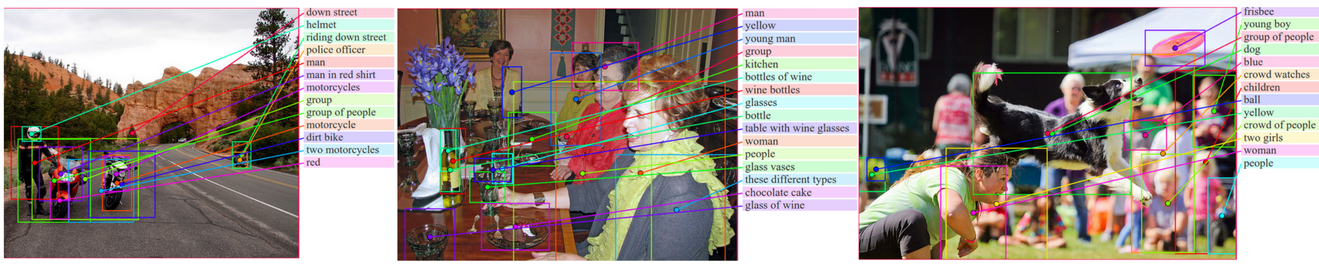

Figure 7. Example region predictions. We use our region-level multimodal RNN to generate text (shown on the right of each image) for some of the bounding boxes in each image. The lines are grounded to centers of bounding boxes and the colors are chosen arbitrarily.

Method of generating text	B-1	B-2	B-3
Human agreement	0.54	0.33	0.16
Ranking: Nearest Neighbor	0.14	0.03	**0.07**
Generating: Full frame model	0.12	0.03	0.01
Generating: Region level model	**0.17**	**0.05**	0.01

Table 3. BLEU score evaluation of image region annotations.

of that length. This ensures that the ranking baseline does not accumulate any brevity penalty in its BLEU scores.

We report the results in Table 3, and show example predictions in Figure 7. To reiterate the difficulty of the task, consider that the phrase "table with wine glasses" that is generated on the middle image in Figure 7 only occurs in the training set 30 times. Each time it may have a different appearance and each time it may occupy a few (or none) of the bounding boxes. To generate this string for the image, the model had to correctly infer the correspondence and then learn to generate this string.

There are several takeaways from Table 3. First, the human agreement baseline displays stronger performance relative to our RNN models on the region-level task than the full image task. Additionally, the performance of the ranking baseline is now competitive with the RNN model. One possible explanation is that the snippets of text are shorter in this dataset, which makes it easier to find a good match in the training sentences. We believe that these results are an encouraging first step towards the task of dense scene descriptions, and we release our annotations so that future work can compare to these results.

4.3. Limitations

Although our results are encouraging, the RNN model is subject to multiple limitations. First, the model can only generate a description of one input array of pixels at a fixed resolution. A more sensible approach might be to use multiple saccades around the image to identify all entities, their mutual interactions and wider context before generating a description. Additionally, the RNN (as formulated in Equation 13) couples the visual and language domains in the hidden representation only through additive interactions, which are known to be less expressive than more complicated multiplicative interactions [41]. Lastly, going directly from an image-sentence dataset to region-level annotations as part of a single model that is trained end-to-end with a single objective remains an open problem.

5. Conclusions

We introduced a model that generates free-form descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hardcoded assumptions. Our approach relied on a novel structured objective that aligned the visual and textual modalities through a common, multimodal embedding. We showed that this approach leads to consistent state of the art performance on ranking experiments across three datasets. We then described a multimodal Recurrent Neural Network architecture that generates textual descriptions based on image regions, and evaluated its performance with fullframe and region-level experiments. We showed that in both cases the multimodal RNN outperforms retrieval baselines.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003. 2

[2] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*. Springer, 2006. 2

[3] 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. 3

[4] D. Elliott and F. Keller. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302, 2013. 2

[5] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. 5

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1

[7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*. 2010. 1, 2

[8] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10, 2007. 1

[9] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 2

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2, 6

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[12] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1–8. IEEE, 2009. 2

[13] A. Gupta and P. Mannem. From image annotation to image description. In *Neural information processing*. Springer, 2012. 2

[14] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013. 1, 2, 5, 6

[15] R. JeffreyPennington and C. Manning. Glove: Global vectors for word representation. 2

[16] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[17] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011. 2

[18] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014. 2, 3, 4, 5, 6

[19] R. Kiros, R. S. Zemel, and R. Salakhutdinov. Multimodal neural language models. *ICML*, 2014. 2

[20] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 2

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3

[22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 1, 2, 3, 6

[23] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, 2012. 2

[24] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2(10):351–362, 2014. 2

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[26] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*. IEEE, 2009. 2

[27] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *CoNLL*, 2011. 2

[28] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. 2014. 2

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 1, 5

[30] C. Matuszek*, N. FitzGerald*, L. Zettlemoyer, L. Bo, and D. Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012. 2

[31] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, 2010. 2, 5

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 2, 3

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

[34] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 6

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein,

A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014. 1, 2, 3

[37] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 1997. 3

[38] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 2

[39] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2014. 2, 6

[40] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012. 2

[41] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *ICML*, 2011. 2, 5, 8

[42] T. Tieleman and G. E. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude., 2012. 5

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

[44] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. 2

[45] M. Yatskar, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. *Lexical and Computational Semantics*, 2014. 2

[46] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 1, 5

[47] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 5

[48] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. *ICCV*, 2013. 2