Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy, *Member, IEEE* and Li Fei-Fei, *Member, IEEE*

Abstract—We present a model that generates natural language descriptions of images and their regions. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Our alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks (RNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. We demonstrate that our alignment model produces state of the art results in retrieval experiments on Flickr8K, Flickr30K and MSCOCO datasets. We then show that the generated descriptions outperform retrieval baselines on both full images and on a new dataset of region-level annotations. Finally, we conduct large-scale analysis of our RNN language model on the Visual Genome dataset of 4.1 million captions and highlight the differences between image and region-level caption statistics.

Index Terms—Image captioning, deep neural networks, visual-semantic embeddings, recurrent neural network, language model

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 [1]. 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 [2], [3]. However, while closed vocabularies of visual concepts constitute a convenient modeling assumption, they are 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 [4], [5]. However, these models often rely on hard-coded visual concepts and explicitly defined sentence templates, which limits their variety. Moreover, the focus of these works has been on reducing complex visual scenes into a single sentence, which we view as an unnecessary restriction.

In this work, we strive to take a step towards the goal of generating dense descriptions of images (See concept Fig. 1). The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason 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 learn from patterns present in the raw training data in an end-to-end fashion. The second, practical challenge is that datasets of image captions are available in large quantities on the internet [6], [7], [8], 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 in a language modeling framework. Concretely, our contributions are two fold:

- 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 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.

Our code, data and annotations are publicly available.[1]

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. [9] and Socher et al.

• *The authors are with the Computer Science Department, Stanford University, Stanford, CA 94305. E-mail: {karpathy, feifeili}@cs.stanford.edu.*

Manuscript received 16 Dec. 2015; revised 16 June 2016; accepted 25 July 2016. Date of publication 4 Aug. 2016; date of current version 2 Mar. 2017. Recommended for acceptance by K. Grauman, A. Torralba, E. Learned-Miller, and A. Zisserman.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2598339

1. cs.stanford.edu/people/karpathy/deepimagesent/

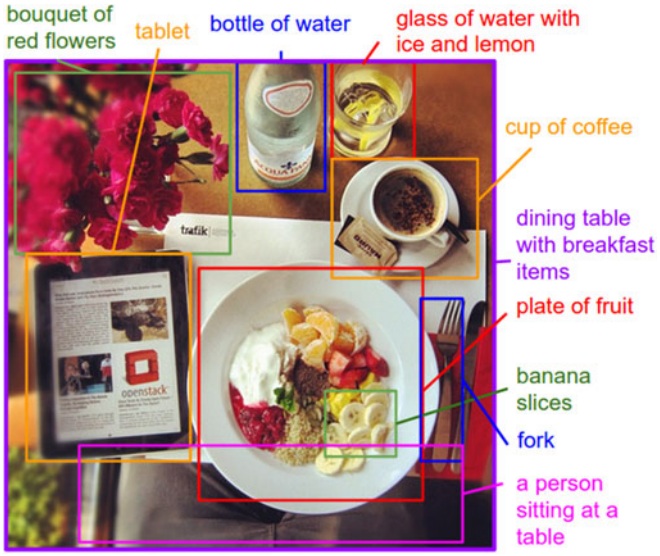

Fig. 1. Motivation/Concept figure: Our model treats language as a rich label space and generates descriptions of image regions.

[10] studied the multimodal correspondence between words and images to annotate segments of images. Several works [11], [12], [13], [14] 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 Descriptions. The task of describing images with sentences has also been explored. 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 [5], [6], [15], [16], [17], or where training annotations are broken up and stitched together [18], [19], [20]. Several approaches generate image captions based on fixed templates that are filled based on the content of the image [4], [5], [21], [22], [23], [24], [25] or generative grammars [26], [27], but this approach limits the variety of possible outputs. More closely related to us is the approach of Kiros et al. [28], [29] who developed a log-bilinear model that generates captions based on a finite length context window. Several approaches that alleviate the finite context constraint have been developed simultaneously with this work based on a Recurrent Neural Network language model [30], [31], [32], [33], [34]. Our RNN is simpler than most of these approaches but also suffers slightly in performance. We quantify this comparison in our experiments.

Grounding Natural Language in Images. A number of approaches have been developed for grounding text in the visual domain [35], [36], [37], [38], [39], [40]. Our approach is inspired by Frome et al. [41] who associate words and images through a semantic embedding. More closely related is the work of Karpathy et al. [42], 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) [43], [44] have recently emerged as a powerful class of models for image classification and object detection [3]. On the sentence side, our work takes advantage of pretrained word vectors [45], [46], [47] to obtain low-dimensional representations of words. Finally, Recurrent Neural Networks have been previously used or proposed in the context of language modeling [48], [49], [50], 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. We first present a model that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. We then treat these correspondences as training data for a second, multimodal Recurrent Neural Network model that learns to generate the snippets (refer to Fig. 2 for visual overview).

3.1 Learning to Align Visual and Language Data

Our alignment model assumes an input dataset of images and their sentence descriptions. Our key insight is that sentences written by people make frequent references to some particular, but unknown location in the image. For example, in Fig. 2, the words *"Tabby cat is leaning"* refer to the cat, the words *"wooden table"* refer to the table, etc. We would like to infer these latent correspondences, with the eventual goal of later learning to generate these snippets from image regions. We build on the approach of Karpathy et al. [42], who learn to ground dependency tree relations to image regions with a ranking objective. Our contribution is in the use of bidirectional recurrent neural network (BRNN) 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

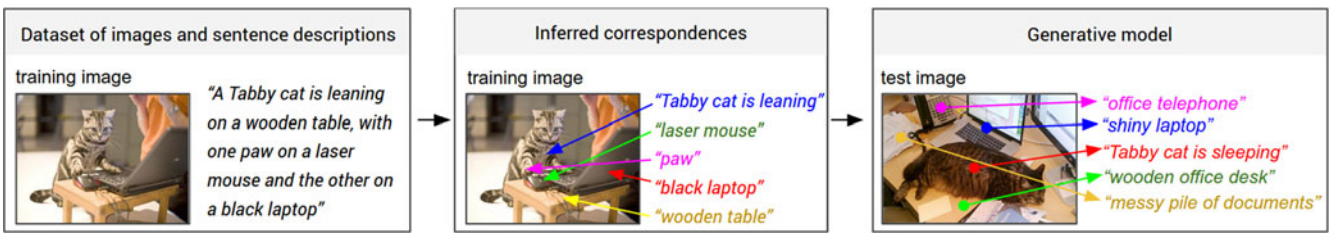

Fig. 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle, Section 3.1) and then learns to generate novel descriptions (right, Section 3.2).

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 can be found in nearby regions of the embedding space.

3.1.1 Representing Images

Following prior work [4], [42], we observe that sentence descriptions make frequent references to objects and their attributes. Thus, we follow the method of Girshick et al. [51] to detect objects in every image with a Region-based Convolutional Neural Network (R-CNN). The CNN is pre-trained on ImageNet [52] and finetuned on the 200 classes of the ImageNet Detection Challenge [3]. Following Karpathy et al. [42], we use the top 19 detected locations in addition to 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 4,096-dimensional activations of the fully connected layer immediately before the classifier. The CNN parameters θ_c contain approximately 60 million parameters. The matrix W_m has dimensions $h \times 4{,}096$, where h is the size of the multimodal embedding space (h ranges from 1,000-1,600 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 [42]. 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 [53] to compute the word representations. 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 looks as follows:

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

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

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

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

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

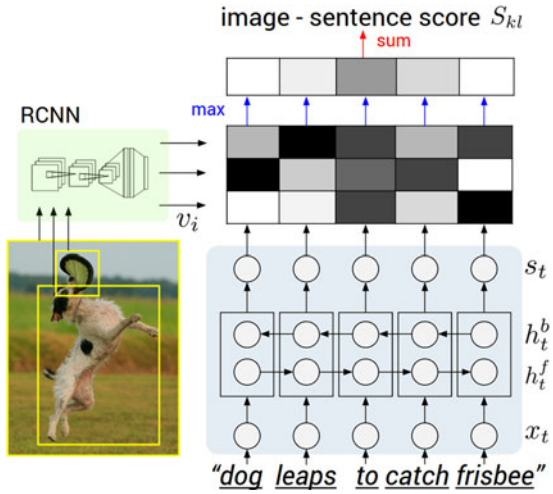

Fig. 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally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

Here, $\mathbb{I}_t$ is an indicator column vector that has a single one at the index of the tth word in a word vocabulary. The weights W_w specify a word embedding matrix that we initialize with 300-dimensional word2vec [47] weights and keep fixed due to overfitting concerns. However, in practice we find little change in final performance when these vectors are trained, even from random initialization. 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 Fig. 3 for diagram). The final h-dimensional representation s_t for the tth 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 empirical finding by manual inspection is that the final word representations (s_t) align most strongly to the visual concept of the word at that location ($\mathbb{I}_t$).

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 the supervision is at the level of entire images and sentences, our strategy is to formulate an image-sentence score as a function of the individual region-word scores. Intuitively, a sentence-image pair should have a high matching score if its words have a confident support in the image. The model of Karpathy et al. [42] interprets the dot product $v_i^T s_t$ between the ith region and tth word as a measure of similarity and use it 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 (i.e., words) 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

$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big] + \alpha \|\theta\|_2^2. \tag{9}$$

This objective encourages aligned image-sentences pairs to have a higher score than misaligned pairs, by a margin. Note that similar to the SVM objective, it is safe to use 1 as a (fixed) value instead of an additional margin size hyperparameter because the L2 weight decay regularization on the weights will encourage the weights to simultaneously shrink towards zero while also meeting the required margins. Therefore, the choice of the margin can be absorbed into the choice of the regularization strength α.

3.1.4 Decoding Text Snippet Alignments to Images

Recall that our goal is to use the ranking model to create an intermediate dataset of regions annotated with snippets of text taken from the captions. These region-snippet correspondences will then form the training data for the Multimodal Recurrent Neural Network, which will learn to generate these snippet descriptions given regions. Now, consider an image from the training set and its corresponding sentence. Given a ranking model trained with the objective discussed in the previous section (Section 3.1.3), we can interpret the quantity $v_i^T s_t$ as the unnormalized log probability of the tth word in the caption as describing (or aligning to) the ith bounding box in the image. Note that at this point the naïve solution that greedily assigns each word independently to its highest-scoring region is insufficient because it leads to words getting scattered inconsistently to different regions. This process would, in other words, produce an intermediate dataset of regions annotated with snippets of text that might very often only be one word long.

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

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

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

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

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 aligning the entire sentence to a single, maximally scoring region when β is large. We minimize the energy to find the best alignments $\mathbf{a}$ using dynamic programming [54]. 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 captioned images. These could be full images and their sentence descriptions, or regions and text snippets, as inferred by the ranking model and described in the previous section. The key challenge is in the design of a model that can predict a variable-sized sequence of words given an image. In previously developed language models based on Recurrent Neural Networks (RNNs) [49], [50], [55], 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 language model's generative process on the content of an input image.

More formally, during training our Multimodal RNN takes the raw image pixels I and a sequence of input vectors $(x_1, \dots, x_T)$. The vectors x_t may typically be 300-dimensional and represent each word in an input caption. These vectors can be learned with backpropagation for each word, or if the data size is a concern they can be set according to a separate criterion (e.g., word2vec [47]). The Multimodal RNN then computes a sequence of hidden states $(h_1, \dots, h_T)$ and a sequence of output vectors $(y_1, \dots, y_T)$ by iterating the following recurrence relation for $t = 1$ to T:

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

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

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

In the equations above, $W_{hi}, W_{hx}, W_{hh}, W_{oh}$ and b_h, b_o are learnable parameters, $CNN_{\theta_c}(I)$ is the last feature layer of a CNN (after non-linearity), and f is the activation function (we use ReLU). The output vectors y_t are interpreted as holding the (unnormalized) log probabilities of all words in the dictionary and an additional special END token. Note

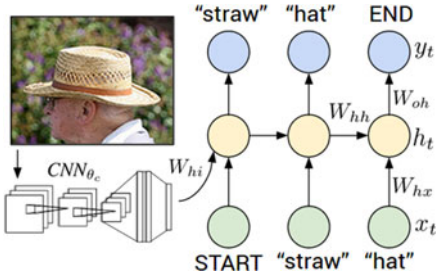

Fig. 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

that we provide the image context vector b_v to the RNN only at the first iteration (modulated by a multiplicative interaction with the delta function $\mathbb{1}(t=1)$), which we found to work better than at each time step (possibly due to easier overfitting). In practice we also found that it can help to also pass $W_{hx}x_t$ through the activation function. A typical size of the hidden layer of the RNN is 512 neurons.

RNN Training. The RNN is trained to take a word (x_t) and the previous context (h_{t-1}) to predict the next word in the sequence (y_t). The RNN's predictions are conditioned on the image information (b_v) via bias interactions on the first step.

The training proceeds as follows (refer to Fig. 4): We set $h_0 = \vec{0}$, x_1 to a special START vector, and the desired label for y_1 to be the first word in the ground truth sequence. Analogously, at the second time step $t=2$, 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 (normalized) log probability assigned to the target labels (i.e., the Softmax classifier, or the *cross-entropy* loss).

Finally, note that the number of time steps T is one greater than the number of words in the ground truth caption for the image due to the offset introduced by the START and END tokens. For example, a caption with seven words would require $T=8$ applications of the recurrence. In practice we upper bound T to 16. When constructing mini-batches of data during optimization we keep track of the length of each individual caption in the minibatch. We then forward the RNN for 16 timesteps on each example in the minibatch in parallel and backpropagate the gradients only in the "occupied" parts of captions that contain a word. This approach wastes some computation but results in faster convergence because minibatches can be processed more efficiently than individual examples.

RNN at Test Time. To predict a caption for a test image we compute the image representation b_v, set $h_0 = 0$, x_1 to the START vector and compute the distribution over the first word in the caption, y_1. We normalize the distribution and sample a word (or pick the argmax), set x_2 to be its embedding vector, and repeat this process until the END token is generated. In practice we found that beam search decoding [56] (e.g., with beam size 3) can improve results since it normally produces a more globally likely caption for an input image, which might otherwise not be produced in a greedy manner one word at a time.

3.3 Optimization

To optimize the alignment model we use SGD with mini-batches of 100 image-sentence pairs and momentum of 0.9. We cross-validate the learning rate and the L2 weight decay. We also use dropout regularization in all layers except in the recurrent layers [57] and clip gradients elementwise at five (assuming that the loss/gradients are normalized across a batch but not across time).

The Multimodal RNN is more difficult to optimize, party due to the word frequency disparity between rare words and common words (e.g., "a" or the END token). We achieved good results with *RMSprop* [58], which is an adaptive step size method that scales the update of each weight by a running average of its gradient norm. We also experimented with SGD, SGD+Momentum, Adadelta and Adagrad, but found these to work worse. In later experiments we also used Adam, which slightly outperformed RMSProp. We found that *clipping the gradients* (we only experimented with simple per-element clipping) at an appropriate value provided consistent but small improvements. Since the distribution of the words in English is highly non-uniform, the model spends the first few iterations mostly learning the biases for the Softmax classifier such that it is predicting every word with the appropriate dataset frequency. We found that we could obtain faster convergence early in the training (and nicer-looking loss curves) by explicitly *initializing the biases* of all words in the dictionary (in the Softmax classifier) to log probability of their occurrence in the training data. Therefore, with small weights and biases set appropriately the model right away predicts word at random according to their chance distribution. Lastly, we experimented with initializing word representations x_i with word2vec vectors [47], but found that it was sufficient to train these vectors from random initialization without changes in the final performance. Moreover, we found that the word2vec vectors have some unappealing properties when used in multimodal language-visual tasks. For instance, all colors (e.g., red, blue, green) are clustered nearby in the word2vec representation because they are relatively interchangeable in most language contexts. However, their visual instantiations are very different.

4 Experiments

Datasets. The image captioning datasets we use in our experiments are the Flickr8K [6], Flickr30K [8] and MSCOCO [7]. These datasets contain 8,000, 31,000 and 123,000 images respectively and each is annotated with five sentences using Amazon Mechanical Turk (AMT). For Flickr8K and Flickr30K, we use 1,000 images for validation, 1,000 for testing and the rest for training (consistent with [6], [42]). For MSCOCO we use 5,000 images for both validation and testing.

We also evaluate our captioning model in isolation on the Visual Genome (VG) dataset [59], which contains region-level annotations that consist of a bounding box and a snippet text caption collected with AMT. This dataset contains 94,313 images and 4,100,413 snippets of text (approx. 43.5 per image), each grounded to a bounding box of an image. The vast majority of the images are taken from the MSCOCO dataset, and our splits again contain 5,000 images for both validation and testing, and the rest for training.

TABLE 1
Image-Sentence Ranking Experiment Results

Model	Image Annotation				Image Search			
	R@1	R@5	R@10	Med *r*	R@1	R@5	R@10	Med *r*
Flickr8K								
Hodosh et al. [6]	8.3	21.6	30.3	34	7.6	20.7	30.1	38
Kiros et al. [28]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
Mao et al. [33]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
Our implementation of DeFrag [42]	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								
SDT-RNN (Socher et al. [17])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [28]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [33]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Donahue et al. [31]	17.5	40.3	50.8	9	-	-	-	-
DeFrag (Karpathy et al. [42])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [42]	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**
Vinyals et al. [34] (more powerful CNN)	23	-	63	5	17	-	57	8
MSCOCO								
Our model: 1K test images	38.4	69.9	80.5	1.0	27.4	60.2	74.8	3.0
Our model: 5K test images	16.5	39.2	52.0	9.0	10.7	29.6	42.2	14.0

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 five 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.*

Data Preprocessing. We convert all sentences to lowercase, discard non-alphanumeric characters. We filter words to those that occur at least 5 times in the training set, which results in 2,538, 7,414, and 8,791 words for Flickr8k, Flickr30K, and MSCOCO datasets respectively.

For the VG dataset we use the count threshold of 15 instead, leading to a dictionary of 10,497 words. Unlike the other three datasets, the VG region captions dataset has only a single caption for each bounding box. However, some of these bounding boxes can heavily overlap when they describe an identical visual concept. To improve the caption evaluation process which relies on matching a prediction to words in a reference caption, we try to merge all captions that refer to the same region. We do so by grouping all captions that heavily overlap (based on an intersection over union (IoU) threshold of 0.7), and assign them to a new box obtained by taking the average across all grouped boxes.

4.1 Image-Sentence Alignment Evaluation

We first investigate the quality of the inferred image alignments (produced from the ranking model in Section 3.1) with ranking experiments. We consider a withheld set of images and sentences and retrieve items in one modality given a query from the other by sorting based on the image-sentence score S_{kl} (Section 3.1.3). We report the median rank of the closest ground truth result in the list and Recall@K, which measures the fraction of times a correct item was found among the top K results. The result of these experiments can be found in Table 1, and example retrievals in Fig. 5. We now highlight some of the takeaways.

Our Full Model Outperforms Previous Work. First, our full model ("Our model: BRNN") outperforms Socher et al. [17] who trained with a similar loss but used a single image representation and a Recursive Neural Network over the sentence. A similar loss was adopted by Kiros et al. [28], who use an LSTM [60] to encode sentences. We list their performance with a CNN that is equivalent in power (AlexNet [43]) to the one used in this work, though similar to [34] they outperform our model with a more powerful CNN (VGGNet [61], GoogLeNet [62]). Our performance is also significantly higher than Hodosh et al. [6] who used image and sentence features and a Kernel Canonical Correlation Analysis (KCCA) loss to align feature vectors of the two domains. However, it is important to note that our use of ConvNets is likely responsible for most of this difference. Inspired by this work, it is possible that a CCA alignment loss could additionally improve our results if we swapped it for our simpler and more common max-margin loss, but we do not investigate this extension in this work. Finally, "DeFrag" are the results reported by Karpathy et al. [42]. Since we use different word vectors, dropout for regularization and different cross-validation ranges and larger embedding sizes, we also re-implemented their loss for a fair comparison ("Our implementation of DeFrag"). In summary, compared to other previous work that uses AlexNets, our full model shows consistent improvements.

Our Simpler Cost Function Improves Performance. We strive to better understand the source of our performance. First, we removed the BRNN and used dependency tree relations exactly as described in Karpathy et al. [42] ("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

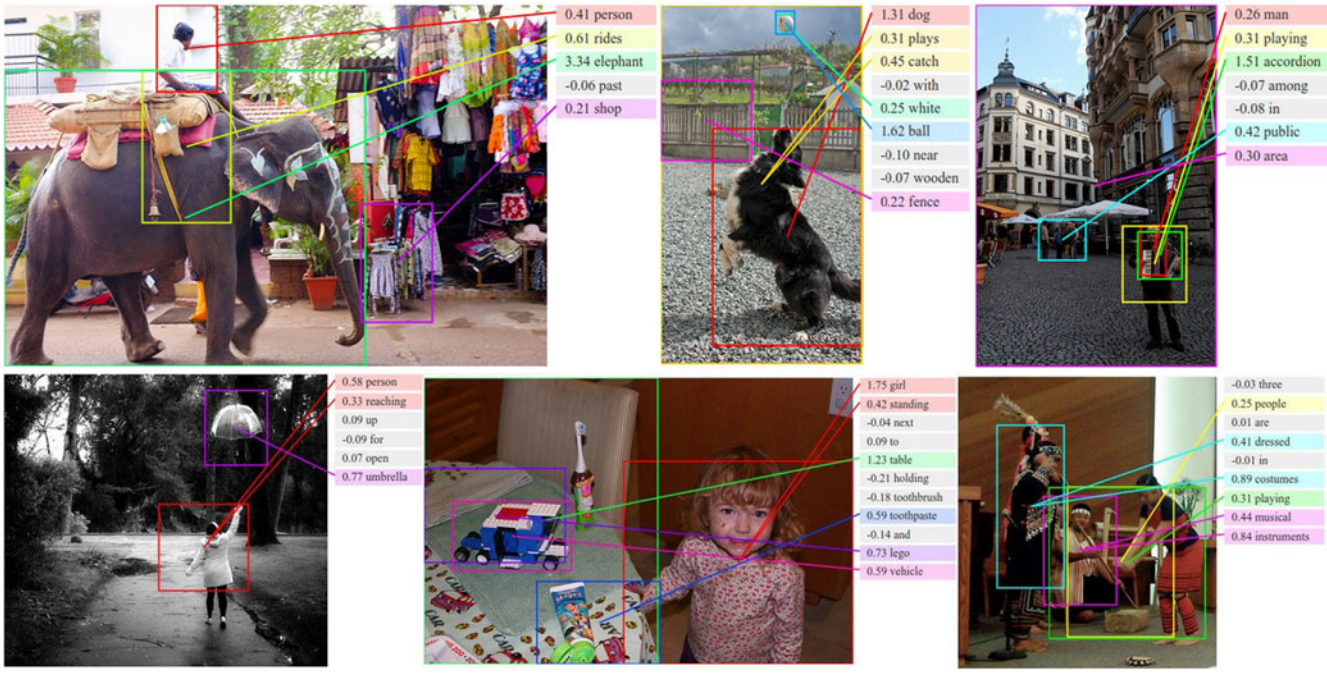

Fig. 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring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

better than single words and bigrams [42], 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.

MSCOCO Results. For MSCOCO we report results on a subset of 1,000 images and the full set of 5,000 test images. Note that the 5,000 images numbers are lower since Recall@K is a function of test set size (e.g., it is more difficult to retrieve the correct image in top 10 from a pool of 5,000 candidates than from a pool of 1,000 candidates).

Qualitative. As can be seen from example groundings in Fig. 5, the model discovers interpretable visual-semantic correspondences, even for small or relatively rare objects such as an *"accordion"*. These would be likely missed by models that only reason about full images. Note that one limitation of our model is that it does not explicitly handle or support counting. For instance, the last example we show contains the phrase *"3 people"*. These words should align to the 3 people in the image, but our model puts the bounding box around two of the people. In doing so, the model may be taking advantage of the BRNN structure to modify the "people" vector to preferentially align to regions that contain multiple people. However, this is still unsatisfying because such spurious detections only exist as a result of an error in the RCNN inference process, which presumably failed to localize the individual people.

In another qualitative retrieval experiment we consider a query text snippet and then retrieve image regions across the entire test set that has the highest average score with each word in the query. We show examples of such queries in Fig. 7. Notice that the model is sensitive to compound words and modifiers. For example, *"red bus"* and *"yellow bus"* give very different results. Additionally, it can be seen that the quality of the results deteriorates for less frequently occurring concepts, such as *"straw hat"*. However, we emphasize that the model learned these visual appearances of text snippets from raw data of full images and sentences, without any explicit correspondences. We have additionally published a web demo that displays our alignment results for all images in the MSCOCO test set.[2]

Learned Region and Word Vector Magnitudes. An appealing feature of our alignment model is that it learns to modulate the importance of words and regions by scaling the magnitude of their embedding vectors s, v. To see this, recall that we compute the image-sentence similarity between image k and sentence l as follows:

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t. \tag{16}$$

Disciminative Words. As a result of this formulation, we observe that representations of visually discriminative words such as *"kayaking, pumpkins"* tend to have higher magnitude in the embedding space, which translates to a higher influence on the final image-sentence scores due to the inner product. Conversely, the model learns to map stop words such as *"now, simply, actually, but"* near the origin, which reduces their influence. Table 2 show the top 20 words with highest and lowest magnitudes $\|s_t\|$.

Discriminative Regions. Similarly, image regions that contain discriminative entities are assigned vectors of higher magnitudes by our model. This can be be interpreted as a measure of visual saliency, since these regions would produce large scores if their textual description was present in a corresponding sentence. We show the regions with high magnitudes in Fig. 6. Notice the common occurrence of often described and easily visually identifiable regions such as balls, bikes, helmets.

2. http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

TABLE 2
This Table Shows the Top Magnitudes of Vectors ($\|s_t\|$) for Words in Flickr30K

Magnitude	Word	Magnitude	Word
0.42	now	2.61	kayaking
0.42	simply	2.59	trampoline
0.43	actually	2.59	pumpkins
0.44	but	2.58	windsurfing
0.44	neither	2.56	wakeboard
0.45	then	2.54	acrobatics
0.45	still	2.54	sousaphone
0.46	obviously	2.54	skydivers
0.47	that	2.52	wakeboarders
0.47	which	2.52	skateboard
0.47	felt	2.51	snowboarder
0.47	not	2.51	wakeboarder
0.47	might	2.50	skydiving
0.47	because	2.50	guitar
0.48	appeared	2.50	snowboard
0.48	therefore	2.48	kitchen
0.48	been	2.48	paraglider
0.48	if	2.48	ollie
0.48	also	2.47	firetruck
0.48	only	2.47	gymnastics

Since the magnitude of individual words in our model is also a function of their surrounding context in the sentence, we report the average magnitude.

4.2 Generated Descriptions: Fulframe Evaluation

We now evaluate the ability of our RNN model to describe images and regions. We first trained our Multimodal RNN to generate sentences on full images with the goal of verifying that the model is rich enough to support the mapping from image data to sequences of words. For these full image experiments we use the more powerful VGGNet image features [61]. We report the BLEU [63], METEOR [64] and CIDEr [65] scores computed with the `coco-caption` code [66].[3] Each method evaluates a *candidate* sentence by measuring how well it matches a set of five *reference* sentences written by humans.

Qualitative. The model generates sensible descriptions of images (see Fig. 8), although we consider the last two images failure cases. The first prediction *"man in black shirt is playing a guitar"* does not appear in the training set. 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. In general, we find that a relatively large portion of generated sentences (60 percent with beam size 7) can be found in the training data. This fraction decreases with lower beam size; For instance, with beam size 1 this falls to 25 percent, but the performance also deteriorates (e.g., from 0.66 to 0.61 CIDEr).

The model often gets the right gist of the scene, but sometimes guesses specific fine-grained words incorrectly. We find one example result (*"woman in bikini is jumping over a hurdle"*) to be especially illuminating. This sentence does not occur in the training data. Our general qualitative impression of the model is that it learns certain templates, e.g., *"<noun>in <noun>is <verb>in <noun>"*, and then fills these in based on textures in the image. In this particular case, the volleyball net has the visual appearance of a hurdle, which may have caused the model to insert it as a noun (along with the woman) into one of its learned sentence templates.

3. `https://github.com/tylin/coco-caption`

Fig. 6. Flickr30K test set regions with high vector magnitude, indicating a strong influence on the image-sentence score.

Multimodal RNN Outperforms Retrieval Baseline. Our first comparison is to a nearest neighbor retrieval baseline. Here, we annotate each test image with a sentence of the most similar training set image as determined by L2 norm over VGGNet [61] fc7 features. Table 3 shows that the Multimodal RNN confidently outperforms this retrieval method. Hence, even with 113,000 train set images in MSCOCO the retrieval approach is inadequate. It is worth noting that a more sophisticated nearest neighbor approach based on multiple neighbors and a consensus voting strategy can be very competitive with state of the art CNN-RNN approaches similar to the one explored in this work [67]. However, unlike retrieval approaches, our parametric approach takes only a fraction of a second to evaluate per image.

Comparison to Other Work. Several related models have been developed in parallel to this work. We include these in Table 3 for comparison. Most similar to our model is the model of Vinyals et al. [34]. Unlike this work where the image information is communicated through a bias term on the first step, they incorporate it as a first word, they use a more powerful but more complex sequence learner (LSTM [60]), a different CNN (GoogLeNet [62]), and report results of a model ensemble. Donahue et al. [31] use a two-layer factored LSTM (similar in structure to the RNN in Mao et al. [33]). Both models appear to work worse than ours, but this is likely in large part due to their use of the less powerful AlexNet [43] features. Compared to these approaches, our model prioritizes simplicity and speed at a slight cost in performance.

4.3 Generated Descriptions: Region Evaluation

We now train the Multimodal RNN on the correspondences between image regions and snippets of text, as inferred by the alignment model. To support the evaluation, we used Amazon Mechanical Turk to collect a new dataset of region-level annotations that we only use at test time. The labeling interface displayed a single image and asked annotators (we used nine per image) to draw five bounding boxes and annotate each with text. In total, we collected 9,000 text snippets for 200 images in our MSCOCO test split (i.e., 45 snippets per image). The snippets have an average length of 2.3 words. Example annotations include *"sports car"*, *"elderly couple sitting"*, *"construction site"*, *"three dogs on leashes"*, *"chocolate cake"*. We noticed that asking annotators for

Fig. 7. Examples of highest scoring regions for queried snippets of text, on 5,000 images of our MSCOCO test set.

grounded text snippets induces language statistics different from those in full image captions. Our region annotations are more comprehensive and feature elements of scenes that would rarely be considered salient enough to be included in a single sentence about the full image, such as *"heating vent"*, *"belt buckle"*, and *"chimney"*.

Qualitative. We show example region model predictions in Fig. 9. To reiterate the difficulty of the task, consider for

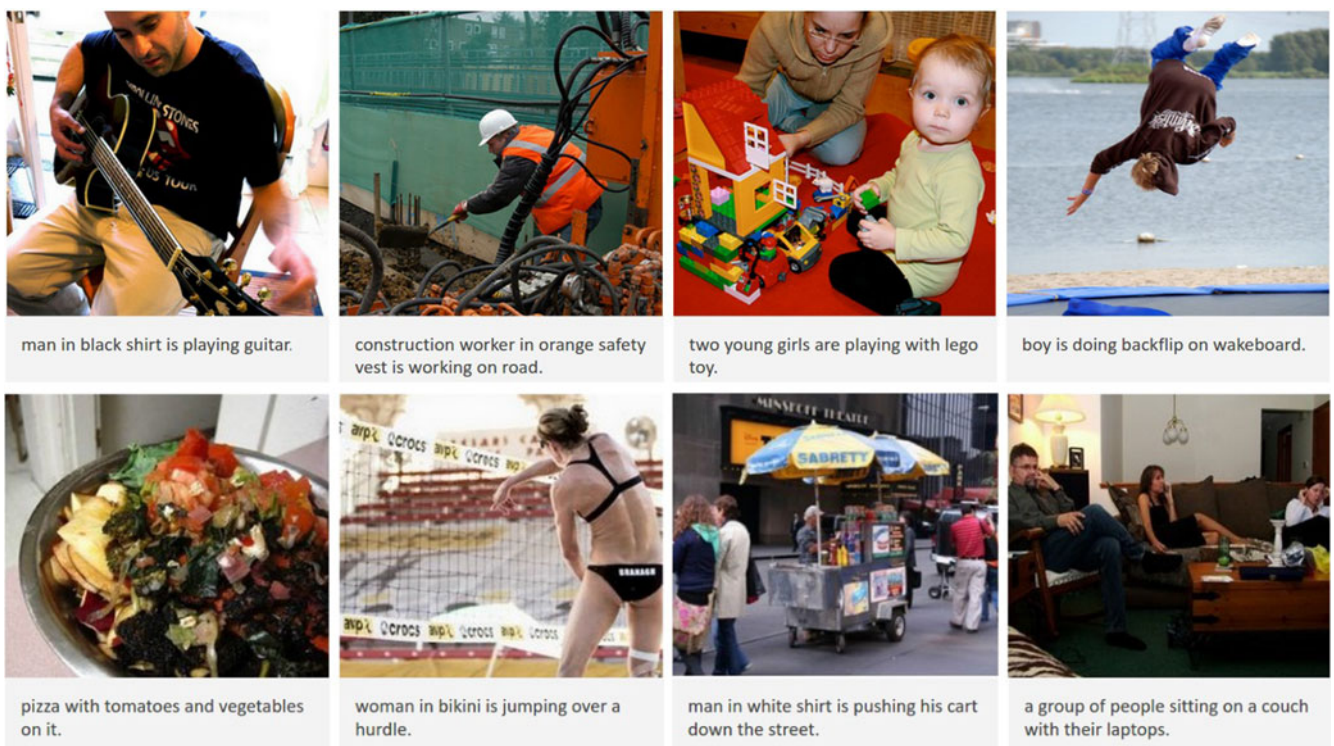

Fig. 8. Example sentences generated by the multimodal RNN for test images.

TABLE 3
Evaluation of Full Image Predictions on 1,000 Test Images

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	0.157	0.38
Mao et al. [33]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [34]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [31]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [32]	—	—	—	—	—	—	—	—	—	—	—	21.1	0.207	—
Chen and Zitnick [30]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	0.204	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	0.195	0.66

B-n *is BLEU score that uses up to n-grams. High is good in all columns. For future comparisons, our METEOR/CIDEr Flickr8K scores are 16.7/31.8 and the Flickr30K scores are 15.3/24.7.*

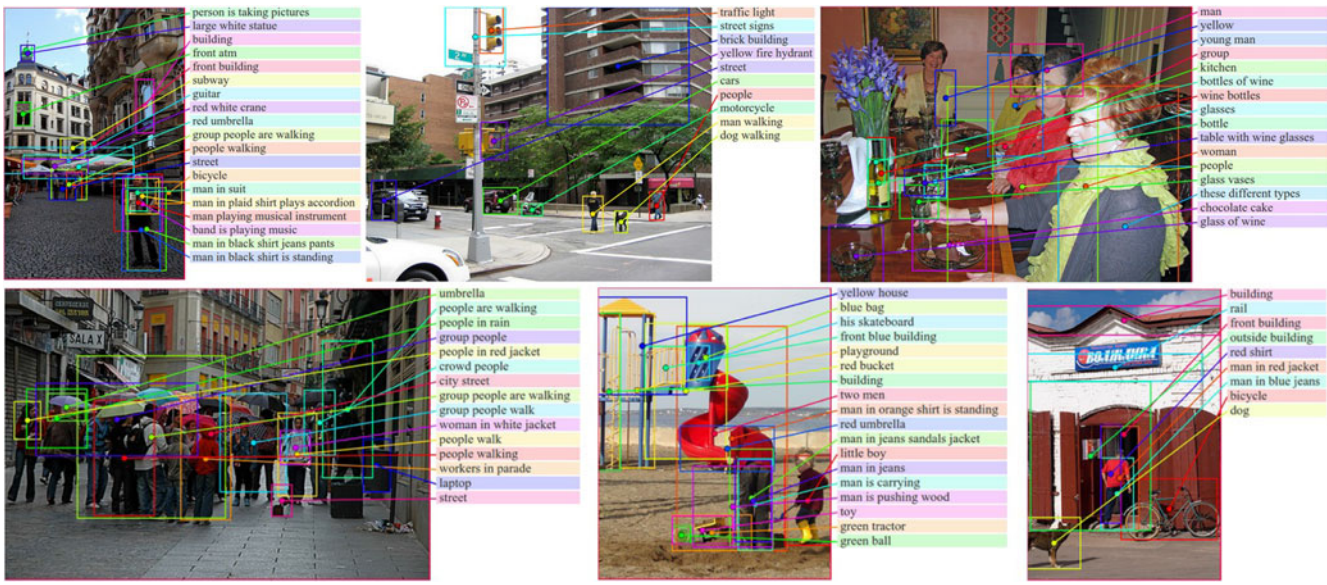

Fig. 9. Example region predictions on MSCOCO. We use our region-level multimodal RNN to generate text (shown on the right of each image) for some of the bounding boxes in each image. The lines are grounded to centers of bounding boxes and the colors are chosen arbitrarily.

example the phrase *"table with wine glasses"* that is generated on the image on the right in Fig. 9. This phrase 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 our object bounding boxes. To generate this string for the region, the model had to first correctly learn to ground the string and then also learn to generate it.

Region Model Outperforms Full Frame Model and Ranking Baseline. Similar to the full image description task, we evaluate this data as a prediction task from a 2D array of pixels (one image region) to a sequence of words and record the BLEU score. The ranking baseline retrieves training sentence substrings most compatible with each region as judged by the BRNN model. Table 4 shows that the region RNN model produces descriptions most consistent with our collected data. Note that the fullframe model was trained only on full images, so feeding it smaller image regions likely deteriorates its performance. However, its sentences are also longer than the region model sentences, which likely negatively impacts the BLEU score. The sentence length is non-trivial to control for with an RNN, but we note that the region model also outperforms the fullframe model on all other metrics: CIDEr 0.62/0.20, METEOR 0.158/0.133, ROUGE 35.1/21.0 for region/fullframe respectively.

TABLE 4
BLEU Score Evaluation of Image Region Annotations

Model	B-1	B-2	B-3	B-4
Human agreement	61.5	45.2	30.1	22.0
Nearest Neighbor	22.9	10.5	0.0	0.0
RNN: Fullframe model	14.2	6.0	2.2	0.0
RNN: Region level model	**35.2**	**23.0**	**16.1**	**14.8**

4.4 Region Captioning with Strong Supervision

With the recent availability of large-scale datasets that have region-level captions it has become possible to train the Multimodal RNN directly on this strongly supervised data without having to infer the approximate sentence snippet-region alignments. In this set of experiments we use the Visual Genome [59] region captions data. Unlike previous sections we use the LSTM [60] instead of an RNN, use the VGG-16 [61] convolutional neural network, and finetune the CNN. The experiments in this section use the publicly available code under the NeuralTalk2 project,[4] which is written in Torch [68].

4. https://github.com/karpathy/neuraltalk2

TABLE 5
Captioning Evaluation on VG Test Image Regions for a Fullframe Model Trained on MSCOCO and a Region-Level Model Trained on VG Regions

Model	METEOR
RNN: Fullframe model	0.209
RNN: Region level model	**0.272**

Similar to the smaller-scale experiment performed in the previous section, our intention is to validate the hypothesis that if one is interested in generating region-level captions then it is not an adequate strategy to simply train a full image captioning model and then run it on individual regions. The large-scale Visual Genome regions dataset allows us to support this conclusion with a more statistically significant result. In particular, we train two separate Multimodal RNN models: a full image captioning model on the MSCOCO dataset, and a region captioning model on the VG dataset. Note that the VG dataset is largely made up of MSCOCO images so both models are trained on images with the same visual statistics, but compared to MSCOCO (with 110 K training images) the VG regions data only contains approximately 90 K training images. On the other hand, the region model is trained on a total of 4.1 million captions, while the full image model is trained only on about 0.55 million captions. Both asymmetries slightly favor one model over the other, but it is difficult to precisely estimate their relative impacts. In both cases we evaluate the two models on all regions in the test set of 5,000 VG images.

Discrepancy Between Region-Level and Fullimage-Level Statistics. The result of this experiment can be found in Table 5. We report the METEOR [64] score since this metric was reported to strongly correlate with human judgements in cases where a small number of reference captions are available [65]. Again, we observe a noticeable improvement on the region-level task (0.272 versus 0.209 METEOR) when training on regions and region captions, as opposed to training a fullframe model with fullframe captions, suggesting that the visual and semantic statistics of the two cases are likely different.

Qualitative. Fig. 10 shows examples of running the Multimodal RNN on regions of test images in the VG dataset. Note that in these particular experiments there is no detection involved-instead the regions are chosen to be the ones provided by AMT workers. Notice that the RNN is capable of generating a variety of snippet descriptions with a high accuracy, such as *"cake on a table"*, *"silver fork on the plate"*, *"white lines of the road"*, or *"traffic light is green"*. These results also illuminate the discrepancy between region and fullimage statistics. For example, the region-level model often correctly describes shoes and their color, but these types of objects would rarely be judged as salient enough to be included in a caption for the full image.

Notice also that the lack of context introduces some mistakes in the generated output. For instance, in the last image in Fig. 10, the wheel of the cart is described as *"wheel of a bike"*, which could be a plausible description of that region in isolation.

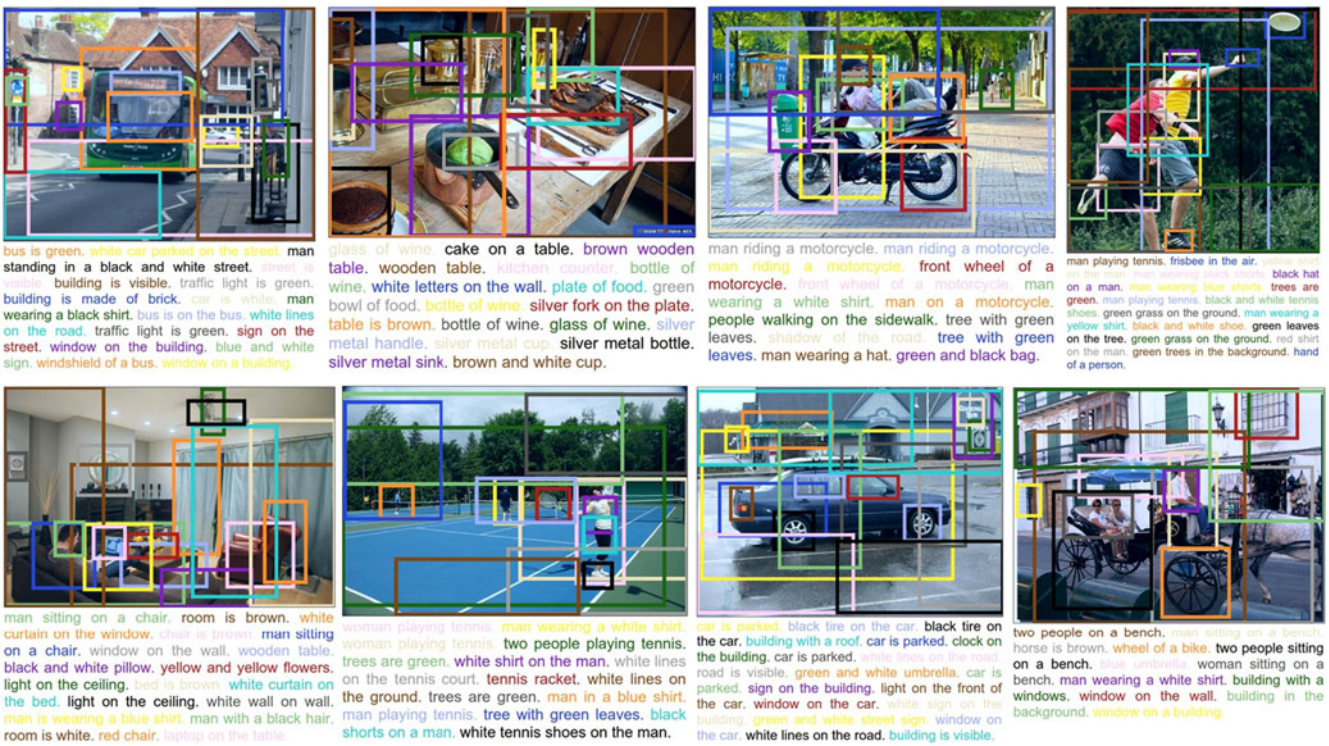

Fig. 10. RNN generated captions on test images, using regions provided by AMT worker annotations in the VG data. Each region is described by our RNN model in isolation, and we sort resulting captions by the conditional log probability of the caption given the region, which can be interpreted as a measure of confidence. This visualization tries to show more generated captions without cluttering the images with correspondence lines; the colors indicate which caption matches which region, but the specific colors are arbitrary.

4.5 Limitations

Although our results are encouraging, the Multimodal 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 receives the image information only through additive bias interactions, which are known to be less expressive than more complicated multiplicative interactions [50], [60]. Evaluating every region in isolation also leads to computational inefficiency because one must forward every individual region of interest separately through the convolutional network.

Lastly, our full proposed approach for producing region-level captions consists of two separate modeling steps, and our detection step relies on the external R-CNN regions, which are trained on a different dataset to predict specifically object-like regions. Going directly from an image-sentence dataset (a setting where data is cheaper to obtain) to high-quality region-level annotations as part of a single model trained end-to-end remains an open problem.

5 Conclusions

We introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hard-coded assumptions. Our approach features a novel ranking model that aligned parts of visual and language modalities through a common, multimodal embedding. We showed that this model provides results superior to all previous work, controlling for the strength of the underlying convolutional network features. Second, we described a Multimodal Recurrent Neural Network architecture that generates descriptions of visual data. We evaluated its performance on both fullframe and region-level experiments and showed that in both cases the Multimodal RNN outperforms simple retrieval baselines.

Acknowledgments

We thank Justin Johnson and Jon Krause for helpful comments and discussions. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research. We would also like to thank the maintainers of Torch 7, and especially Soumith Chintala for his support. This research is partially supported by an ONR MURI grant, and US National Science Foundation ISS-1115313.

References

[1] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *J. Vis.*, vol. 7, no. 1, 2007, Art. no. 10.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[3] O. Russakovsky, et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[4] G. Kulkarni, et al., "Baby talk: Understanding and generating simple image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1601–1608.

[5] A. Farhadi, et al., "Every picture tells a story: Generating sentences from images," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.

[6] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artificial Intell. Res.*, vol. 47, pp. 853–899, 2013.

[7] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *arXiv preprint arXiv:1411.5654*, 2014.

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.

[9] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[10] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 966–973.

[11] S. Fidler, A. Sharma, and R. Urtasun, "A sentence is worth a thousand pixels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, 1995–2002.

[12] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[13] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[14] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 2036–2043.

[15] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2407–2414.

[16] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Advances Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[17] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 207–218, 2014.

[18] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 359–368.

[19] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, no. 10, pp. 351–362, 2014.

[20] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Language Learn.*, 2011, pp. 220–228.

[21] A. Barbu, et al., "Video in sentences out," *arXiv:1204.2742*, 2012.

[22] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proc. Empirical Methods Natural Language Process.*, 2013, pp. 1292–1302.

[23] A. Gupta and P. Mannem, "From image annotation to image description," in *Neural Information Processing*. Berlin, Germany: Springer, 2012.

[24] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 444–454.

[25] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2T: Image parsing to text description," in *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, Aug. 2010.

[26] M. Yatskar, L. Vanderwende, and L. Zettlemoyer, "See no evil, say no evil: Description generation from densely labeled images," in *Proc. 3rd Joint Conf. Lexical Comput. Semantics*, 2014, pp. 110–120.

[27] M. Mitchell, et al., "Midge: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 747–756.

[28] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[29] R. Kiros, R. S. Zemel, and R. Salakhutdinov, "Multimodal neural language models," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 595–603.

[30] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1411.5654

[31] J. Donahue, et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2625–2634.
[32] H. Fang, et al., "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1473–1482.
[33] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *arXiv:1410.1090*, 2014.
[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3156–3164.
[35] T. L. Berg "Names and faces in the news," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 2, pp. II-848–II-854.
[36] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler, "What are you talking about? text-to-image coreference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 351–362.
[37] D. Lin, S. Fidler, C. Kong, and R. Urtasun, "Visual semantic search: Retrieving videos via complex textual queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2657–2667.
[38] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proc. 29th Int. Conf. Mach. Learn.*, Jun. 2012, pp. 1671–1678.
[39] C. L. Zitnick, D. Parikh, and L. Vanderwende, "Learning the visual interpretation of sentences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1681–1688.
[40] Y. Zhu, et al., "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 19–27.
[41] A. Frome, et al., "Devise: A deep visual-semantic embedding model," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.
[42] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[45] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*. Berlin, Germany: Springer, 2006.
[46] R. Socher, J. Pennington, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Empirical Methods Natural Language Process.*, 2014, pp. 1532–1543.
[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[48] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003.
[49] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1045–1048.
[50] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1017–1024.
[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 580–587.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
[53] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[54] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. TIT-13, no. 2, pp. 260–269, Apr. 1967.
[55] J. L. Elman, "Finding structure in time," *Cogn. Science*, vol. 14, no. 2, pp. 179–211, 1990.
[56] W. Zhang, *State-Space Search: Algorithms, Complexity, Extensions, and Applications*. Berlin, Germany: Springer, 1999.
[57] T. Tieleman and G. Hinton, "Lecture 6.5rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.
[58] T. Tieleman and G. E. Hinton, "Lecture 6.5-RmsProp: Divide the gradient by a running average of its recent magnitude," 2012.
[59] R. Krishna, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016. [Online]. Available: http://arxiv.org/abs/1602.07332
[60] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
[62] C. Szegedy, et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1–9.
[63] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[64] M. Denkowski and A. Lavie, "METEOR universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Statistical Mach. Transl.*, 2014, pp. 67–78
[65] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4566–4575.
[66] X. Chen, et al., "Microsoft Coco captions: Data collection and evaluation server," *arXiv:1504.00325*, 2015.
[67] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring nearest neighbor approaches for image captioning," *arXiv:1505.04467*, 2015.
[68] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like environment for machine learning," in *Proc. Big Learn Advances Neural Inf. Process. Syst. Workshop*, 2011, pp. 1681–1688.

Andrej Karpathy received the BSc degree in computer science and physics from the University of Toronto, in 2009, and the MSc degree in computer science from the University of British Columbia, in 2011. He is currently working toward the PhD degree in the Computer Science Department, Stanford University. His research interests include end-to-end training of models that simultaneously process visual and natural language data. He is a member of the IEEE.

Li Fei-Fei received the BA degree (High Honors) in physics from Princeton, in 1999, and the PhD degree in electrical engineering from California Institute of Technology (Caltech), in 2005. She is an associate professor in the Computer Science Department, Stanford, and the director of the Stanford Artificial Intelligence Lab and the Stanford Vis. Lab. She is also the director of the recently established Stanford Toyota Center for Human-Centric AI Research. Her main research areas are in machine learning, computer vision and cognitive, and computational neuroscience. She has published more than 100 scientific articles in top-tier journals and conferences, including the *Nature*, PNAS, the *Journal of Neuroscience*, CVPR, ICCV, NIPS, ECCV, the *International Journal of Computer Vis.*, the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, etc. She joined Stanford in 2009 as an assistant professor, and was promoted to associate professor with tenure, in 2012. Prior to that, she was on faculty with Princeton University (2007-2009) and University of Illinois Urbana-Champaign (2005-2006). She was a speaker at the TED2015 main conference, a recipient of the 2014 IBM Faculty Fellow Award, 2011 Alfred Sloan Faculty Award, 2012 Yahoo Labs FREP award, 2009 NSF CAREER award, the 2006 Microsoft Research New Faculty Fellowship and a number of Google Research awards. Work from her lab have been featured in a variety of popular press magazines and newspapers including the *New York Times*, *Science*, *Wired Magazine*, and *New Scientists*. She is a member of the IEEE.