Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy, Li Fei-Fei
Department of Computer Science, Stanford University.

Introduction. The majority of visual recognition approaches focus on labeling images with a fixed set of categories. Great progress has been achieved in these endeavors, but 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. In this work, we think of natural language as a rich label space, capable of simultaneously representing many visual aspects (e.g. actions, attributes, objects, etc.), and take a step towards the goal of generating natural language descriptions of images and their regions.

Generating text for images. To this end, we develop a model that generates text descriptions of images and their regions (see Figure 1). The primary challenge of our approach is in the design of a model that can simultaneously reason about contents of images and their mapping to variable-sized descriptions. To address this challenge we develop a Multimodal Recurrent Neural Network language model that is conditioned on the image information. During training, our multimodal RNN takes the image pixels I and a sequence of input vectors $(x_1, \dots, x_T)$ that encode the words of the ground truth sentence. It then computes a sequence of hidden states $(h_1, \dots, h_t)$ and a sequence probabilities of the next word in the sequence $(y_1, \dots, y_t)$ by iterating the following recurrence relation for $t = 1$ to T:

$$b_v = W_{hi}[CNN_{\theta_c}(I)] \tag{1}$$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v) \tag{2}$$

$$y_t = softmax(W_{oh}h_t + b_o). \tag{3}$$

In the equations above, $W_{hi}, W_{hx}, W_{hh}, W_{oh}, x_i$ and b_h, b_o are learnable parameters, and $CNN_{\theta_c}(I)$ is the last layer of a Convolutional Network that takes the image pixels I as input. The output vector y_t has the size of the word dictionary and one additional dimension for a special END token. The model is then trained to maximize the log likelihood of generating the next word in a ground truth sentence, as a function of the image and the previous words. To predict a sentence, we compute the representation of the image b_v, set $h_0 = 0$, x_1 to the START vector and compute the distribution over the first word y_1. We sample a word from the distribution (or pick the argmax), set its embedding vector as x_2, and repeat this process until the END token is generated. Our experiments on Flickr30K [3] and MSCOCO [2] datasets demonstrate that the model is capable of generating accurate, novel text descriptions of image data.

Aligning text snippets to image regions. A second practical challenge is that while datasets of image captions are available in large quantities on the internet, these descriptions multiplex mentions of several entities whose locations in the images are unknown. For instance, a sentence such as *"a dog chasing a soccer ball"* is a source of two region annotations: One referring to (*"a dog chasing"*) and the other to a *"soccer ball"*. Our core insight is to treat the sentences as weak labels, in which contiguous segments of words correspond to some particular, but unknown location in the image. We develop a model that learns to align the two modalities based on a unique combination of a Bidirectional Recurrent Neural Network (BRNN), a Convolutional Neural Network (CNN) and a max-margin objective (see details in the full paper). This model allows us to infer the latent correspondences between text snippets and image regions, which we then use as training data for the region-level generation task.

Evaluation: Image and Sentence Retrieval. During the forward pass of the network, our BRNN alignment model computes a compatibility score between any image-sentence pair while inferring the latent alignment between their parts (regions and words respectively). We evaluate the model on image-sentence retrieval experiments and achieve state-of-the-art results. Moreover, qualitative experiments (see Figure 1, left) indicate that the model effectively pairs up words (e.g. *"accordion"*) with their respective regions in the image, despite the fact that its input only consists of entire images and full sentences. The inferred alignments for our full test set can be browsed on our project website .

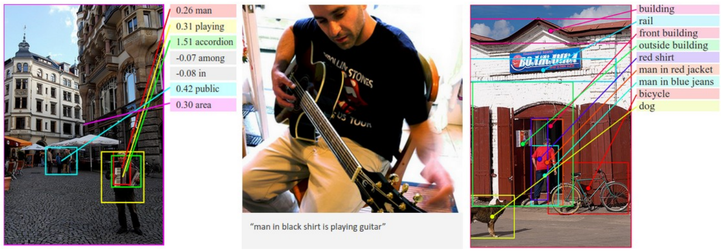

Figure 1: Example alignments produced by our ranking model (left), and example generated text for full images and image regions (middle, right).

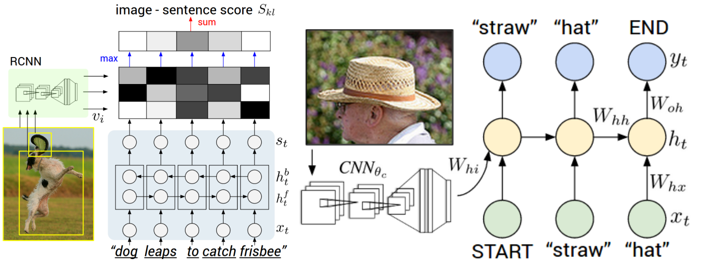

Figure 2: Our alignment Bidirectional RNN model learns to infer the alignment between text snippets and regions (left). Our Multimodal RNN learns to generate text given an input image (right).

Evaluation: Generating Image Captions. We first evaluated the Multimodal RNN on the task of captioning full images. We use the Flickr8K [1], Flickr30K [3] and MSCOCO [2] 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. Our results show significant improvements over baseline ranking models (e.g. 66.0 vs. 38.3 CIDEr score on MSCOCO) and the model is shown to yield comparable performance to recently proposed related models.

Evaluation: Generating Region Captions. We additionally evaluated 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 (AMT) to collect a new dataset of region-level annotations that we only use at test time. In total, we collected 9,000 text snippets for 200 images in our MSCOCO test split. Similar to full-frame experiments, our generation model significantly outperforms retrieval baselines (e.g. 35.2 vs. 22.0 BLEU-1), providing evidence that the model effectively creates novel captions by generalizing from the training data.

Reproducibility. We make our Multimodal RNN Python/numpy code, data, model checkpoints and prediction visualizations available on Github.

[1] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[3] 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.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

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