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 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 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 significantly outperform retrieval baselines on both full images and on 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 [14]. 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 [45, 11]. 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 [29, 13]. 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 to be an unnecessary restriction.

In this work, we strive to take a step towards the goal of generating dense descriptions of images (Figure 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 rely on learning from the training data. The second, practical challenge is that datasets of image captions are available in large quantities on the internet [21, 58, 37], but these descriptions multiplex mentions of several entities whose locations in the images are unknown.

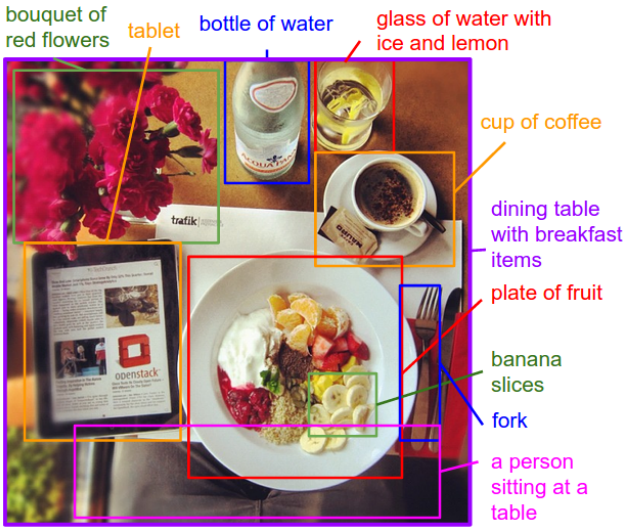

Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

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.