DEEP VISUAL-SEMANTIC ALIGNMENTS FOR GENERATING IMAGE DESCRIPTIONS

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

Presented by: Taslima Akter
Indiana University, Bloomington

Motivation

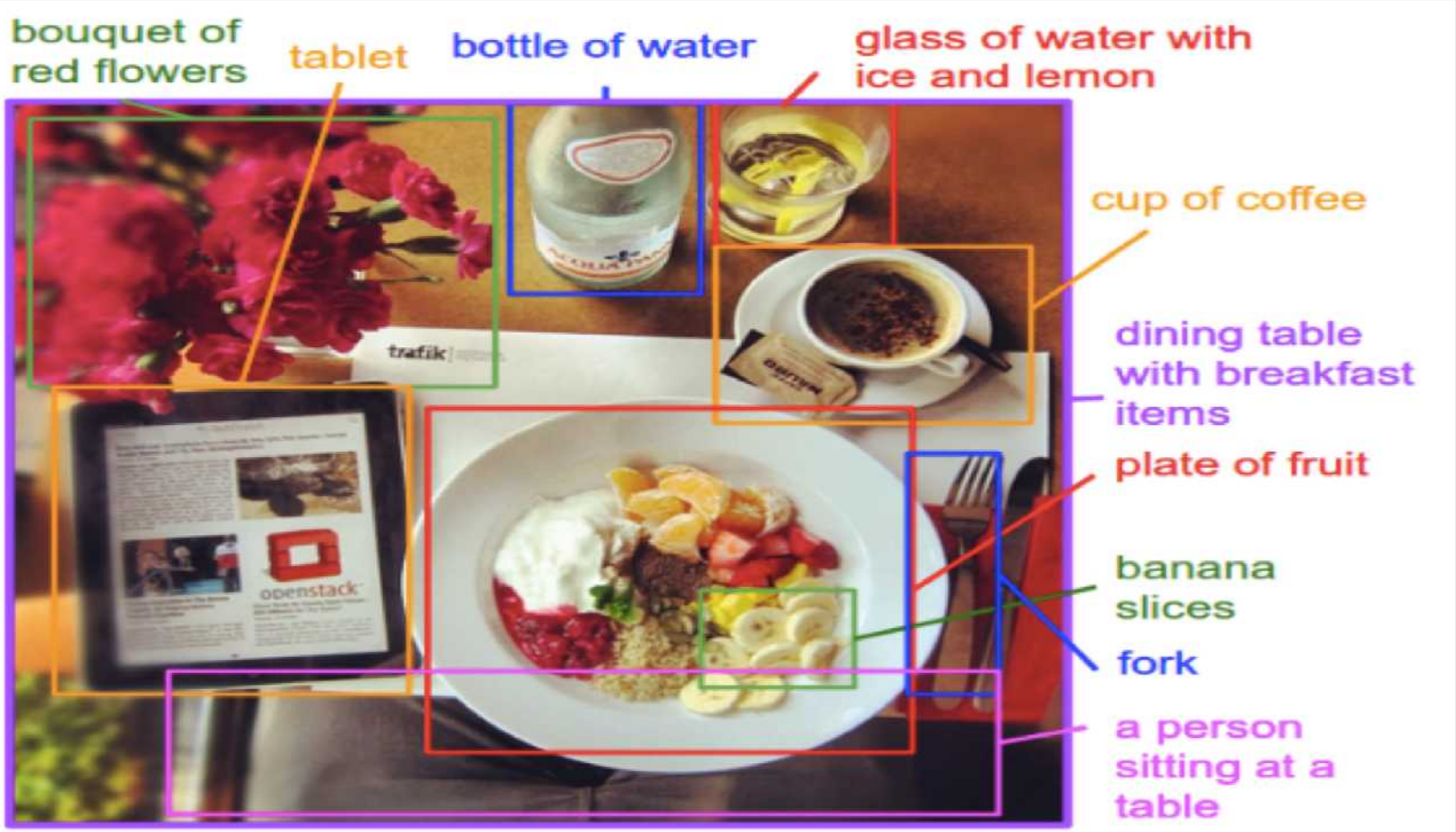

Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

Objective

- Generates natural language description of images and their regions
- Alignment model based on:
  - *Convolutional Neural Networks (CNN) over image regions*
  - *Bidirectional RNN over sentences*
  - *Structured objective that aligns two modalities through a multimodal embedding*
- Multimodal Recurrent Neural Network generates novel descriptions of image regions.

Introduction

- Two Challenges:
  - *Free of specific hard-coded templates, rules or categories and instead rely on learning from the training data.*
  - *Large dataset of image captions are available but locations of the entities in the images are unknown.*

Contributions

- Developed a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe.
- Introduced a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text.

Related Work

- Dense image annotations
  - *Scene type, objects and their spatial support*
- Generating descriptions
  - *Based on content of the image or generative grammars*
- Grounding natural language in images
  - *Based on grounding dependency tree relations*
- Neural networks in visual and language domains
  - *CNN for images and pre trained word vectors for sentences*

Model

- Aligned sentence snippets to the visual regions through multimodal embedding
- Treat these correspondences as training data for a second multimodal Recurrent Neural Network model that learns to generate the snippets.

Model

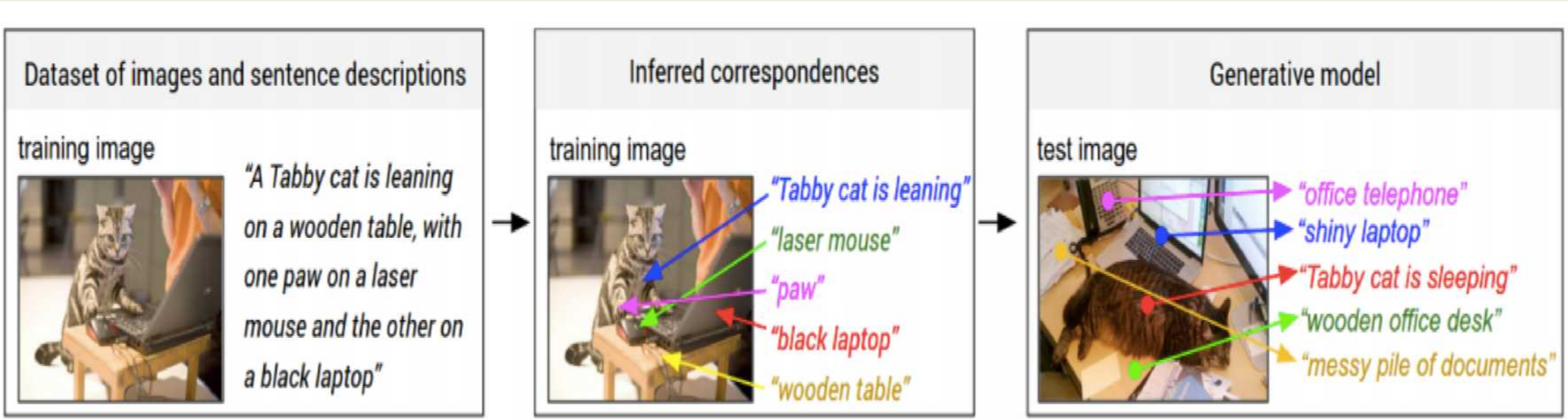

Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle, Section 3.1) and then learns to generate novel descriptions (right, Section 3.2).

Learning to align visual and language data

- Representing images:
  - *Detect objects in every image with a Region Convolutional Neural Network (RCNN)*
  - *The CNN is pre-trained on ImageNet and finetuned on the 200 classes*
  - *Used top 19 locations in addition to the whole image*
  - *Every image is represented as h-dimensional vector*

Learning to align visual and language data

- Representing sentences
  - *Used Bidirectional Recurrent Neural Network (BRNN) to compute the word representations*
  - *The BRNN takes a sequence of N words and transforms each one into an h-dimensional vector.*
  - *The representation of each word is enriched by a variably-sized context around that word.*

Learning to align visual and language data

- Alignment objective
  - g_k *is the set of image fragments in image k*
  - g_l *is the set of sentence fragments in sentence l*
  - *A sentence-image pair should have a high matching score if its words have a confident support in the image.*

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

Learning to align visual and language data

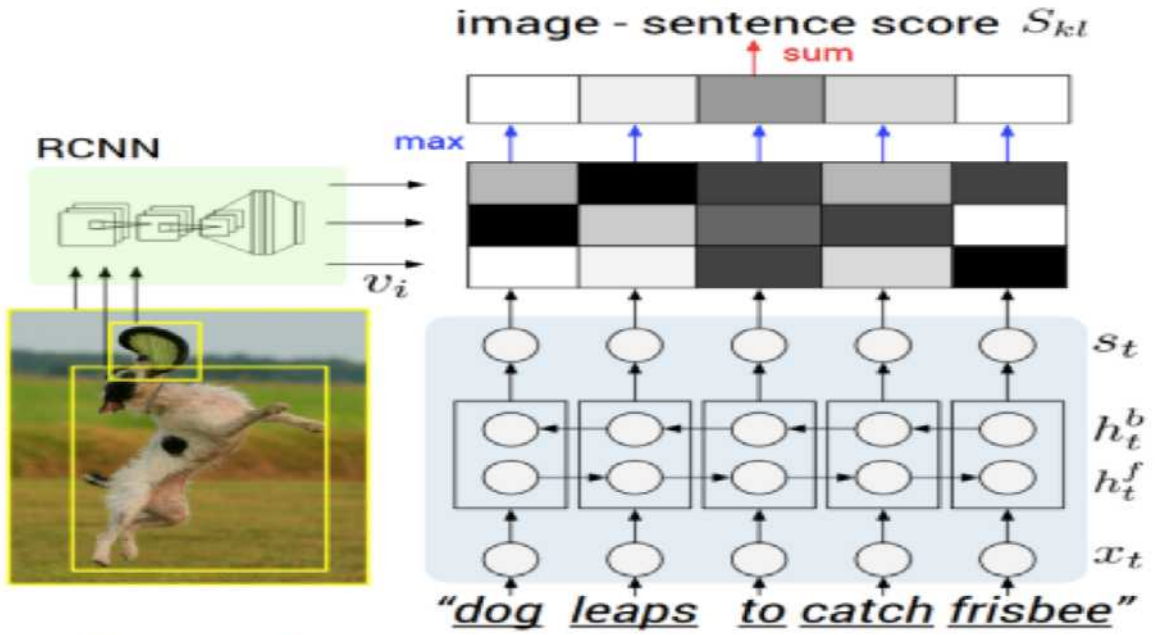

Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and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

Multimodal Recurrent Neural Network for generating descriptions

- RNN Training:
  - *Combined a word (x_t), the previous context (h_t-1) to predict the next word (y_t)*
  - *Conditioned the RNN's predictions on the image information (b_v) via bias interactions on the first step.*
- RNN at test time:
  - *Compute the image representation b_v, setting $h_0 = 0$, x_1 to the START vector and compute the distribution over the first word y_1*

Multimodal Recurrent Neural Network for generating descriptions

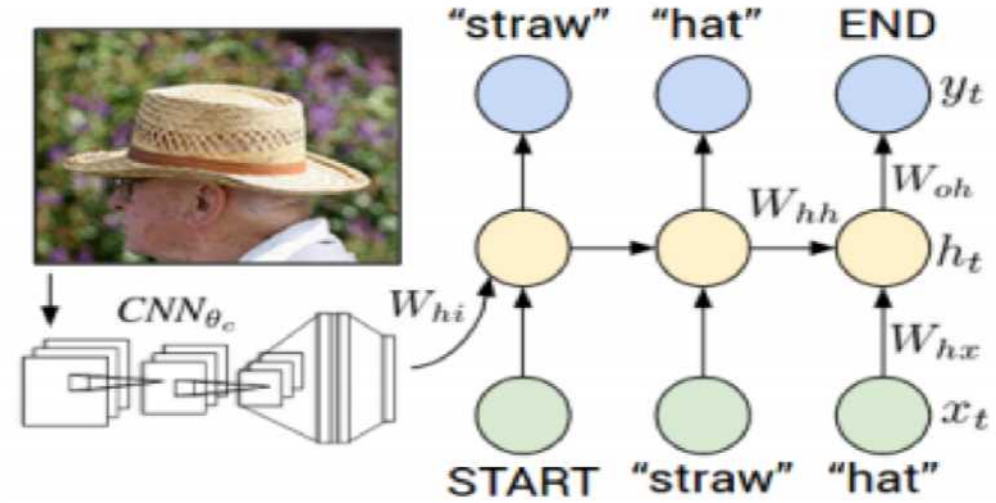

Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

Dataset and Data Preprocessing

- Flickr8K (8,000), Flickr30K (31,000), MSCOCO (123,000)
- Each image is annotated with 5 sentences
- Converted all sentences to lowercase
- Discarded non-alphanumeric characters
- Filtered words occurring at least 5 times in the training set
- 1000 images for validation, 1000 for testing, rest for training

Experiments

- Image-Sentence Alignment Evaluation
- Generated Descriptions: Fulframe evaluation
- Generated Descriptions: Region evaluation

Image-Sentence Alignment Evaluation

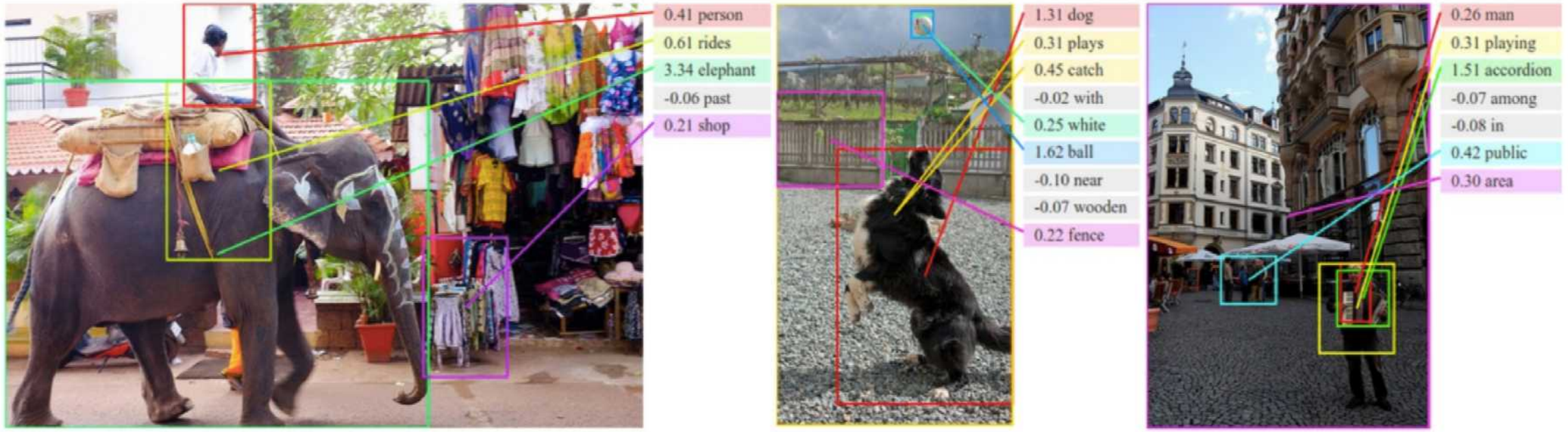

Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring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

Generated Descriptions: Fulframe evaluation

Figure 6. Example sentences generated by the multimodal RNN for test images. We provide many more examples on our project page.

Generated Descriptions: Region evaluation

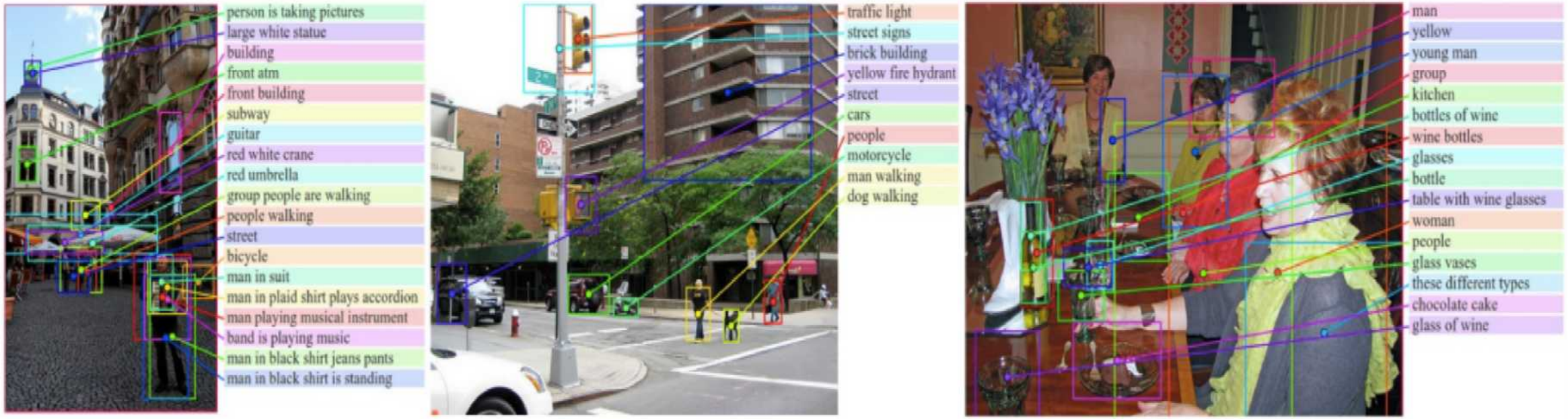

Figure 7. Example region predictions. We use our region-level multimodal RNN to generate text (shown on the right of each image) for some of the bounding boxes in each image. The lines are grounded to centers of bounding boxes and the colors are chosen arbitrarily.

Limitations

- The model can only generate a description of one input array of pixels at a fixed resolution.
- The RNN receives the image information only through additive bias interactions, which are known to be less expressive than more complicated multiplicative interactions
- This approach consists of two separate models

Thank You