Deep Visual-Semantic Alignments for Generating Image Descriptions

Xuewen Yang

May 3. 2018

In this work, we strive to take a step towards the goal of generating dense descriptions of images such as Fig 1.

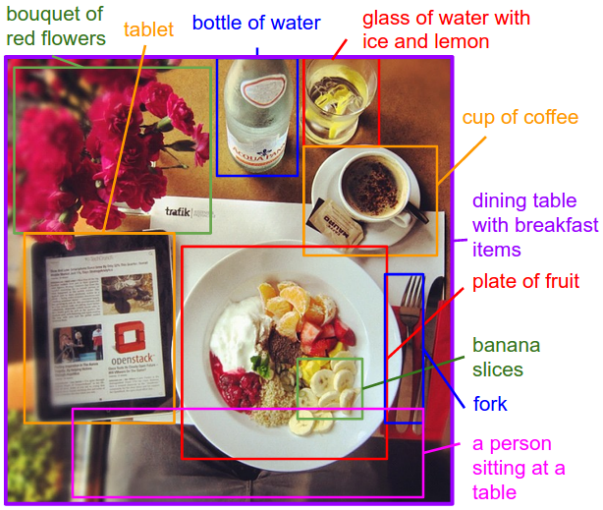

Figure 1: Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

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 domainof 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 data sets of image captions are available in large quantities on the internet[1], 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 imagesentence data sets 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.

References

[1] P. Young M. Hodosh. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 27(1):10, 2014.