文獻研討(二) Extended Abstract

Title: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Authors: Peter Anderson, Xiaodong He, Chirs Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang

Source: IEEE Conference on Computer Vision and Pattern Recognition

Presenter: WenWei Kang. M0706729

Date: 19 Mar. 2019

The problem of image analysis and understanding has gained high prominence over the last decade. There are many tasks about image-text understanding like image caption, video caption and visual question answering. The vast majority work about foregoing tasks are of the **top-down** variety. In this work, the foregoing tasks (image caption, visual question answering(VQA)) are a combined **bottom-up** and **top-down** attention mechanism that enable attention to be calculated at the level of objects and other salient image regions.

Image caption

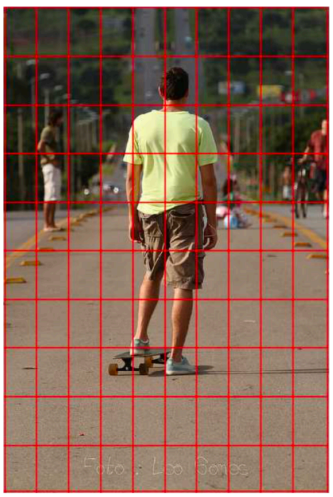
Two men playing frisbee in a dark field.

VQA

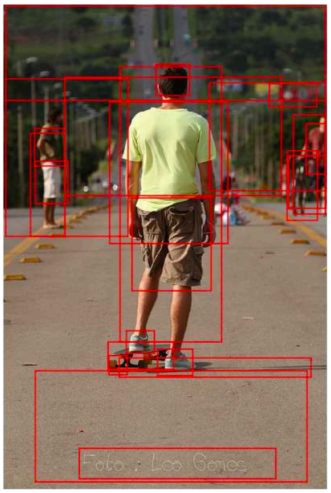
Question: What room are they in? Answer: kitchen

The bottom-up mechanism focus on converting image to feature vector V and top-down mechanism is responsible for predicting feature weightings $\boldsymbol{\alpha}$. In the past, the bottom-up mechanism just apply on vanilla CNN. Thanks to previous contributions. The bottom-up mechanism proposes image regions(attributes) by using *Faster R-CNN*[1]. *Faster R-CNN* is an image object detection algorithm.

Vanilla CNN



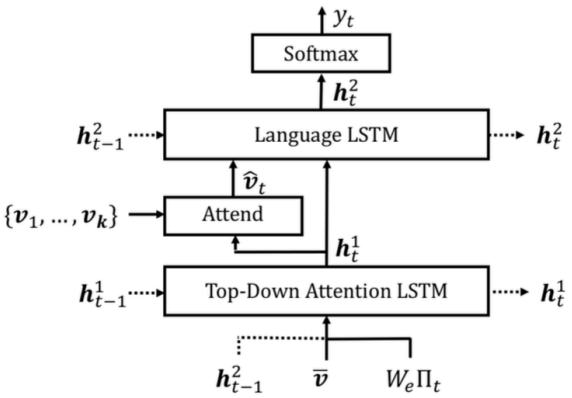
Faster R-CNN



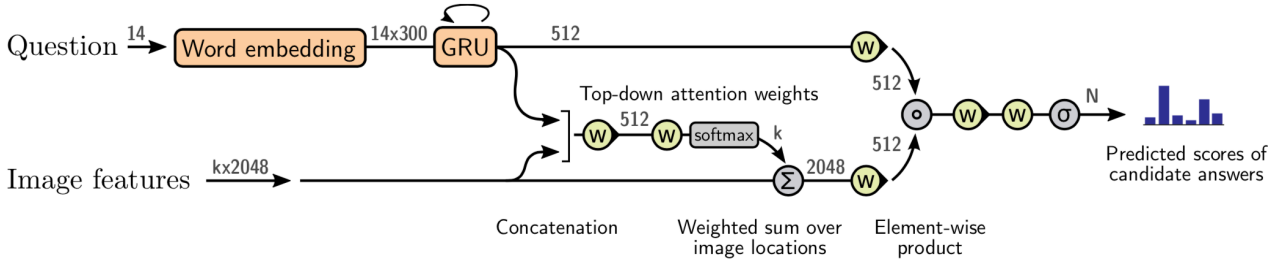
In contrast to vanilla CNN algorithm, the *Faster R-CNN* effectively functions as a 'hard' attention mechanism, because it implicitly determines where the salient region should be attend in an image, but the vanilla CNN just blindly extracts all the regions in an images.

When the bottom-up mechanism proposes a set of salient image regions $V = \{v_1, v_2, \ldots, v_k\}$, the top-down mechanism(based on Top-Down Attention LSTM) determines an attention distribution $\boldsymbol{\alpha} = \{a_1, a_2, \ldots, a_k\}$ over the image regions, with each attention weight α_i represented an importance per region. Then the image is encoded into a context vector $\hat{v} = \sum_{i=1}^{k} a_i v_i$. The foregoing structure is similar between image captioning model and VQA model. Both use the same output of bottom-up attention model and generate attention distribution.

Image caption model



VQA model



Applying this approach to image captioning and VQA, we get a state-of-the-art on MSCOCO image captioning test server, achieving iconic metrics(BLEU4/CIDEr/SPICE) of the 36.9, 117.9, 21.5. In the 2017 VQA Challenge, we obtain first place.

In the past, this bottom-up attention ('hard' attention) idea that apply on image caption just focus on attention distribution[2]. In this paper, we find that image features are extracted well by simply replacing CNN features with bottom-up attention features(object detection) and bottom-up attention model can generate more natural attention distribution for these attributes.

[1]Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.

[2]Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. 2015