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

Chenchen Qiu

College of Information Science and Engineering
Ocean University of China

VISION@OUC

Outline

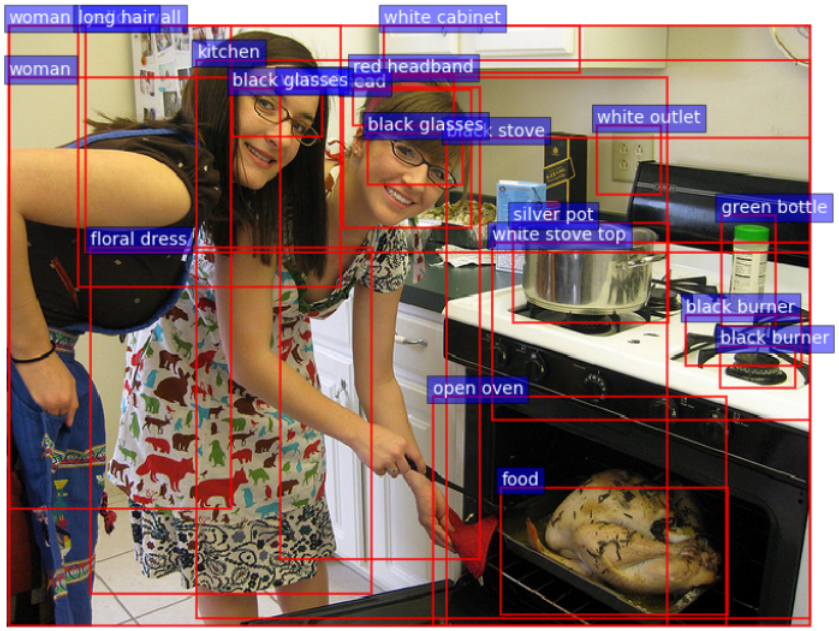
Brief Introduction of VQA

Visual Question Answering: given an image and a natural language question about the image, the task is to provide an accurate natural language answer.

Outline

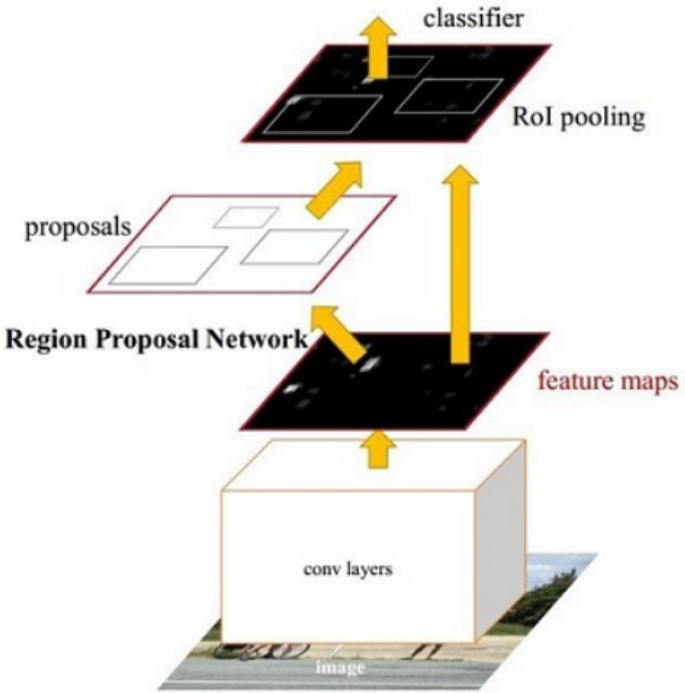
Faster R-CNN

Example output from our Faster R-CNN bottom-up attention model. Each bounding box is labeled with an attribute class followed by an object class. Note however, that in captioning and VQA we utilize only the feature vectors–not the predicted labels.



Structure of Faster R-CNN

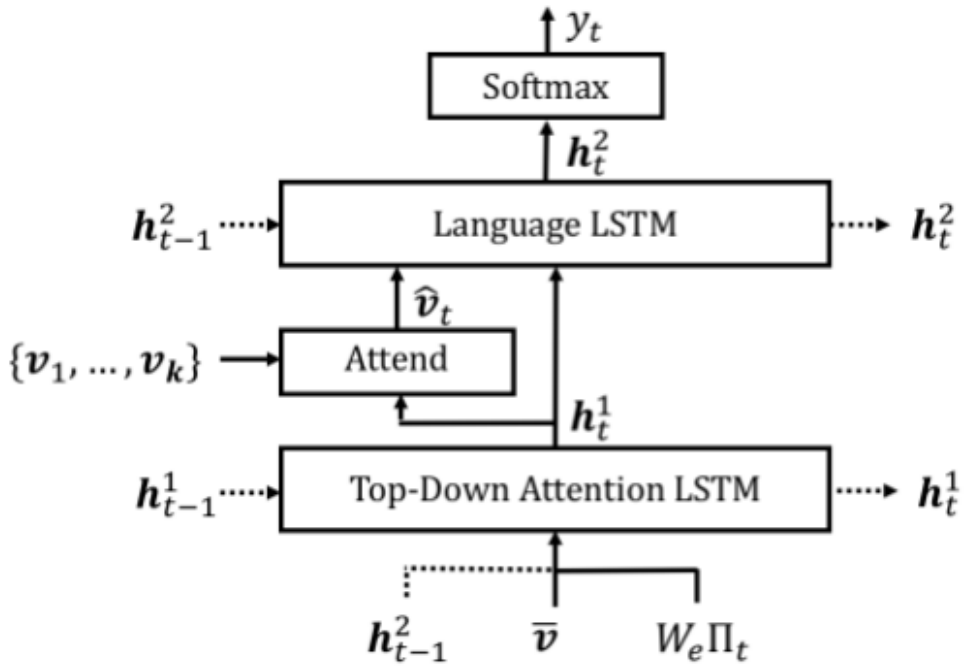
In this work we define spatial regions in terms of bounding boxes and implement bottom-up attention using Faster R-CNN . Faster R-CNN is an object detection model designed to identify instances of objects belonging to certain classes and localize them with bounding boxes.



Outline

Overview of the proposed captioning model

Overview of the proposed captioning model. Two LSTM layers are used to *selectively attend to spatial image features* $v_1, \ldots v_k$. These features can be defined as the spatial output of a CNN, or following our approach, generated using bottom-up attention.



Input and Output of Attention LSTM

Input of Attention LSTM

$x_t^1 = [h_{t-1}^2, \overline{v}, W_e \prod_t]$

- $\overline{v}$ is the mean-pooled image feature.
- $\prod_t$ is one-hot encoding of the input word at timestep t.
- W_e is a word embedding matrix for a vocabulary.

Attention Weight

$$a_{i,t} = w_a^T tanh(W_{va} v_i + W_{ha} h_t^1)$$
$$\alpha_t = softmax(a_t)$$
$$\hat{v}_t = \sum_{i=1}^{K} \alpha_{i,t} v_i$$

Language LSTM

Input of Language LSTM

$x_t^2 = [\hat{v}_t, h_t^1]$

- h_t^1 is the current hidden state of attention LSTM.

Conditonal distribution over possible ouput words

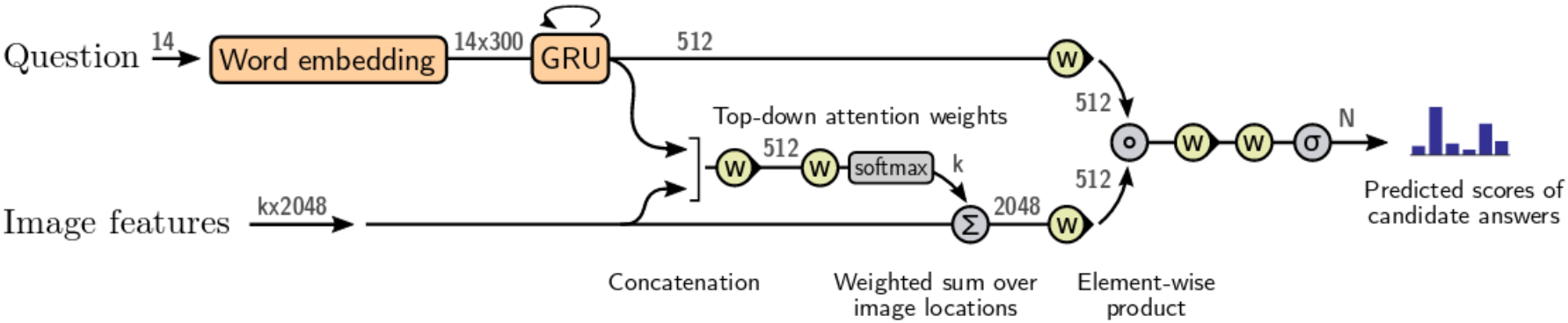
$p(y_t|y_{1:t-1}) = softmax(W_p h_t^2 + b_p)$

- W_p and b_p are learned weights and bias.

$p(y_{1:T}) =$

Outline

VQA Model



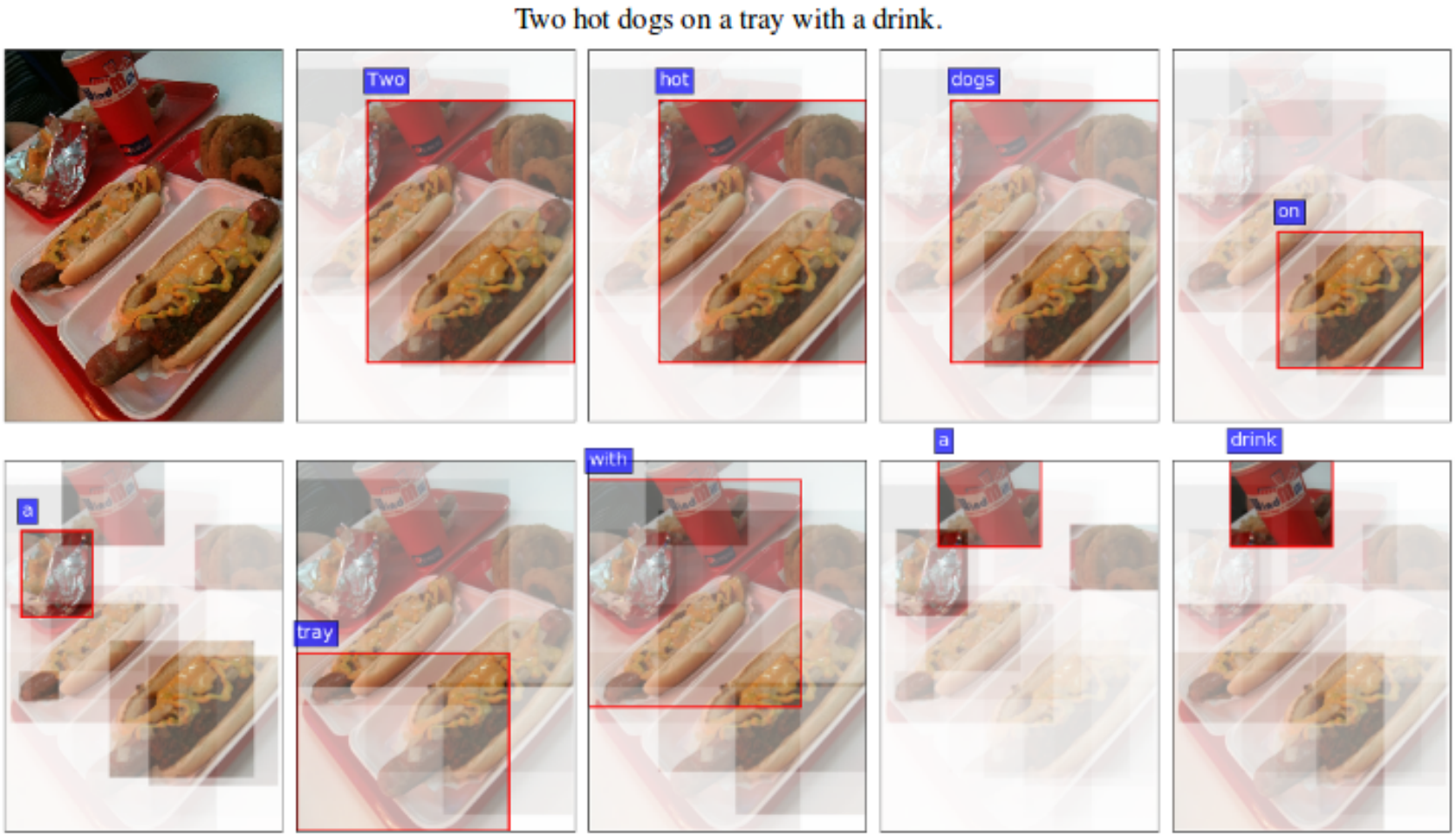
These features can be defined as the spatial output of a CNN, or following our approach, generated using bottom-up attention. Output is generated by a multi-label classifier operating over a fixed set of candidate answers.

Result on VQA v2.0

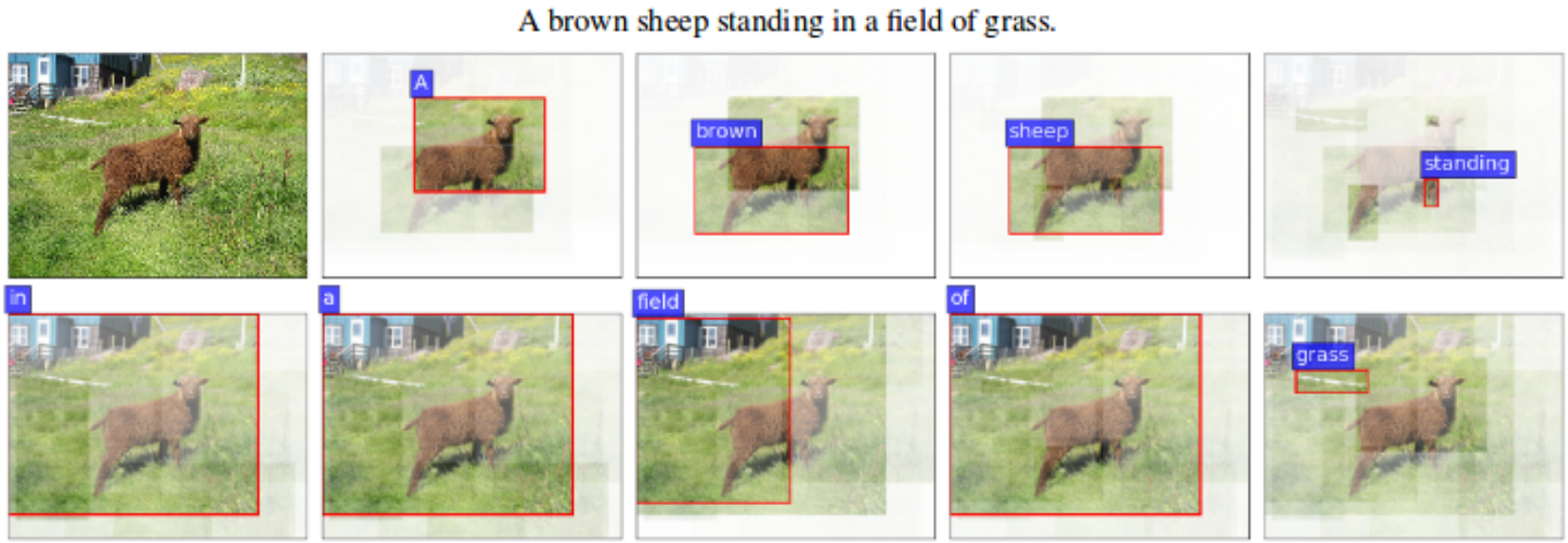
	Yes/No	Number	Other	Overall
Prior [12]	61.20	0.36	1.17	25.98
Language-only [12]	67.01	31.55	27.37	44.26
d-LSTM+n-I [26, 12]	73.46	35.18	41.83	54.22
MCB [11, 12]	78.82	38.28	53.36	62.27
UPMC-LIP6	82.07	41.06	57.12	65.71
Athena	82.50	44.19	59.97	67.59
HDU-USYD-UNCC	84.50	45.39	59.01	68.09
Ours: Up-Down	**86.60**	**48.64**	**61.15**	**70.34**

VQA v2.0 test-standard server accuracy as at 8 August 2017, ranking our submission against published and unpublished work for each question type.

Result on Attention



Result on Attention



Q & A