Generating Visually Descriptive Language from Object Layouts

Author: Xuwang Yin, Vicente Ordonez

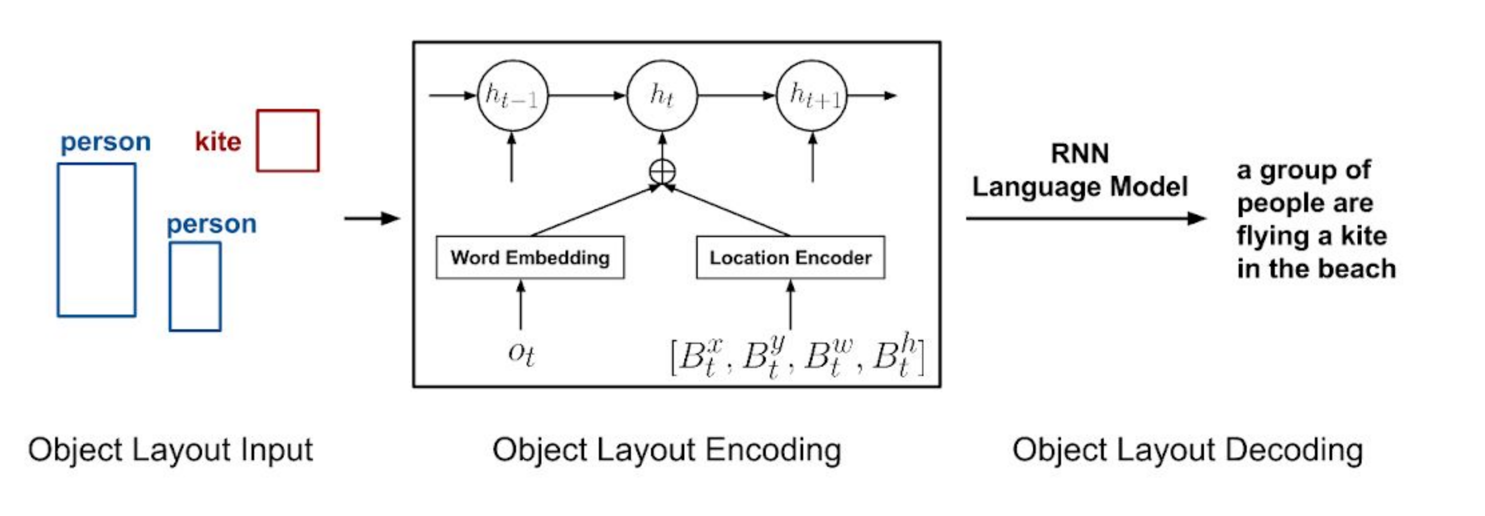
Presenter: Xing Zeng

✓

Motivations

- Image Captioning is still a challenging problem
- Tackle a simpler problem instead: describing object layouts only
- This could be used as a middle stage to better image captioning models

Task


person
kite
person
h_{t-1}
h_t
h_{t+1}
Word Embedding
Location Encoder
o_t
$[B_t^x, B_t^y, B_t^w, B_t^h]$
RNN
Language Model
a group of
people are
flying a kite
in the beach
Object Layout Input
Object Layout Encoding
Object Layout Decoding

Encoder

Each time takes input pair

(o_t, l_t)

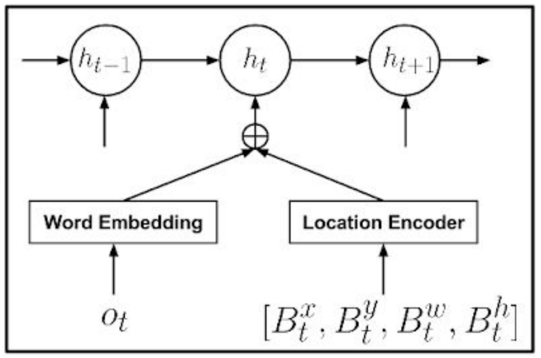
o_t=object category one-hot encoding

l_t=[left-most position (B^x_t),

top-most position (B^y_t),

width of the box (B^w_t),

height of the box (B^h_t)]



Object Layout Encoding

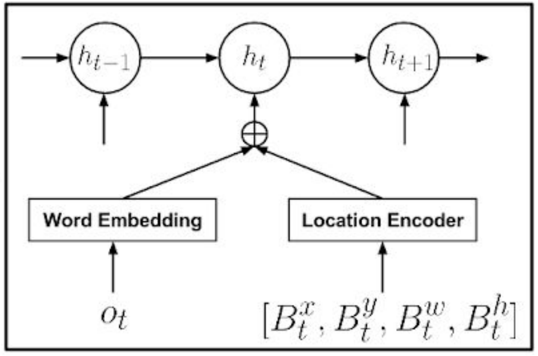
Encoder

Embedding at each time step

$X_t = W_o o_t + W_l [B^x_t, B^y_t, B^w_t, B^h_t]$

Hidden state at each time step

$h^e_t = \mathrm{LSTM}(h^e_{t-1}, x_t | W_{encoder})$



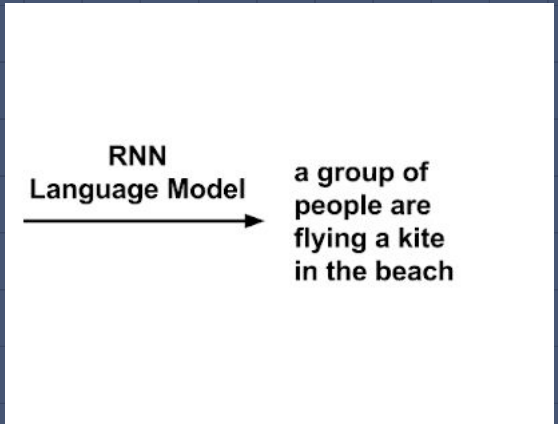
Object Layout Encoding

Decoder

$$p(s|h^{encoder}) = \prod_t p(s_t|h^{encoder}, s_{<t})$$

$$p(s_t|h^{encoder}, s_{<t})$$
$$=\mathrm{softmax}(W_h h^d_{t-1} + b_h)$$

$$h^d_{t-1}$$
$$=\mathrm{LSTM}(h^d_{t-2}, W_s s_{t-1} | W_{decoder})$$

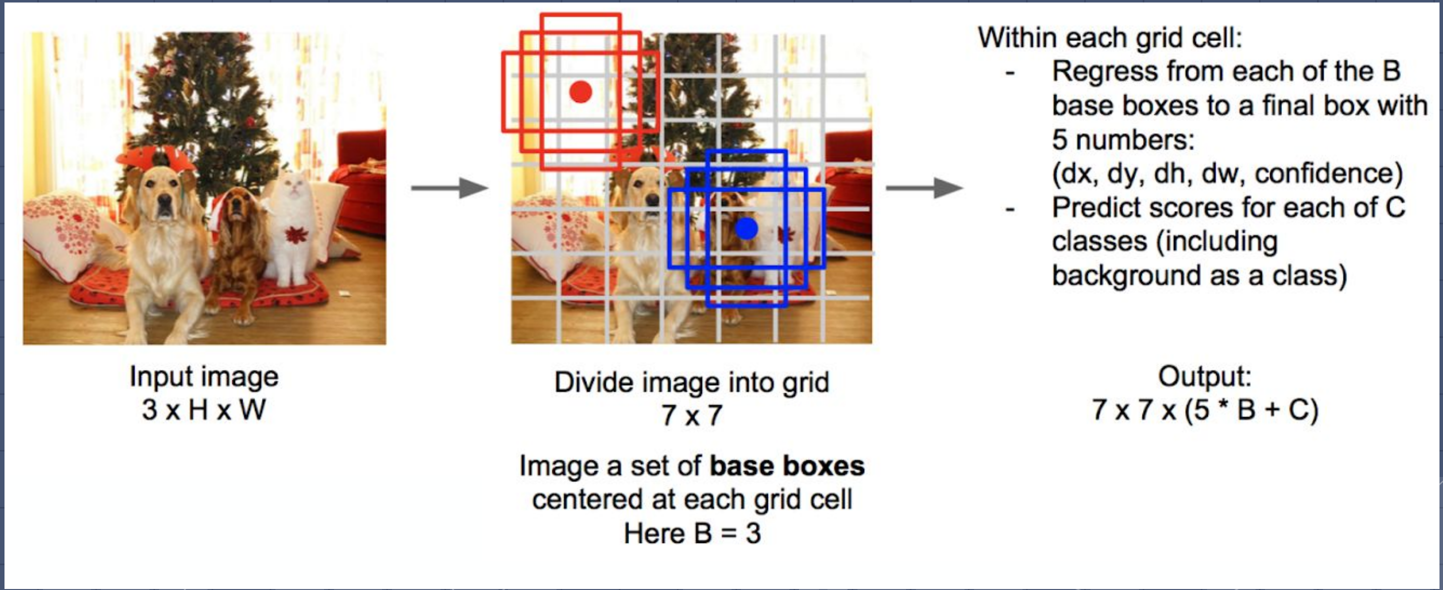


Object Layout Decoding

Variants

- OBJ2TEXT
  - Basic Variant
- OBJ2TEXT-YOLO
  - Object layout are generated from model YOLO instead of taking the ground truth
- OBJ2TEXT-YOLO + CNN-RNN
  - In addition to YOLO , extract visual feature using VGG-16
  - Feed encoded object layouts plus visual feature to the decoder

OBJ2TEXT-YOLO Variant



Redmon et al, CVPR 2016

OBJ2TEXT-YOLO Variant

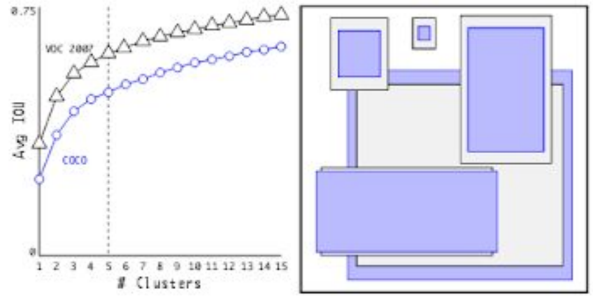


Figure 2: Clustering box dimensions on VOC and COCO. We run k-means clustering on the dimensions of bounding boxes to get good priors for our model. The left image shows the average IOU we get with various choices for k. We find that $k = 5$ gives a good tradeoff for recall vs. complexity of the model. The right image shows the relative centroids for VOC and COCO. Both sets of priors favor thinner, taller boxes while COCO has greater variation in size than VOC.

B^x, B^Y remains the same

$B^w = p_w e^{t1}$

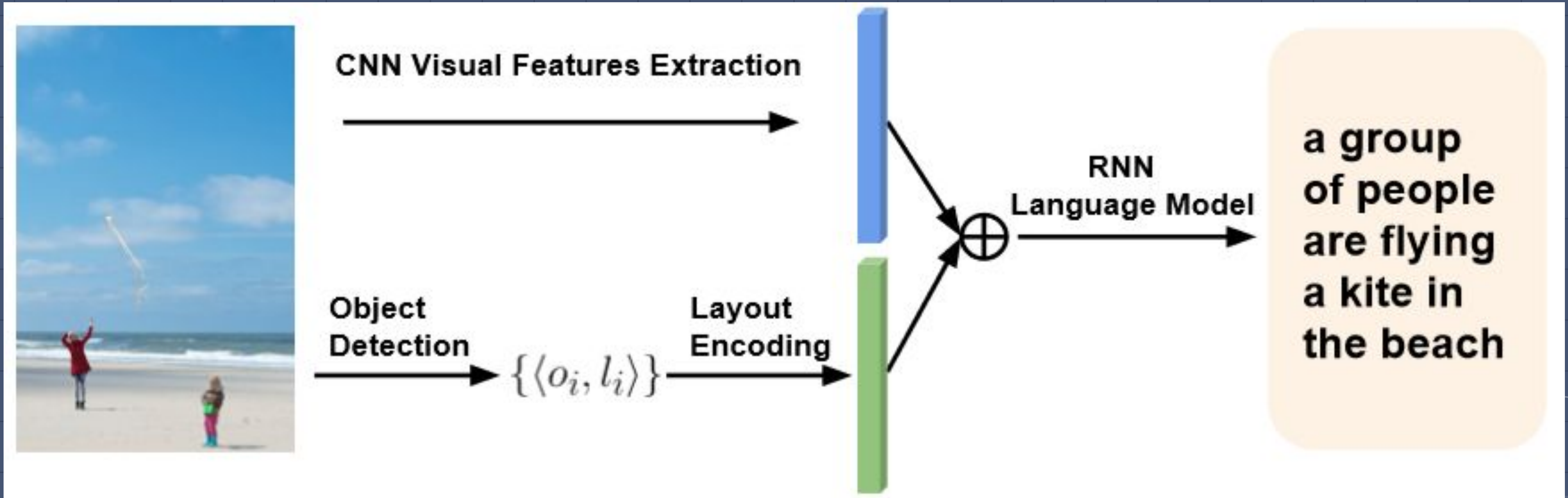
$B^h = p_h e^{t2}$

p_w, p_h are priors

t1, t2 are output from NN

Redmon et al, CVPR 2017

OBJ2TEXT-YOLO + CNN-RNN Variant

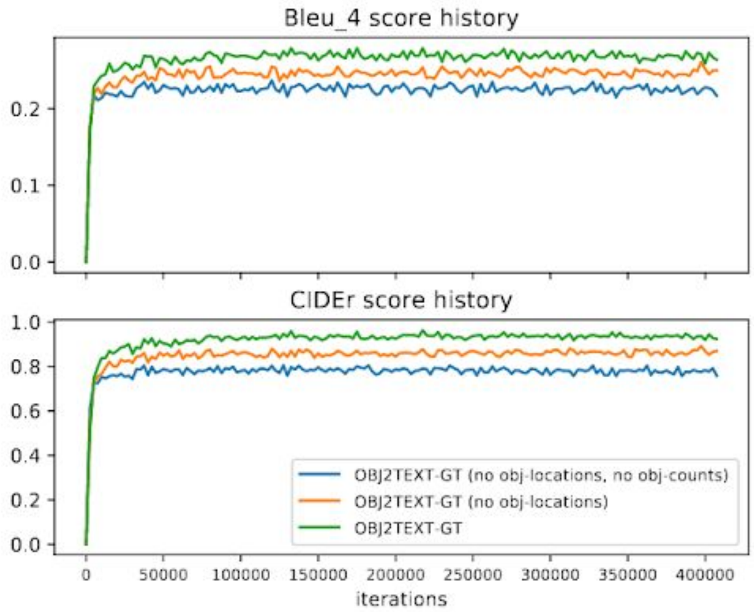


Evaluation

Train & Validation on the MS-COCO training Dataset

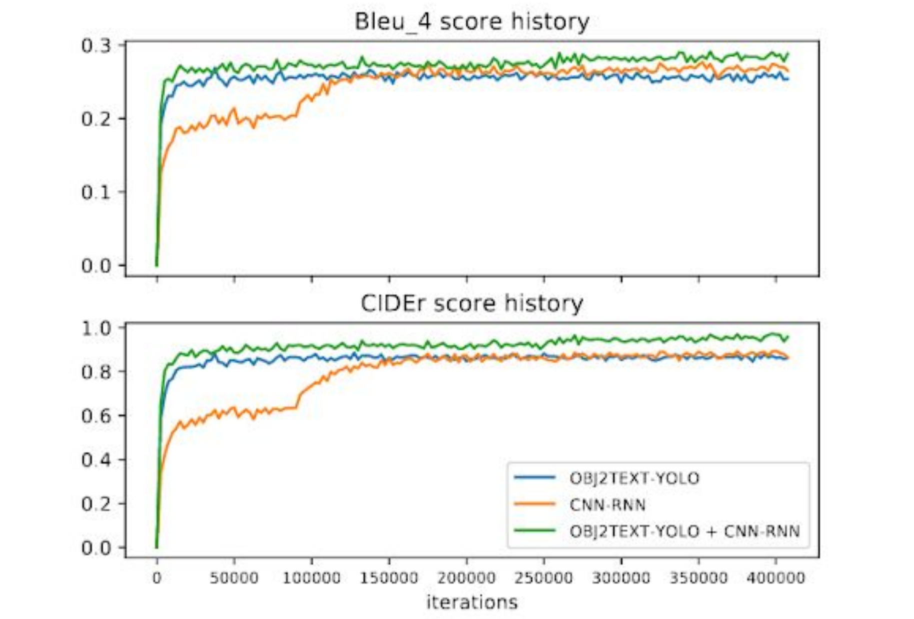
Test on the MS-COCO official test set

Evaluation: Ablation on OBJ2TEXT



(a) Score histories of lesioned versions of the proposed model for the task of object layout captioning.

Evaluation: YOLO-based variants



(b) Score histories of image captioning models. Performance boosts of CNN-RNN and combined model around iteration 100K and 250K are due to fine-tuning of the image CNN model.

Evaluation: Human based

"Two Alternative Forced-Choice Evaluation (2AFC)":

User are presented with one image and two alternatives, choose the best one that describe it.

Done on Amazon Mechanical Turk.

Evaluation: Human based

Alternatives	Choice-all	Choice-agreement	Agreement
OBJ2TEXT-GT vs. OBJ2TEXT-GT (no obj-locations)	54.1%	62.1%	40.6%
OBJ2TEXT-YOLO vs. CNN+RNN	45.6%	40.6%	54.7%
OBJ2TEXT-YOLO + CNN-RNN vs. CNN-RNN	58.1%	65.3%	49.5%
OBJ2TEXT-GT vs. HUMAN	23.6%	9.9%	58.8%

Table 2: Human evaluation results using two-alternative forced choice evaluation. Choice-all is percentage the first alternative was chosen. Choice-agreement is percentage the first alternative was chosen only when all annotators agreed. Agreement is percentage where all annotators agreed (random is 25%).

Choice-All: Percentage of times A was picked over B

Agreement: Percentage of times all 3 user select the same method

Choice-Agreement: Percentage of times A was picked over B and was agreed by all 3 users among all Agreement

Some Sample Output

http://www.cs.virginia.edu/~xy4cm/obj2text/samples/

Potential Extensions

- A better combination with visual features
- Better Seq2Seq model
- Different Training mechanism
- Can this be done in reverse i.e. TEXT2OBJ?