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

Object Layout Input → Object Layout Encoding → Object Layout Decoding

RNN Language Model: a group of people are flying a kite in the beach
Encoder

Each time takes input pair $(o_t, l_t)$

- $o_t = \text{object category one-hot encoding}$
- $l_t = [\text{left-most position } (B^x_t), \text{top-most position } (B^y_t), \text{width of the box } (B^w_t), \text{height of the box } (B^h_t)]$
Encoder

Embedding at each time step
\[ X_t = W_o o_t + W_l [B_{xt}^x, B_{yt}^y, B_{wt}^w, B_{ht}^h] \]

Hidden state at each time step
\[ h^e_t = \text{LSTM}(h^e_{t-1}, x_t | W_{\text{encoder}}) \]
Decoder

\[ p(s|h_{\text{encoder}}) = \prod_t p(s_t|h_{\text{encoder}}, s_{<t}) \]

\[ p(s_t|h_{\text{encoder}}, s_{<t}) = \text{softmax}(W_h h_{t-1}^d + b_h) \]

\[ h_{t-1}^d = \text{LSTM}(h_{t-2}^d, W_s s_{t-1} | W_{\text{decoder}}) \]

RNN Language Model

a group of people are flying a kite in the beach

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**

**Input image**
3 x H x W

**Divide image into grid**
7 x 7

Image a set of **base boxes**
centered at each grid cell
Here B = 3

**Within each grid cell:**
- Regress from each of the B base boxes to a final box with 5 numbers:
  (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

**Output:**
7 x 7 x (5 * B + C)

Redmon et al, CVPR 2016
OBJ2TEXT-YOLO Variant

$B^x, B^Y$ remains the same

$B^w = p_w e^{t_1}$

$B^h = p_h e^{t_2}$

$p_w, p_h$ are priors

t1, t2 are output from NN

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.
OBJ2TEXT-YOLO + CNN-RNN Variant

CNN Visual Features Extraction

Object Detection \{o_i, l_i\} → Layout Encoding

RNN Language Model

a group of people are flying a kite in the beach
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

<table>
<thead>
<tr>
<th>Alternatives</th>
<th>Choice-all</th>
<th>Choice-agreement</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ2TEXT-GT vs. OBJ2TEXT-GT (no obj-locations)</td>
<td>54.1%</td>
<td>62.1%</td>
<td>40.6%</td>
</tr>
<tr>
<td>OBJ2TEXT-YOLO vs. CNN+RNN</td>
<td>45.6%</td>
<td>40.6%</td>
<td>54.7%</td>
</tr>
<tr>
<td>OBJ2TEXT-YOLO + CNN-RNN vs. CNN-RNN</td>
<td>58.1%</td>
<td>65.3%</td>
<td>49.5%</td>
</tr>
<tr>
<td>OBJ2TEXT-GT vs. HUMAN</td>
<td>23.6%</td>
<td>9.9%</td>
<td>58.8%</td>
</tr>
</tbody>
</table>

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?