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

Anderson et al.

Presented by: Abhinav Deep Singh
Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
Motivation

● Previous approaches only used top-down visual attention to attend to image regions with salient features

● **Issue:** These determined regions have no underlying basis; They do not represent anything on their own.

● **Natural basis** of attention: Humans can attend to certain areas using
  ○ **Top Down signals** determined by the current task
  ○ **Bottom up signals** associated with unexpected, novel or salient stimuli
Motivation

Only Top-Down Attention

Top-Down & Bottom-Up Attention
Outline

1. Problem Motivation
2. **Bottom-Up Attention Model**
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
Bottom-Up Attention Model

- “Hard” Attention
- **Rough Aim:** Generate Good Object Regions

**Final output:**
- Variably sized set of $k$ image features
- $\mathbf{V} = \{v_1, ..., v_k\}, v_i \in \mathbb{R}^D$
- Here, $D = 2048$

- Faster RCNN model with ResNet-101
Bottom-Up Attention Model

- Refresher: Faster RCNN

  - Mean Pooled Layer with dimension D set to 2048

  - Attribute Class Training
    - Input: Mean Features “D”
    - Input: Learned embedding of GT class
    - Softmax

  - Initialized with ImageNet weights

  - Then, trained on Visual Genome Dataset

Source: Ren et al, 2016 [2]
Bottom-Up Attention Model

- blue sky
- man
- mountains
- green trees
- black hair
- red bridge
- blue water
- black jacket
- hand
- blue motorcycle
- blue vest
- hand
- green grass
- black wheel
- smiling man
- hill
- blue jeans
- woman
- long hair
- all
- kitchen
- red headband
- black hair
- black glasses
- black glasses
- stove
- white cabinet
- white outlet
- silver pot
- white stove top
- green bottle
- black burner
- open oven
- food
Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. **Captioning Model**
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
Captioning Model

- Uses two LSTM layers:
  - 1st layer: Top-Down Visual Attention Model
  - 2nd layer: Language Model
Captioning Model

Layer 1: Top-Down Visual Attention LSTM

- "Soft" Attention

- Input:
  - $h_{t-1}^2$: Previous output of language LSTM
  - $ar{v} = \frac{1}{k} \sum v^i$: (mean image feature)
  - $W_e$: Word embedding matrix
  - $\Pi_t$: One-hot encoding of word at timestep $t$

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

Layer 2: Language LSTM

- **Input:**
  - $h^1_t$: Output of Top-Down Attention
  - $\hat{v}_t$: Attended image feature

- **Output:**
  - $p(y_t \mid y_{1:t-1}) = \text{softmax}(W_{\theta} h^2_t + b_p)$
  - $p(y_{1:T}) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1})$

- **Objective:**
  - Cross Entropy Loss
  - CIDEr Optimisation (Sampling based optimization)
A man sitting on a couch in a bathroom
Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. **Captioning Results**
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
## Captioning Results on MSCOCO Dataset

- Single model tested on MSCOCO Karpathy Splits

<table>
<thead>
<tr>
<th></th>
<th>Cross-Entropy Loss</th>
<th>CIDEr Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-1</td>
<td>BLEU-4</td>
</tr>
<tr>
<td>SCST: Att2in [34]</td>
<td>-</td>
<td>31.3</td>
</tr>
<tr>
<td>SCST: Att2all [34]</td>
<td>-</td>
<td>30.0</td>
</tr>
<tr>
<td>Ours: ResNet</td>
<td>74.5</td>
<td>33.4</td>
</tr>
<tr>
<td>Ours: Up-Down</td>
<td>77.2</td>
<td>36.2</td>
</tr>
<tr>
<td>Relative Improvement</td>
<td>4%</td>
<td>8%</td>
</tr>
</tbody>
</table>

- **BLEU-1**: BiLingual Evaluation Understudy (unigram)
- **BLEU-4**: BiLingual Evaluation Understudy (4 gram)
- **METEOR**: Metric for Evaluation of Translation with Explicit ORDERing
- **ROUGE-L**: Recall-Oriented Understudy for Gisting Evaluation (Longest Common Subsequence)
- **CIDEr**: Consensus-based Image Description Evaluation
- **SPICE**: Semantic Propositional Image Caption Evaluation
Resnet – A man sitting on a *toilet* in a bathroom.

Up-Down – A man sitting on a *couch* in a bathroom.
## Captioning Results on MSCOCO Dataset

- An ensemble of 4 models tested on MSCOCO online test server

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-1 c5</th>
<th>BLEU-1 c40</th>
<th>BLEU-2 c5</th>
<th>BLEU-2 c40</th>
<th>BLEU-3 c5</th>
<th>BLEU-3 c40</th>
<th>BLEU-4 c5</th>
<th>BLEU-4 c40</th>
<th>METEOR c5</th>
<th>METEOR c40</th>
<th>ROUGE-L c5</th>
<th>ROUGE-L c40</th>
<th>CIDEr c5</th>
<th>CIDEr c40</th>
<th>SPICE c5</th>
<th>SPICE c40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Net [48]</td>
<td>72.0</td>
<td>90.0</td>
<td>55.0</td>
<td>81.2</td>
<td>41.4</td>
<td>70.5</td>
<td>31.3</td>
<td>59.7</td>
<td>25.6</td>
<td>34.7</td>
<td>53.3</td>
<td>68.6</td>
<td>96.5</td>
<td>96.9</td>
<td>18.5</td>
<td>64.9</td>
</tr>
<tr>
<td>Adaptive [27]</td>
<td>74.8</td>
<td>92.0</td>
<td>58.4</td>
<td>84.5</td>
<td>44.4</td>
<td>74.4</td>
<td>33.6</td>
<td>63.7</td>
<td>26.4</td>
<td>35.9</td>
<td>55.0</td>
<td>70.5</td>
<td>104.2</td>
<td>105.9</td>
<td>19.7</td>
<td>67.3</td>
</tr>
<tr>
<td>PG-BCM [24]</td>
<td>75.4</td>
<td>-</td>
<td>59.1</td>
<td>-</td>
<td>44.5</td>
<td>-</td>
<td>33.2</td>
<td>-</td>
<td>25.7</td>
<td>-</td>
<td>55</td>
<td>-</td>
<td>101.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SCST:Att2all [34]</td>
<td>78.1</td>
<td>93.7</td>
<td>61.9</td>
<td>86.0</td>
<td>47.0</td>
<td>75.9</td>
<td>35.2</td>
<td>64.5</td>
<td>27.0</td>
<td>35.5</td>
<td>56.3</td>
<td>70.7</td>
<td>114.7</td>
<td>116.7</td>
<td>20.7</td>
<td>68.9</td>
</tr>
<tr>
<td>LSTM-A3 [49]</td>
<td>78.7</td>
<td>93.7</td>
<td>62.7</td>
<td>86.7</td>
<td>47.6</td>
<td>76.5</td>
<td>35.6</td>
<td>65.2</td>
<td>27</td>
<td>35.4</td>
<td>56.4</td>
<td>70.5</td>
<td>116</td>
<td>118</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours: Up-Down</td>
<td><strong>80.2</strong></td>
<td><strong>95.2</strong></td>
<td><strong>64.1</strong></td>
<td><strong>88.8</strong></td>
<td><strong>49.1</strong></td>
<td><strong>79.4</strong></td>
<td><strong>36.9</strong></td>
<td><strong>68.5</strong></td>
<td><strong>27.6</strong></td>
<td><strong>36.7</strong></td>
<td><strong>57.1</strong></td>
<td><strong>72.4</strong></td>
<td><strong>117.9</strong></td>
<td><strong>120.5</strong></td>
<td><strong>21.5</strong></td>
<td><strong>71.5</strong></td>
</tr>
</tbody>
</table>
Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
VQA Model

- “Soft” attention
- Joint multimodal embedding of question and image
- Regression of scores over set of candidate answers
VQA Model

- Uses Glove embeddings for questions
- Output vocabulary: 3129 for VQA v2.0 dataset

Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
VQA Results on VQA v2.0 Dataset

- Comparison with Resnet-200 and different spatial size
- Better performance, even though Resnet baselines have almost double the number of convolution layers

<table>
<thead>
<tr>
<th></th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: ResNet (1×1)</td>
<td>76.0</td>
<td>36.5</td>
<td>46.8</td>
<td>56.3</td>
</tr>
<tr>
<td>Ours: ResNet (14×14)</td>
<td>76.6</td>
<td>36.2</td>
<td>49.5</td>
<td>57.9</td>
</tr>
<tr>
<td>Ours: ResNet (7×7)</td>
<td>77.6</td>
<td>37.7</td>
<td>51.5</td>
<td>59.4</td>
</tr>
<tr>
<td>Ours: Up-Down</td>
<td>80.3</td>
<td>42.8</td>
<td>55.8</td>
<td>63.2</td>
</tr>
<tr>
<td>Relative Improvement</td>
<td>3%</td>
<td>14%</td>
<td>8%</td>
<td>6%</td>
</tr>
</tbody>
</table>
## VQA Results on VQA v2.0 Dataset

- An ensemble of 30 models tested on VQA v2.0 test-standard server

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

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
Critique

+ Natural Approach
  ○ More interpretable attention weights
+ Complementary to other models
  ○ Just need to replace the attention candidates
+ Gives benefit over simple attention for selecting number of image regions
  ○ As no trade-off between coarse vs fine features for details
+ Code and models are available
  ○ https://www.panderson.me/up-down-attention/
Critique

- Failure Cases:
  
  ○ Image Captioning: Poor salient region cropping

  **Prediction:** A dog laying in the grass with a frisbee
  **Ground Truth:** A dog jumping in the grass with a frisbee
Critique

- Failure Cases:
  ○ VQA: Counting capability

Critique

- Model decisions not consistent (1): Word Embeddings
  - Used GloVe embedding only in VQA,
  - But used randomly initialized embedding is used in Image Captioning

- Model decisions not consistent (2): RNN Unit Choice
  - Used GRUs for VQA
  - But only used LSTMs in Image Captioning
Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
Future Work

- Apply to any other language-vision tasks, and see the performance improvement if any.
- Use other techniques for encoding language: GRUs, Bidirectional attention, Transformers etc.
- Other detection networks which have better/faster performance can be used to generate candidates
- Maybe work with Multi-class Instance Segmentation over Detection
  - Even finer/purer image features
  - Can prevent poor region cropping from Faster-RCNN
Outline

1. Problem Motivation
2. Bottom-Up Attention Model
3. Captioning Model
4. Captioning Results
5. VQA Model
6. VQA Results
7. Critique
8. Future Work
9. References
References


Note: All images and tables in the slides are taken from the paper [1], unless otherwise stated.
Thank You