Attention Based VQA Methods

Nick Jiang
Papers

1. Where to Look
2. Ask, Attend, and Answer
Where to Look: Focus Regions for Visual Question Answering
Shih et al., 2016

1 Goal and Approach
- Problem Overview
- Attention-Based VQA

2 Model Architecture
- Model Overview
- Image Features
- Language Representation
- Region Selection Layer
- High-Level and Training

3 Experiments/Results
- Testing Overview
- Ablation/Comparison Tests
- Qualitative Analysis
- Region Evaluation
1 Goal and Approach

- Problem Overview
- Attention-Based VQA
Goal and Approach: Problem Overview

**Problem Setting:** Answer multiple-choice natural language questions about images where the answers rely on specific visual information contained in the images.

- Is it raining?
- What color is the walk light?
Goal and Approach: Attention-Based VQA

Is it raining?

What color is the walk light?
**Intuition:** When answering visual questions there are often specific areas of the image relevant to figuring out the answer.

**Insight:** Improve performance by use of attention, figuring out “where to look” and explicitly incorporating this information into the model.
Attention: Given the same image, the model should vary its focus based on the query.
2 Model Architecture

- Model Overview
- Image Features
- Language Representation
- Region Selection Layer
- High-Level and Training
Model Architecture: Model Overview

**Input:**
Fixed-size image & Word2Vec encoding of variable-length question/answer pair

**Output:**
Score of how well the input answer answers the input question correctly
Model Architecture: Image Features

Edge boxes

CNN Features for Each Region

Region₁

Region₂

... ⋱

Regionₙ
Model Architecture: Image Features (cont.)

Use Edge Boxes to extract 99 candidate regions + 1 region containing the whole image
Run each region through the VGG-s network

Concatenate the last fully connected layer (4096-d) and the pre-softmax layer (1000-d) to get the region feature vector (5096-d)
Model Architecture: Language Representation

Word2Vec
Encoding on Q/A pair
“What color is the fire hydrant?”/“yellow”

Text Q/A 3 layer net embedding

Text Features
Stanford Parser-based Bins of Word2Vec Averages

1. Question type (first 2 words)
2. Nominal Question Subject
3. Other Question Nouns
4. Other Question Words
5. Answer Words

Concatenate each 300-d bin to get a single 1500-d question/answer representation then embed to 1024-d vector

**Insight:** captures important components of a variable length question/answer while maintaining a fixed-length representation
Model Architecture:
Region Selection Layer
Model Architecture:
Region Selection Layer (cont.)
1. Inputs:
   a. 5096-d image features for each of 100 image regions
   b. 1024-d language features for the question/answer pair
2. Project the image and language features into the same 900-d space
3. Region weights as dot product of language and image vectors
4. **Insight:** identify relevant image regions given text features
Dot product explicitly embodies attention as region weights are high when region content and question/answer pair embeddings are similar.

q1: [Is it raining?] [Yes]
\[<q_1, b_1> = 0.3, <q_1, b_2> = 0.05, <q_1, b_3> = 0.07\]

q2: [What color is the walk light?] [Green]
\[<q_2, b_1> = 0.04, <q_2, b_2> = 0.03, <q_2, b_3> = 0.05\]
Model Architecture:
Region Selection Layer

CNN Features for Each Region
- Region_1
- Region_2
- ... Region_N

Per-Region Vision/Text Features
- Region_1 + Text
- Region_2 + Text
- ... Region_N + Text

Embedded W on concatenated region/text

Weighted Average Features
For each region concatenate the 5096-d region image features and 1024-d text features

Use a linear projection to embed them into a 2048-d vision-text feature vector for each region
Model Architecture: Region Selection Layer

CNN Features for Each Region

Region_1
Region_2
... Region_N

Text Features

Per-Region Weights

s_1
s_2
... s_N

Embedding A
dot product, softmax

Embedding B

Per-Region Vision/Text Features

Region_1 + Text
Region_2 + Text
... Region_N + Text

Weighted Average Features

Region Selection Layer

Embedding W on concatenated region/text
Using the per-region weights, compute a weighted sum of the per-region vision-text features to obtain a single 2048-d weighted average feature vector.

**Insight:** This vector represents the information captured in the image and text when focusing on the relevant regions.
Model Architecture: High-Level and Training

- Weighted average features run through small network to generate final score.
- For both training and testing, the question and each candidate answer are run through the network, generating a final score for each candidate answer.
Complete model architecture
Model Architecture: High-Level and Training (cont.)

Training & Testing

- The loss function used for training is a maximum margin/structured hinge loss over the scores for each answer, requiring that the score of the correct answer be above the highest scoring incorrect answer by a margin equal to the annotator margin

- Ex: If 6/10 annotators answer $p$ and 4/10 answer $n$ then $y_p$ should outscore $y_n$ by a margin of $\geq 0.2$

\[
\mathcal{L}(y) = \max_{\forall n \neq p} \left( 0, y_n + (a_p - a_n) - y_p \right)
\]

- **Insight:** answers could be acceptable to varying degrees since correctness is determined by consensus of 10 annotators.
3 Experiments/Results

- Testing Overview
- Ablation/Comparison Tests
- Qualitative Analysis
- Region Evaluation
Experiments/Results:
Testing Overview

- Tested on MS COCO VQA dataset
  - ~83k Training, ~41k Validation, ~81k Testing
  - 3 questions per image
  - 10 free-response answers per question
  - 18-way multiple choice

- **Insight:** chose this dataset due to the open-ended nature of the language in both question and answers and chose multiple choice tasks as evaluation is much less ambiguous than open-ended answer verification
Experiments/Results: Ablation/Comparison Tests

- **Ablation Testing**: removing parts of the model to test whether those parts are actually beneficial to performance.

- This model was tested against 3 separate baselines:
  - **Language-only**: baseline to demonstrate improvement due to image features.
  - **Word+Whole image**: baseline to demonstrate improvement due to selecting image regions.
  - **Word+Uniform averaged region features**: baseline to demonstrate improvement due to weighting of image regions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Only</td>
<td>53.98</td>
</tr>
<tr>
<td>Word+Whole Image</td>
<td>57.83</td>
</tr>
<tr>
<td>Word+ave. reg.</td>
<td>57.88</td>
</tr>
<tr>
<td>Word+Region Sel.</td>
<td>58.94</td>
</tr>
<tr>
<td>Category</td>
<td>Region</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>overall</td>
<td>58.94</td>
</tr>
<tr>
<td>is/are/was</td>
<td>75.42</td>
</tr>
<tr>
<td>identify: what kind/type/animal</td>
<td>52.89</td>
</tr>
<tr>
<td>how many</td>
<td>33.38</td>
</tr>
<tr>
<td><strong>what color</strong></td>
<td><strong>53.96</strong></td>
</tr>
<tr>
<td>interpret: can/could/does/has</td>
<td>75.73</td>
</tr>
<tr>
<td>none of the above</td>
<td>45.40</td>
</tr>
<tr>
<td>where</td>
<td>42.11</td>
</tr>
<tr>
<td>why/how</td>
<td>26.31</td>
</tr>
<tr>
<td>relational: what is the man/woman</td>
<td>70.15</td>
</tr>
<tr>
<td>relational: what is in/on</td>
<td>54.78</td>
</tr>
<tr>
<td>which/who</td>
<td>43.97</td>
</tr>
<tr>
<td>reading: what does/number/name</td>
<td>33.31</td>
</tr>
<tr>
<td><strong>identify scene: what room/sport</strong></td>
<td><strong>86.21</strong></td>
</tr>
<tr>
<td>what time</td>
<td>41.47</td>
</tr>
<tr>
<td>what brand</td>
<td>45.40</td>
</tr>
</tbody>
</table>

Ablation testing results across question categories
Ablation testing reveals clear improvements in specific question categories and continued failures in others.

<table>
<thead>
<tr>
<th></th>
<th>region</th>
<th>image</th>
<th>text</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>why/how</td>
<td>26.31</td>
<td>28.18</td>
<td>29.24</td>
<td>2.2%</td>
</tr>
<tr>
<td>how many</td>
<td>33.38</td>
<td>36.84</td>
<td>34.05</td>
<td>10.3%</td>
</tr>
<tr>
<td>what color</td>
<td>53.96</td>
<td>43.52</td>
<td>32.59</td>
<td>9.8%</td>
</tr>
<tr>
<td>identify scene: what room/sport</td>
<td>86.21</td>
<td>76.65</td>
<td>61.26</td>
<td>0.9%</td>
</tr>
</tbody>
</table>
### Experiments/Results:
#### Ablation/Comparison Tests (cont.)

<table>
<thead>
<tr>
<th>Model</th>
<th>All</th>
<th>Y/N</th>
<th>Num.</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>test-dev</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM Q+I (Antol et al. 2015)</td>
<td>57.17</td>
<td>78.95</td>
<td>35.80</td>
<td>43.41</td>
</tr>
<tr>
<td>Q+I (Antol et al. 2015)</td>
<td>58.97</td>
<td>75.97</td>
<td>34.35</td>
<td>50.33</td>
</tr>
<tr>
<td>iBOWIMG (Zhou et al. 2015)</td>
<td>61.68</td>
<td>76.68</td>
<td>37.05</td>
<td>54.44</td>
</tr>
<tr>
<td>DPPnet (Noh et al 2016)</td>
<td>62.48</td>
<td><strong>80.79</strong></td>
<td><strong>38.94</strong></td>
<td>52.16</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>63.3</strong></td>
<td>78.09</td>
<td>34.22</td>
<td><strong>57.23</strong></td>
</tr>
<tr>
<td><strong>test-standard</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deeperLSTM_NormalizeCNN (Antol et al. 2016)</td>
<td>63.09</td>
<td>80.59</td>
<td>37.7</td>
<td>53.64</td>
</tr>
<tr>
<td>iBOWIMG (Zhou et al. 2015)</td>
<td>61.97</td>
<td>76.86</td>
<td>37.30</td>
<td>54.60</td>
</tr>
<tr>
<td>DPPnet (Noh et al 2016)</td>
<td>62.69</td>
<td><strong>80.35</strong></td>
<td><strong>38.79</strong></td>
<td>52.79</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>63.53</strong></td>
<td>78.08</td>
<td>34.26</td>
<td><strong>57.43</strong></td>
</tr>
<tr>
<td>Model</td>
<td>All</td>
<td>Y/N</td>
<td>Num.</td>
<td>Others</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>test-dev</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM Q+I (Antol et al. 2015)</td>
<td>57.17</td>
<td>78.95</td>
<td>35.80</td>
<td>43.41</td>
</tr>
<tr>
<td>Q+I (Antol et al. 2015)</td>
<td>58.97</td>
<td>75.97</td>
<td>34.35</td>
<td>50.33</td>
</tr>
<tr>
<td>iBOWIMG (Zhou et al. 2015)</td>
<td>61.68</td>
<td>76.68</td>
<td>37.05</td>
<td>54.44</td>
</tr>
<tr>
<td>DPPnet (Noh et al 2016)</td>
<td>62.48</td>
<td><strong>80.79</strong></td>
<td><strong>38.94</strong></td>
<td>52.16</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>63.3</strong></td>
<td>78.09</td>
<td>34.22</td>
<td><strong>57.23</strong></td>
</tr>
<tr>
<td><strong>test-standard</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deeperLSTM_NormalizeCNN (Antol et al. 2016)</td>
<td>63.09</td>
<td>80.59</td>
<td>37.7</td>
<td>53.64</td>
</tr>
<tr>
<td>iBOWIMG (Zhou et al. 2015)</td>
<td>61.97</td>
<td>76.86</td>
<td>37.30</td>
<td>54.60</td>
</tr>
<tr>
<td>DPPnet (Noh et al 2016)</td>
<td>62.69</td>
<td><strong>80.35</strong></td>
<td><strong>38.79</strong></td>
<td>52.79</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>63.53</strong></td>
<td>78.08</td>
<td>34.26</td>
<td><strong>57.43</strong></td>
</tr>
</tbody>
</table>
In addition to quantitative comparisons, the paper also presents qualitative results where the attention can be visualized given an image and question/answer pair.
What room is this?
Answer: Kitchen

Kitchen: 22.3

Attention weights visualization
What color on the stop light is lit up?  

![Image of a traffic light]

T: red (-0.1)  
I: red (-0.8)  
R: green (1.1)  
Ans: green

What color is the light?  

![Image of a street sign]

T: red (1.0)  
I: red (0.3)  
R: red (1.7)  
Ans: red

What color is the street sign?  

![Image of a street sign]

T: gray (-0.2)  
I: gray (-0.4)  
R: yellow (0.4)  
Ans: yellow

T: text-only, I: whole image+text, R: region-selection. Margin shown in parenthesis (ground truth confidence - top incorrect)
What color on the stop light is lit up?  
L: red (-0.1)  
I: red (-0.8)  
R: green (1.1)  
Ans: green

What color is the light?  
L: red (1.0)  
I: red (0.3)  
R: red (1.7)  
Ans: red

What color is the street sign?  
L: gray (-0.2)  
I: gray (-0.4)  
R: yellow (0.4)  
Ans: yellow

What color is the fence?  
L: black (-0.7)  
I: gray (-0.6)  
R: white (0.1)  
Ans: white

What animal is that?  
L: sheep (1.1)  
I: sheep (2.5)  
R: sheep (0.0)  
Ans: sheep

How many birds are in the sky?  
L: 1 (-0.7)  
I: several (-0.1)  
R: 9600 (-0.2)  
Ans: 5

What is the woman flying over the beach?  
L: goose (-1.1)  
I: kite (1.4)  
R: kite (5.3)  
Ans: kite

What color is the walk light?  
L: red (-0.3)  
I: red (-0.3)  
R: green (1.1)  
Ans: green
How many birds are in the sky?

L: 1 (-0.7)
I: several (-0.1)
R: 9600 (-0.2)
Ans: 5

Attention for a counting question shows focus on the correct object despite incorrect final answer.
How many people?
L: 4 (0.0)
I: 3 (-0.1)
R: 2 (-0.2)
Ans: 8

What is on the ground?
L: airplane(-0.9)
I: snow (2.9)
R: snow (3.7)
Ans: snow

What room is this?
L: bathroom(0.1)
I: bathroom (2.6)
R: bathroom (6.8)
Ans: bathroom

Is the faucet turned on?
L: no (3.6)
I: no (3.1)
R: no (5.1)
Ans: no

What is behind the man?
L: dog(0.0)
I: dog (0.0)
R: dog (1.4)
Ans: dog

What is the man doing?
L: surfing (2.5)
I: blue (3.7)
R: surfing (9.7)
Ans: surfing

Where is the shampoo?
L: on shelf (-1.4)
I: on shelf (-0.7)
R: on tub (-0.1)
Ans: windowsill

Is there a lot of pigeons in the picture?
L: yes (1.5)
I: yes (0.5)
R: yes (1.0)
Ans: yes
Experiments/Results: Region Evaluation

- Compare predicted weights to annotated relevant regions
- 72% of the images showed higher weights within annotated regions than outside
- The difference was often much greater than 0 and rarely much smaller
Conclusions

- Attention through weighted region selection shows significant improvements over other VQA methods.
- The performance gains are particularly large for questions that require focusing on specific regions such as “What is the woman holding?”, “What color…?”, “What room…?”
Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Xu et al., 2016

1 Goal and Approach
- Problem Overview
- Spatial Memory Network

2 Model Architecture
- Model Overview
- Image/Word Embeddings
- Word-Guided Attention
- First Hop
- Second Hop

3 Experiments/Results
- Testing Overview
- Exploring Attention on Synthetic Data
- Experiments on Standard Datasets
1 Goal and Approach

- Problem Overview
- Spatial Memory Network
Goal and Approach: Problem Overview

**Problem Setting:** Answer open-ended natural language questions about images where the answers rely on specific visual information contained in the images.

What is the child standing on?
Goal and Approach: Spatial Memory Network

- Recurrent neural network that stores the spatial arrangement of scenes in its visual memory
- Extends the idea of memory networks from NLP which stored information from specific locations in the input to attend over
2 Model Architecture

- Model Overview
- Image/Word Embeddings
- Word-Guided Attention
- First Hop
- Second Hop
Model Architecture: Model Overview

- **Input:**
  - Fixed-size image & Variable-length question

- **Output:**
  - Softmax over all possible answers
Image Embeddings

- GoogLeNet spatial features at $L$ gridpoints from the last convolutional layer
- Two separate linear embeddings
- **Attention embedding** maps features to the shared attention space in $\mathbb{R}^N$
- **Evidence embedding** maps to output that captures visual information in each region also in $\mathbb{R}^N$
The words in the question are converted into word vectors $v_j$ in the attention embedding space in $\mathbb{R}^N$.

The model also computes $Q$, a weighted average of the individual word embeddings that acts as a full question embedding.
Model Architecture: Image/Word Embeddings (cont.)

Dimensions:
- $M$: size of visual features for each region extracted from GoogLeNet
- $L$: number of spatial regions
- $T$: length of (padded) question
- $N$: size of attention and evidence embedding space

Objects:
- **Visual features from GoogLeNet**: features extracted from last convolutional layer of GoogleLeNet for each region forming a matrix in $\mathbb{R}^{L \times M}$
- **Attention-embedded visual features**: embedding of each spatial region into shared attention space in $\mathbb{R}^{N}$ collectively forming a matrix in $\mathbb{R}^{L \times N}$
- **Evidence-embedded visual features**: embedding of each spatial region to capture visual semantic information in $\mathbb{R}^{N}$ collectively forming a matrix in $\mathbb{R}^{L \times N}$
- **Embedded individual word vectors**: individual word vectors $v_j$ in $\mathbb{R}^{N}$ representing each question word, collectively forming a matrix $V$ in $\mathbb{R}^{T \times N}$
- **Full question embedding**: weighted average of individual word vectors $Q$ in $\mathbb{R}^{N}$
Model Architecture: Word-Guided Attention

- Attention weights for each region based on highest similarity to any single word in the question

- **Insight:** using individual word vectors instead of a BOW representation leads to more fine-grained attention
What is the child standing on?
Model Architecture: Word-Guided Attention (detailed)

- Take the dot product of each region’s attention-embedded visual features and each word’s embedded features to obtain a correlation matrix in $\mathbb{R}^{L \times T}$
- Take the highest correlation value for each region and softmax to obtain attention weights in $\mathbb{R}^L$
Model Architecture: First Hop

- Attention weighted average of evidence vectors produces selected visual evidence vector $S_{\text{att}}$
- $S_{\text{att}}$ added to question vector $Q$ to get $O_{\text{hop1}}$
- In single hop architecture, $O_{\text{hop1}}$ is directly used to predict answer
Model Architecture: Second Hop

- Output from first hop $O_{hop1}$ is combined with the evidence space embeddings to form new attention weights.
- **Insight:** second hop refines attention based on whole image-question understanding gained from $O_{hop1}$.
**Model Architecture:**

**Second Hop (cont.)**

- $W_{\text{att2}}$ used to calculate new selected visual evidence vector which is used along with $O_{\text{hop1}}$ to generate final predictions.

- **Insight:** second hop adds new information to previous understanding $O_{\text{hop1}}$ to generate better answer.
A second visual evidence embedding is created and weighted according to \( W_{att2} \) to generate \( S_{att2} \).

\( S_{att2} \) and \( O_{hop1} \) are summed and passed through nonlinear + softmax layer to generate final output predictions over possible output space.
Is there a cat in the basket?

Complete model architecture.
3 Experiment/Results

- Testing Overview
- Exploring Attention on Synthetic Data
- Experiments on Standard Datasets
Testing Goals:

- Extensively test ability to form spatial inference
- Compare to existing VQA models by testing on standard datasets
Experiments/Results: Exploring Attention on Synthetic Data

- Create and test on a synthetic dataset specifically designed to evaluate the performance of the spatial attention mechanism.
- Overcomes variation and difficulty associated with standard datasets as well as bias present in question text that makes text-only models a generally strong predictor.
Experiments/Results: Exploring Attention on Synthetic Data (cont.)

Absolute Position Recognition

- One-hop model achieves 100% accuracy while iBOWIMG achieves 75% accuracy (same as always answering “no”)
- Attention learned 2 logical rules:
  - Look at question position for square
  - Look at square and compare to question position
Relative Position Recognition

- One-hop model achieves 96% accuracy while iBOWIMG again achieves 75% accuracy
- Same 2 logical rules learned but this time the position is relative to the cat
- Confused by multiple cats
Is there a red square on the right of the cat?
GT: yes  Prediction: yes

Is there a red square on the right of the cat?
GT: no  Prediction: no

Is there a red square on the left of the cat?
GT: no  Prediction: no

Is there a red square on the top of the cat?
GT: no  Prediction: no

**Left:** original image, **Center:** evidence embedding, **Right:** attention weights
### Experiments/Results:

#### Experiments on Standard Datasets

#### Results on DAQUAR

- Both one-hop and two-hop model outperform all baselines
- Second hop greatly increases performance

<table>
<thead>
<tr>
<th></th>
<th>DAQUAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-World [1]</td>
<td>12.73</td>
</tr>
<tr>
<td>Neural-Image-QA [10]</td>
<td>29.27</td>
</tr>
<tr>
<td>Question LSTM [10]</td>
<td>32.32</td>
</tr>
<tr>
<td>VIS+LSTM [11]</td>
<td>34.41</td>
</tr>
<tr>
<td>IMG+BOW [11]</td>
<td>34.17</td>
</tr>
<tr>
<td>SMem-VQA One-Hop</td>
<td>36.03</td>
</tr>
<tr>
<td>SMem-VQA Two-Hop</td>
<td>40.07</td>
</tr>
</tbody>
</table>
### Experiments/Results:

#### Experiments on Standard Datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>test-dev</th>
<th>yes/no</th>
<th>number</th>
<th>others</th>
<th>Overall</th>
<th>test-standard</th>
<th>yes/no</th>
<th>number</th>
<th>others</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM Q+I [2]</td>
<td>53.74</td>
<td>78.94</td>
<td>35.24</td>
<td>36.42</td>
<td>-</td>
<td>54.06</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ACK* [26]</td>
<td>55.72</td>
<td>79.23</td>
<td>36.13</td>
<td>40.08</td>
<td>-</td>
<td>55.98</td>
<td>79.05</td>
<td>36.10</td>
<td>40.61</td>
<td>-</td>
</tr>
<tr>
<td>DPPnet* [27]</td>
<td>57.22</td>
<td>80.71</td>
<td>37.24</td>
<td>41.69</td>
<td>-</td>
<td>57.36</td>
<td>80.28</td>
<td>36.92</td>
<td>42.24</td>
<td>-</td>
</tr>
<tr>
<td>iBOWIMG [3]</td>
<td>55.72</td>
<td>76.55</td>
<td>35.03</td>
<td>42.62</td>
<td>-</td>
<td>55.89</td>
<td>76.76</td>
<td>34.98</td>
<td>42.62</td>
<td>-</td>
</tr>
<tr>
<td>SMem-VQA One-Hop</td>
<td>56.56</td>
<td>78.98</td>
<td>35.93</td>
<td>42.09</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SMem-VQA Two-Hop</td>
<td><strong>57.99</strong></td>
<td><strong>80.87</strong></td>
<td><strong>37.32</strong></td>
<td><strong>43.12</strong></td>
<td>-</td>
<td><strong>58.24</strong></td>
<td><strong>80.8</strong></td>
<td><strong>37.53</strong></td>
<td><strong>43.48</strong></td>
<td></td>
</tr>
</tbody>
</table>

#### Results on VQA

- Two-hop model shows \(~2.25\%\) performance increase over iBOWIMG
- Two-hop model even outperforms DPPnet model pre-trained on large scale text corpus
Visualization of spatial attention weights for the one-hop and two-hop models.

**Question:** What electrical appliance is the woman using?  
*GT: blender*  
*One Hop: wine*  
*Two Hop: blender*

**Question:** What is the colour of the object near the bed?  
*GT: pink*  
*One Hop: bed*  
*Two Hop: pink*
Conclusions

- Multi-hop model allows combining of fine-grained attention with global knowledge to obtain refined results.
- Attention allows these models to easily represent and learn spatial relationships enabling them to tackle new types of VQA problems.
- Performance is still far from human level especially for certain categories such as counting questions and abstract reasoning questions (“Why/How...?”).
print('end')