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

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?

This and Following Slides Images Credit: *Shih et al.*

Goal and Approach: Attention-Based VQA

Is it raining?

What color is the walk light?

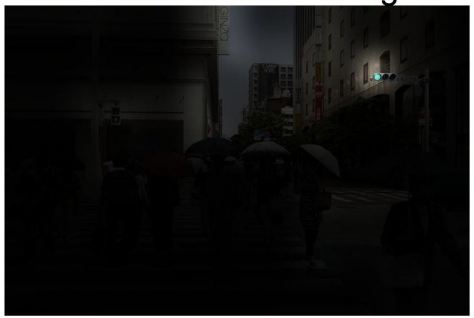
Goal and Approach: Attention-Based VQA (cont.)

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

Is it raining?

What color is the walk light?

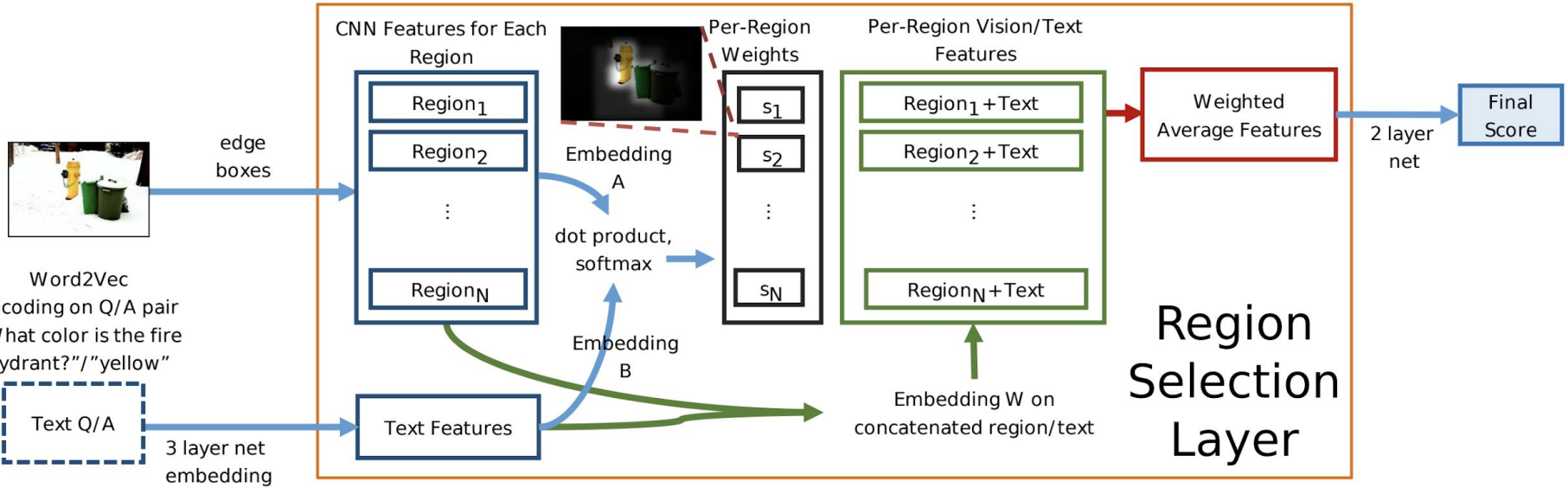


Attention: Given the same image, the model should vary its focus based on the query

2 Model Architecture

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

Model Architecture: Model Overview



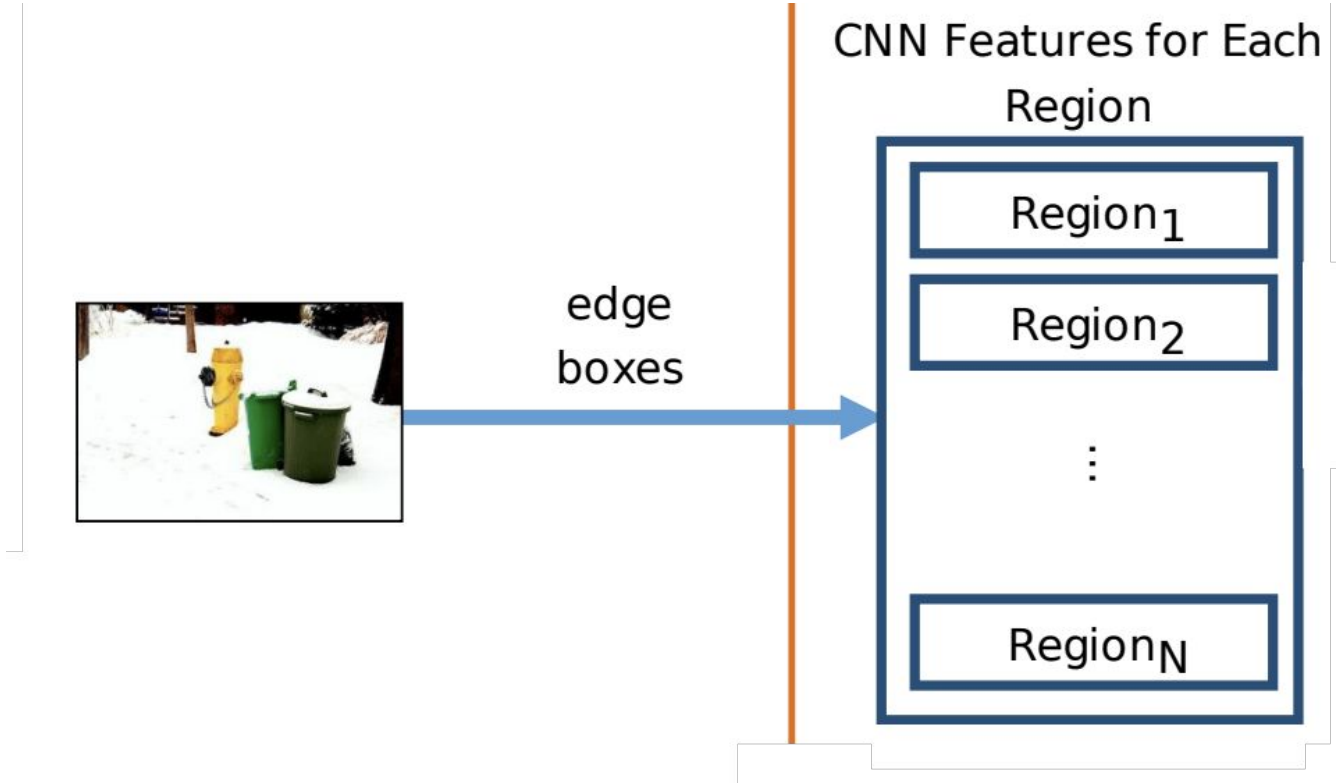
- **Input:**

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

- **Output:**

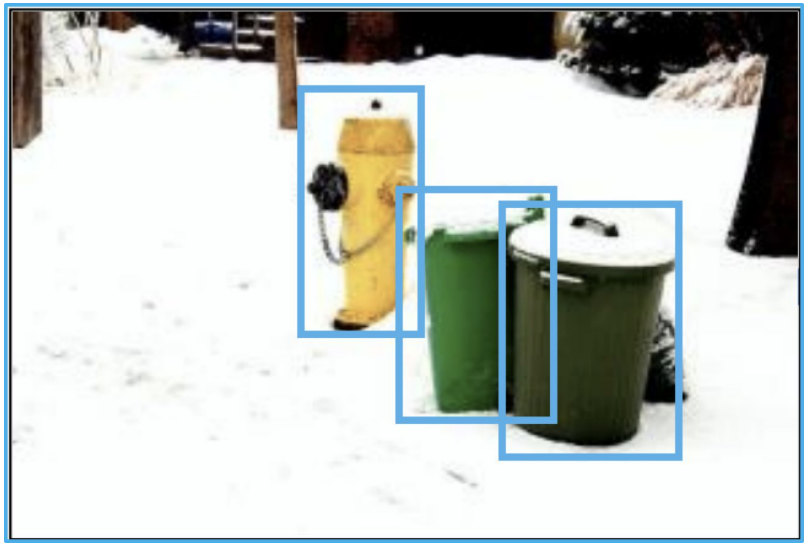
  Score of how well the input answer answers the input question correctly

Model Architecture: Image Features



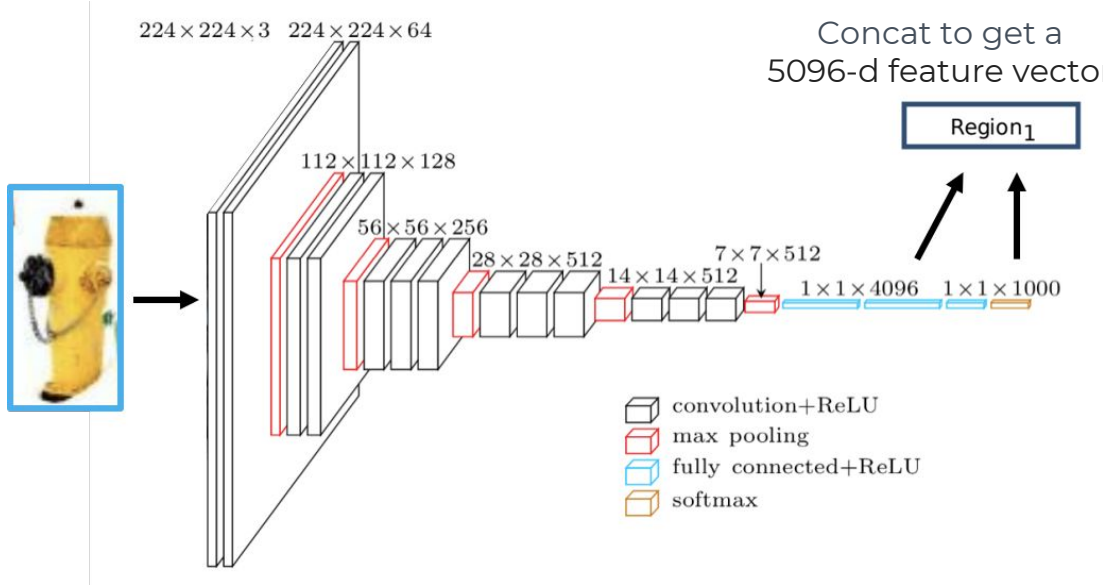
Model Architecture: Image Features (cont.)

Use Edge Boxes to extract 99 candidate regions + 1 region containing the whole image



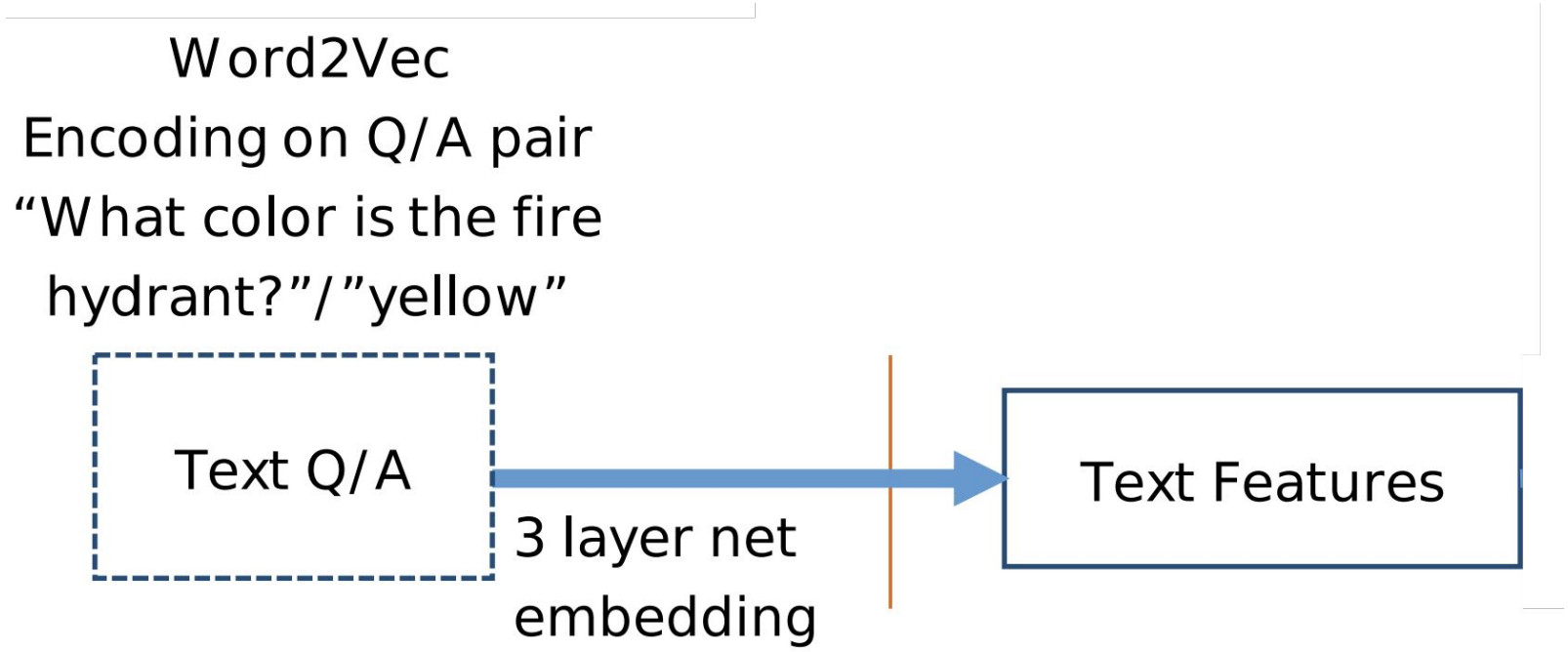
Model Architecture: Image Features (cont.)

- 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)



Additional Image Credit: *Heuritech Blog*

Model Architecture: Language Representation

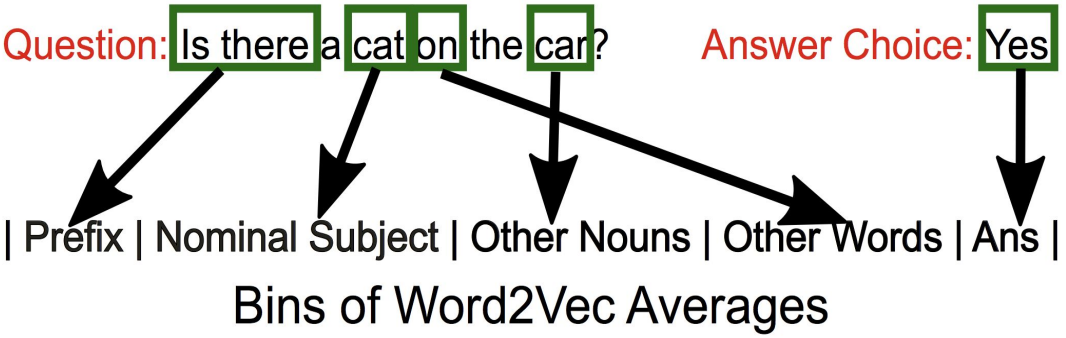


Model Architecture: Language Representation (cont.)

Stanford Parser-based Bins of Word2Vec Averages

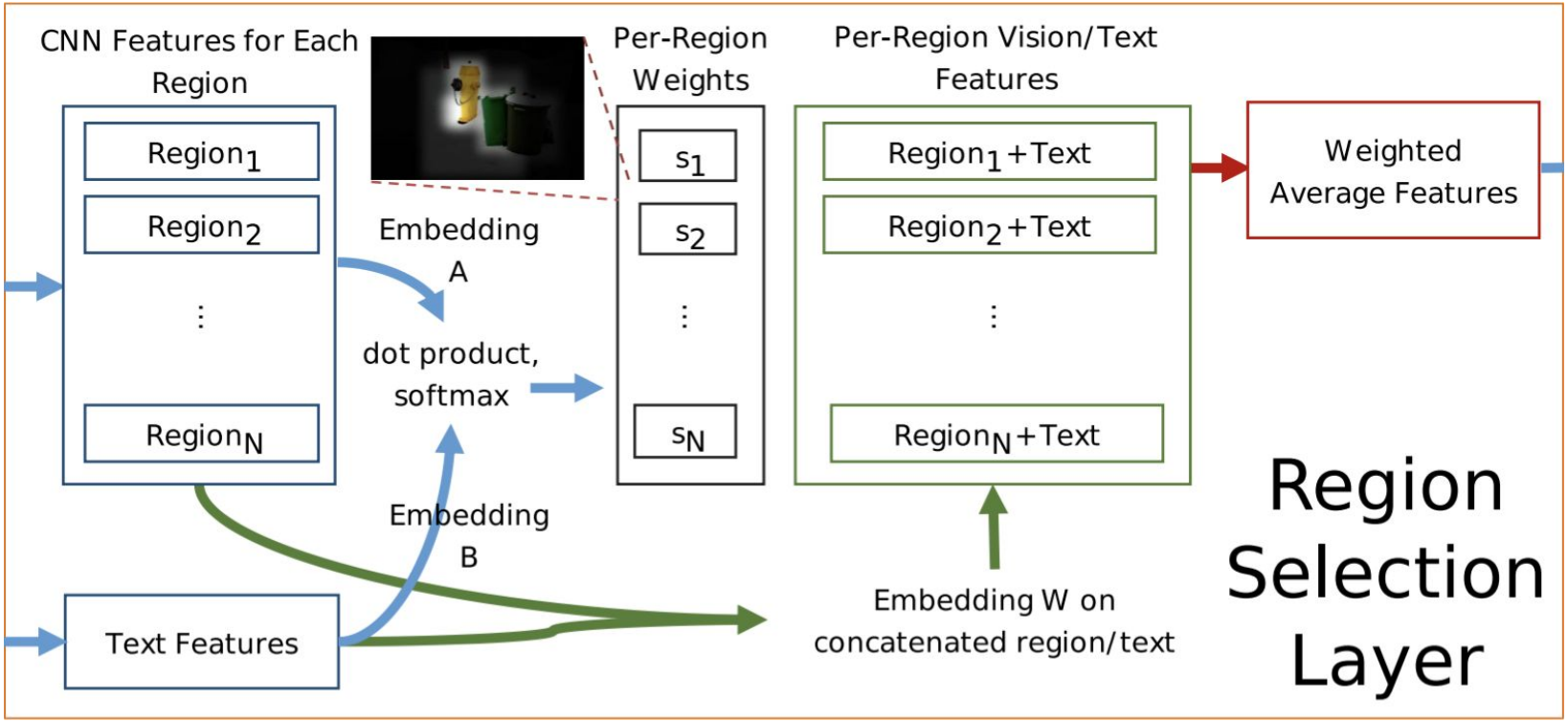
1. Question type (first 2 words)
2. Nominal Question Subject
3. Other Question Nouns
4. Other Question Words
5. Answer Words

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

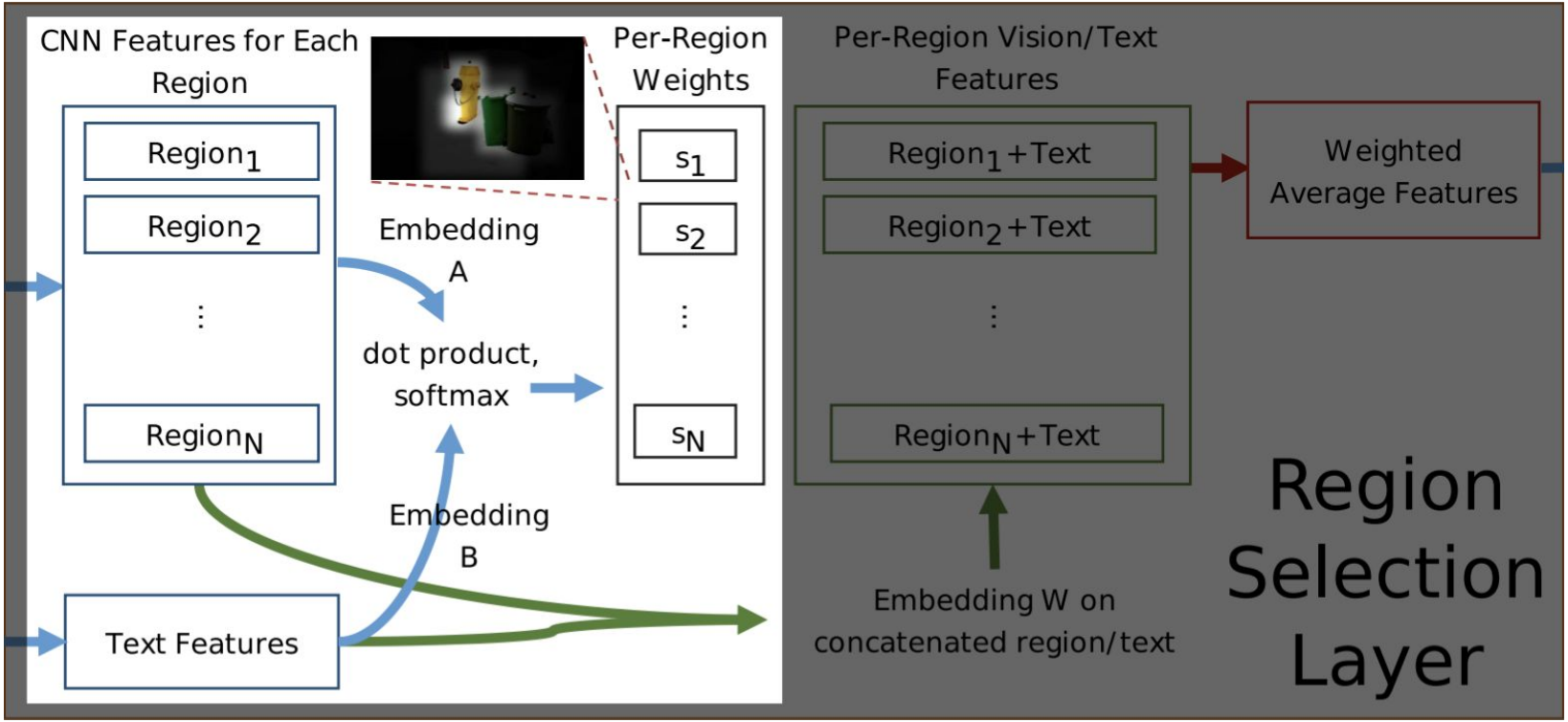


Insight: captures important components of a variable length question/answer while maintaining a fixed-length representation

Model Architecture: Region Selection Layer

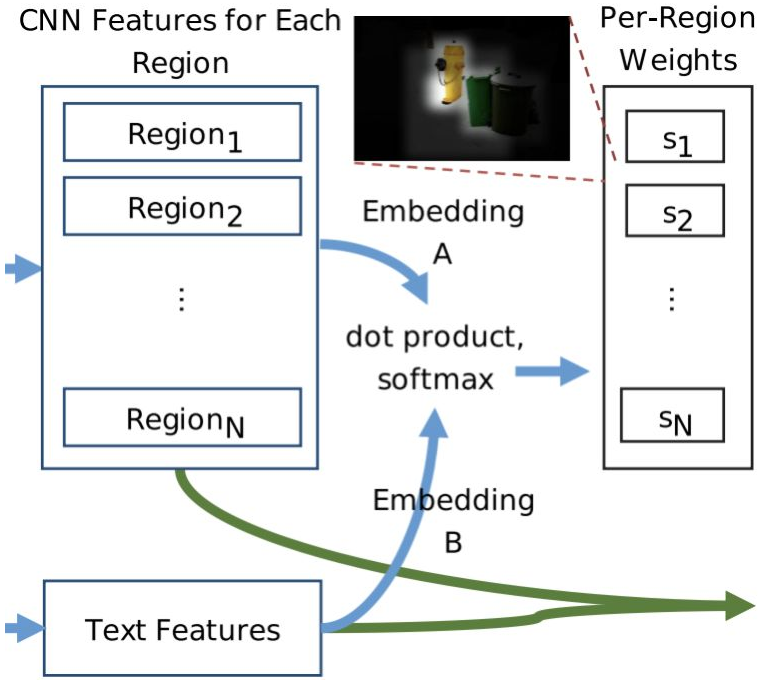


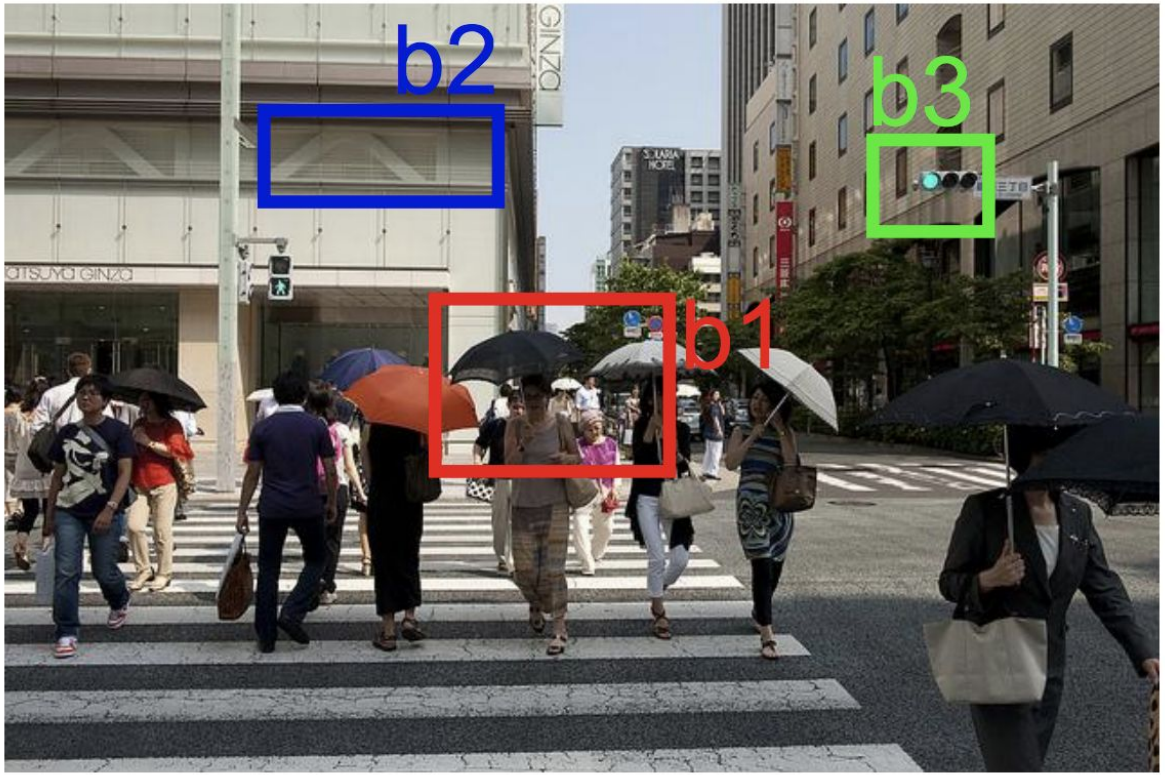
Model Architecture: Region Selection Layer (cont.)



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





q1: [Is it raining?][Yes]

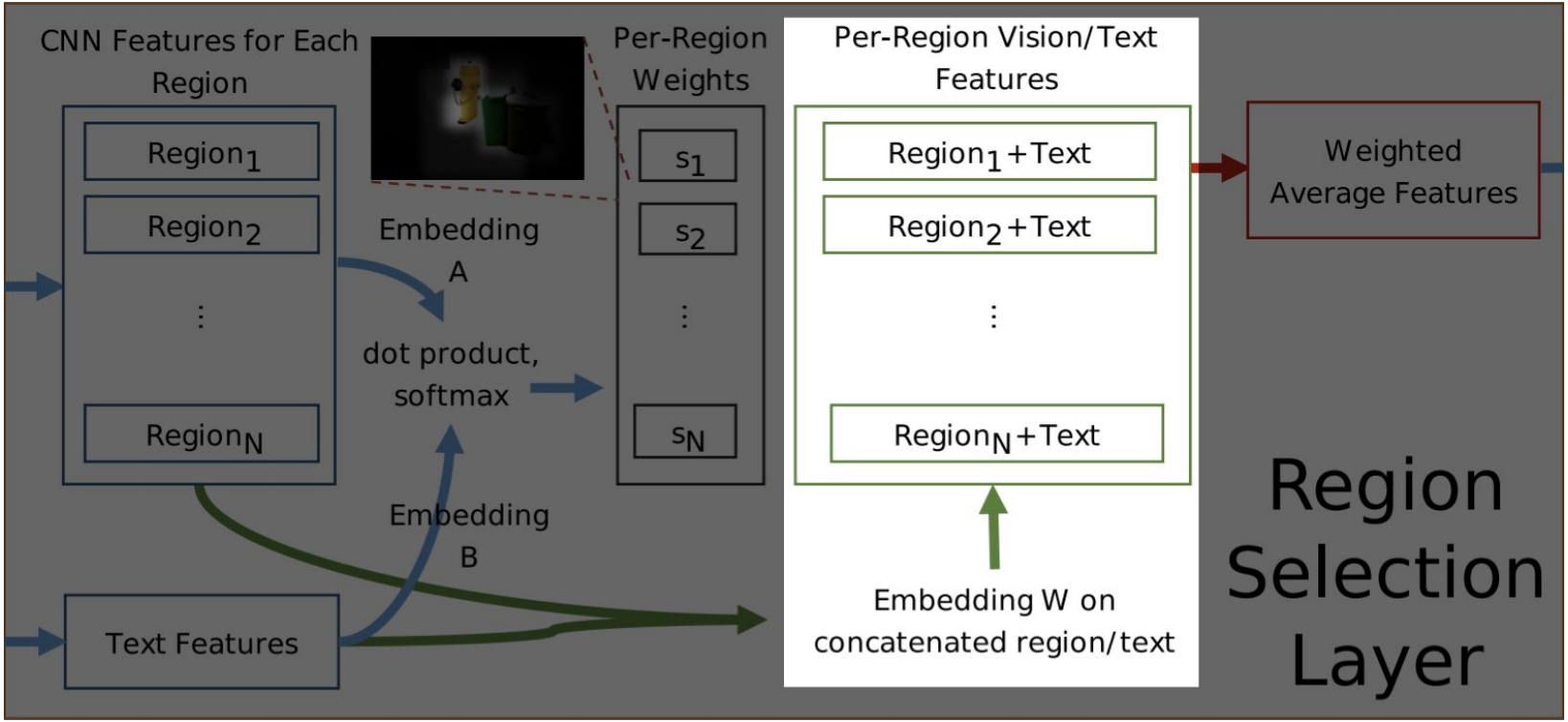
<q1,b1>=0.3, <q1,b2>=0.05,
<q1,b3>=0.07

q2: [What color is the walk light?][Green]

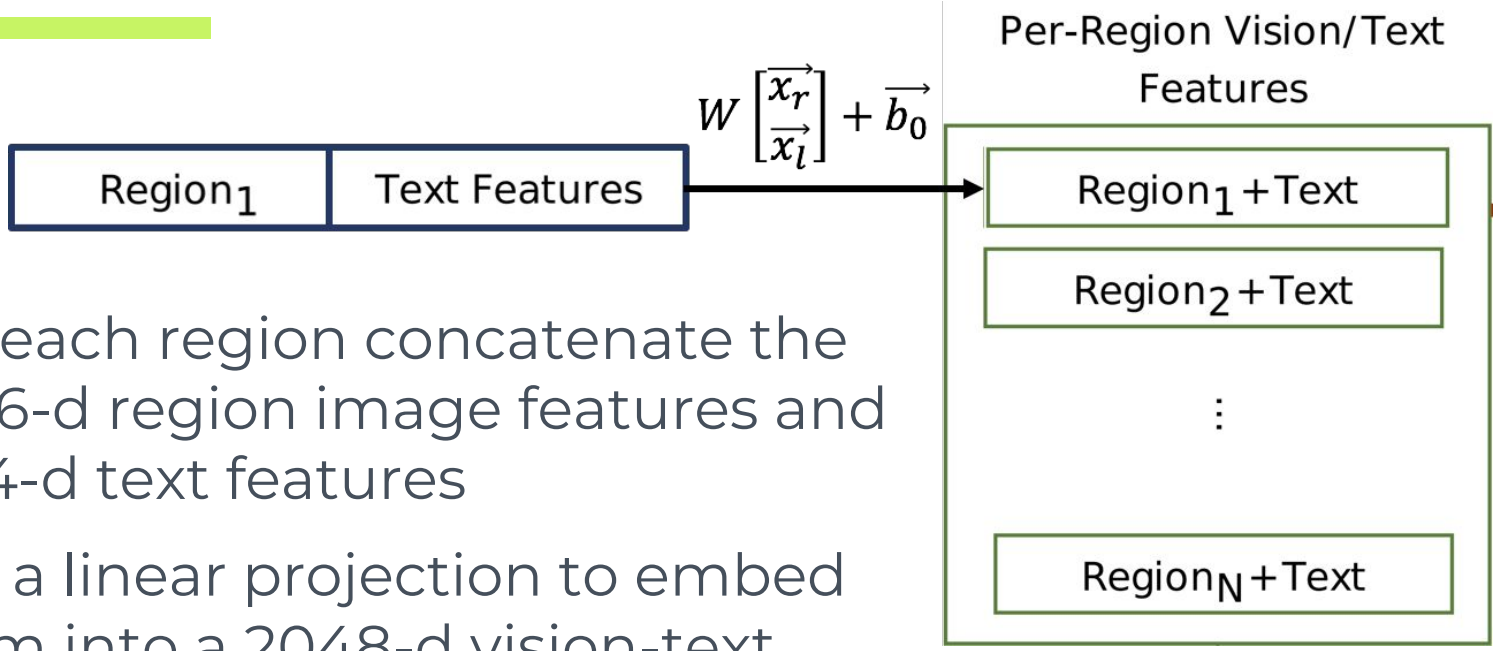
<q2,b1>=0.04, <q2,b2>=0.03,
<q2,b3>=0.5

Dot product explicitly embodies attention as region weights are high when region content and question/answer pair embeddings are similar

Model Architecture: Region Selection Layer

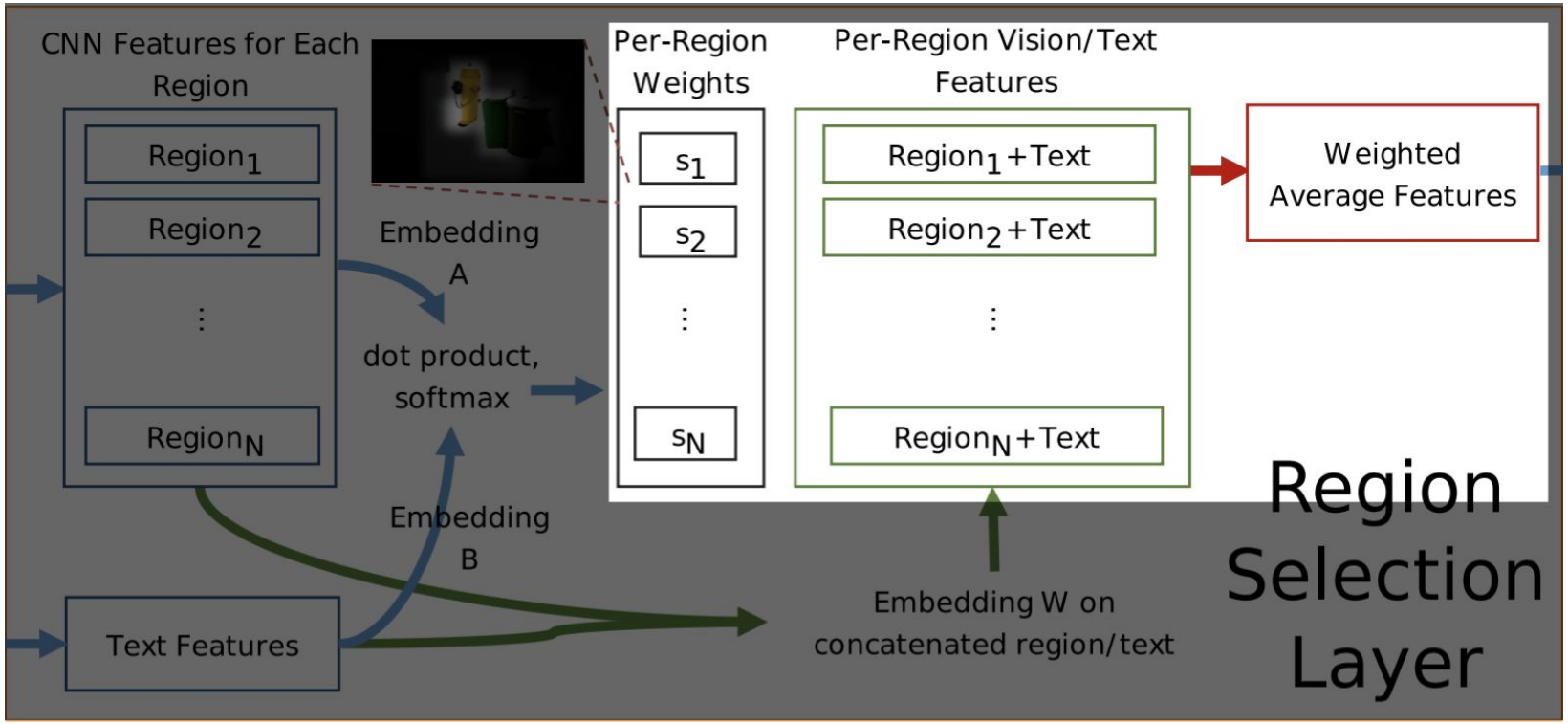


Model Architecture: Region Selection Layer (cont.)



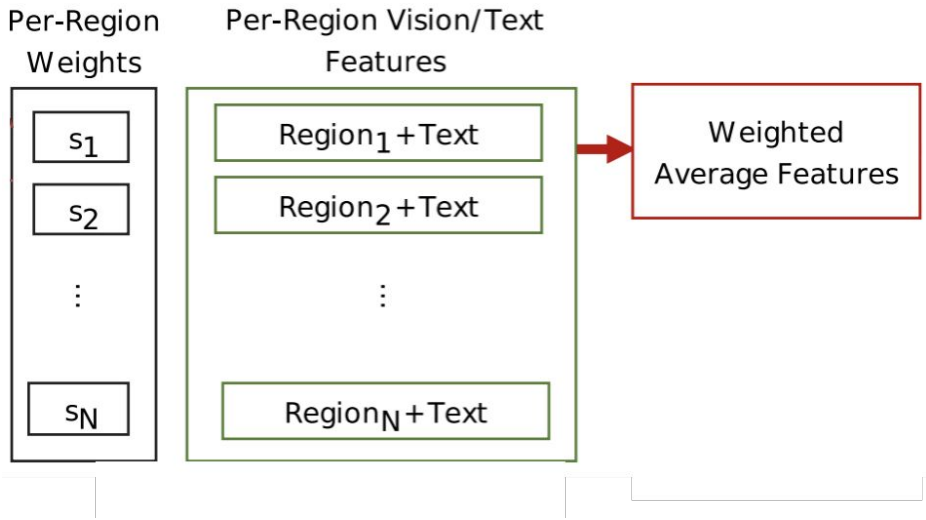
- For each region concatenate the 5096-d region image features and 1024-d text features
- Use a linear projection to embed them into a 2048-d vision-text feature vector for each region

Model Architecture: Region Selection Layer



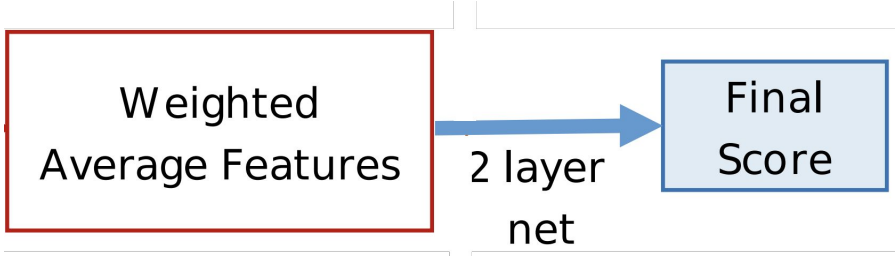
Model Architecture: Region Selection Layer (cont.)

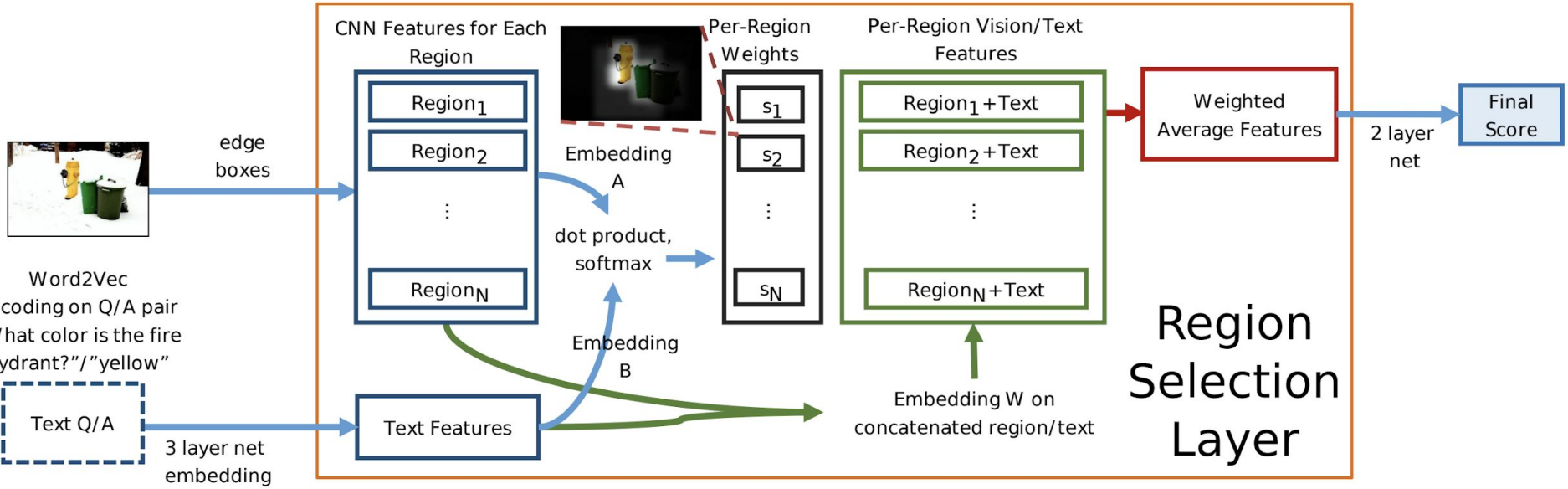
- Using the per-region weights, compute a weighted sum of the per-region vision-text features to obtain a single 2048-d weighted average feature vector
- **Insight:** This vector represents the information captured in the image and text when focusing on the relevant regions



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 ≥ 0.2

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

- **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 that 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

Model	Overall (%)
Language Only	53.98
Word+Whole Image	57.83
Word+ave. reg.	57.88
Word+Region Sel.	**58.94**

	region	image	text	freq
overall	**58.94**	57.83	53.98	100.0%
is/are/was	**75.42**	74.63	75.00	33.3%
identify: what kind/type/animal	**52.89**	52.10	45.11	23.8%
how many	33.38	**36.84**	34.05	10.3%
what color	**53.96**	43.52	32.59	9.8%
interpret:can/could/does/has	**75.73**	74.43	**75.73**	4.6%
none of the above	45.40	44.04	**48.23**	4.1%
where	42.11	**42.43**	37.61	2.5%
why/how	26.31	28.18	**29.24**	2.2%
relational: what is the man/woman	**70.15**	67.48	56.64	2.0%
relational: what is in/on	54.78	**54.80**	45.41	1.8%
which/who	**43.97**	42.70	38.62	1.7%
reading: what does/number/name	**33.31**	31.54	30.84	1.6%
identify scene: what room/sport	**86.21**	76.65	61.26	0.9%
what time	**41.47**	37.74	38.64	0.8%
what brand	45.40	44.04	**48.23**	0.4%

Ablation testing results across question categories

	region	image	text	freq
why/how	26.31	28.18	**29.24**	2.2%
how many	33.38	**36.84**	34.05	10.3%
what color	**53.96**	43.52	32.59	9.8%
identify scene: what room/sport	**86.21**	76.65	61.26	0.9%

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

Experiments/Results: Ablation/Comparison Tests (cont.)

Model	All	Y/N	Num.	Others
test-dev				
LSTM Q+I (Antol et al. 2015)	57.17	78.95	35.80	43.41
Q+I (Antol et al. 2015)	58.97	75.97	34.35	50.33
iBOWIMG (Zhou et al. 2015)	61.68	76.68	37.05	54.44
DPPnet (Noh et al 2016)	62.48	**80.79**	**38.94**	52.16
Ours	**63.3**	78.09	34.22	**57.23**
test-standard				
deeperLSTM_NormalizeCNN (Antol et al. 2016)	63.09	80.59	37.7	53.64
iBOWIMG (Zhou et al. 2015)	61.97	76.86	37.30	54.60
DPPnet (Noh et al 2016)	62.69	**80.35**	**38.79**	52.79
Ours	**63.53**	78.08	34.26	**57.43**

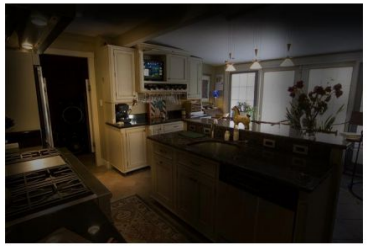
Experiments/Results: Ablation/Comparison Tests (cont.)

Model	All	Y/N	Num.	Others
test-dev				
LSTM Q+I (Antol et al. 2015)	57.17	78.95	35.80	43.41
Q+I (Antol et al. 2015)	58.97	75.97	34.35	50.33
iBOWIMG (Zhou et al. 2015)	61.68	76.68	37.05	54.44
DPPnet (Noh et al 2016)	62.48	**80.79**	**38.94**	52.16
Ours	**63.3**	78.09	34.22	**57.23**
test-standard				
deeperLSTM_NormalizeCNN (Antol et al. 2016)	63.09	80.59	37.7	53.64
iBOWIMG (Zhou et al. 2015)	61.97	76.86	37.30	54.60
DPPnet (Noh et al 2016)	62.69	**80.35**	**38.79**	52.79
Ours	**63.53**	78.08	34.26	**57.43**

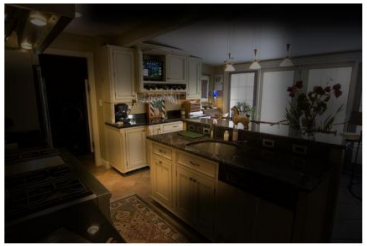
Experiments/Results: Qualitative Analysis

What room is this?
Answer: Kitchen

Kitchen: 22.3



Living room: 5.8



Bathroom: 4.8

Blue: 1.5

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

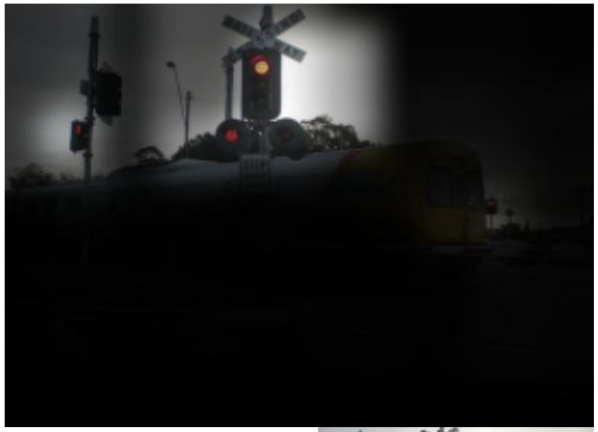
Kitchen: 22.3

Attention weights visualization

What color on the stop light is lit up?

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

What color is the light?



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

What color is the street sign?

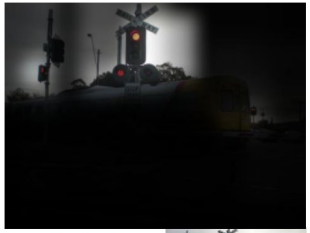
T:gray(-0.2)
I:gray(-0.4)
R:yellow(0.4)
Ans: yellow

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?

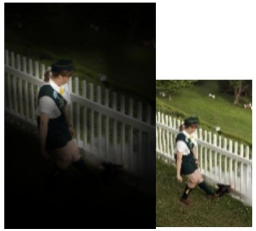


L: red (1.0)
I: red (0.3)
R: red (1.7)
Ans: red

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?

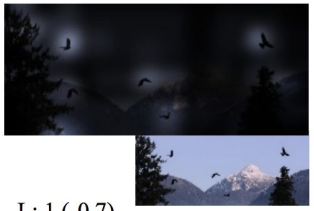


L: black (-0.7)
I: gray (-0.6)
R: white (0.1)
Ans: white

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?

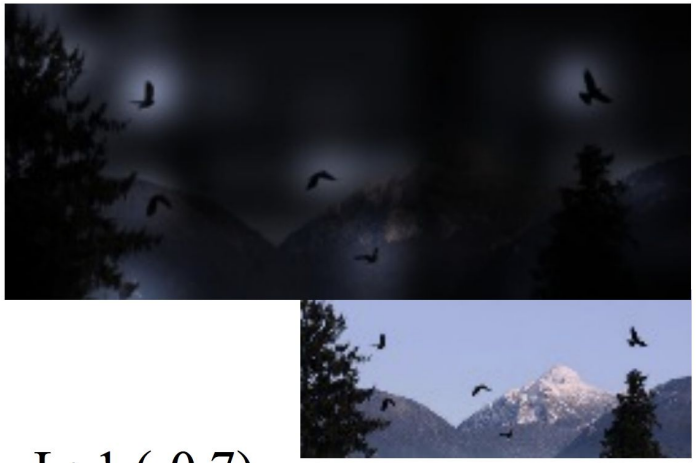
L: goose (-1.1)
I: kite (1.4)
R: kite (5.3)
Ans: kite

What color is the walk light?

L: red (-0.3)
I: red (-0.3)
R: green (1.1)
Ans: green

T: text-only, I: whole image+text, R: region-selection. Margin shown in parenthesis (ground truth confidence - top incorrect)

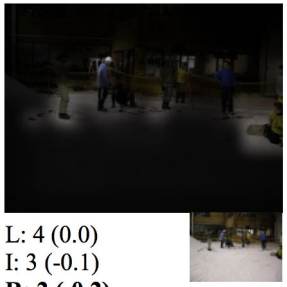
How many birds are in the sky?



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

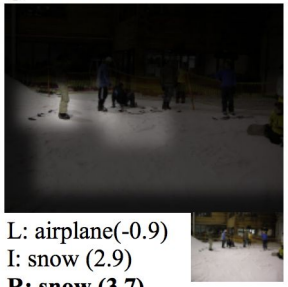
Attention for a counting question shows focus on the correct object despite incorrect final answer

How many people?



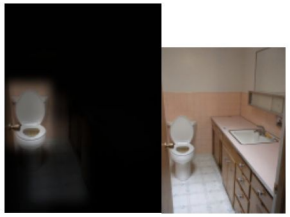
L: 4 (0.0)
I: 3 (-0.1)
R: 2 (-0.2)
Ans: 8

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

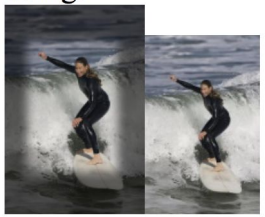
Is the faucet turned on?

L: no(3.6)
I: no (3.1)
R: no (5.1)
Ans: no

What is behind the man?

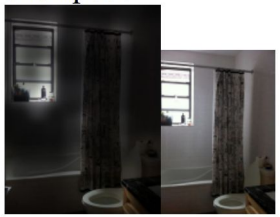
L: dog(0.0)
I: dog (0.0)
R: dog (1.4)
Ans: dog

What is the man doing?



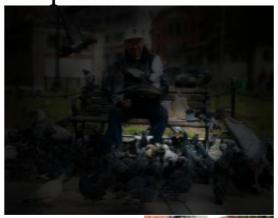
L: surfing (2.5)
I: blue (3.7)
R: surfing (9.7)
Ans: surfing

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?

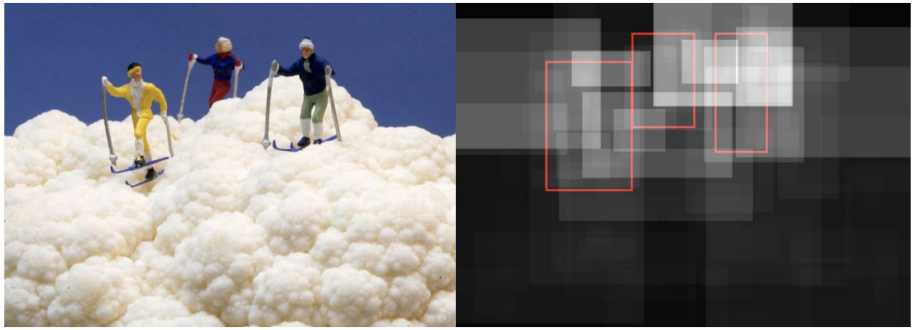


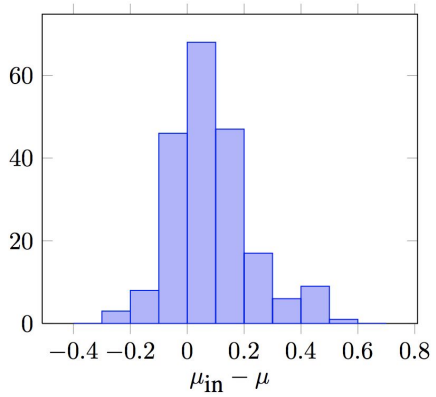
L: yes (1.5)
I: yes (0.5)
R: yes (1.0)
Ans: yes

T: text-only, I: whole image+text, R: region-selection. Margin shown in parenthesis (ground truth confidence - top incorrect)

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





Top: predicted region weights and ground truth annotation of relevant regions

Left: Histogram of normalized differences between mean pixel weight inside annotated regions and across the whole image

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

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?

This and Following Slides Images Credit: *Xu et al.*

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

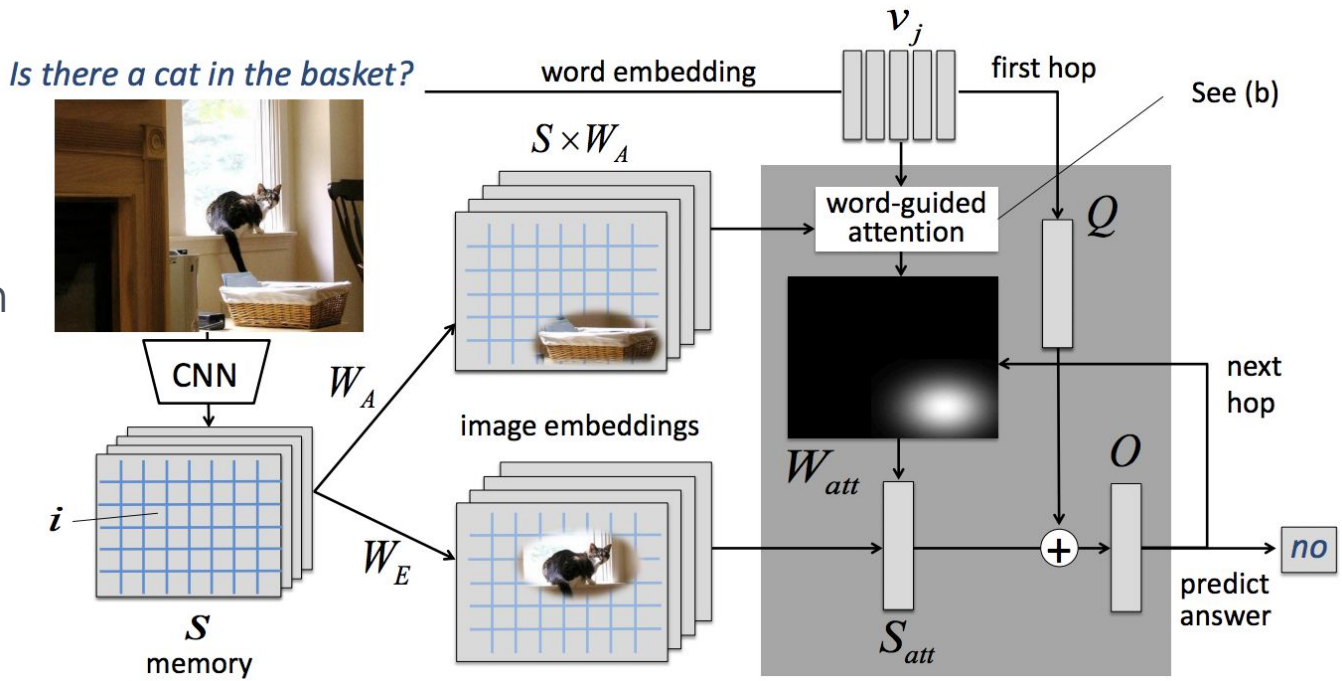
Model Architecture: Model Overview

- **Input:**

  Fixed-size image & Variable-length question

- **Output:**

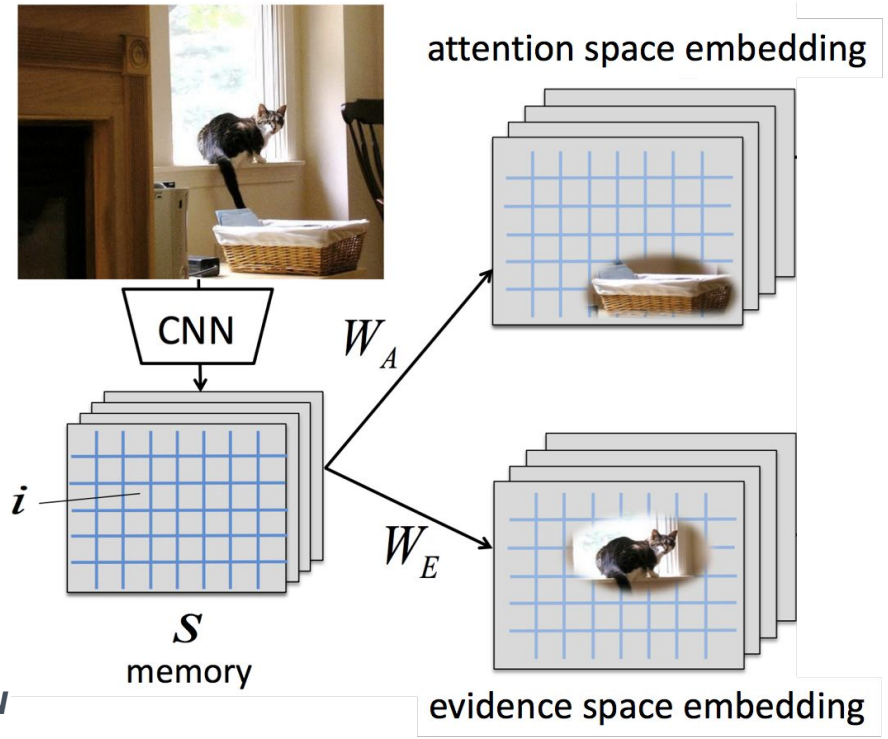
  Softmax over all possible answers



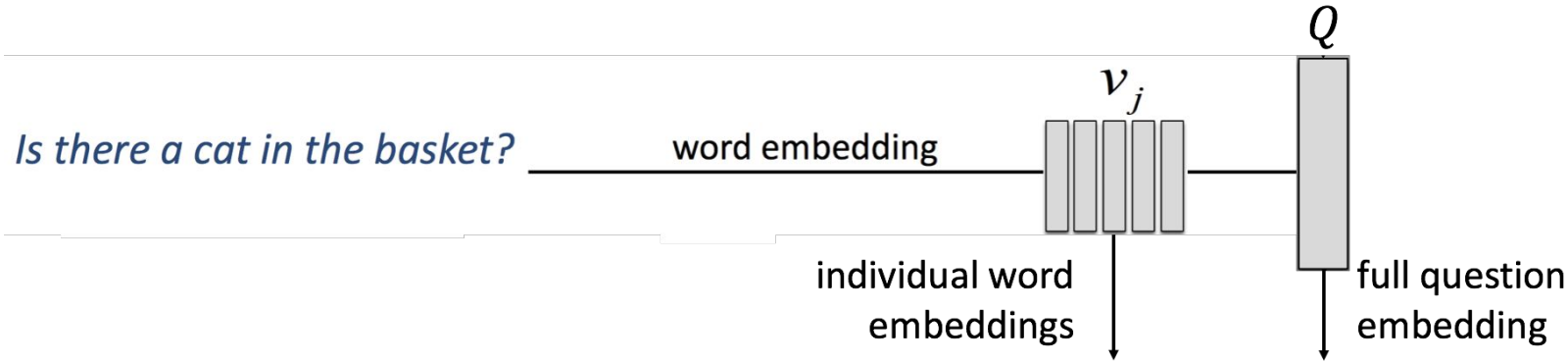
Model Architecture: Image/Word Embeddings

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 R^N
- **Evidence embedding** maps to output that captures visual information in each region also in R^N



Model Architecture: Image/Word Embeddings (cont.)



- The words in the question are converted into word vectors $\boldsymbol{v_j}$ in the attention embedding space in 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:

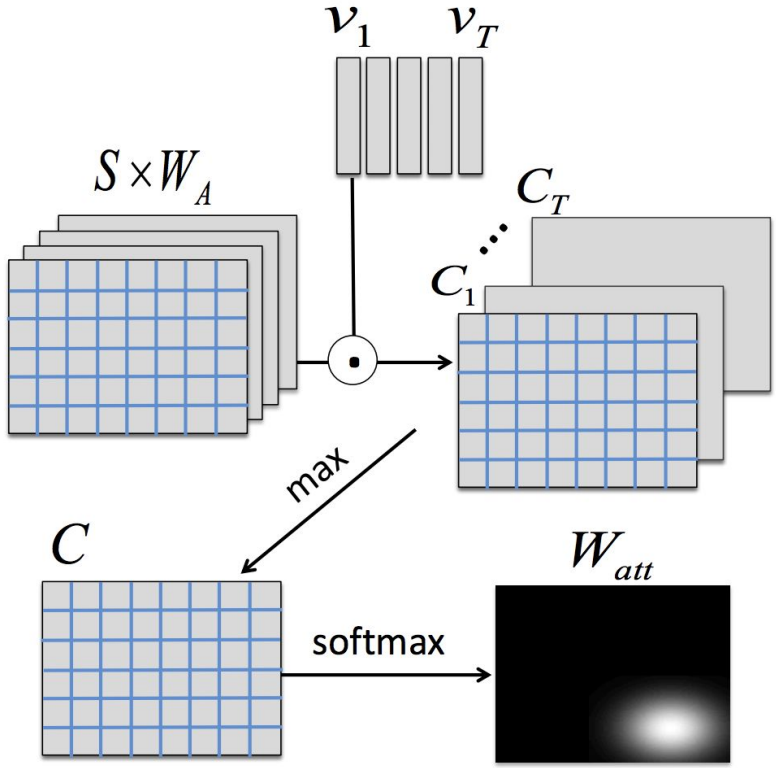
- ***M*:** size of visual features for each region extracted from GoogLeNet
- ***L*:** number of spatial regions
- ***T*:** length of (padded) question
- ***N*:** size of attention and evidence embedding space

Objects:

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

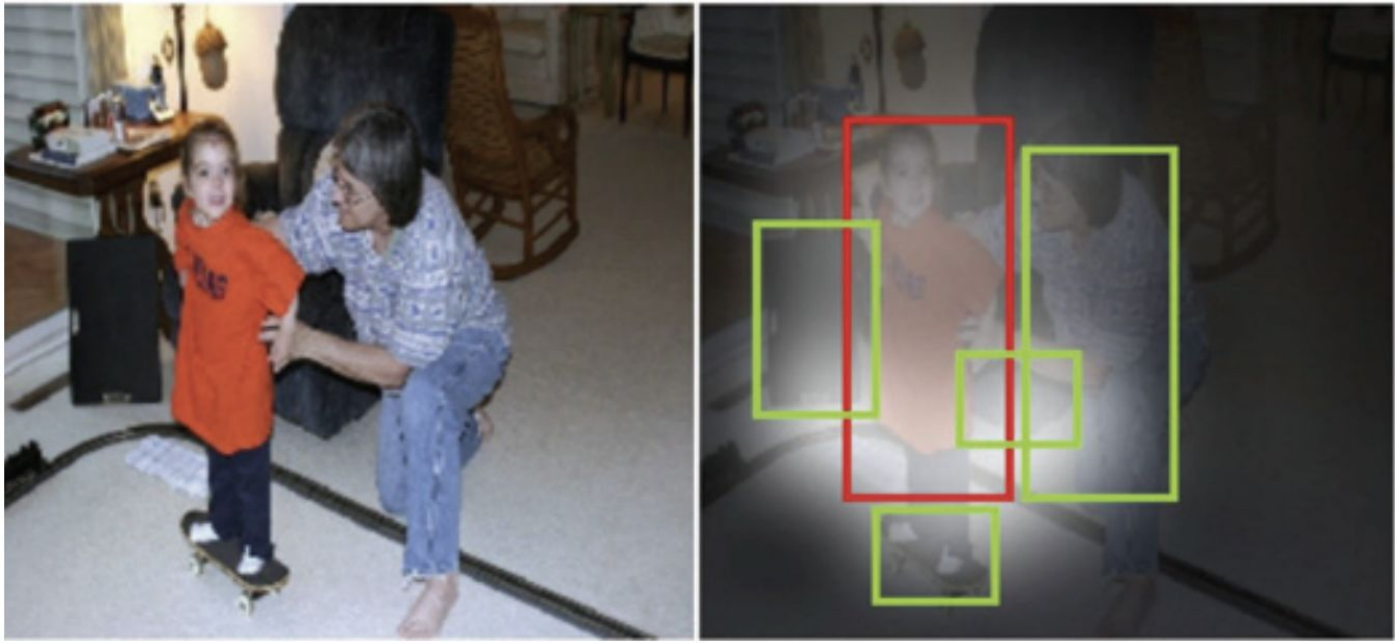
Model Architecture: Word-Guided Attention

- 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



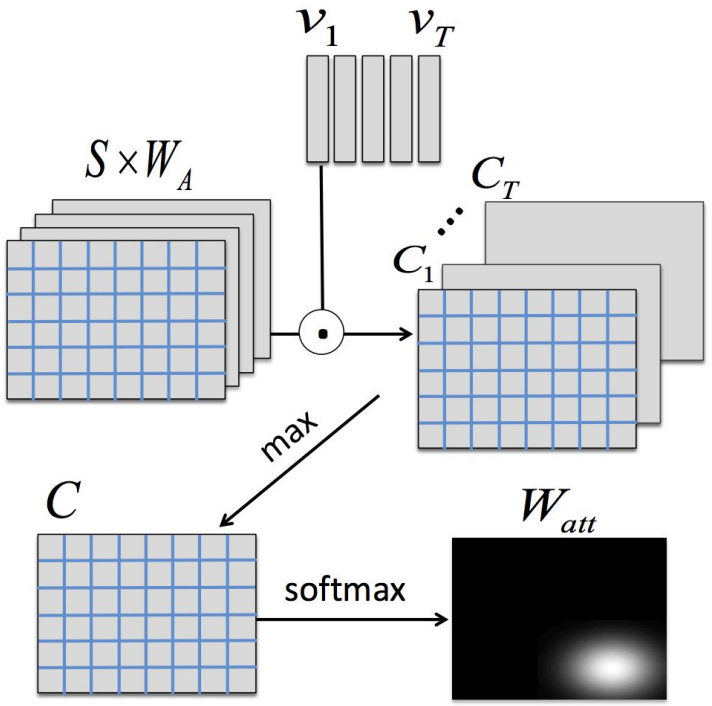
Model Architecture: Word-Guided Attention (cont.)

What is the child standing on?



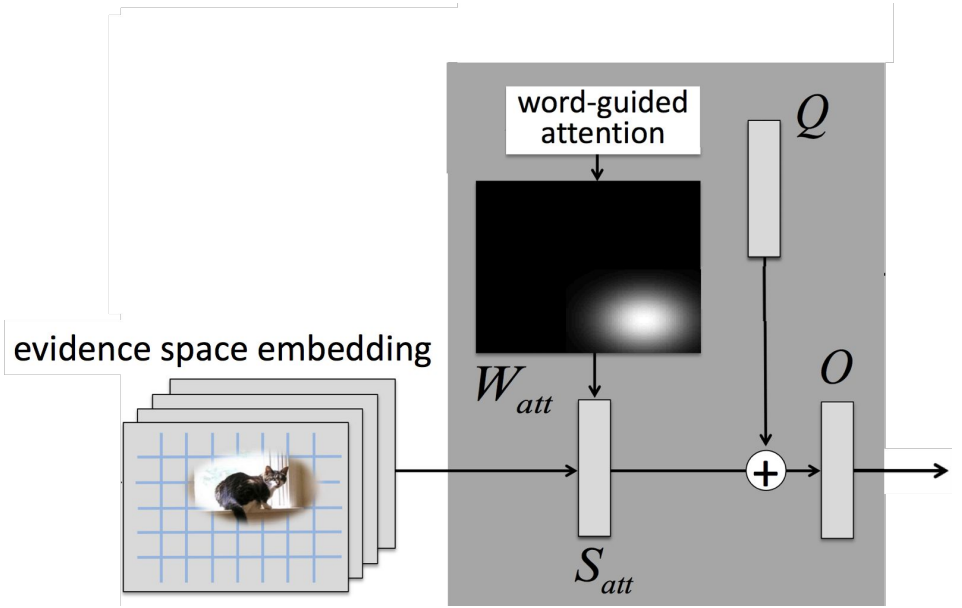
Model Architecture: Word-Guided Attention (detailed)

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



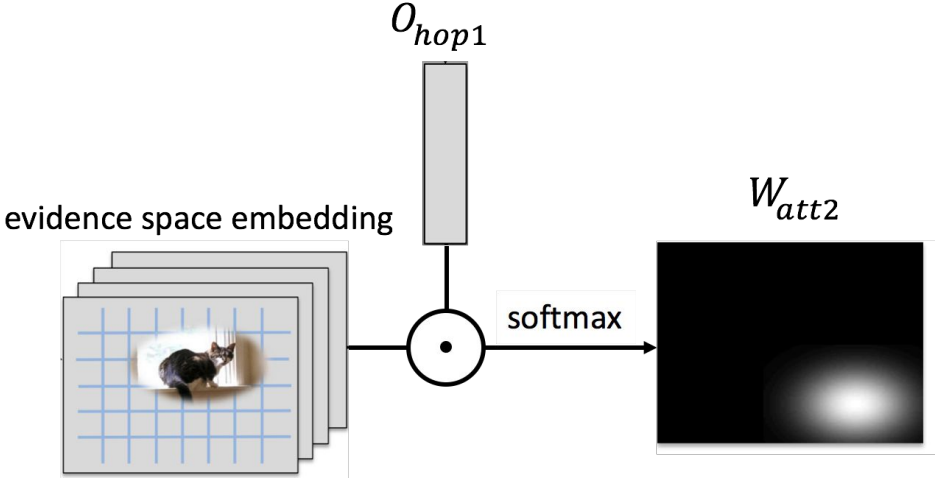
Model Architecture: First Hop

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



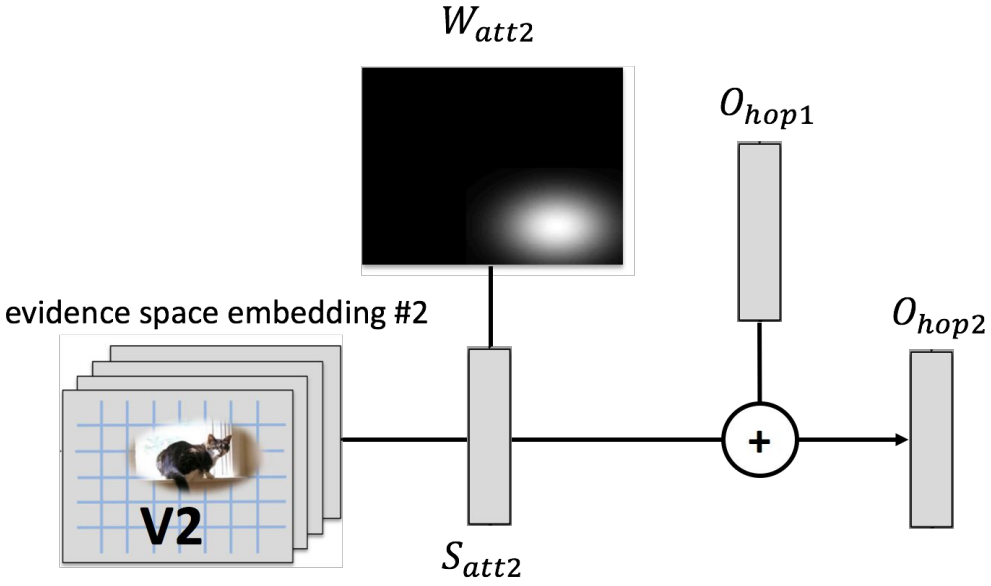
Model Architecture: Second Hop

- Output from first hop $\mathbf{O_{hop1}}$ 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 $\mathbf{O_{hop1}}$



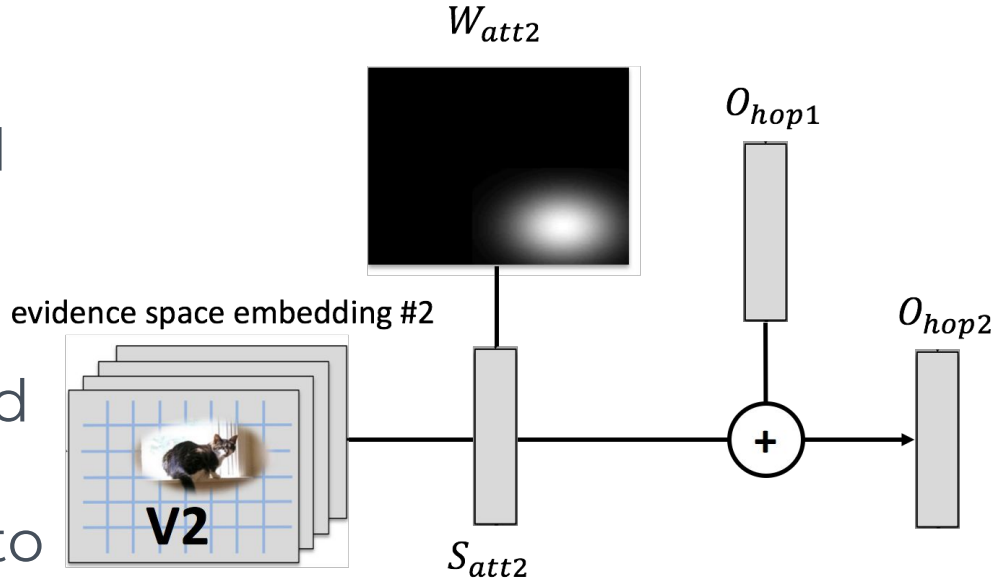
Model Architecture: Second Hop (cont.)

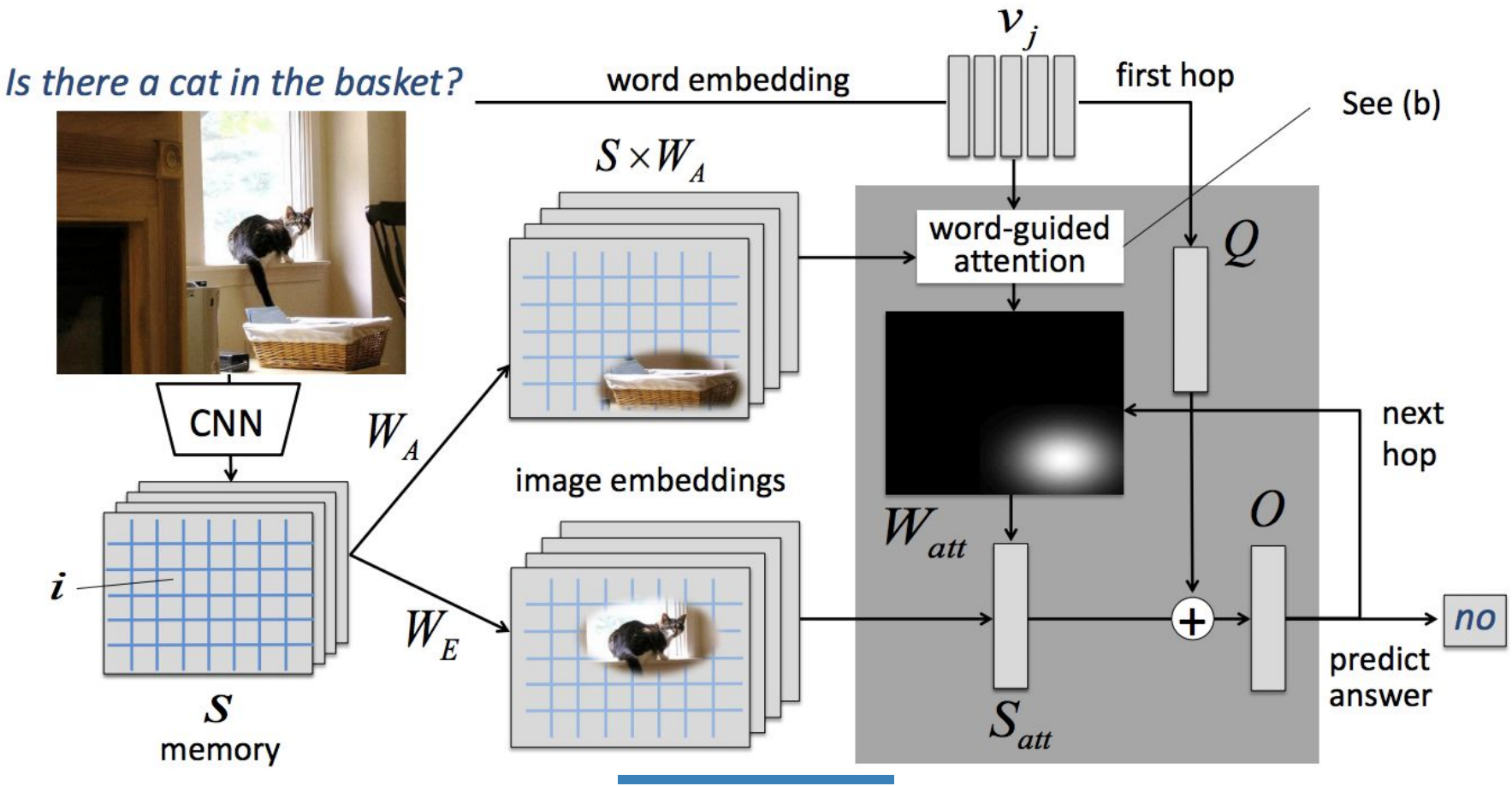
- **W_{att2}** used to calculate new selected visual evidence vector which is used along with **O_{hop1}** to generate final predictions
- **Insight:** second hop adds new information to previous understanding **O_{hop1}** to generate better answer



Model Architecture: Second Hop (cont.)

- A second visual evidence embedding is created and weighted according to $\mathbf{W_{att2}}$ to generate $\mathbf{S_{att2}}$
- $\mathbf{S_{att2}}$ and $\mathbf{O_{hop1}}$ are summed and passed through nonlinear + softmax layer to generate final output predictions over possible output space





Complete model architecture

3 Experiment/Results

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

Experiments/Results: Testing Overview

Testing Goals:

- Extensively test ability to form spatial inference
- Compare to existing VQA models by testing on standard datasets

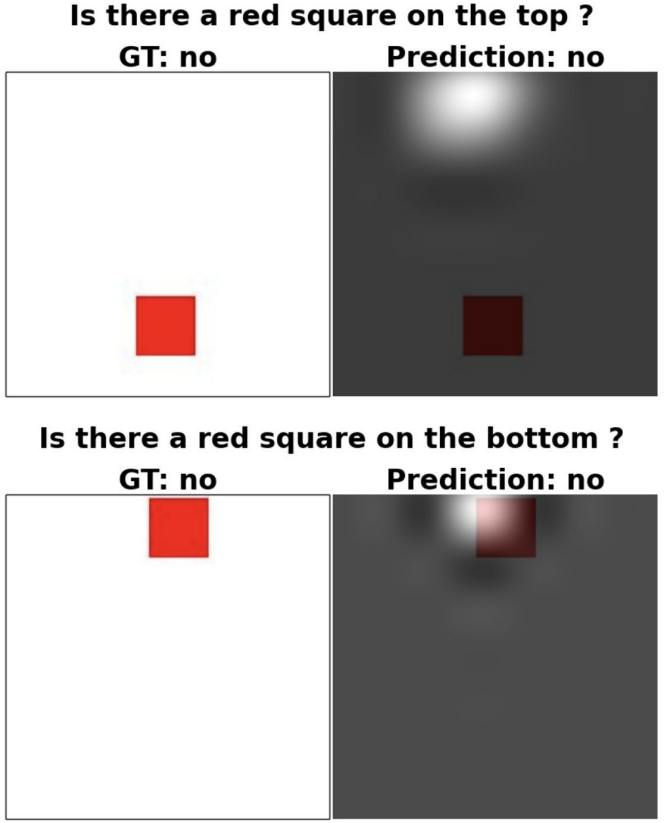
Experiments/Results:
Exploring Attention on Synthetic Data

- 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

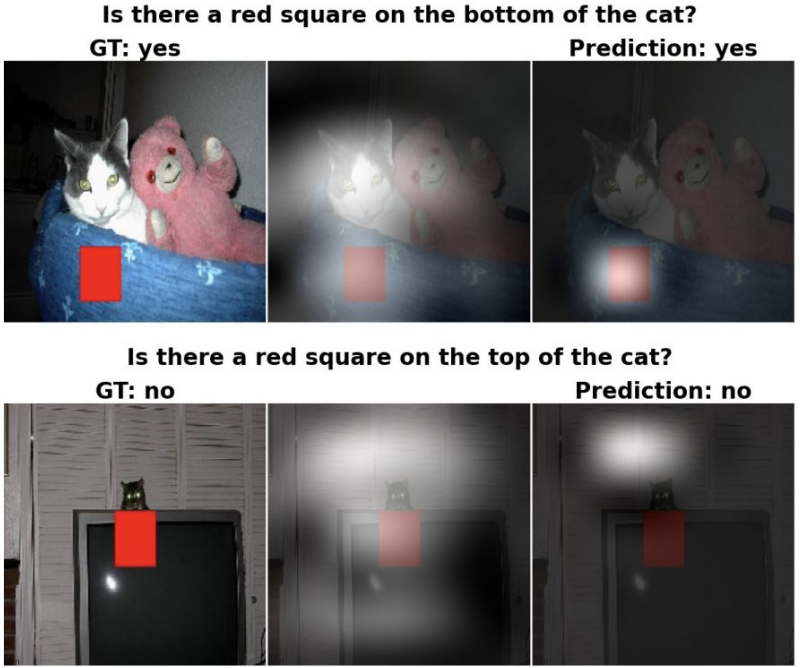
- One-hop model achieves **100%** accuracy while iBOWIMG achieves **75%** accuracy (same as always answering “no”)
- Attention learned 2 logical rules:
  - Look at question position for square
  - Look at square and compare to question position

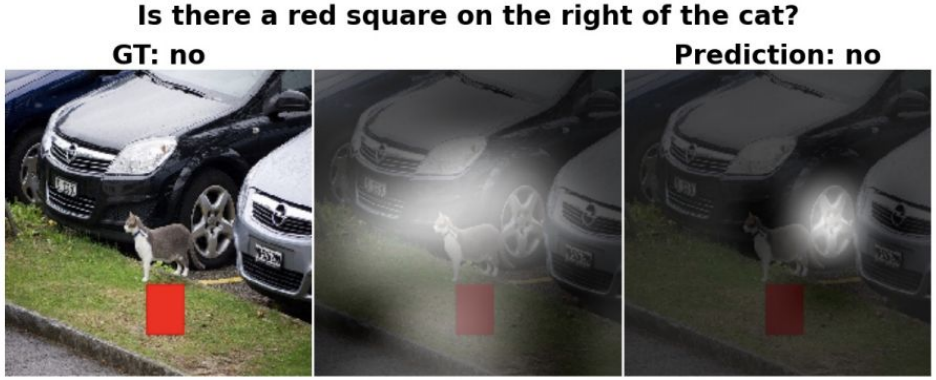


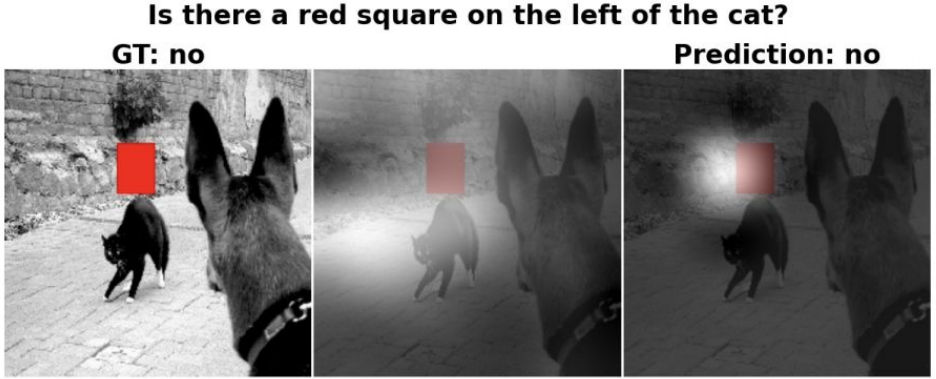
Experiments/Results:
Exploring Attention on Synthetic Data (cont.)

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







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

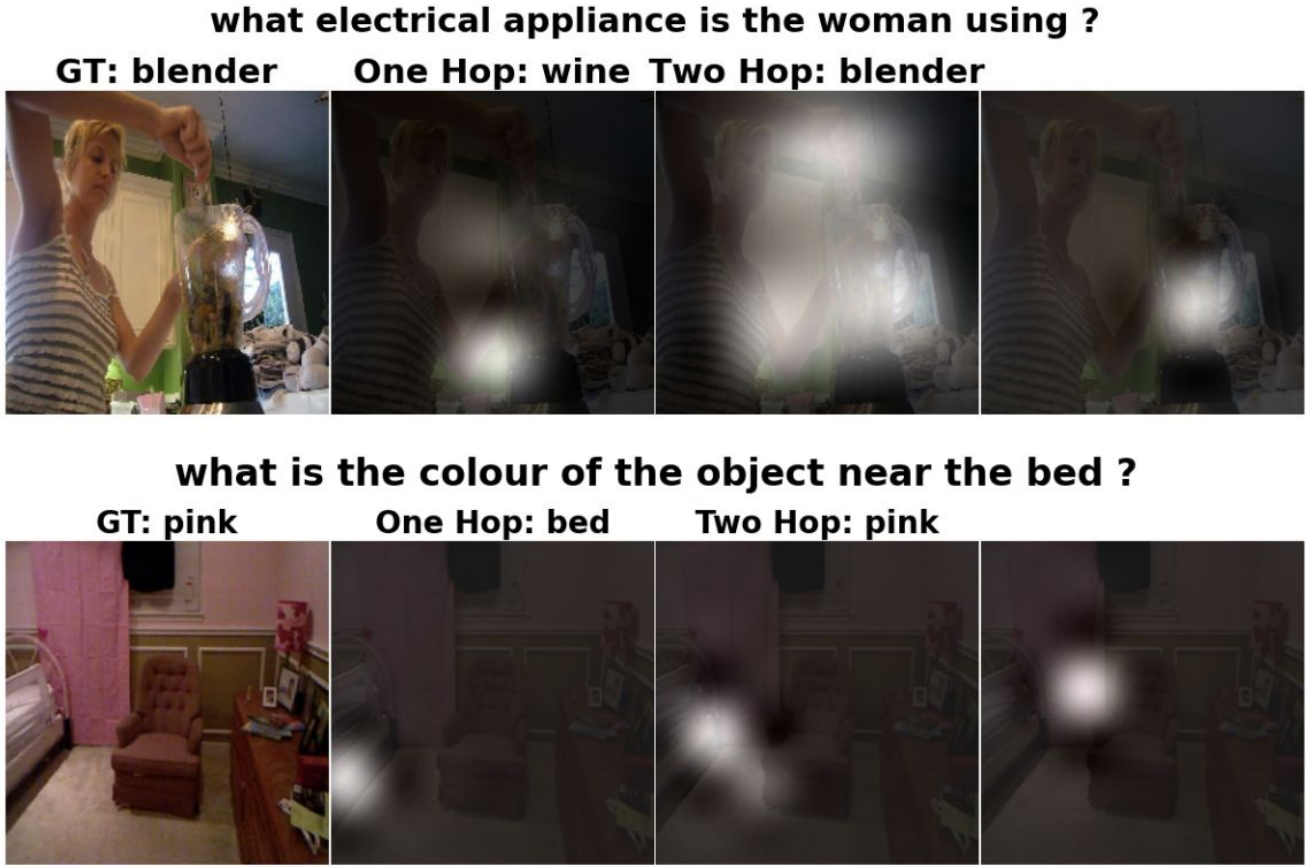
	DAQUAR
Multi-World [1]	12.73
Neural-Image-QA [10]	29.27
Question LSTM [10]	32.32
VIS+LSTM [11]	34.41
Question BOW [11]	32.67
IMG+BOW [11]	34.17
SMem-VQA One-Hop	36.03
SMem-VQA Two-Hop	**40.07**

Experiments/Results: Experiments on Standard Datasets

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

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

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')
```