Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Huijuan Xu, Kate Saenko
Department of Computer Science, Boston University, USA
{hxu, saenko}@bu.edu

GOAL

- Task
  - Visual Question Answering (VQA)
  - Answer a question about a given photograph
- Applications
  - Assist the visually impaired
  - Automatically query surveillance video

CONTRIBUTIONS

- Existing Methods
  - End-to-end deep VQA networks adapted from captioning models: utilize a recurrent LSTM network, which takes the question and CNN image features as input and outputs the answer. [6, 7]
- Problems
  - Do not have any explicit notion of object position
  - Use the whole question encoding to infer the answer, without considering fine-grained information from the question
- Contributions
  - Propose Spatial Memory Network VQA (SMem-VQA)
  - Incorporate explicit spatial attention based on memory networks
  - Use fine-grained word embeddings to collect visual evidence for each word in the question

SCHEMATIC DIAGRAM

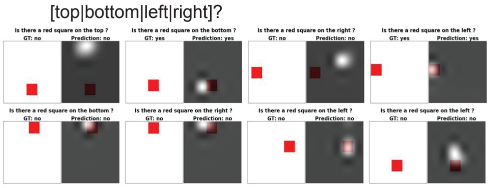
Attention is applied in two steps (hops):

What is the child standing on ? skateboard

SYNTHETIC EXPERIMENTS

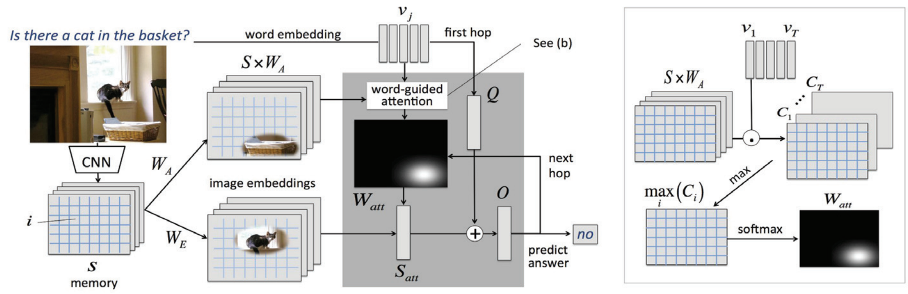
By visualizing attention, we can figure out how the network learns to answer questions.

- Absolute Position Recognition
  - Input image: a red square appears in one of the four regions of a white-background image
  - Question: Is there a red square on the [top|bottom|left|right]?



- Network learned two logic rules
  - Look at the position specified in question (top|bottom|right|left), if it contains a square, then answer "yes"; if not, then answer "no".
  - Look at the region where there is a square, then answer "yes" for the question about that position and "no" for the questions about the other three positions.
- Relative Position Recognition

SMem-VQA NETWORK ARCHITECTURE



(a) Overview

(b) Word-guided attention

hop1:

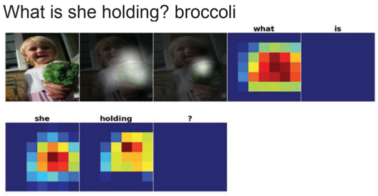
$$C = V \cdot (S \cdot W_A + b_A)^T$$
$$W_{att} = \text{softmax}(\max_{i=1,\cdots,T}(C_i)),\ C_i \in \mathbb{R}^L$$
$$S_{att} = W_{att} \cdot (S \cdot W_E + b_E)$$
$$Q = W_Q \cdot V + b_Q$$
$$P = \text{softmax}(W_P \cdot f(S_{att} + Q) + b_P)$$

hop2:

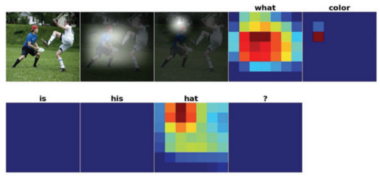
$$O_{hop1} = S_{att} + Q$$
$$C_{hop2} = (S \cdot W_E + b_E) \cdot O_{hop1}$$
$$W_{att2} = \text{softmax}(C_{hop2})$$
$$S_{att2} = W_{att2} \cdot (S \cdot W_{E_2} + b_{E_2})$$
$$P = \text{softmax}(W_P \cdot f(O_{hop1} + S_{att2}) + b_P)$$

- Visualization of attention weights S_{att}, S_{att2}, and correlation matrix C:

What is she holding? broccoli



What color is his hat? red



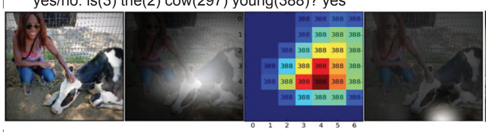
EXPERIMENTAL RESULTS

Test-dev and test-standard results on Open-Ended VQA dataset [1] (accuracy). Models with ∗ use extra training data in addition to the VQA dataset.

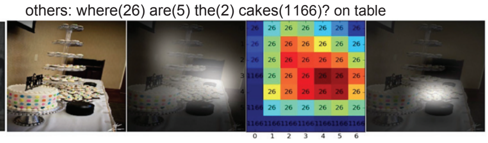
methods	test-dev				test-standard			
	Overall	yes/no	number	others	Overall	yes/no	number	others
LSTM Q+I [1]	53.74	78.94	35.24	36.42	54.06	-	-	-
ACK* [2]	55.72	79.23	36.13	40.08	55.98	79.05	36.10	40.61
DPPnet* [3]	57.22	80.71	37.24	41.69	57.36	80.28	36.92	42.24
iBOWIMG [4]	55.72	76.55	35.03	42.62	55.89	76.76	34.98	42.62
SMem-VQA 1-Hop	56.56	78.98	35.93	42.09	-	-	-	-
SMem-VQA 2-Hop	57.99	80.87	37.32	43.12	58.24	80.8	37.53	43.48

- 0-1 accuracy result on the reduced DAQUAR dataset [5] is 40.07%.
- Per-answer category attention weight visualization analysis:

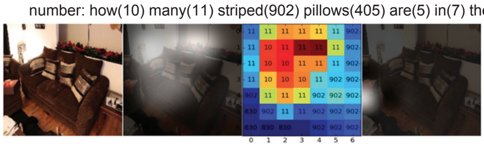
yes/no: is(3) the(2) cow(297) young(388)? yes



others: where(26) are(5) the(2) cakes(1166)? on table



number: how(10) many(11) striped(902) pillows(405) are(5) in(7) the(2) sofa(830)? 4



REFERENCES

[1] Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[2] Wu, Qi, et al. "Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources." *arXiv preprint arXiv:1511.06973*(2015).

[3] Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *arXiv preprint arXiv:1511.05756* (2015).

[4] Zhou, Bolei, et al. "Simple Baseline for Visual Question Answering." *arXiv preprint arXiv:1512.02167* (2015).

[5] Malinowski, Mateusz, and Mario Fritz. "A multi-world approach to question answering about real-world scenes based on uncertain input." *Advances in Neural Information Processing Systems*. 2014.

[6] Malinowski, Mateusz, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A neural-based approach to answering questions about images."*Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[7] Ren, Mengye, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering." *Advances in Neural Information Processing Systems*. 2015.

This research was supported by NSF Award IIS-1212928 and a Google Faculty Research Award.

SMem-VQA network architecture animation on YouTube: https://www.youtube.com/watch?v=FjpRwVKYJQ8&feature=youtu.be

SMem-VQA code: https://github.com/VisionLearningGroup/Ask_Attend_and_Answer