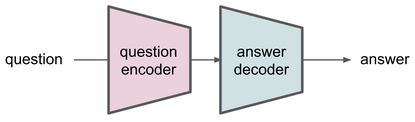
Knowledge Acquisition for Visual Question Answering via Iterative Querying

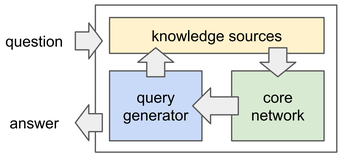
Yuke Zhu, Joseph J. Lim, Li Fei-Fei
Computer Science Department, Stanford University

Introduction

- Most of today's VQA methods make their predictions based on a predefined set of information
- These VQA models have been shown to be "myopic" (tend to fail on novel instances) [1]
- Current state-of-the-art VQA models have indicated that they could benefit from better visually grounded evidence [4]
- We push one step further by enabling a VQA model to ask for and collect "clues" – in particular, visually grounded evidence.



Standard VQA model



Our iterative VQA model

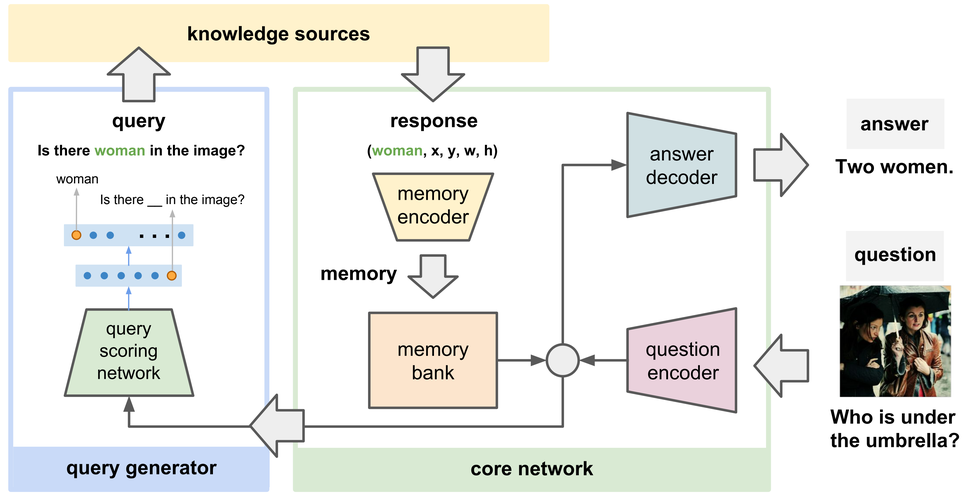
Iterative Training Algorithm

procedure
  Generate random query rollouts $R^{(0)}$
  Train initial core network $\mathcal{C}^{(0)}$ with rollout $R^{(0)}$
  Generate training samples $S^{(0)}$ for query scoring network with $\mathcal{C}^{(0)}$
  Train initial query scoring network $\mathcal{G}^{(0)}$ with $S^{(0)}$
  for $t = 1, \ldots,$ N **do**
    Generate query rollouts $R^{(t)}$ with query scoring network $\mathcal{G}^{(t-1)}$
    Finetune core network $\mathcal{C}^{(t)}$ from $\mathcal{C}^{(t-1)}$ with rollout $R^{(t)}$
    Generate training samples $S^{(t)}$ for query scoring network from $\mathcal{C}^{(t)}$
    Finetune query scoring network $\mathcal{G}^{(t)}$ from $\mathcal{G}^{(t-1)}$ with $S^{(t)}$
  end for
  return $\{\mathcal{G}^{(N)}, \mathcal{C}^{(N)}\}$
end procedure

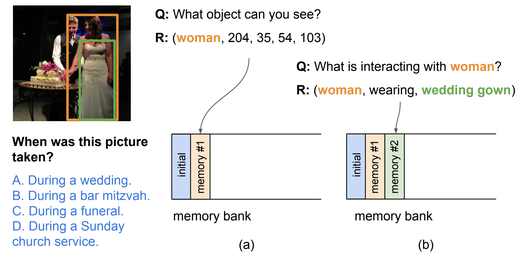
References

[1] Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. EMNLP, 2016
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. ICCV, 2015
[3] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. EMNLP, 2016.
[4] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. ECCV, 2016.
[5] R. Krishna,et al.. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV, 2016.
[6] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. NIPS, 2016
[7] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. CVPR, 2016.

Model Overview



Workflow of the iterative VQA model



An example iterative query process

Query types and templates	Response formats
What object can you see?	(**object**, x, y, w, h)
Is there **object** in the image?	(**object**, x, y, w, h)
How does **object** look?	(**object**, **attribute**)
What is interacting with **object1**?	(**object2**, **relation**)

Query types and response formats

Experiments

Datasets
Visual7W telling task [7]
VQA Real MultipleChoice Challenges [1]

Knowledge Sources
Visual Genome scene graphs [5]
Faster R-CNN detectors

Results
New state-of-the-art on Visual7W
On par with VQA challenge winning model

Method	Visual7W
LSTM-Attention [6]	0.543
MCB [3]	0.622
MLP [4]	0.648
MLP + all knowledge	0.658
MLP + uniform sampling	0.653
MLP + query generator	0.679

Method	VQA (dev)	VQA (standard)
Two-layer LSTM [2]	0.627	0.631
Co-Attention [6]	0.658	0.661
MCB + Att. + GloVe [3]	0.691	–
MCB Ensemble + Genome [3]	0.702	0.701
MLP [4]	0.659	–
MLP + query generator	0.691	0.689