

IQA: Visual Question Answering in Interactive Environments

Daniel Gordon, Aniruddha Kembhavi, Mohammad
Rastegari, Joseph Redmon, Dieter Fox, Ali Farhadi

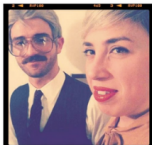
Presenter: Fanbo Xiang

06/02/2020

Visual Question Answering

Who is wearing glasses?

man



woman

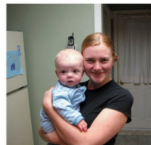


Where is the child sitting?

fridge



arms

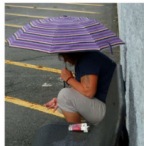


Is the umbrella upside down?

yes



no

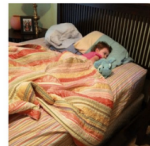


How many children are in the bed?

2




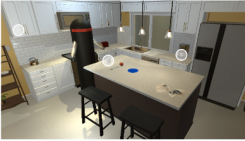




1



Answer question using vision

Interactive Question Answering

Question and answer	Initial Image	Scene View
Q: Is there bread in the room? A: No		
Q: How many mugs are in the room? A: 3		
Q: Is there a tomato in the fridge? A: Yes		

Answer question with vision and interaction

IQA Setting

Challenges

- Navigate the environment
- Understanding the objects
- Interact with the objects
- Plan actions conditioned on the question

Environment

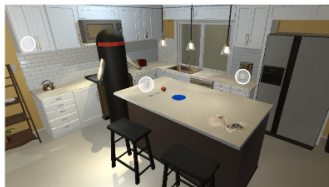
- AI2-THOR
- A Unity based environment
- Discrete actions, no dynamics

IQA Dataset

Interactive Question Answering Dataset Statistics		
	Train	Test
Existence	25,600	640
Counting	25,600	640
Spatial Relationships	25,600	640
Rooms	25	5
Total scene configurations (s.c.)	76,800	1,920
Avg # objects per (s.c.)	46	41
Avg # interactable objects (s.c.)	21	16
Vocabulary Size	70	70

Now let's solve the problem

Q: How many mugs
are in the room?
A: 3



How do we approach this problem?

- **Go** to the counter
- **Look for** mugs
- **Look for** drawers and doors
- **Open** drawers and doors one by one
- **Keep a counter** in our head
- **Answer** the question

Now let's solve the problem

How do we approach this problem?

- **Go** to the counter
- **Look for** mugs
- **Look for** drawers and doors
- **Open** drawers and doors one by one
- **Keep a counter** in our head
- **Answer** the question

How do we decide which action to take?

- We think

Navigator

Detector

Scanner

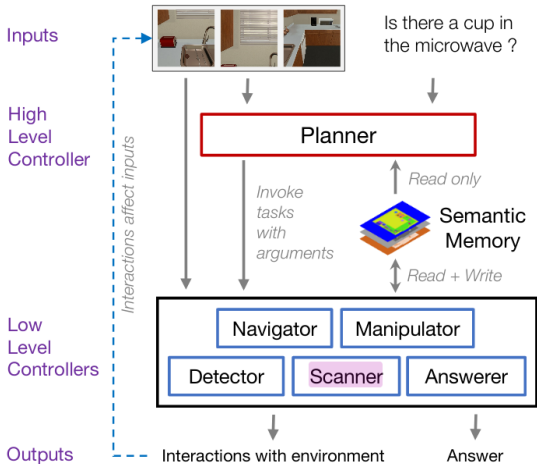
Manipulator

Memory

Answerer

Planner

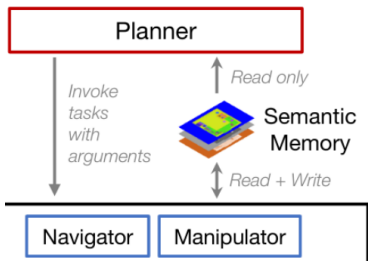
HIMN: Hierarchical Interactive Memory Network



Related works

- VQA (**Answerer**)
- (Hierarchical) Reinforcement learning
- Visual navigation (**Navigator**)
- Visual pladnning (**Planner**)
- Visual learning by Simulation

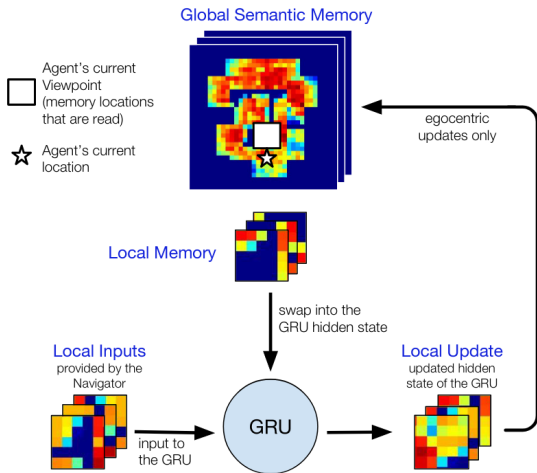
Semantic Memory



Since there are counting tasks, we want the agent to hold on the information of specific locations for long.

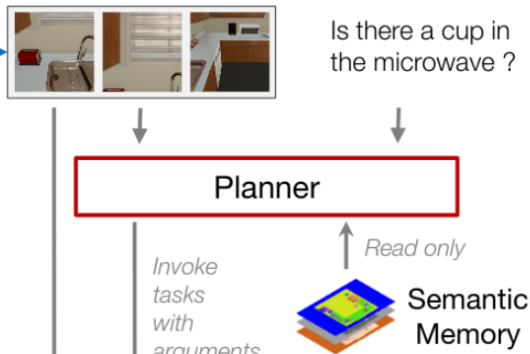
Idea: use explicit memory that encodes a representation for each location in the scene.

Semantic Memory



Egocentric Spatial GRU

Planner



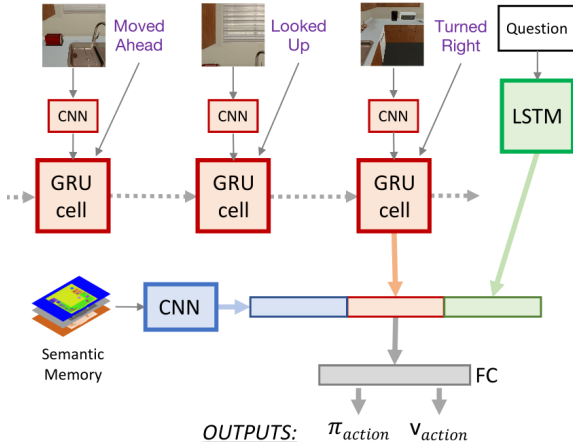
RL-like module

Given state (visual input, language input, memory)

Produce action (which low-level controller to use)

Planner

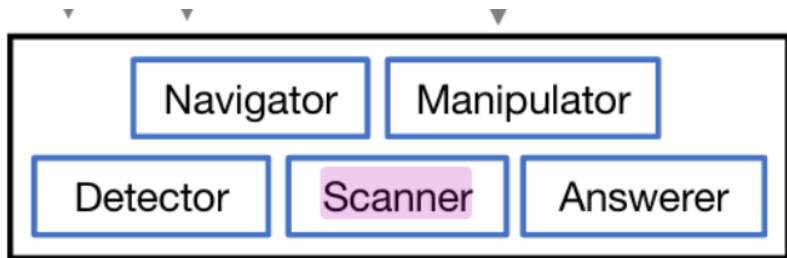
INPUTS: Images + Previous Actions + Question



Trained with rewards.

It also predicts which actions are possible.

Low level controllers



Navigator

Planner decides a position (5×5 in front of the agent)

Planner(state) \rightarrow (“navigator”, (x, y))

- The navigator takes the coordinate and current estimated occupancy grid and runs A^* .
- The navigator uses the esGRU to produce a new estimation of local occupancy given current visual input (supervised learning).
- The navigator invokes the **scanner** to find objects on its way.
- The navigator also predicts if the position is not reachable.

Scanner

Scanner simply rotates the camera up, down, left, and right. Each time it calls the **detector**.

Detector

- Object detection: YOLOv3 fine tuned on AI2-THOR.
- Depth estimation: FCRN (Fully Convolutional Residual Networks).
- It also learns what objects have interactions.
(Microwave can be opened)

Manipulator

The manipulator is invoked by the planner
Planner(state) \rightarrow (“manipulator”, “microwave”)

- Manipulator fails if the object cannot be interacted or it is too far away or out of view.

Answerer

It answers the question.

- It uses current image, the full spatial memory, and question embedding vector to predict the probability of each answer.
- The episode ends when the question is answered.

Training

- Planner: ground truth detector and navigator.
- Navigator: train by random start and random goal.
- Answerer: Pretrained with partial semantic maps which contains enough information for a correct answer.
- Detector: YOLOv3.
- Scanner and Manipulator: hard-coded.
- Joint training in the end.

Experimental Results

Model	Existence		Counting		Spatial Relationships	
	Accuracy	Length	Accuracy	Length	Accuracy	Length
Most Likely Answer Per Q-type (MLA)	50	-	25	-	50	-
A3C with ground truth (GT) detections	48.59	332.41	24.53	998.32	49.84	578.71
HIMN with YOLO [56] detections	68.47	318.33	30.43	926.11	58.67	516.23
Human (small sample)	90	58.40	80	81.90	90	43.00

A3C with ground truth detection actually performs no better than trivial solution.

Abalation study

Model	Existence		Counting		Spatial Relationships	
	Accuracy	Length	Accuracy	Length	Accuracy	Length
HIMN with YOLO [56] detections	68.47	318.33	30.43	926.11	58.67	516.23
HIMN with GT detection	86.56	679.70	35.31	604.79	70.94	311.03
HIMN with GT detection and oracle navigator (HIMN-GT)	88.60	618.63	48.44	871.12	72.50	475.55
HIMN-GT Question not given to planner	50.00	150.60	24.50	293.33	50.25	118.09
HIMN-GT No loss on invalid actions	49.84	659.28	24.84	911.46	50.00	613.50

Generalization

	Existence		Counting		Spatial Relationships	
Model	S	U	S	U	S	U
HIMN with YOLO [56] detections	73.68	68.47	36.26	30.43	60.71	58.67
HIMN with GT detections	94.00	86.56	42.38	35.31	73.38	70.94

Takeaways

- Long horizon planning problems are very hard for End-to-end RL even with some ground truth.
- A solution: compose low-level learning tasks and high-level learning tasks.
- Introducing human designed procedure can help get the task done.