AIProbe

Akash Yadav Muniraju
Chelsi Jain
Nirali Mehta
Rahil Piyush Mehta
Shivani Jinger

Introduction

1. The problem statement revolves around evaluating the hypothesized capability (HC) of a SDM agent within a constrained environment governed by specific rules.
2. The focus lies on assessing whether the SDM possesses the hypothesized capability, and if so, determining the precise instructions required to elicit the desired behavior.
3. To achieve this, the action space is defined, which is pivotal for generating valid directives within the environment.
4. The objective entails not only determining the presence of the hypothesized capability but also identifying the corresponding instruction set necessary for the SDM to perform the hypothesized behavior.
5. This investigation bridges the gap between theoretical capability and actionable implementation within a defined environment, offering insights into the functionality and adaptability of the SDM within its operational constraints.

Input/Output

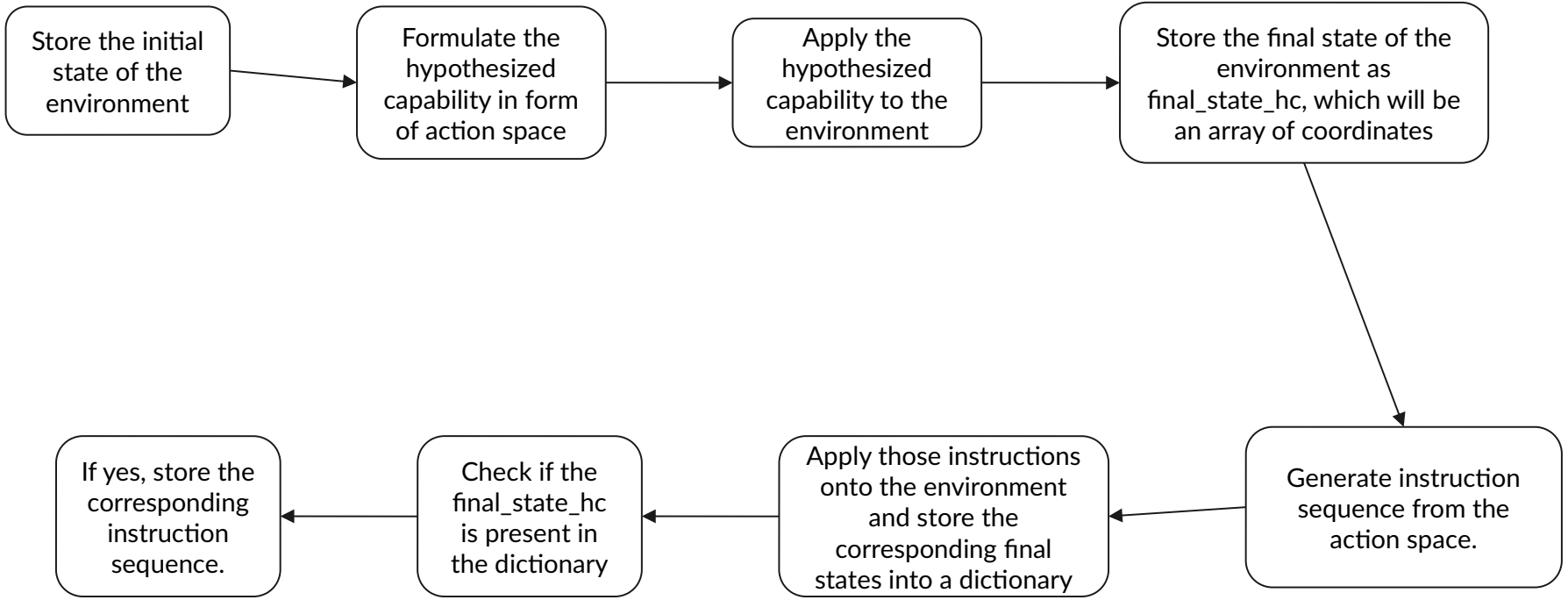
Input:-

1. Hypothesized Capability - formulated using the action space
2. SDM + environment(constrained)
3. Action Space (For example - [Pick, Move, Forward, Left, Right])

Output:-

4. Does SDM agent is able to perform this hypothesized capability (Yes/No)
5. If yes, then what instruction will make SDM perform that HC.

Project Methodology



Project Timeline

Task	Time	Assignee
Defining the environment and agent	1 week	Rahil Piyush Mehta, Nirali Mehta, Chelsi Jain
Generating instructions from the action space	1 week	Chelsi Jain
Applying the instructions to the environment	2 days	Rahil Piyush Mehta, Shivani Jinger
Formulating Hypothesized capability	1 week	Akash Yadav Muniraju
Comparing the final states, documentation	3 weeks	Chelsi Jain, Shivani Jinger, Rahil Piyush Mehta

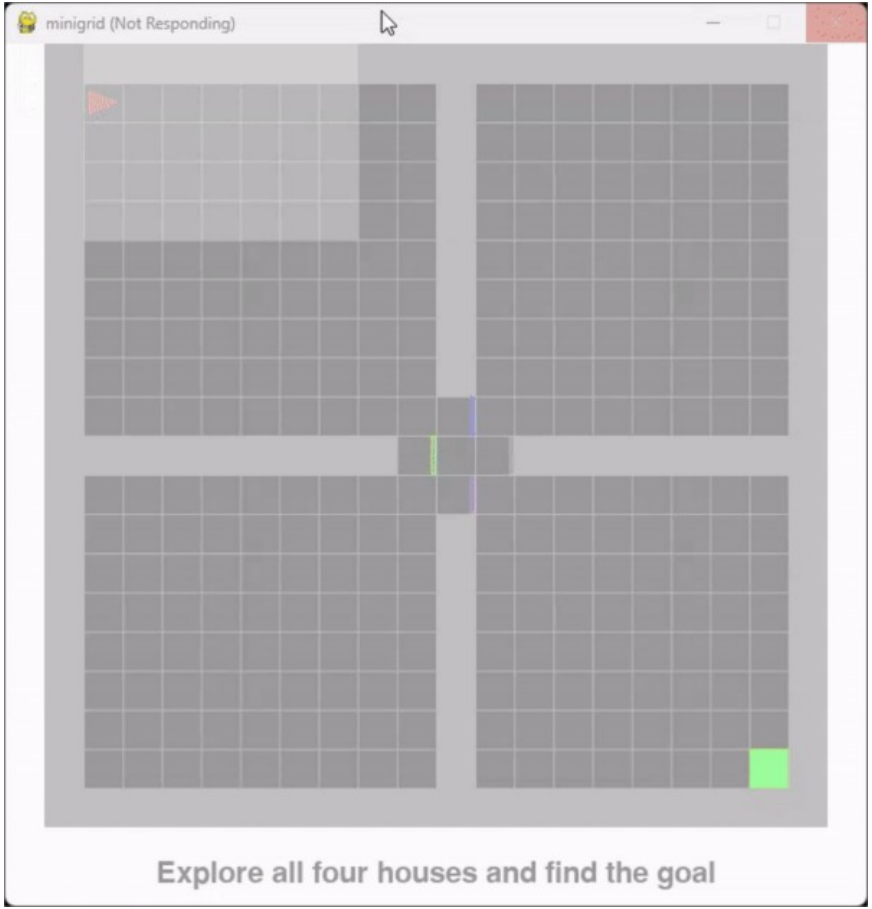
Action Space Example

Num	Name	Action
0	left	Turn left
1	right	Turn right
2	forward	Move forward
3	pickup	Unused
4	drop	Unused
5	toggle	Unused
6	done	Unused

Progress Update

1) Figured out how the environment is defined - https://minigrid.farama.org/environments/minigrid/FourRoomsEnv/
2) Figured out how the agent is defined
3) Generated instructions from the action space. - Done using brute force method
4) Applied instructions to the environment - https://minigrid.farama.org/environments/minigrid/FourRoomsEnv/
5) Created hypothesized capability from the action space for the classic four room mini grid environment.

Environment



Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Turn right ---> Turn right ---> Move forward ---> Turn right ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Turn right ---> Turn right ---> Move forward ---> Turn right ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Turn right ---> Turn right ---> Move forward ---> Turn right ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Move forward ---> Turn right ---> Turn right ---> Move forward ---> Turn right

Instructions from action space

```
Turn left
Turn left ---> Turn left
Turn left ---> Turn left ---> Turn left
Turn left ---> Turn left ---> Turn left ---> Turn left
Turn left ---> Turn left ---> Turn left ---> Turn left ---> Turn left
Turn left ---> Turn left ---> Turn left ---> Turn left ---> Turn right
Turn left ---> Turn left ---> Turn left ---> Turn left ---> Move forward
Turn left ---> Turn left ---> Turn left ---> Turn right
Turn left ---> Turn left ---> Turn left ---> Turn right ---> Turn left
Turn left ---> Turn left ---> Turn left ---> Turn right ---> Turn right
Turn left ---> Turn left ---> Turn left ---> Turn right ---> Move forward
Turn left ---> Turn left ---> Turn left ---> Move forward
Turn left ---> Turn left ---> Turn left ---> Move forward ---> Turn left
Turn left ---> Turn left ---> Turn left ---> Move forward ---> Turn right
Turn left ---> Turn left ---> Turn left ---> Move forward ---> Move forward
```

Hypothesized Capability

1) The agent traverse all the four in the direction room 1-2-3-4.

LEFT FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD RIGHT FORWARD FORWARD LEFT FORWARD FORWARD FORWARD FORWARD FORWARD RIGHT FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD—FORWARD—RIGHT FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD RIGHT FORWARD FORWARD FORWARD FORWARD FORWARD FORWARD.

2) The Agent fail to traverse the room and collide with the wall while traversing to room 3

FORWARD → FORWARD → LEFT → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → RIGHT → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD → FORWARD

Next Steps

1) We will create our own dataset with 10-15 mini grid environments and hypothesized capabilities.
2) Store the initial state of the environment.
3) Apply the hypothesized capability to the environment.
4) Store the final state as final_state_hc of the environment after applying.
5) Apply the generated instructions on to the environment and the agent.
6) Store the final state as final_state_2 of the environment.
7) Compare the final_state_hc and final_state_2, if they are the same, then return the output as True, else False.
8) Explore more techniques to optimize the algorithm in each of the above step.

Research Questions

1. What instructions lead any agent to perform a particular query?
2. How do the negative hypothesized capability within agent systems affect their overall performance, adaptability, and usability across diverse task domains and environmental conditions?