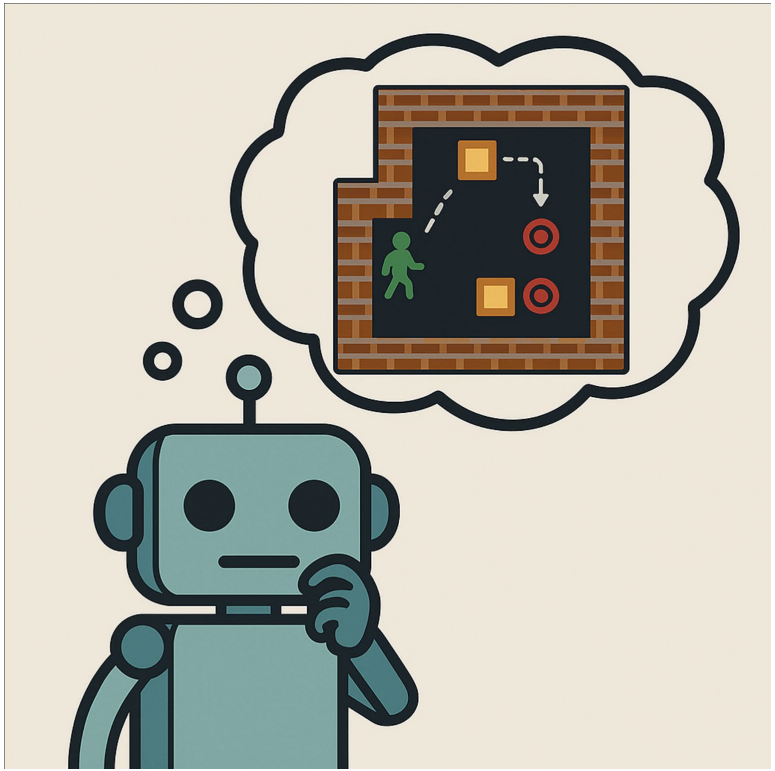
Interpreting Emergent Planning in Model-Free Reinforcement Learning

Thomas Bush[1], Stephen Chung[1†], Usman Anwar[1†], Adria Garriga-Alonso[2], David Krueger[3]

[1]University of Cambridge, [2]FAR AI, [3]Mila, University of Montreal / [†]Equal Contribution



Background

Planning in RL

- **Decision-time planning** ~= selecting actions by predicting and evaluating their long-run consequences.
- Typically associated with agents that possess explicit world models.
- This leads to the question we seek to answer:

Can reinforcement learning agents without explicit world models learn to perform planning?

Setting - The Agent

We mechanistically study a **Deep Repeated ConvLSTM (DRC)** agent as introduced by Guez et al. [1].

- Generic model-free training
- Recurrent, with 3D hidden states
- Performs N=3 recurrent computations per time step

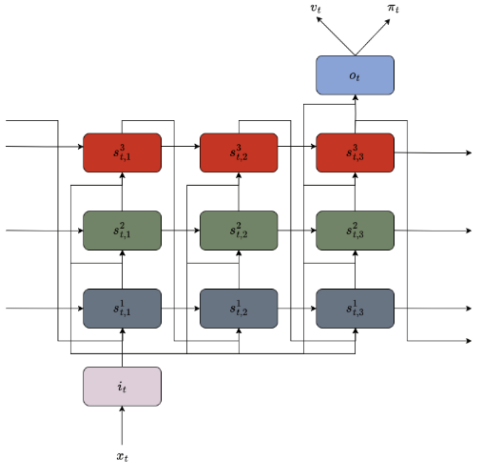
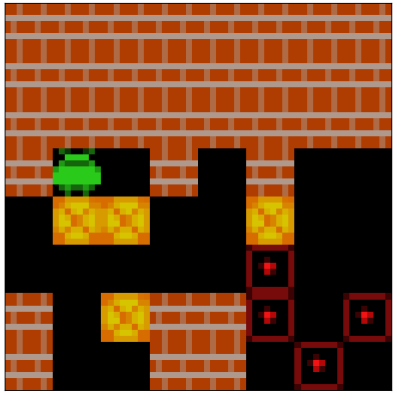


Figure 44: Illustration of DRC(3,3) architecture. For each time step, the architecture encodes the input x_t as a convolutional encoding i_t, passes it to a stack of 3 ConvLSTMs which perform three ticks of recurrent computation and then outputs policy logits π_t and a value estimate v_t.

Setting - The Environment

Specifically, we analyse a DRC agent trained to play the game of **Sokoban**.

- Agent navigates around walls in 8x8 grid
- Aim is to push 4 boxes on to 4 targets
- Actions have irreversible consequences (e.g. making level unsolvable).

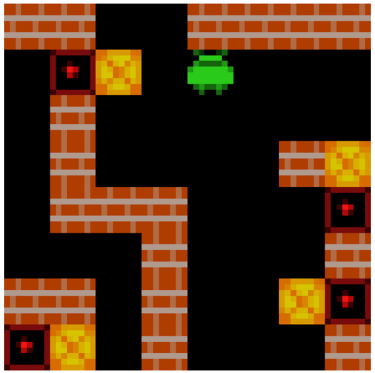


Planning-Like Behaviour

- Past work [1,2] has found that, in Sokoban, DRC agents **behave as though they planning**.
- For instance, DRC agents can solve harder levels if given “thinking time”.
- “Thinking time” = forced stationary steps at start of episodes before acting.

Trajectory 1

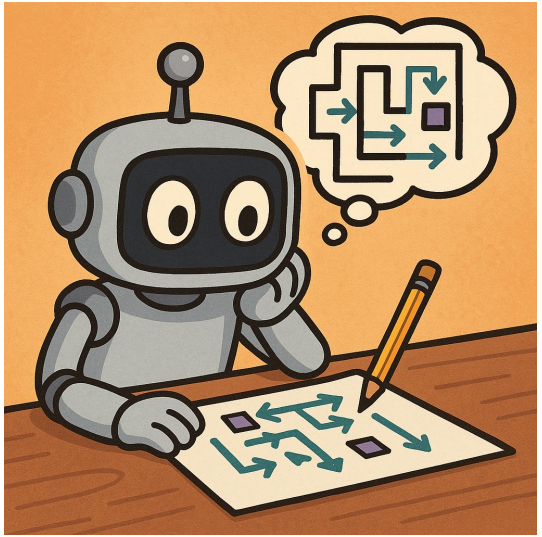
agent starts acting at time step 0



Step 1 - Probing For Planning-Relevant Concepts

How Might The Agent Plan?

We hypothesise the agent will plan by **representing planning-relevant concepts** e.g. concepts relating to future actions and their consequences.

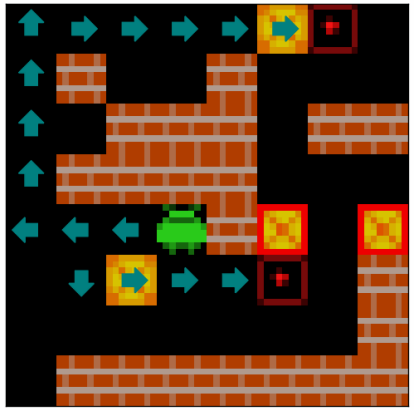


What Concepts Might The Agent Use?

We study two “square-level” concepts i.e. concepts that map each grid square to one of five classes: {UP, DOWN, LEFT, RIGHT, NEVER}

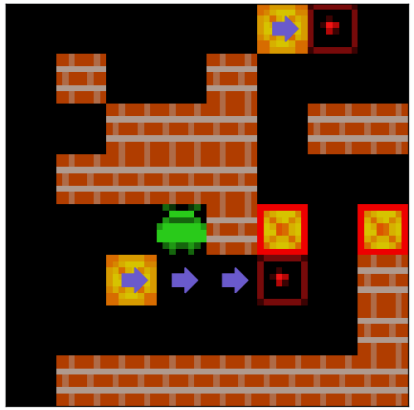
Concept 1: Agent Movements

What squares does the agent move onto in the rest of this episode, and from what direction?



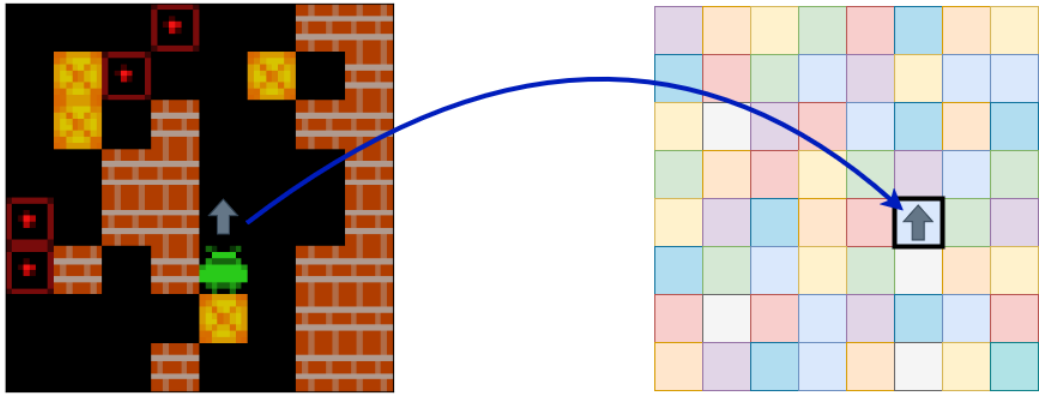
Concept 2: Box Movements

What squares does the agent push boxes off of, and in what direction?



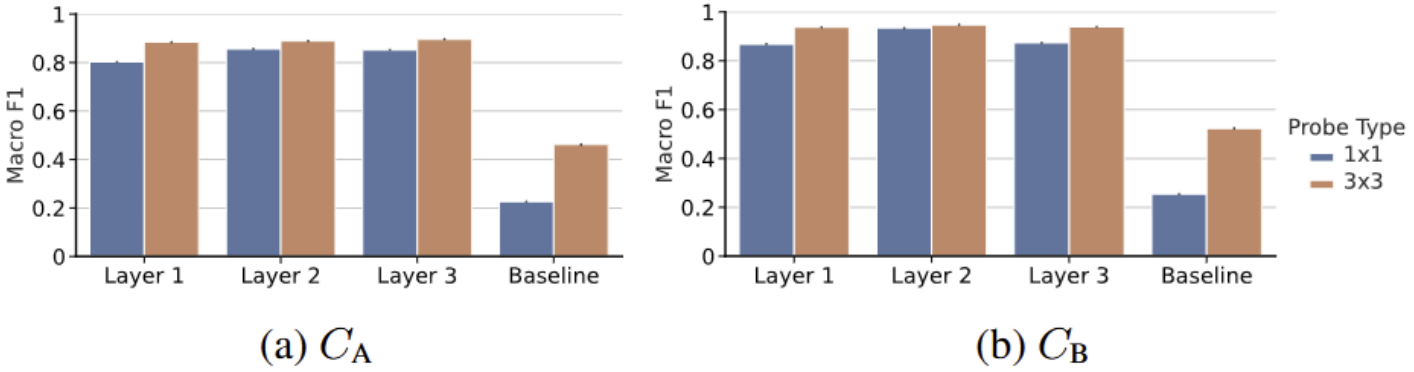
How Could The Agent Represent Concepts?

- We hypothesise that the agent will learn a **spatial bijection** i.e. the agent will represent concepts regarding square (x,y) at cell state position (x,y).
- We use **spatially-local linear probes** that predict concept class for square (x,y) using 1x1 or 3x3 patch of activations around (x,y).



Probing Results

1x1 linear probes can **accurately predict future agent movements and future box movements** from the agent's cell state.

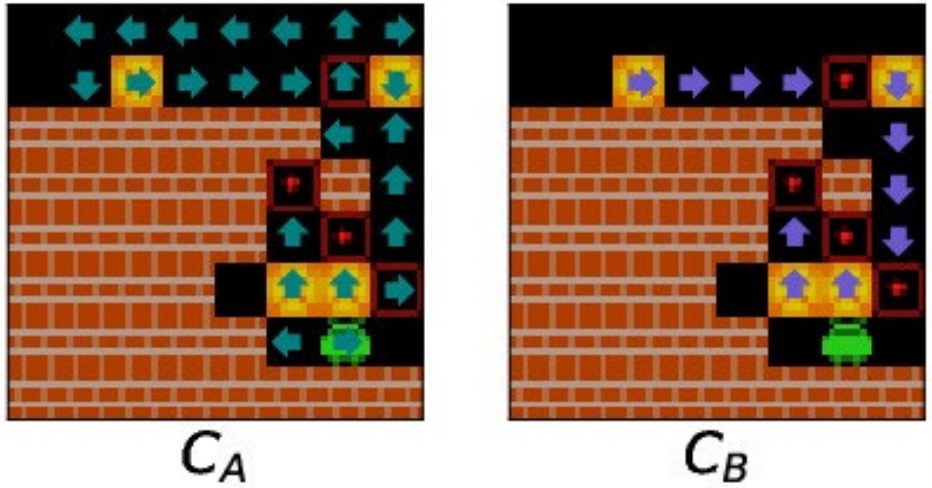


(a) C_{A} (b) C_{B}

Step 2 - How Are These Concepts Used?

How Does The Agent Use These Concepts?

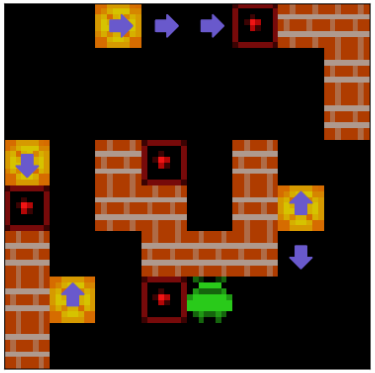
When applying linear probes to all positions of the agent's cell state, we find **internal plans** formulated by the agent.



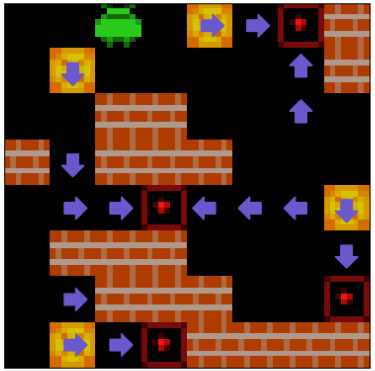
How Do These Plans Form?

The agent internally performs a search-based planning procedure.

- Extends plans **forward** from boxes.
- Extends plans **backward** from targets.
- **Evaluates** intermediate plans.
- **Adapts** plan when errors arise.
- Constructs sub-plans in **parallel**.



step 0, tick 1

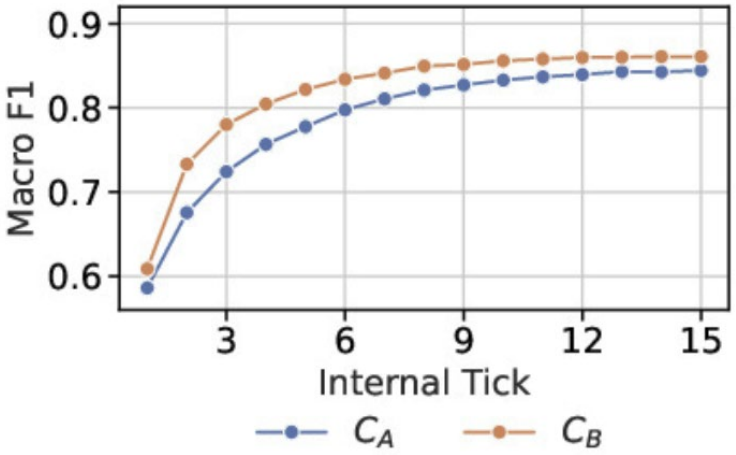


step 0, tick 1

Why Does The Agent Benefit From “Thinking Time”?

The agent’s plans become iteratively more accurate when given time to “think” before acting.

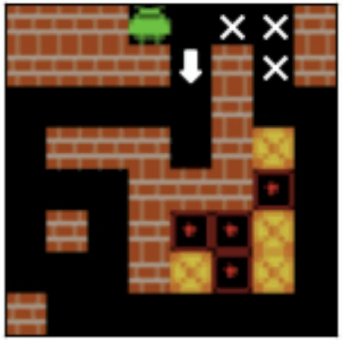
- Force agent to pause and “think”.
- Apply linear probes to agent’s cell state at each tick of “thinking” time.
- Measure correctness of internal plan at each tick.



Step 3 - Do Plans Influence Behaviour?

Do These Plans Causally Influence Behaviour?

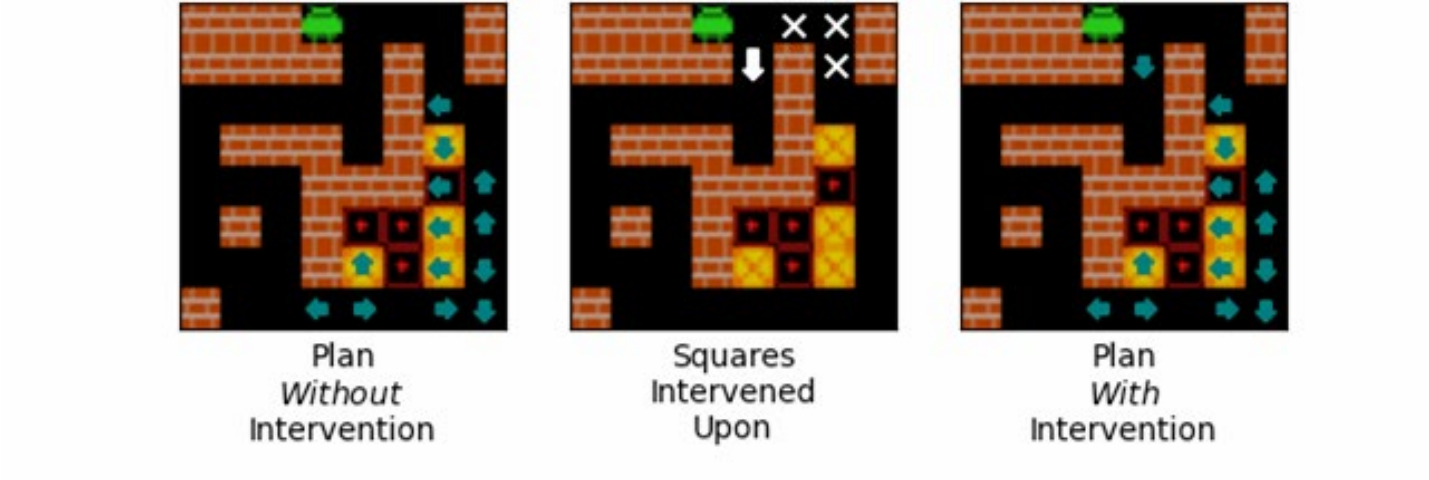
We intervene on the agent's cell state by adding the vectors learned by linear probes to encourage the agent to step onto certain squares (using {UP, DOWN, LEFT, RIGHT} vectors) and avoid other squares (using {NEVER} vector).



Squares Intervened Upon

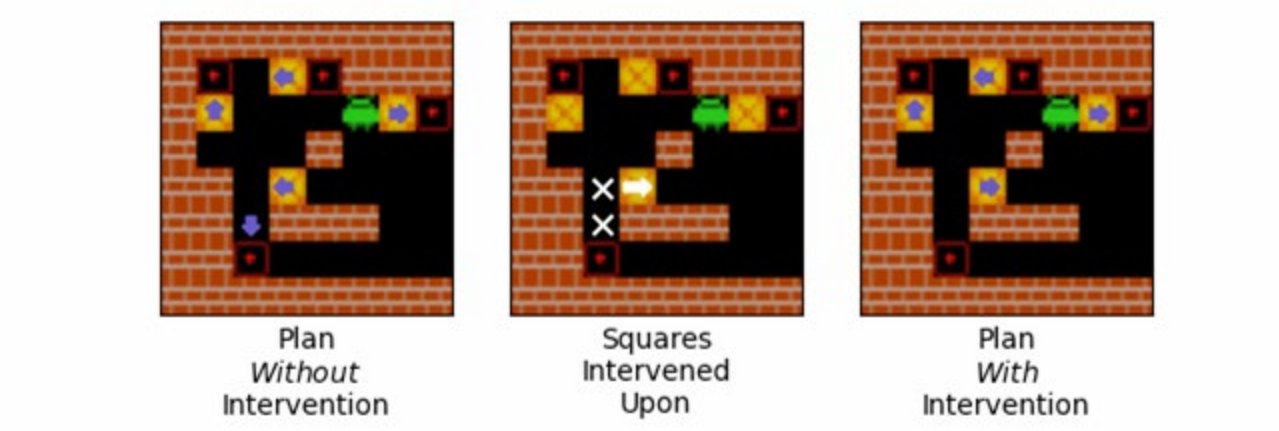
Do These Plans Causally Influence Behaviour?

We can intervene on the agent's cell state to cause the agent to **form and execute alternate long-term plans** involving taking sub-optimally long paths.



Do These Plans Causally Influence Behaviour?

This also works with intervening to encourage the agent to push boxes sub-optimal routes!



Summary

Summary

- We provide the first line of non-behavioural evidence that RL agents can learn to plan without an explicit world model.
- We demonstrate that a model-free RL agent has spontaneously learned an internal planning procedure with similarities to bidirectional search.

References

[1] Guez, Arthur, et al. "An investigation of model-free planning." *International conference on machine learning*. PMLR, 2019.

[2] Taufeeque, Mohammad, et al. "Planning in a recurrent neural network that plays Sokoban." *arXiv preprint arXiv:2407.15421* (2024).

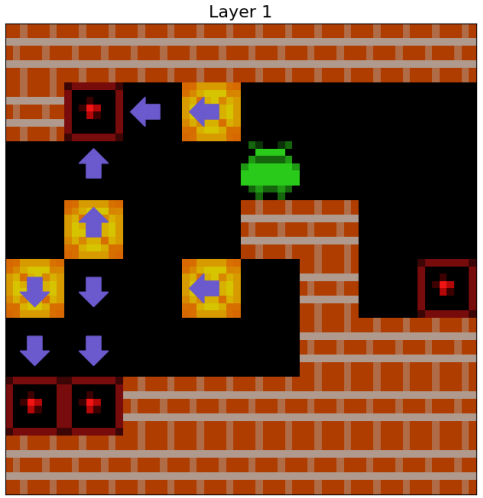
Thanks for listening!

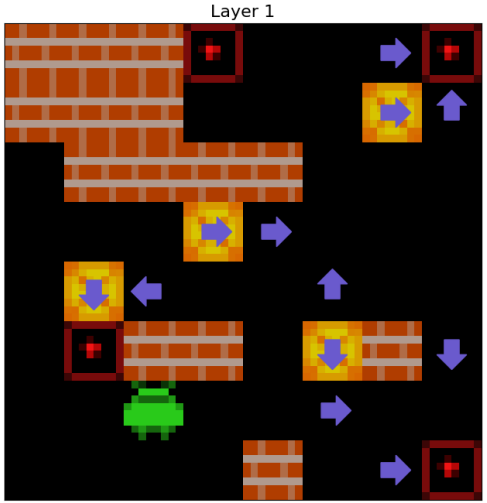
Questions?

Appendix

Planning in a ResNet

Using the same methodology, we find preliminary evidence of a **non-recurrent ResNet agent** planning iteratively across layers.





Planning in Mini Pacman

We also find evidence indicating that a DRC agent does an alternate type of (dynamic-horizon forward-facing) planning in **Mini PacMan**