Re-understanding Finite-State Representations of Recurrent Policy Networks

Mohamad H. Danesh, Anurag Koul, Alan Fern, Saeed Khorram

Oregon State University
College of Engineering

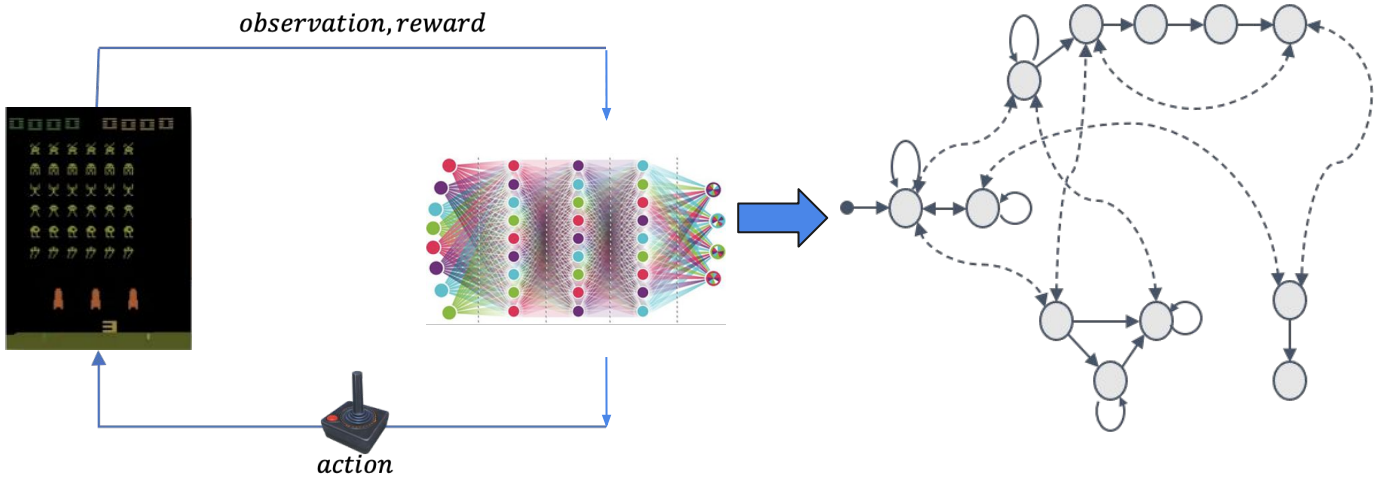
Prior Works: Attention Maps

- Highlighting salient parts of the input

Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. "Visualizing and understanding atari agents." *ICML'18*.

Prior Work: Extracting FSMs

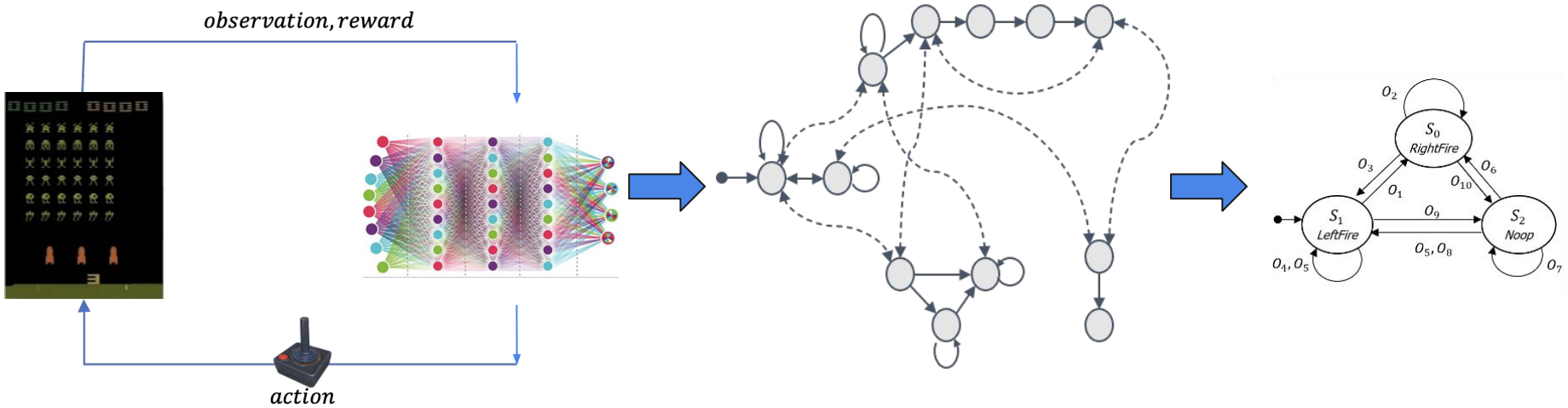
- Extracting finite-state machine representations



Anurag Koul, Samuel Greydanus, and Alan Fern. "Learning finite state representations of recurrent policy networks." *ICLR'19*.

Prior Work: Extracting FSMs

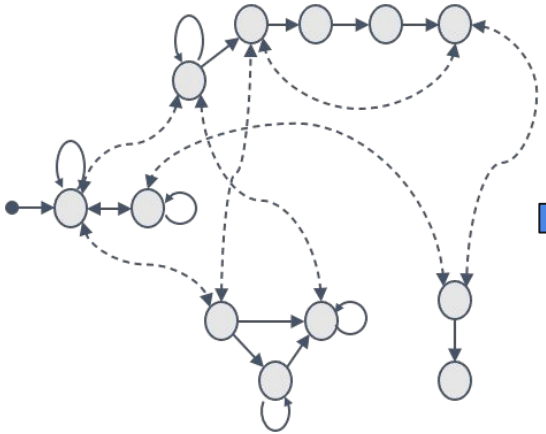
- Extracting finite-state machine representations



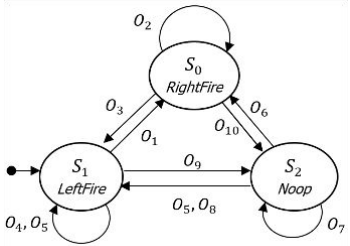
Anurag Koul, Samuel Greydanus, and Alan Fern. "Learning finite state representations of recurrent policy networks." *ICLR'19*.

Minimal May Not Be Most Interpretable

- Minimization may result in cryptic representations

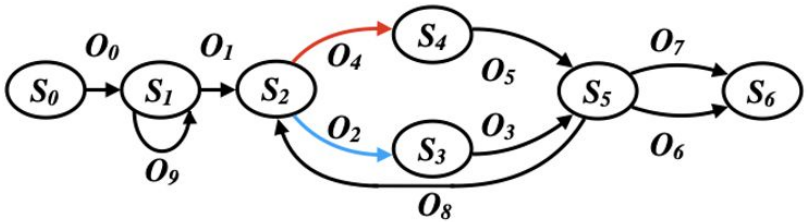


often very large, difficult analysis



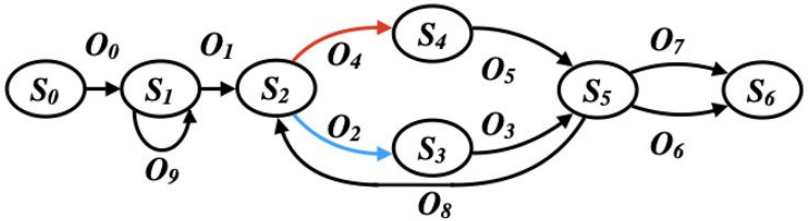
Interpretable Reductions

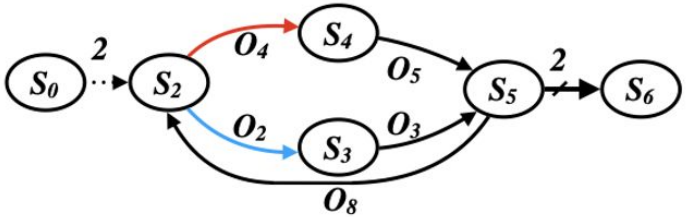
- Focusing on extracting decision points



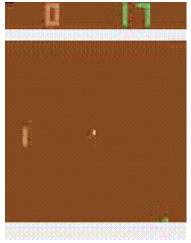
Interpretable Reductions

- Focusing on extracting decision points

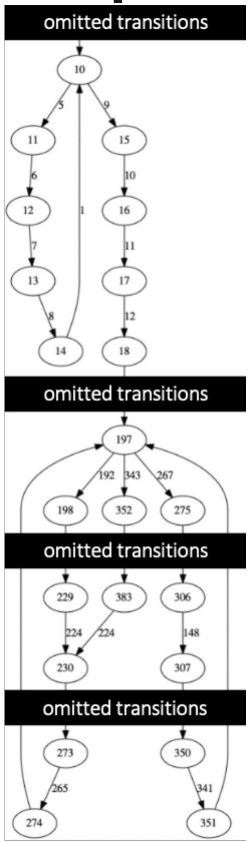




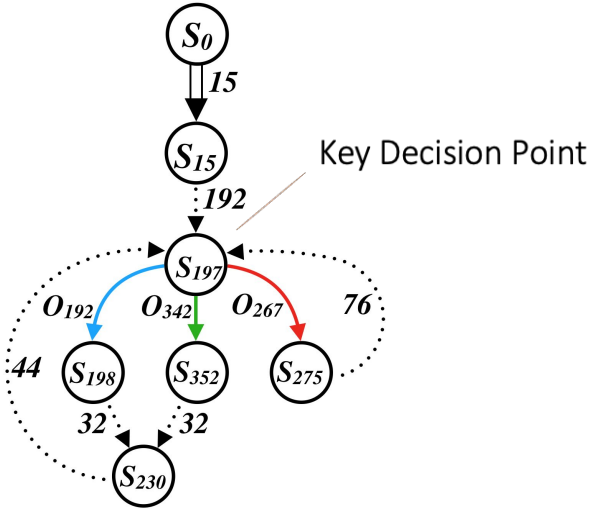
Example of Interpretable Reductions



Extracting FSM

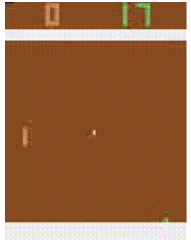


Interpretable reductions

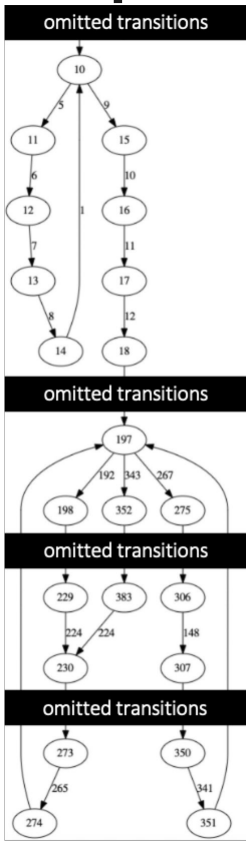


Key Decision Point

Example of Interpretable Reductions

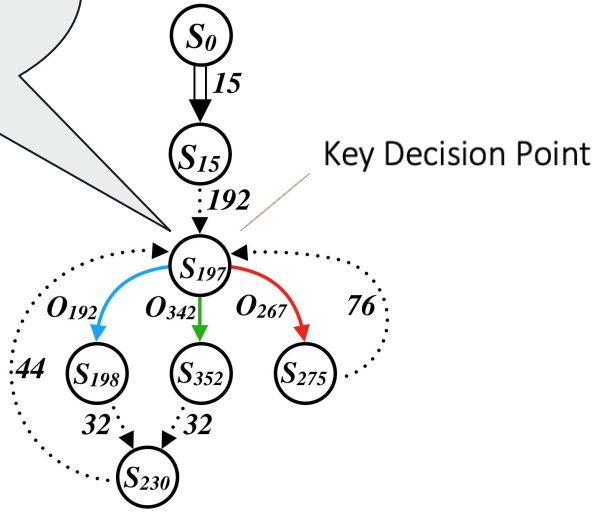


Extracting FSM



Interpretable reductions

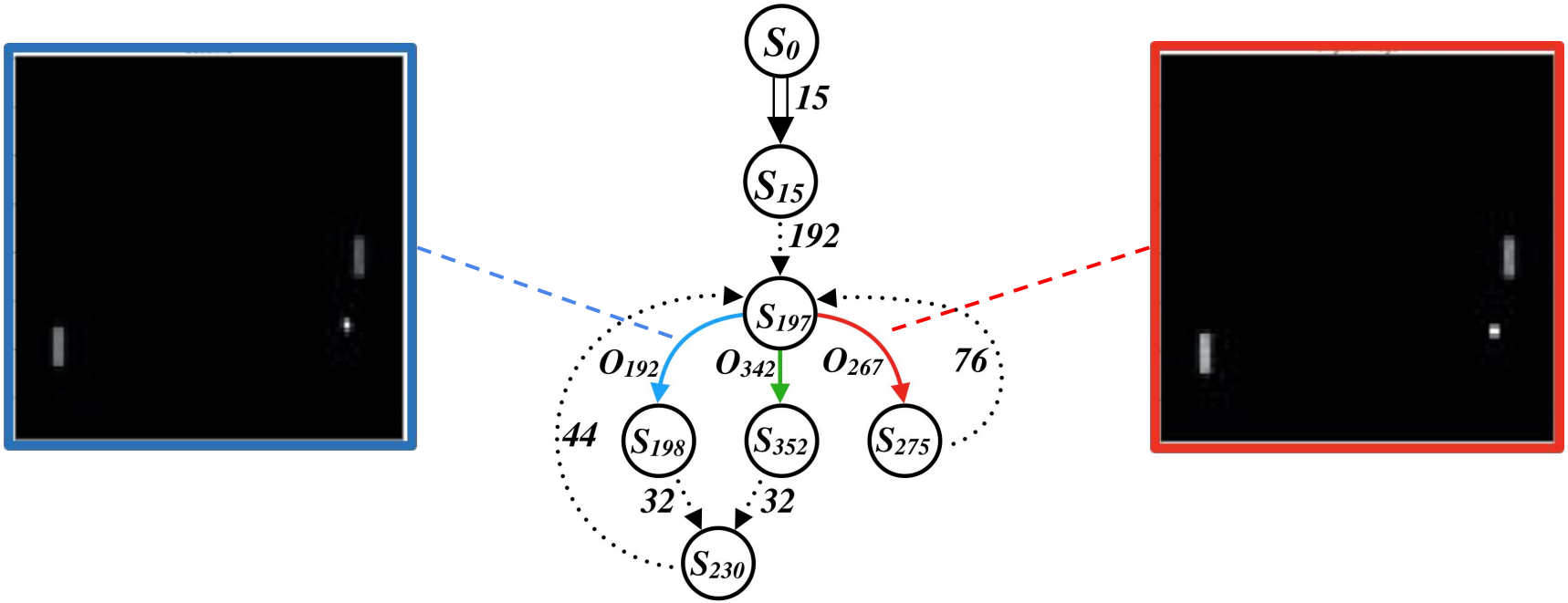
What is the basis for decision at 197?



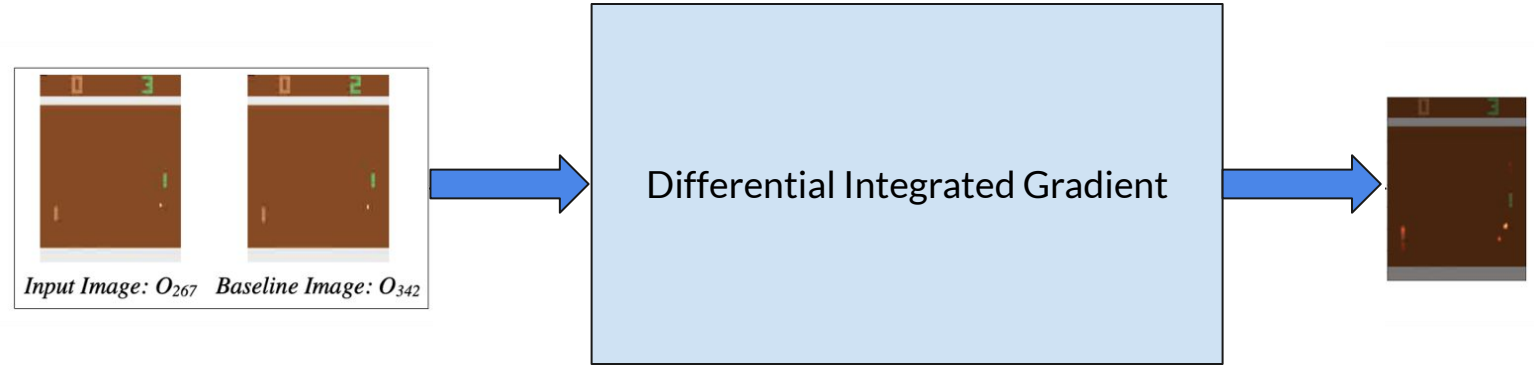
Key Decision Point

Differential Saliency for Decision Points

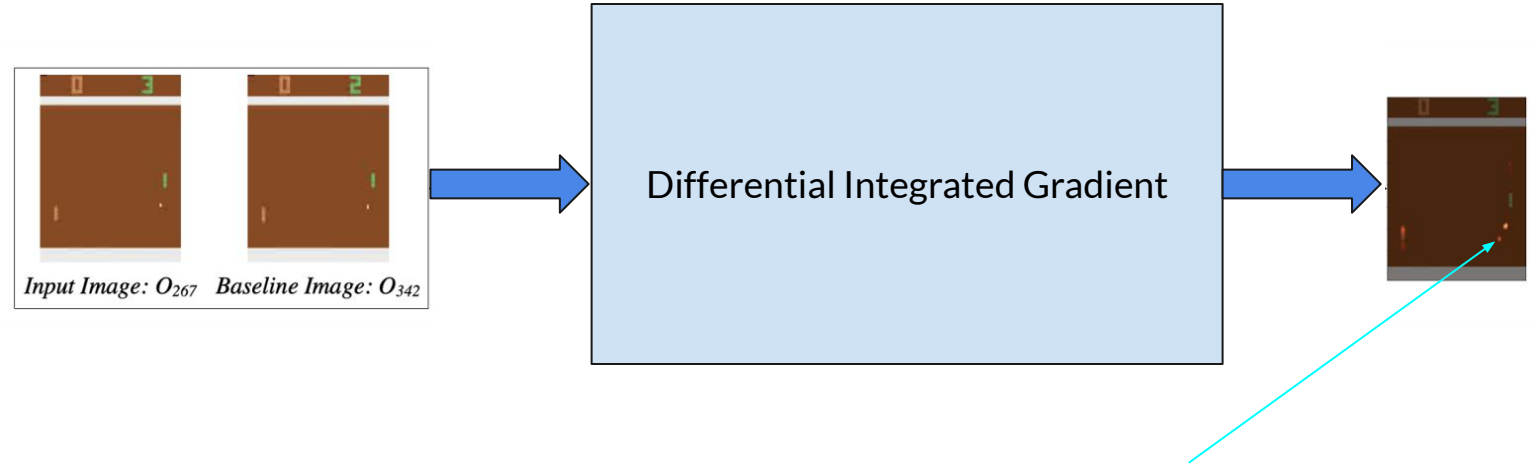
- What pixels make blue decision preferred over red?



Differential Saliency for Decision Points

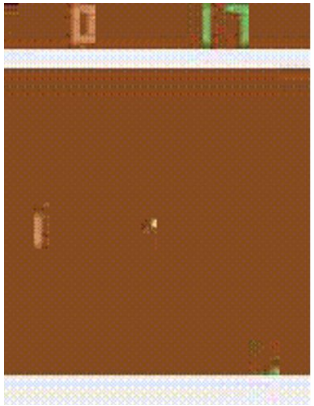


Differential Saliency for Decision Points



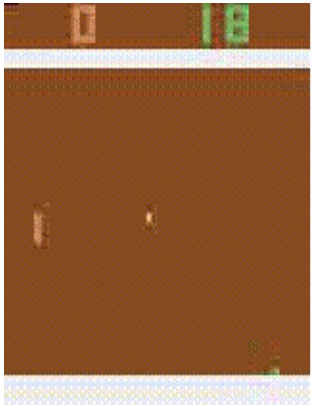
Differential Saliency for Decision Points

- Initial Screens for rounds 17 vs. 18:
- Decision point is effectively conditioning on whether game is at even vs. odd round!



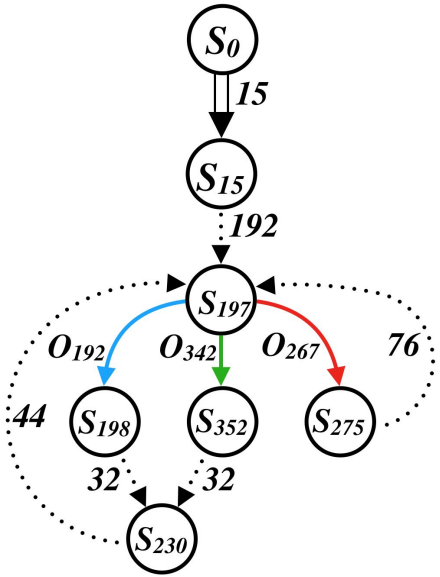
Differential Saliency for Decision Points

- Initial Screens for rounds 17 vs. 18:
- Decision point is effectively conditioning on whether game is at even vs. odd round!



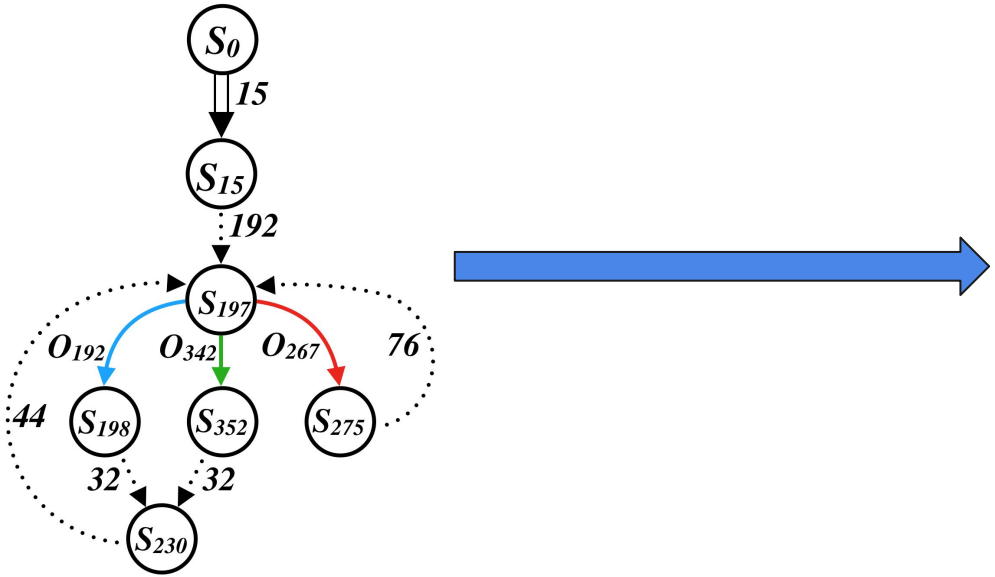
Functional Pruning

- What if we force machine to always go one way?



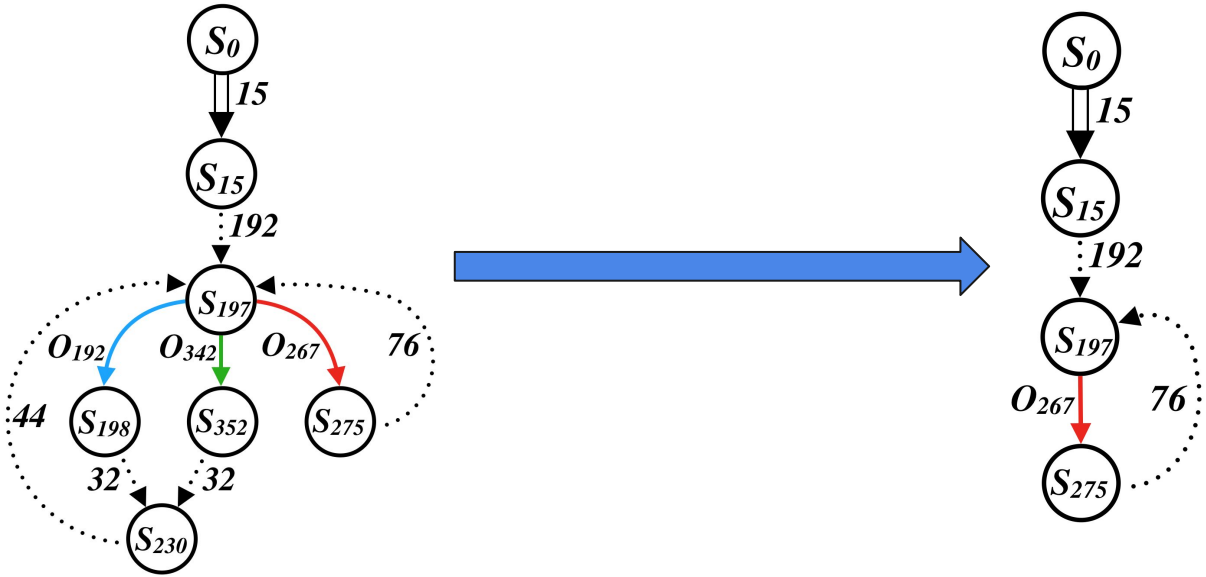
Functional Pruning

- What if we force machine to always go one way?



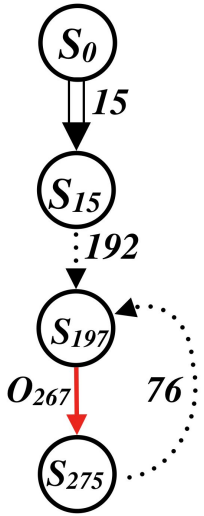
Functional Pruning

- What if we force machine to always go one way?



Pruned Open-Loop Controllers

- Decision points are not strategic
- For all 7 Atari games we considered the state machines were pruned open-loop controllers!



Comparing to Prior Insights

- Machines use observations in unintuitive ways
- Observations are not actually needed

Summary

- Explanations should be well-defined

Summary

- Explanations should be well-defined
- We are good at (mis-) interpreting explanations

Summary

- Explanations should be well-defined
- We are good at (mis-) interpreting explanations
- To use explanations to build trust:
  - They must be trustworthy and sound

Thank you!