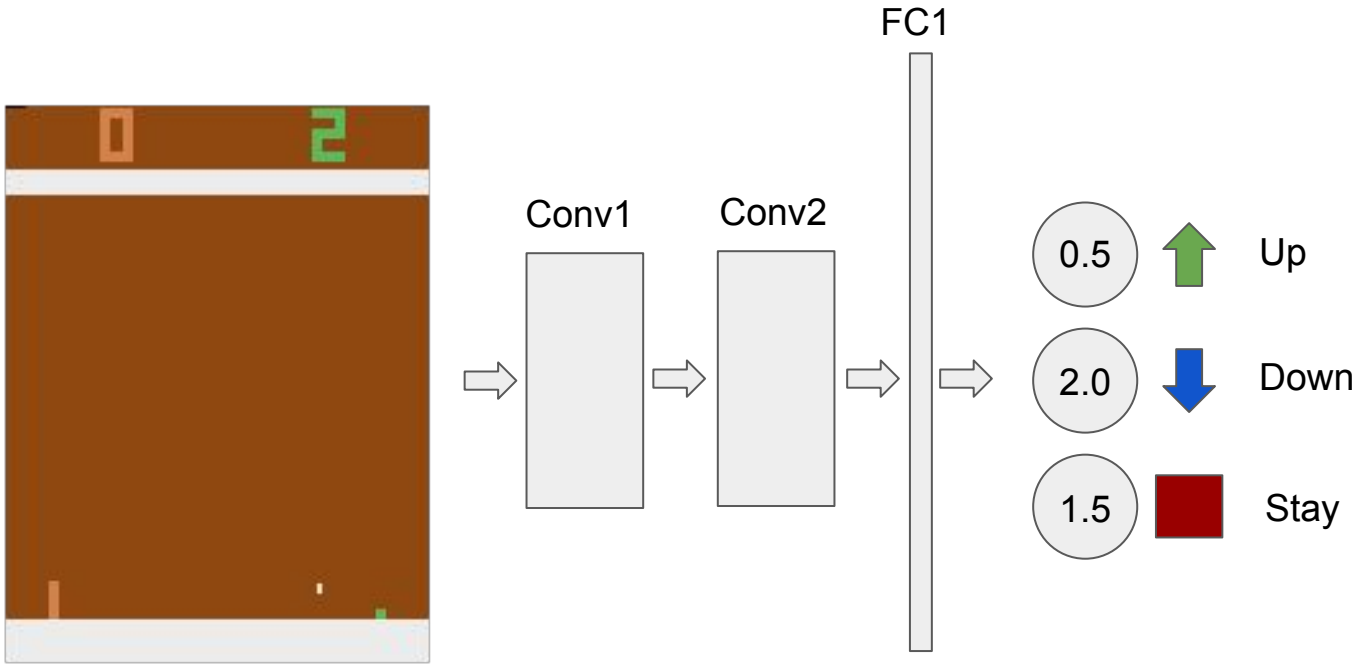
Surprising Negative Results for Generative Adversarial Tree Search

Kamyar Azizzadenesheli[1,2,5], Brandon Yang[2], Weitang Liu[3], Emma Brunskill[2], Zachary C Lipton[4], Animashree Anandkumar[5]

[1]UC Irvine, [2]Stanford University, [3]UC Davis, [4]Carnegie Mellon University, [5]Caltech

Introduction: Deep Q-Network (DQN)

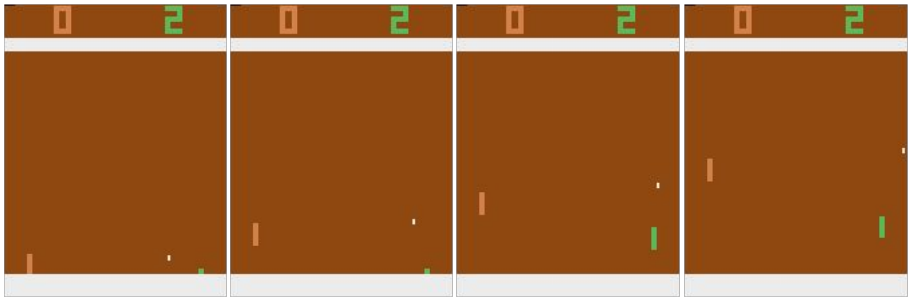


Introduction: DQN

$$Q - \hat{Q} = bias + variance$$

The DQN estimation of the Q-function can be arbitrarily biased (Thrun & Schwartz 1993, Antos et al. 2008)

We empirically observe this phenomenon in DQN for Pong

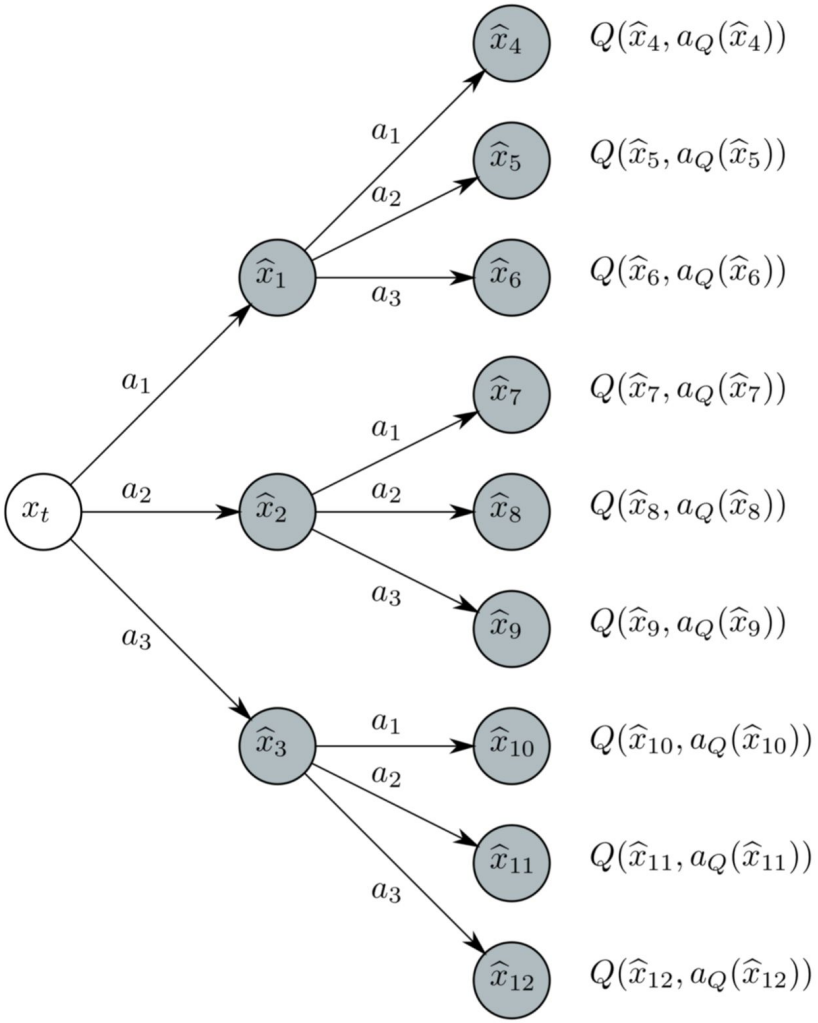


	Action space		
Steps	stay	up	down
t	4.3918	4.3220	4.3933
$t+1$	2.7985	2.8371	2.7921
$t+2$	2.8089	2.8382	2.8137
$t+3$	3.8725	3.8795	3.8690

Generative Adversarial Tree Search

Given a model of the environment:

1. Do Monte-Carlo Tree Search (MCTS) for a limited horizon
2. Bootstrap with the Q function at the leaves

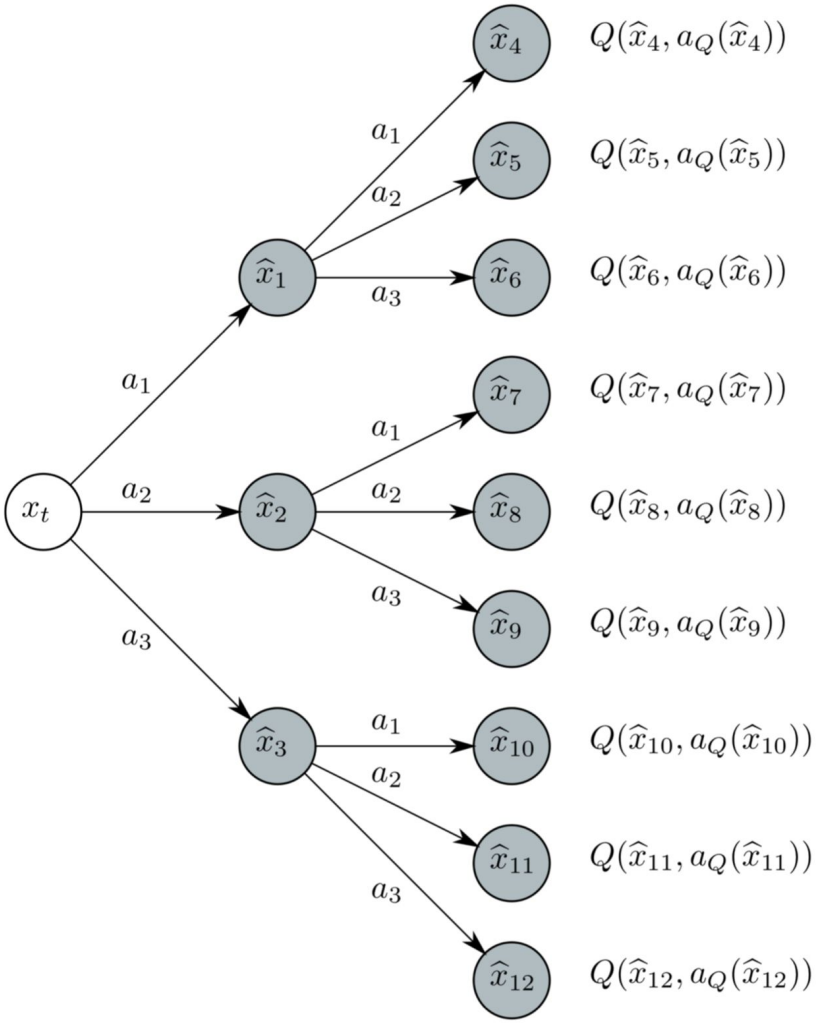


Generative Adversarial Tree Search

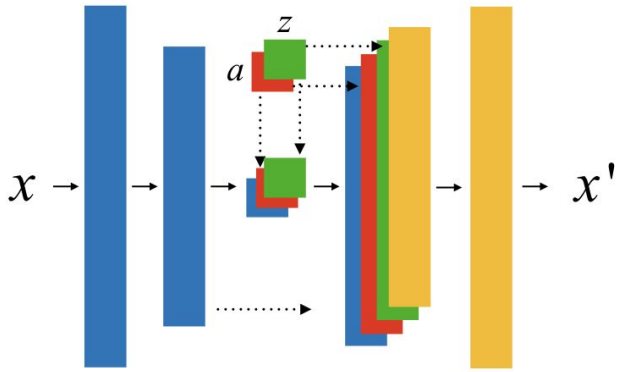
Given a model of the environment:

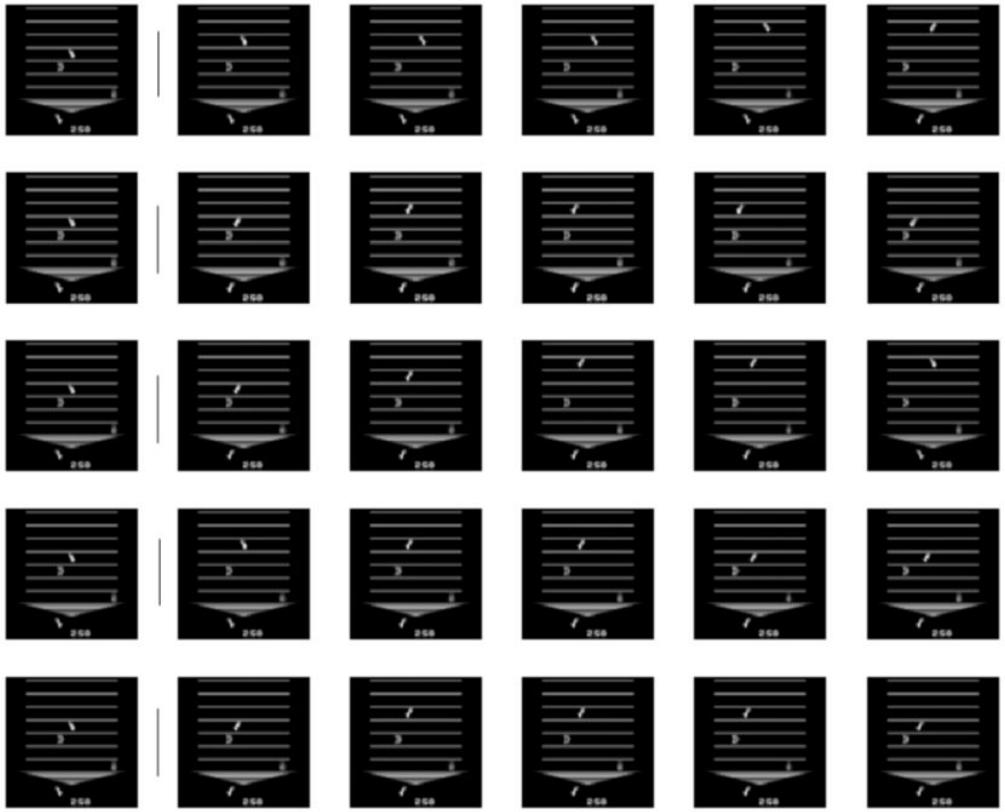
1. Do Monte-Carlo Tree Search (MCTS) for a limited horizon
2. Bootstrap with the Q function at the leaves

[Prop. 1] Let e_Q be the upper bound on the error in estimation of the Q-function. In GATS with roll-out horizon H, it contributes to the error in estimation of the return as $\gamma^H e_q$.



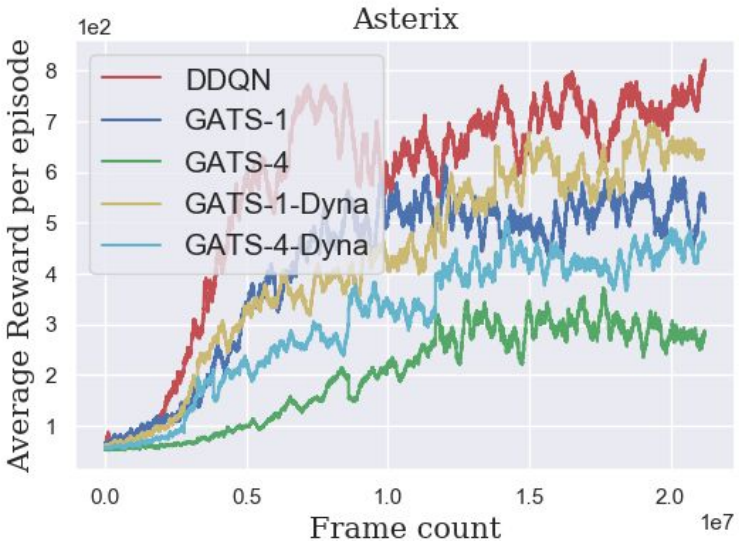
Generative Dynamics Model

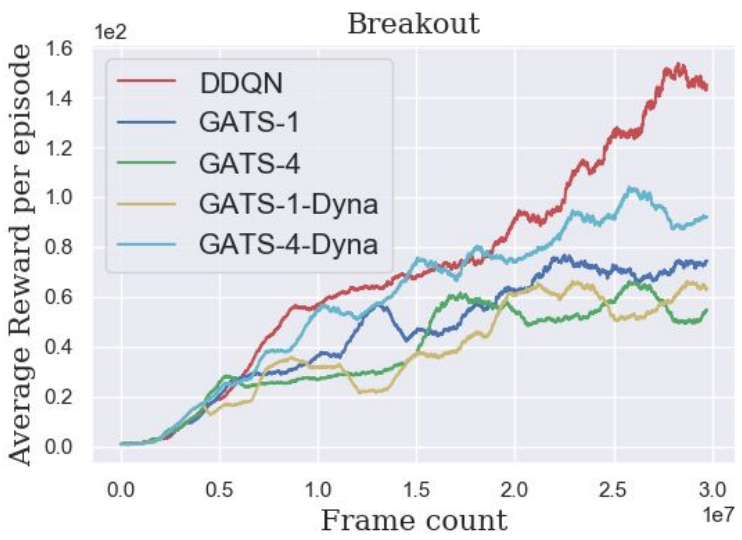




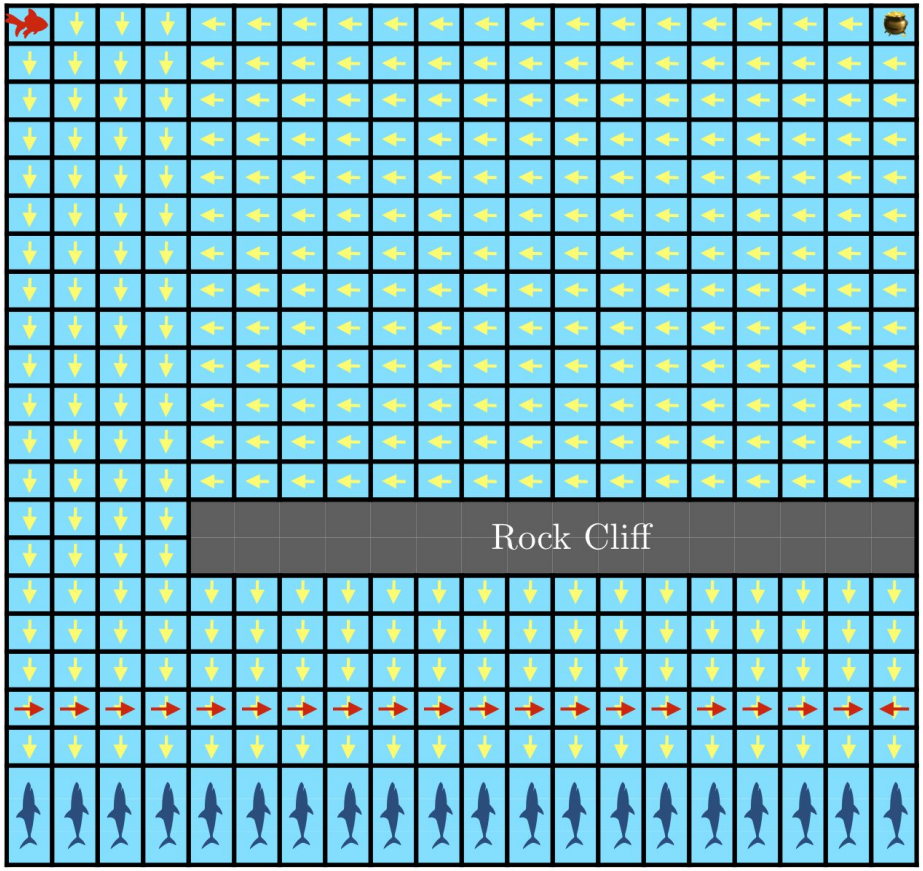
Generates next frames conditioned on the current frames and actions

Negative Results

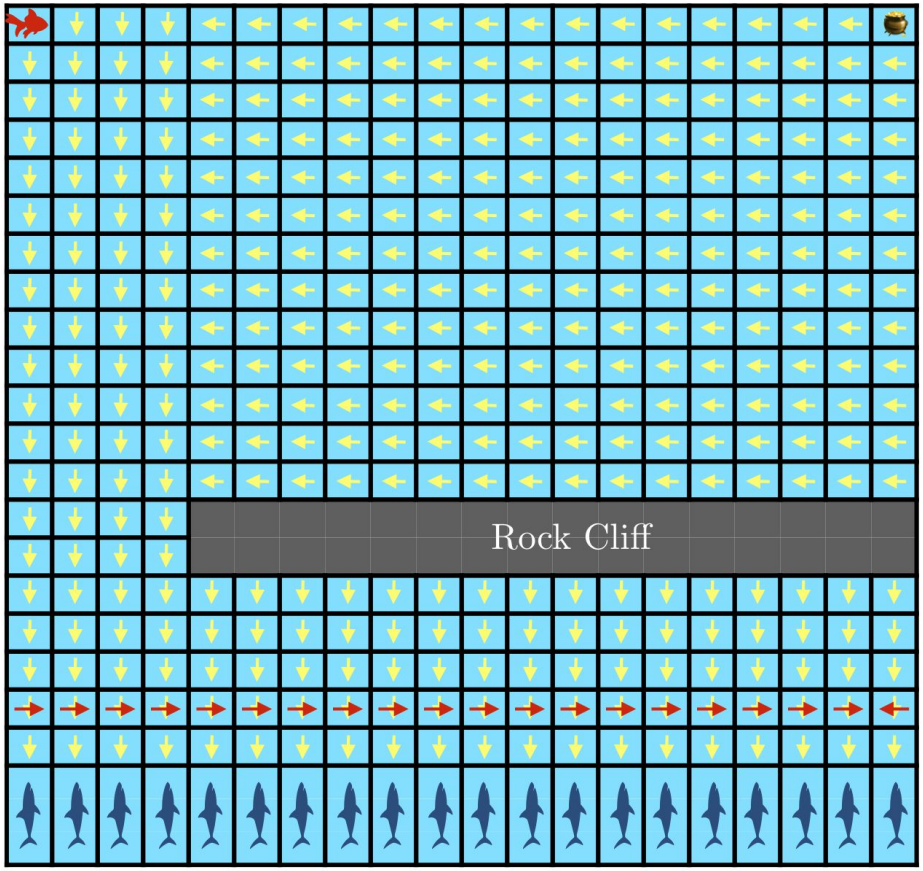


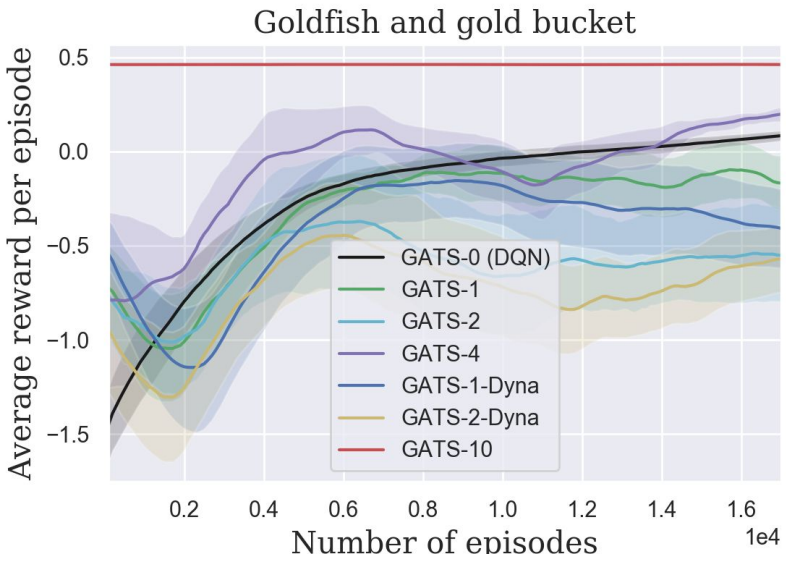


The Goldfish and the Gold Bucket



The Goldfish and the Gold Bucket





Conclusions

We develop a sample-efficient generative model for RL using GANs

Given a fixed Q-function, GATS reduces the worst-case error in estimation from the Q-function exponentially in roll-out depth as $\gamma^H e_q$.

Even with perfect modeling, GATS can impede learning of the Q-function.

This study of GATS highlights important considerations for combining model-based and model-free reinforcement learning.