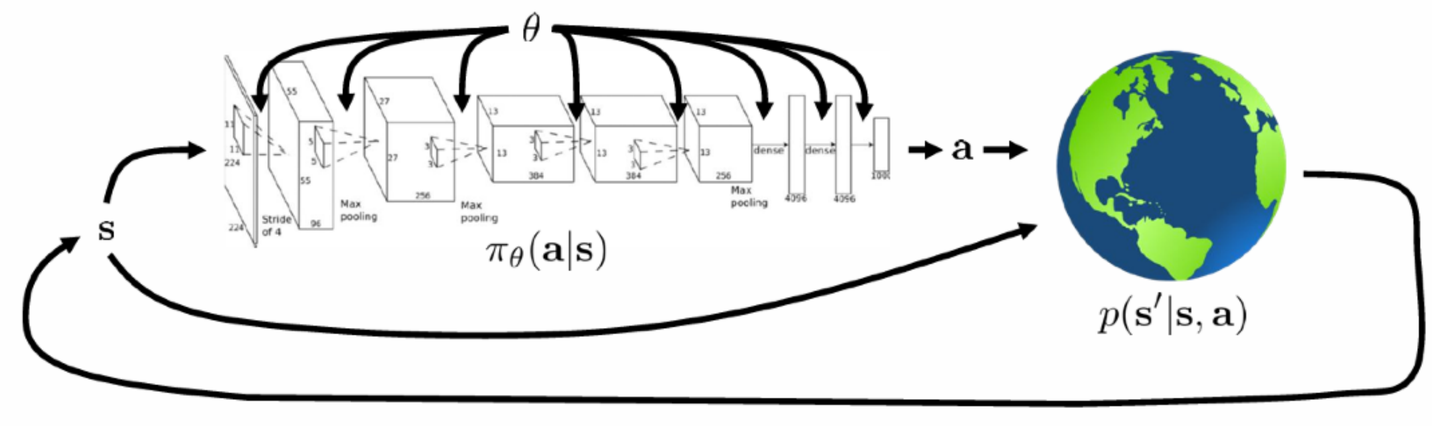
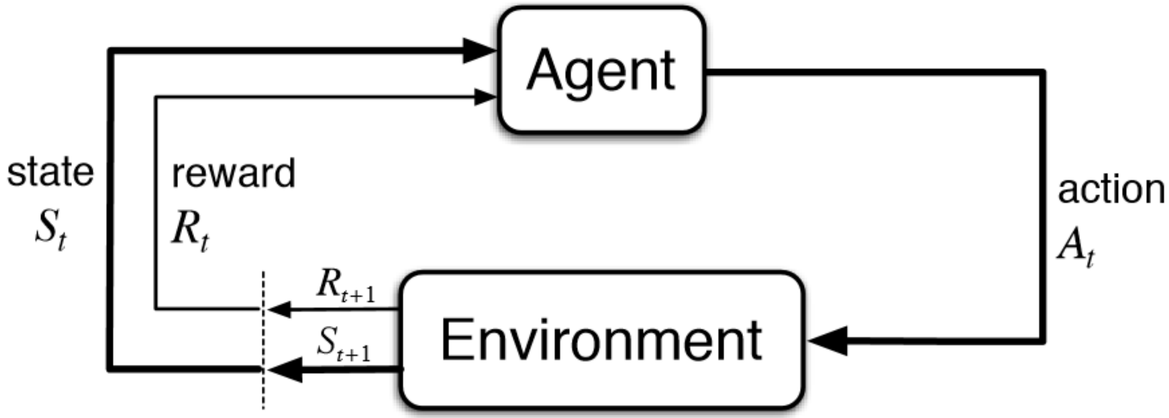
Reinforcement learning



- Observe initial state s_0
- For every round $t = 1, \ldots, T$
  - Take action a_t
  - Observe next state $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$ and reward r_t

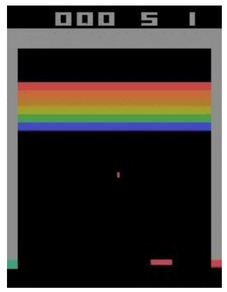
Goal: Find policy π which is a mapping from state to action such that expected reward over T rounds is maximized

$$\sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)} [\gamma^{t-1} r_t(s_t, a_t) \mid s_0]$$

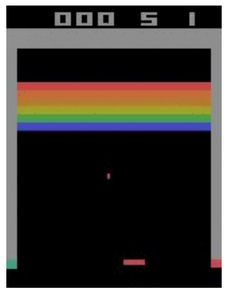


Atari Breakthrough (2013, DeepMind)

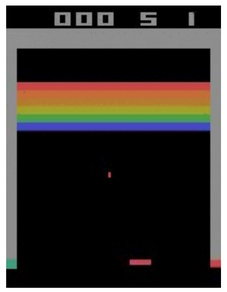
Breakout



Initial performance

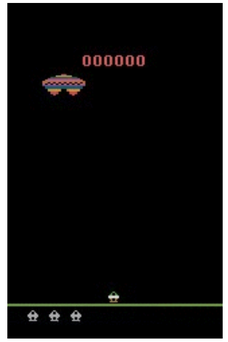


After 15 mins of training

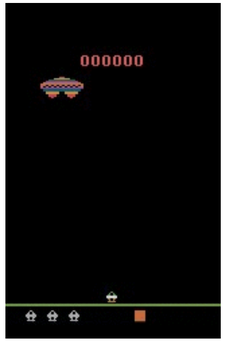


After 30 mins of training

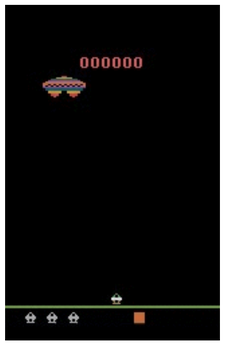
Assault



Initial performance



After 15 mins of training



After 30 mins of training