Real-time Attacks in Deep Reinforcement Learning

Can we effectively fool deep reinforcement learning (RL) policies in real-time ?

The setting:

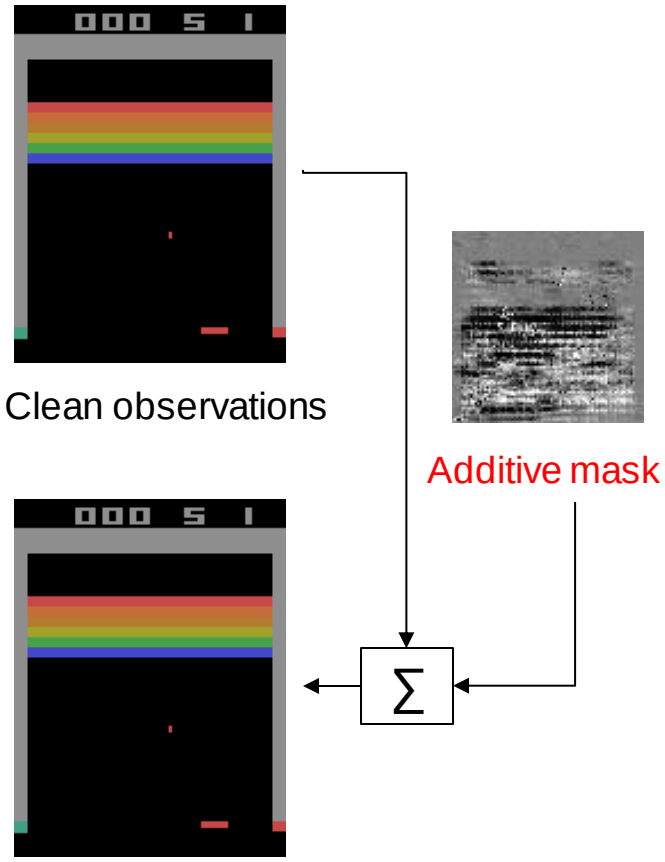
- Goal: Force an RL agent's policy to output a wrong action
- Real-time, untargeted, white-box attacks

The problem:

- Current methods depend on input, too slow to be real-time

Our solution:

- Pre-computed uniform perturbations added to observed input
- Works in real-time but with comparable attack success
- Can evade known defenses



In progress