Title: Adversarial Attacks on Neural Network Policies

Authors: Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, & Pieter Abbeel

Publication Year: 2017

Key Ideas

The article focuses on deep reinforcement learning (DRL), where neural network policies are vulnerable to adversarial attacks, just like in supervised learning. It shows how small perturbations can cause a robot to make incorrect decisions, even if it was previously well-trained. The paper outlines two types of attacks, white-box attacks, where the attacker has full knowledge of the policy network and can create targeted perturbations, and black-box attacks, where the attacker has no internal knowledge but exploits transferability to generate adversarial examples that still mislead the policy.

Contributions

This article systematically demonstrates how adversarial examples can significantly affect neural network policies trained with reinforcement learning. It shows that even agents trained in controlled environments like games can fail due to minor input changes. The paper introduces the Fast Gradient Sign Method (FGSM) as a way to create adversarial observations both during training and test time. It also provides experimental analysis across different RL algorithms (DQN, A3C, TRPO) and multiple environments to show how performance drops under adversarial conditions.

Limitations

While the results are compelling, the experiments were only conducted in simulated environments, not in real-world settings. The study focuses on short-term technical performance and doesn't explore human-robot trust implications. It highlights vulnerabilities but doesn't provide defensive strategies. The attacks also only apply to visual input policies, without addressing more complex sensor fusion scenarios common in real-world robotics.

Extend/Improvement

My work will focus on building defensive strategies for adaptive robots in unpredictable, real-world environments, beyond simulations. I plan to develop methods for detecting and mitigating adversarial inputs during runtime for both white-box and black-box attacks. Since the article emphasizes the vulnerability of neural policies, I will explore learning pipelines that enable robots to adapt to these attacks and maintain reliable performance.