ROBUSTNESS AGAINST RELATIONAL ADVERSARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adversarial attacks have posed serious challenges to the robustness of machine-learning models, and in many settings the adversarial perturbation is bounded by small ℓ_p-norms. Motivated by semantics-preserving attacks in vision and security domain, we investigate relational adversaries, a broad class of attackers who create adversarial examples that are in a reflexive-transitive closure of a logical relation. We analyze the conditions for robustness and propose *normalize-and-predict* – a learning framework with provable robustness guarantee. We compare our approach with adversarial training and derive an unified framework that provides benefits of both approaches. Guided by our theoretical findings, we apply our framework to image classification and malware detection. Results of both tasks show that attacks using relational adversaries frequently fool existing models, but our unified framework can significantly enhance their robustness.