THE ROBUSTNESS AND ACCURACY TRADEOFF MAY ARISE FROM INTRA-CLASS SPATIAL INVARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

In adversarial machine learning, the robustness and accuracy tradeoff has been the focus of much previous work, yet the cause of this tradeoff remains poorly understood. While previous work studies tradeoffs between ℓ_∞ robustness and natural accuracy as well as ℓ_∞ and spatial (rotation-translation, or RT) robustness, they consider each tradeoff independently and fail to consider the nuanced relationships between ℓ_∞ robustness, RT robustness, and natural accuracy. In the following, we harmonize tradeoffs between all combinations of these settings on a single realistic distribution. Our theoretical analysis leads to an intriguing finding: the ℓ_∞ robustness and natural accuracy tradeoff is exacerbated by a fundamental tension between ℓ_∞ robustness and intra-class spatial invariance. In fact, we demonstrate that the tradeoff provably worsens as our distribution admits more intra-class spatial invariance and show empirically that this is the case for robust classifiers trained on variants of the CIFAR10 dataset. Importantly, our findings shed light on the difficulties of training naturally-accurate robust models on distributions with this property (e.g., natural images), frequently used to evaluate adversarial robustness.