1 Introduction

1.1 Motivation

Current regarding adversarial examples mostly focus on regularizers, model geometry or studying intrinsic robustness and features. But there are few works focusing on the effect of loss function. Our initial guess was that training a self-supervised(SSL) model could reduce the adversarial susceptibility by removing spurious correlation between images and labels. And that an SSL model would retain more features. To test this hypothesis I trained supervised and different SSL models on the CIFAR10 classification task which disproved the hypothesis. To gain more insight I used a synthetic dataset to gain more control over the causal factors that affect the dataset.

complete

1.2 Contributions/Findings

The initial experiments on CIFAR10 showed that SSL is more susceptible to adversarial examples than a standard supervised model. Further experiments with the synthetic dataset showed that ...

complete