NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Improving Model Robustness

with Adversarial Regularization for Image Classification

Student: Guo Wanyao Supervisor: Dr Zhang Hanwang

Abstract

This project demonstrates how adversarial regularization can be harnessed to improve an image classifier's robustness against adversarial examples. We implement such regularization technique via the Neural Structured Learning framework. The results of our baseline model and the adversarially regularized model are in stark contrast: the accuracy of our baseline model decreases significantly by 67% on the adversarially perturbed datasets, while the adversarially regularized model only has 8% drop in accuracy.

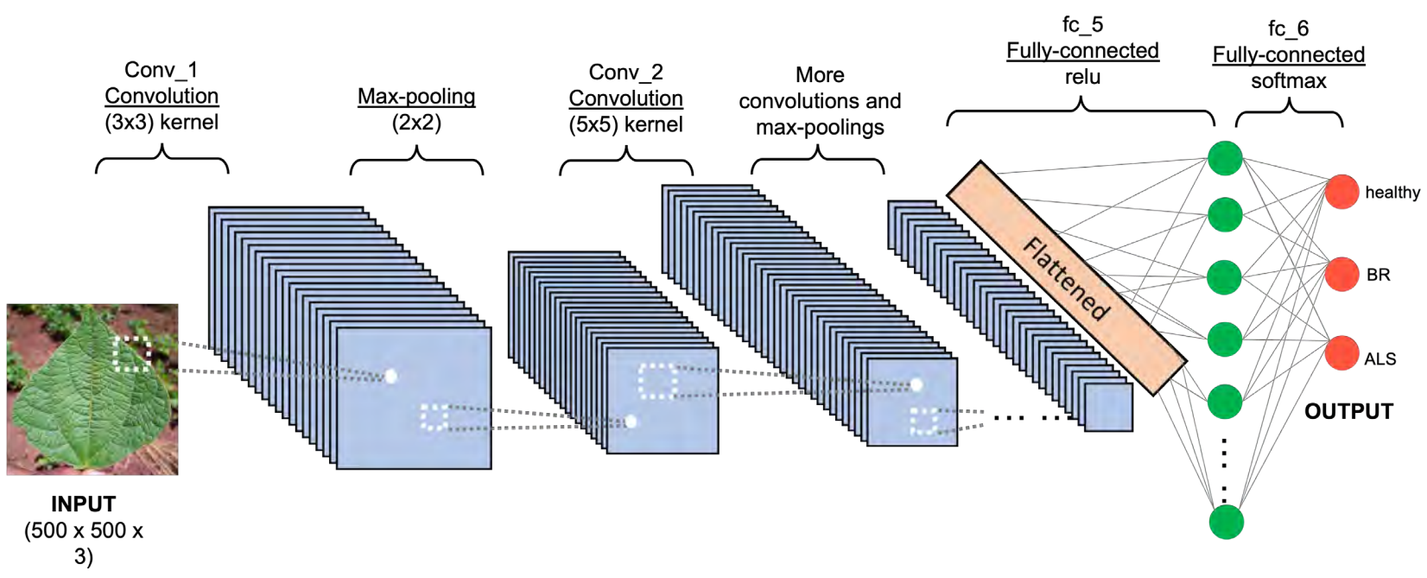
Overview of Results

Accuracy	Baseline model	Regularized model
Accuracy before perturbation	0.89	0.91
Accuracy after perturbation	0.22	0.83
Reduction in accuracy	0.67	0.08

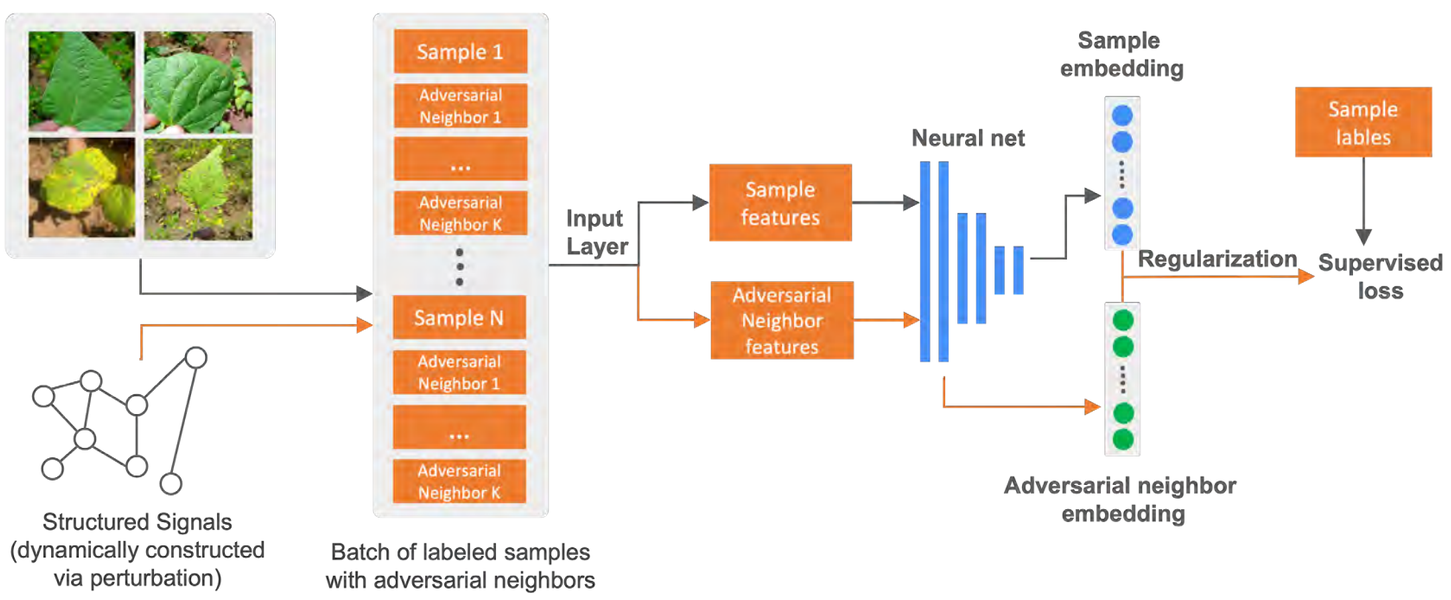
This suggests that back-feeding adversarial examples to training might improve generalization of the resulting model.

Model Architecture

Baseline Model

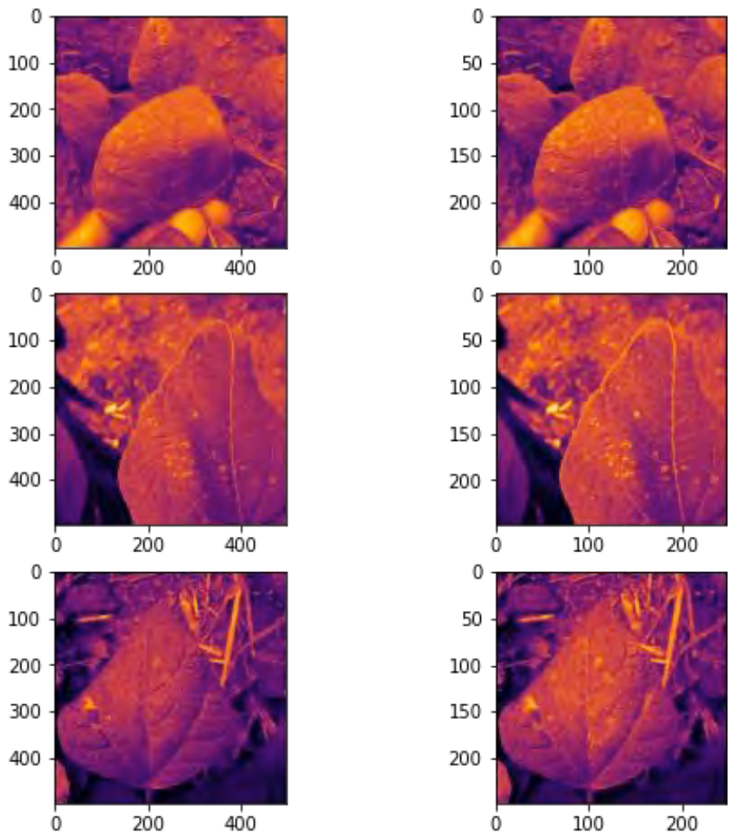


Adversarially Regularized Model



Results Analysis

Baseline model appears to have identified the relevant distinguishing features that define each class:



The model struggles to represent the raw pixels in each image. The representative features are considerably more abstract due to the perturbation: