Inverse Problems Symposium 2025

Name: Boyang Deng
Organization: MSU

Abstract Title: Enhanced Robustness via Adversarial Training for Weed Classification

Authors: Boyang Deng, Yuzhen Lu

Enhanced Robustness via Adversarial Training for Weed Classification

Boyang Deng, Yuzhen Lu*
Department of Biosystems & Agricultural Engineering, Michigan State University, East Lansing, MI 48824
*Correspondence: luyuzhen@msu.edu

Abstract

Weeds pose a significant biological challenge to global agricultural productivity. While AI-based weed recognition systems show promise, they often struggle with the differentiation of plant species in real-world conditions. Previous studies have largely overlooked the analysis and interpretability of feature representations learned in crop and weed classification. Addressing the interpretability gap presents an inverse problem that involves estimating and analyzing model parameters by inferring meaningful input patterns from internal features. Adversarial training, an extension of standard training, promotes the development of robust and human-aligned feature representations. This study investigates the use of adversarially robust optimization to enhance feature learning in weed classification. Standard and adversarially trained ResNet-50 classification models are evaluated on a 10-class weed dataset. As a result, although adversarial training lowers top 1 accuracy on standard test set, from 92.86% to 82.49%, it substantially enhances model robustness under adversarial perturbed data, rising from 0.05% to 68.54%. Additionally, robust models enable semantic visualization in terms of image inversion, interpolation, and direct feature visualization without priors or post-processing, yielding semantically meaningful and interpretable outputs. These results highlight the potential of adversarial robustness in enhancing AI model transparency.

Keywords: Deep learning, Weed classification, Adversarial robustness, Interpretability.