Benchmarking Adversarial Fairness in Image Classification Models

TEAM PROJECT

SUPERVISION : MARGRET KEUPER AND TEJASWINI MEDI

DEPARTMENT : MACHINE LEARNING

Motivation

Deep Neural Networks (DNNs) have been successful in computer vision—classification, segmentation, detection.

But why not in safety-critical applications?

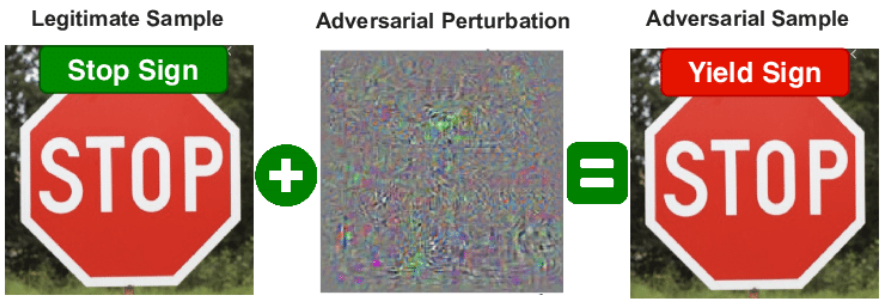
- Lack of robustness to perturbations, domain shifts, corruptions.
- We measure overall robustness but ignore class-specific fairness.

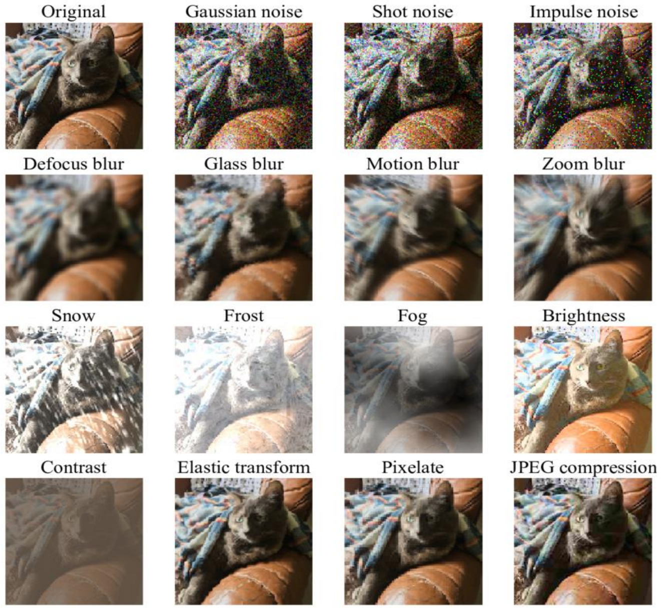
To build reliable AI, we must address both.

Robustness

The ability of a model to maintain performance when faced with adversarial inputs or attacks

1. Adversarial Attacks
2. Common Corruptions
3. Affine Transformations



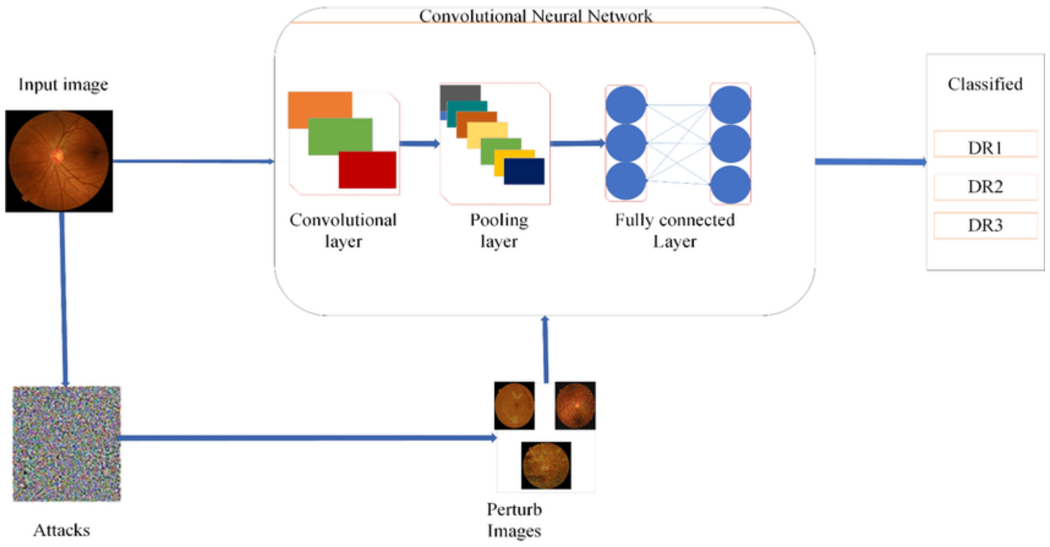


Adversarial Training (AT)

Adversarial Training involves training the model with clean and perturbed images.

Goal is to make model robust towards these perturbations.

Many works focus on improving the training recipe of adversarial training.



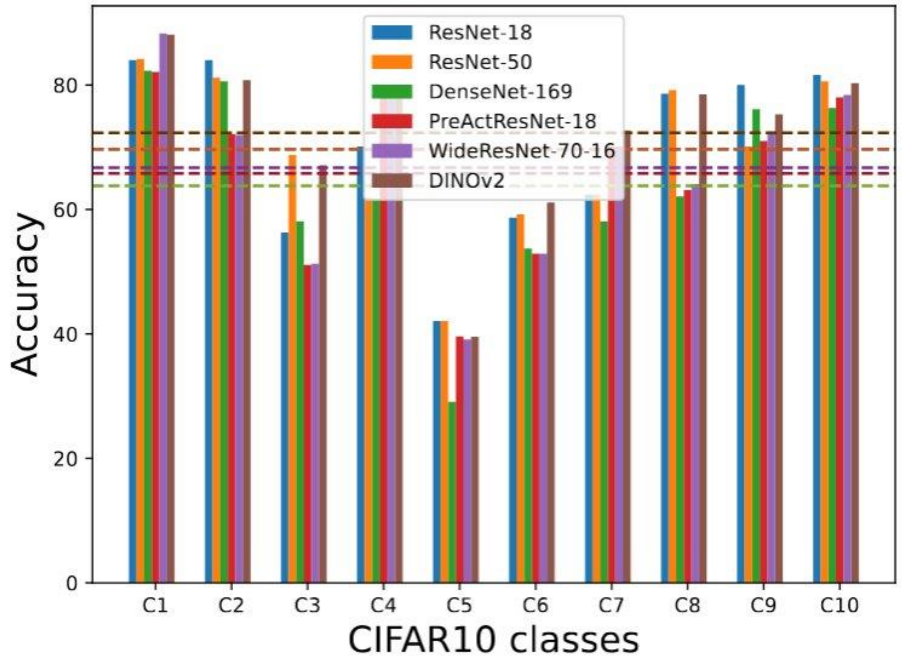
Fairness

Adversarial Training (AT) enhances overall model robustness.

Class-wise Disparity: Strong classes remain robust, while weaker ones suffer.

This disparity arises from inherent biases in representations and the classifier model, affecting harder-to-classify classes.

This **Fairness** issue become very important to explore to build reliable ML models.



TEAM PROJECT PLAN

Robustness and Fairness analysis on adversarial trained classification models

Benchmarking the works which focuses on improving the adversarial fairness

Analyzing the factors contributing to class-wise disparity performance

Key Takeaways & Requirements

- **Robustness & Fairness in ML:** Understanding challenges in building reliable and unbiased models.
- **Adversarial Training Research:** Insights into techniques for defending models against adversarial attacks.
- **Hands-on Model Training:** Practical experience with diverse training strategies in AT.
- **Teamwork & Project Management:** Collaborating effectively.
- Requirements for the Project : Programming in python and basic usage of pytorch, deep learning or high-level computer vision lecture.

Thank you!

Contact : tejaswini.medi@uni-mannheim.de