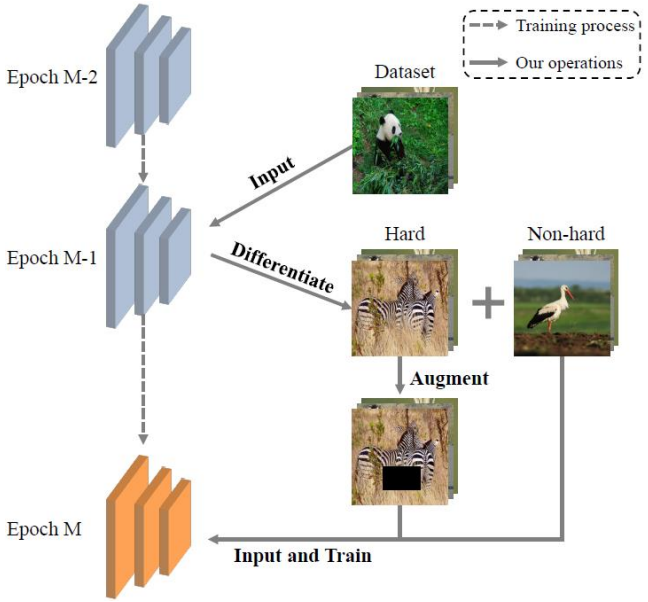
Fairness is Essential for Robustness: Fair Adversarial Training by Identifying and Augmenting Hard Examples

Ningping MOU, Xinli YUE, Lingchen ZHAO, Qian WANG

Frontiers of Computer Science, DOI: 10.1007/s11704-024-3587-1

Problems & Ideas

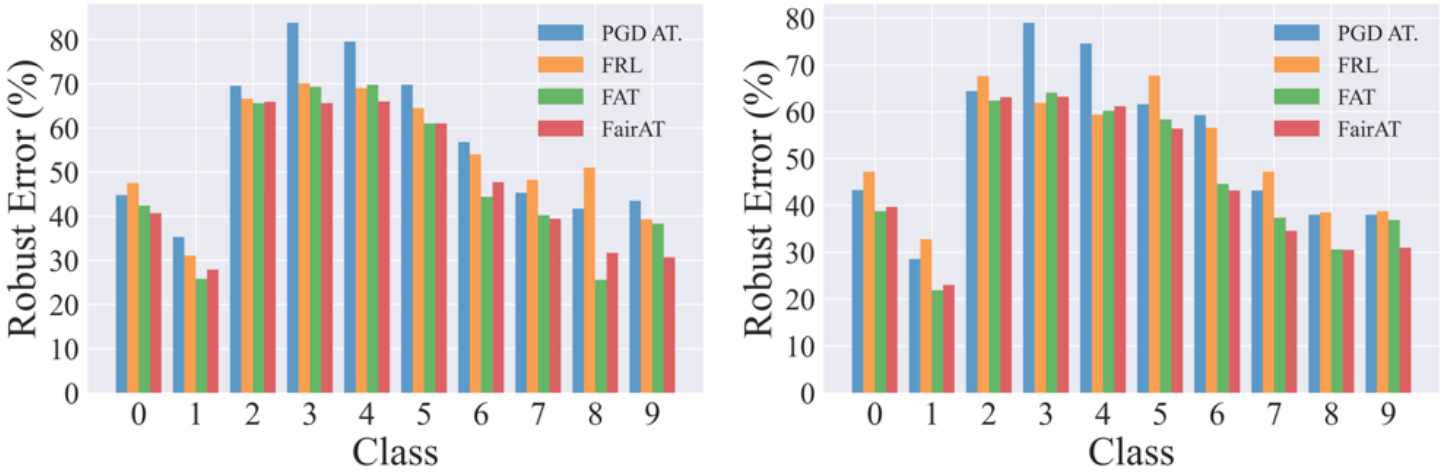
- Problems of adversarial training:
  - A large discrepancy exists in the class-wise robustness of adversarial training (AT), known as robust fairness.
  - The discrepancy may undermine the security of AT and lead to ethical issues, hindering the application of AT.
- Ideas: A fair adversarial training method that dynamically identifies hard examples and augments them to improve the robust fairness.



Training the model of epoch M for FairAT. It firstly inputs the original training dataset into the model of epoch M-1 to calculate the cross-entropy values and differentiate hard examples from non-hard examples. After augmenting hard examples, it combines augmented examples and non-hard examples as the training dataset to train the model of epoch M.

Main Contributions

- Contributions:
  - A novel discovery that the hard examples of higher uncertainty might be a more fine-grained indicator of robust fairness than previous class-level metrics;
  - A fair adversarial training method (FairAT) that dynamically augments hard examples and makes the model more focused on them;
  - Extensive experiments demonstrate that FairAT outperforms state-of-the-art methods in terms of both overall robustness and fairness.



Comparison of different adversarial training algorithms with regard to class-wise robust errors. FairAT significantly reduces the robust errors of hard classes whose robust errors are high.