Review: On the Tradeoff Between Robustness and Fairness (NeurIPS, 2022)

Dongyoon Yang

Seoul National University

January 30, 2023

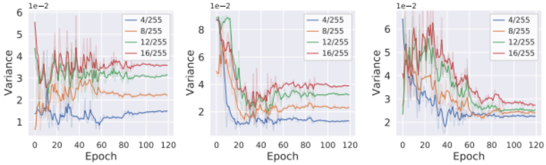
Introduction

- A robust model well-trained by AT exhibits a remarkable disparity of standard accuracy and robust accuracy among different classes compared with natural training.
- Is there a tradeoff between average robustness and robust fairness; specifically, as the perturbation radius increases, will stronger adversarially trained models lead to a larger class-wise disparity of robust accuracy among different classes?

Contribution

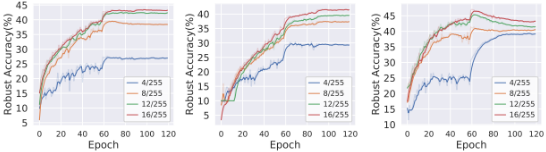
- Authors empirically find the relation between the variance of class-wise robust accuracy and perturbation radius in AT.
- Authors theoretically analyze this new phenomenon above and provide a potential explanation for it through linear model with mixture Gaussian distribution.
- Authors propose FAT to mitigate the tradeoff between robustness and fairness.

Observation



(a) ResNet18 (b) ResNet50 (c) WRN-32-10

Figure 1: The variance of class-wise robust accuracy for Madry using ResNet18, ResNet50 and WRN-32-10 on CIFAR-10. The perturbation radii for AT are chosen from $\epsilon_{train} = \{4/255, 8/255, 12/255, 16/255\}$. The adversarial testing examples are generated by FGSM with testing perturbation radius $\epsilon_{test} = 16/255$.



(a) ResNet18 (b) ResNet50 (c) WRN-32-10

Figure 2: The average robust accuracy for Madry using ResNet18, ResNet50 and WRN-32-10 on CIFAR-10. The perturbation radii for AT are chosen from $\epsilon_{train} = \{4/255, 8/255, 12/255, 16/255\}$. The adversarial testing examples are generated by FGSM with testing perturbation radius $\epsilon_{test} = 16/255$.

Definition 5.1. (Mixture Gaussian Distribution). Let $\mu_+, \mu_- > 0$ be the per-class mean parameter and $\sigma_+, \sigma_- > 0$ be variance parameter of two classes. The $(\mu_+, \mu_-, \sigma_+, \sigma_-)$-Gaussian mixture distribution $\mathcal{D}^*$ can be then defined by the following distribution over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$:

$$y = \begin{cases} +1, & p = \alpha \\ -1, & p = 1 - \alpha, \end{cases} \quad x \sim \begin{cases} \mathcal{N}\left(\boldsymbol{\mu}_+, \sigma_+^2 I\right) & \text{if } y = +1 \\ \mathcal{N}\left(-\boldsymbol{\mu}_-, \sigma_-^2 I\right) & \text{if } y = -1 \end{cases} \tag{1}$$

where α is the prior probability of class "+1" and $\boldsymbol{\mu}_+ = \mu_+ \mathbf{1}$, $\boldsymbol{\mu}_- = \mu_- \mathbf{1}$, $\mathbf{1} = (\overbrace{1, \ldots, 1}^{\text{dim d}})'$, I is a d- dimension identity matrix.

- $(X, Y) \sim \mathcal{D}^*$ and $f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$

Main Theorem

$$f_{\text{adv}} = \underset{f}{\text{argmin}}\ \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})} \max_{\boldsymbol{X}' \in \mathcal{B}_p(\boldsymbol{X},\epsilon)} 1(f(\boldsymbol{X}') \neq \boldsymbol{Y}) \tag{1}$$

$$\text{VCRA}(f) = \frac{1}{C} \sum_{c=1}^{C} (p_{\text{adv}}(c) - \bar{p}_{\text{adv}})^2 \tag{2}$$

where $p_{\text{adv}}(c) = 1 - \mathbb{E}_{(\boldsymbol{X}|Y=c)}\{\max_{\boldsymbol{X}' \in \mathcal{B}_p(\boldsymbol{X},\varepsilon)} 1(f(\boldsymbol{x}') \neq c)|Y = c\}$ and

$\bar{p}_{\text{adv}} = \frac{1}{C} \sum_{c=1}^{C} p_{\text{adv}}(c)$

Theorem

Given an adversarially trained linear model f_{adv} in Equation (1), the variance of class-wise robust accuracy $\text{VCRA}(f_{adv})$ is increasing with respect to ε_{train}.

Algorithm

Theorem

Under appropriate conditions on the loss $\ell(\cdot)$, parameter space Θ, with probability of at least $1-\delta$, the following holds for all $\boldsymbol{\theta} \in \Theta$:

$$\mathcal{R}_{adv}(f) \leq \widehat{\mathcal{R}}_{adv}(f) + \sqrt{\frac{\text{VCAR(f)}}{n \cdot \delta}} + \frac{C}{n} \tag{3}$$

where $\ell(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}_i), y_i)$ is empirical risk of the robust risk,

$\hat{R}_{adv}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}_i), y_i\right)$, $\text{VCAR}(f) = \frac{1}{C}\sum_{c=1}^{C}\left(R_{adv}(f,c) - \bar{R}_{adv}(f)\right)^2$,

$R_{adv}(f,c) = \mathbb{E}_{\boldsymbol{x}|y=c} \max_{x' \in \mathcal{B}_p(\boldsymbol{x},\epsilon)} \ell\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}'), y\right)$

Algorithm

Motivated from Theorem, authors proposes the Fairly Adversarial Training (FAT) which minimizes the following empirical risk:

$$\widehat{\mathcal{R}}_{\text{adv}}(f) + \lambda \widehat{\text{VCAR}}(f) \tag{4}$$

$$:= \sum_{i=1}^{n} \left\{ \ell(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}_i), y_i) + \lambda \frac{1}{C} \sum_{c=1}^{C} \left(\widehat{\mathcal{R}}_{\text{adv}}(f, c) - \bar{\widehat{\mathcal{R}}}_{\text{adv}}(f) \right) \right\} \tag{5}$$

where $\widehat{\boldsymbol{x}}$ is an adversarial example and
$\widehat{\mathcal{R}}_{\text{adv}}(f, c) = \frac{1}{\sum_{i=1}^{n} 1(y_i = c)} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}_i), y_i) 1(y_i = c)$.

Summary

- Authors empirically find the relation between the variance of class-wise robust accuracy and perturbation radius in AT.
- Authors theoretically analyze this new phenomenon above and provide a potential explanation for it through linear model with mixture Gaussian distribution.
- Authors propose FAT to mitigate the tradeoff between robustness and fairness.