Journal of Computational Mathematics
Vol.39, No.6, 2021, 880–896.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2105-m2020-0310

ACHIEVING ADVERSARIAL ROBUSTNESS REQUIRES AN ACTIVE TEACHER*

Chao Ma[1)] and Lexing Ying
Department of Mathematics, Stanford University, California, USA
Email: chaoma@stanford.edu, lexing@stanford.edu

Abstract

A new understanding of adversarial examples and adversarial robustness is proposed by decoupling the data generator and the label generator (which we call the teacher). In our framework, adversarial robustness is a conditional concept—the student model is not absolutely robust, but robust with respect to the teacher. Based on the new understanding, we claim that adversarial examples exist because the student cannot obtain sufficient information of the teacher from the training data. Various ways of achieving robustness is compared. Theoretical and numerical evidence shows that to efficiently attain robustness, a teacher that actively provides its information to the student may be necessary.

Mathematics subject classification: 68T07, 68T99.
Key words: Adversarial robustness, Decoupled supervised learning, Active teacher.

1. Introduction

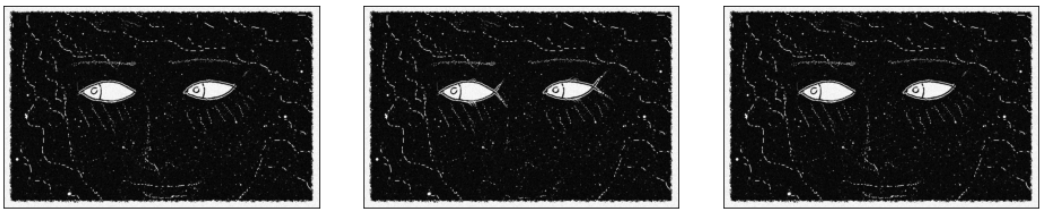
The existence of adversarial examples restricts the application of deep learning in many fields with high demand on the robustness and security, such as autonomous driving and health care. Hence, improving adversarial robustness of deep neural networks has experienced extensive study, both theoretically and practically [1, 15]. Originally, adversarial examples are found to be perturbed images whose perturbations are imperceptible to humans but cause huge error to the neural networks [2,37]. In most existing works, however, adversarial robustness is defined as robustness with respect to perturbations measured by the l_p distance (e.g. [12,37]). Specifically, a model $f_\theta(\cdot)$ is considered to be robust if the adversarial loss

$$L_{\text{adv}}(f_\theta) = \mathbb{E}_{(\mathbf{x},y)} \max_{\|\delta\|_p \le \epsilon} l(f_\theta(\mathbf{x}+\delta), y) \tag{1.1}$$

is small, where ϵ is a pre-defined value and l is some loss function [25]. This simplification helps analysis and implementation. In spite of this, the robustness with small l_p perturbations is very different from the robustness with respect to human-imperceptible perturbations [32]. A human-imperceptible perturbation may not have small l_p norm [5,46], and a perturbation with small l_p norm may also not necessarily be imperceptible to humans [35]. In Figure 1.1, inspired by optical illusions, we show an example of difference between some l_p distances and human perception. This difference makes current "adversarially robust" models easily broken by newly-designed attacks. Besides l_p distances, other measures, such as Wasserstein distance [43] and structural similarity (SSIM) [41], are also shown to be different from human perception [32].

* Received March 23, 2021 / Revised version received May 5, 2021 / Accepted May 7, 2021 / Published online October 12, 2021 /
1) Corresponding author

In this paper, we propose a conditional explanation of adversarial robustness, which highlights the role of human labeler in defining the adversarial examples. Specifically, we decouple the data generator with the labeler, and make two definitions: the teacher is an object or a mechanism that assigns true labels to data points, and the student is a machine learning model used to learn from the data and labels. Within our framework, adversarial robustness is not a universal concept defined unconditionally for any learning problem (like l_p robustness), but rather a relative concept conditioned on a certain teacher. The teacher is usually human, but can also be other objects such as physical processes or neural networks. A student model is said to be (strongly) adversarially robust with respect to a teacher if it can correctly classify any data the teacher can classify with certainty. This is possible because in our framework the teacher has an "uncertain set", and it does not assign labels to data within this set. Hence a robust student model does not need to have the same decision boundary as the teacher. A weaker version of adversarial robustness is also defined by considering the data produced by an "attack", instead of all the data that the teacher can classify. This weak definition of adversarial robustness can cover the l_p robustness, but in a more proper way. We show that our definitions of adversarial robustness are not equivalent with the l_p robustness by simple illustrative examples—l_p robust classifier may not be adversarially robust, vice versa.



(a) Face (b) Fish (c) Face

Fig. 1.1. Difference between human perception and l_2 distances illustrated by an optical illusion. **(a)** The images looks like a face; **(b)** The image looks like two fishes; **(c)** The image in (a) adding a noise. The l_0, l_1 and l_2 distances between (a) and (b) are 15037, 2534.44 and 43.02, respectively. The l_0, l_1 and l_2 distances between (a) and (c) are 812311, 63413.43 and 89.17, respectively. Though the images in (a) and (c) are perceptually the same, their l_p distances are greater than the distances between (a) and (b), which are perceptually different. (The original image is taken from https://pixabay.com/illustrations/fairy-tale-fish-portrait-1077859/)

Based on this new understanding, we point out two reasons that cause adversarial examples: (1) Some features the student uses to make classification are imperceptible to the teacher. (2) The training data do not provide sufficient information of the classification mechanism of the teacher, e.g. which feature the teacher uses to make classification. Combining the two reasons above, we argue that the adversarial examples are caused by insufficient (out-of-distribution) information of the teacher provided by the training data. Without necessary information, the student model cannot select the robust solution among many solutions that perform well on the original data distribution. Therefore, to achieve adversarial robustness, or at least alleviate adversarial vulnerability, more teacher information should be provided to the student model. This can be achieved in two ways:

1. **An active student:** The student model asks information from the teacher, and the teacher passively answers the student's questions, and does not provide extra information.