ROBUSTNESS MAY BE AT ODDS WITH ACCURACY

Reproduced by: Shohaib Mahmud, Zhiming Fan, Zetian Liu, Jiechao Gao

Tsipras, Dimitris, et al. "Robustness may be at odds with accuracy." ICLR 19.
Contribution of group members

We equally distribute all the work: coding, slides making, etc.

All members are highly involved in this project.
Background

Adversarial examples has garnered significant attention recently and resulted in a number of approaches both to finding these perturbations, and to training models that are robust to them. (Goodfellow et al., 2014b; Nguyen et al., 2015; Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2016; Sharif et al., 2016; Kurakin et al., 2016a; Evtimov et al., 2017; Athalye et al., 2017)

However, building such adversarially robust models has proved to be quite challenging. In particular, many of the proposed robust training methods were subsequently shown to be ineffective. Only recently, has there been progress towards models that achieve robustness that can be demonstrated empirically and, in some cases, even formally verified. (Madry et al., 2017; Kolter & Wong, 2017; Sinha et al., 2017; Tjeng & Tedrake, 2017; Raghunathan et al., 2018; Dvijotham et al., 2018a; Xiao et al., 2018b)
Motivation

The paradigm of adversarially robust learning is different from the classic learning setting.

In particular, we know that robustness comes at a cost. For example:
1. Computationally expensive training methods (more training time).
2. The potential need for more training data.

Questions:
Are these the only costs of adversarial robustness? And, if so, once we choose to pay these costs, would it always be preferable to have a robust model instead of a standard one?

Goals:
The goal of this work is to explore these questions and thus, in turn, to bring us closer to understanding the phenomenon of adversarial robustness.
Related Work

F. Fawzi et al. (2018) prove upper bounds on the robust of classifiers and exhibit a standard vs. robust accuracy trade-off for a specific classifier families on a synthetic task.

Ross & Doshi-Velez (2017) propose regularizing the gradient of the classifier with respect to its input. They find that the resulting classifiers have more interpretable gradients and targeted adversarial examples resemble the target class for digit and character recognition tasks.

Wang et al. (2017) analyze the adversarial robustness of nearest neighbor classifiers. Schmidt et al. (2018) study the generalization aspect of adversarially robustness. Gilmer et al. (2018) demonstrate a setting where even a small amount of standard error implies that most points provably have a misclassified point close to them.
What is the relationship between standard and adversarially robust accuracy?

What are the properties of the standard training and adversarial training?
An Intuitive Figure Showing WHY Claim

Epsilon - degree of adversarial training

Adversarial training:
1. strengthen generalization
2. lower standard accuracy
Proposed Solution & Implementation

1. robust features align well with human perception
   Visualize Features that affect classifier most
   Output the loss gradient with respect to input pixels

2. Adversarial examples exhibit salient data characteristics
   Visualize adversarial examples

3. Smooth cross-class interpolations
   Visualize the adversarial examples over training epochs
Data Summary

Images from

1. MNIST
2. CIFAR-10
3. ImageNet
Figure 2: Visualization of the loss gradient with respect to input pixels
Figure 3: Visualizing large-ε adversarial examples for standard and robust (l-2/l-infinity adversarial training) models.
Experimental Results in the paper

Figure 4: Interpolation between original image and large-ε adversarial example
Figure 2: Standard Accuracy and Robust Accuracy Comparison
Reproduce Experimental Results - MNIST

**Figure 3a:** Visualization of Gradient with Respect to Input

**Figure 3b:** Visualization of Adversarial Example
Reproduce Experimental Results - CIFAR10

Figure 4: Standard Accuracy and Robust Accuracy Comparison
Figure 3a: Visualization of Gradient with Respect to Input

Figure 3b: Visualization of Adversarial Example
Reproduce Experiments - IMAGENET

<table>
<thead>
<tr>
<th>ori</th>
<th>standard</th>
<th>l2</th>
<th>linf</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="ori" /></td>
<td><img src="image2" alt="standard" /></td>
<td><img src="image3" alt="l2" /></td>
<td><img src="image4" alt="linf" /></td>
</tr>
</tbody>
</table>

Visualization of loss gradient with respect to the input image
Reproduce Experiments - IMAGENET

ori standard linf l2

Visualization of adversarial examples
Reproduce Experiments - IMAGENET

Smooth cross-class interpolation
Conclusion and Future Work

- ML model has an intrinsic tension between robust accuracy and standard accuracy
- Theoretical bounds on the accuracies depend on the correlation of features
- Robust model emphasizes different features compared to a standard one
- Future work can explore the connection between GANs and adversarial robustness
References


