UVA CS 6316: Machine Learning : 2019 Fall
Course Project: Deep2Reproduce @
https://github.com/qiyanjun/deep2reproduce/tree/master/2019Fall

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.

Claim / Target Task

What is the relationship between standard and adversarially robust accuracy?

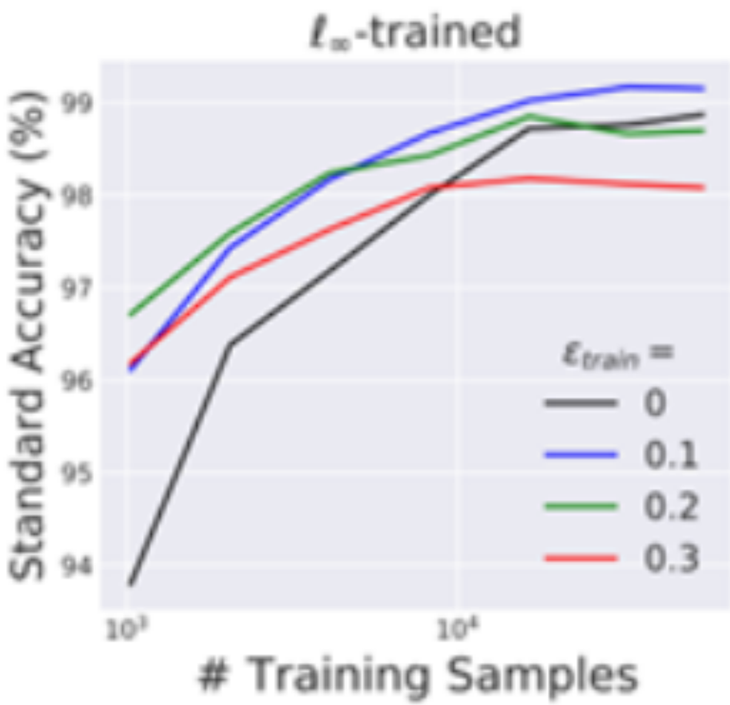
What are the properties of the standard training and adversarial training?

An Intuitive Figure Showing WHY Claim

Epsilon - degree of adversarial training

Adversarial training:

1. strengthen generalization
2. lower standard accuracy



(a) MNIST

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

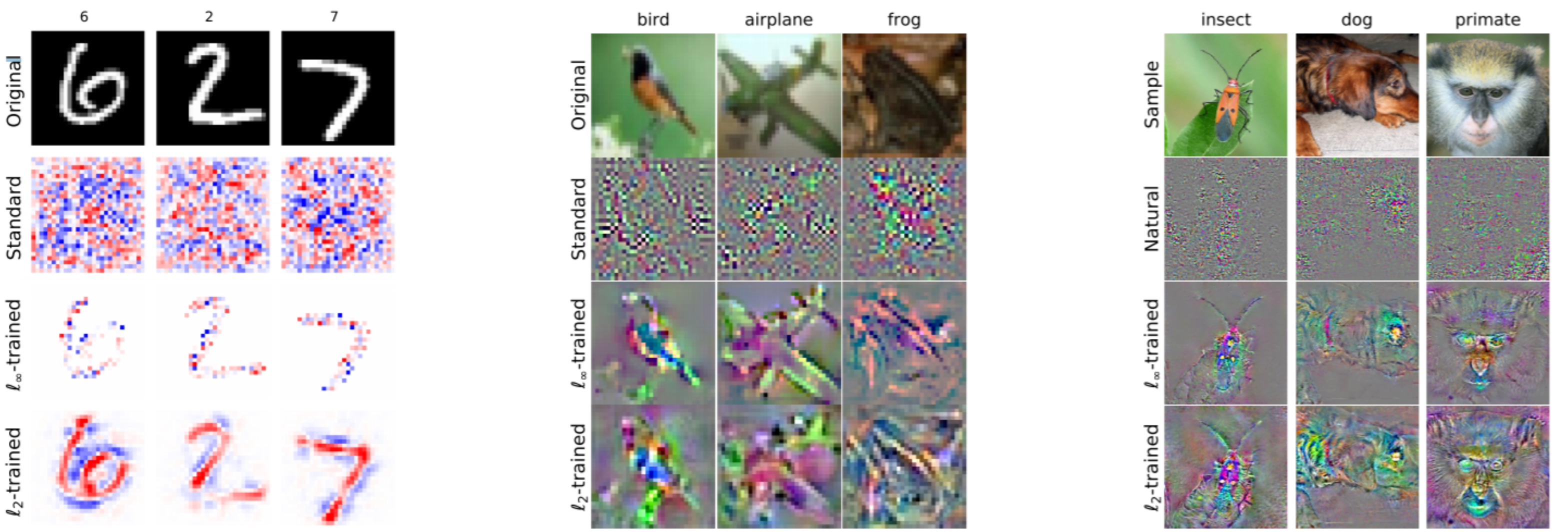
3. Smooth cross-class interpolations
   Visualize the adversarial examples over training epochs

Data Summary

Images from

1. MNIST
2. CIFAR-10
3. ImageNet

Experimental Results in the paper



(a) MNIST (b) CIFAR-10 (c) Restricted ImageNet

Figure 2: Visualization of the loss gradient with respect to input pixels

Experimental Results in the paper

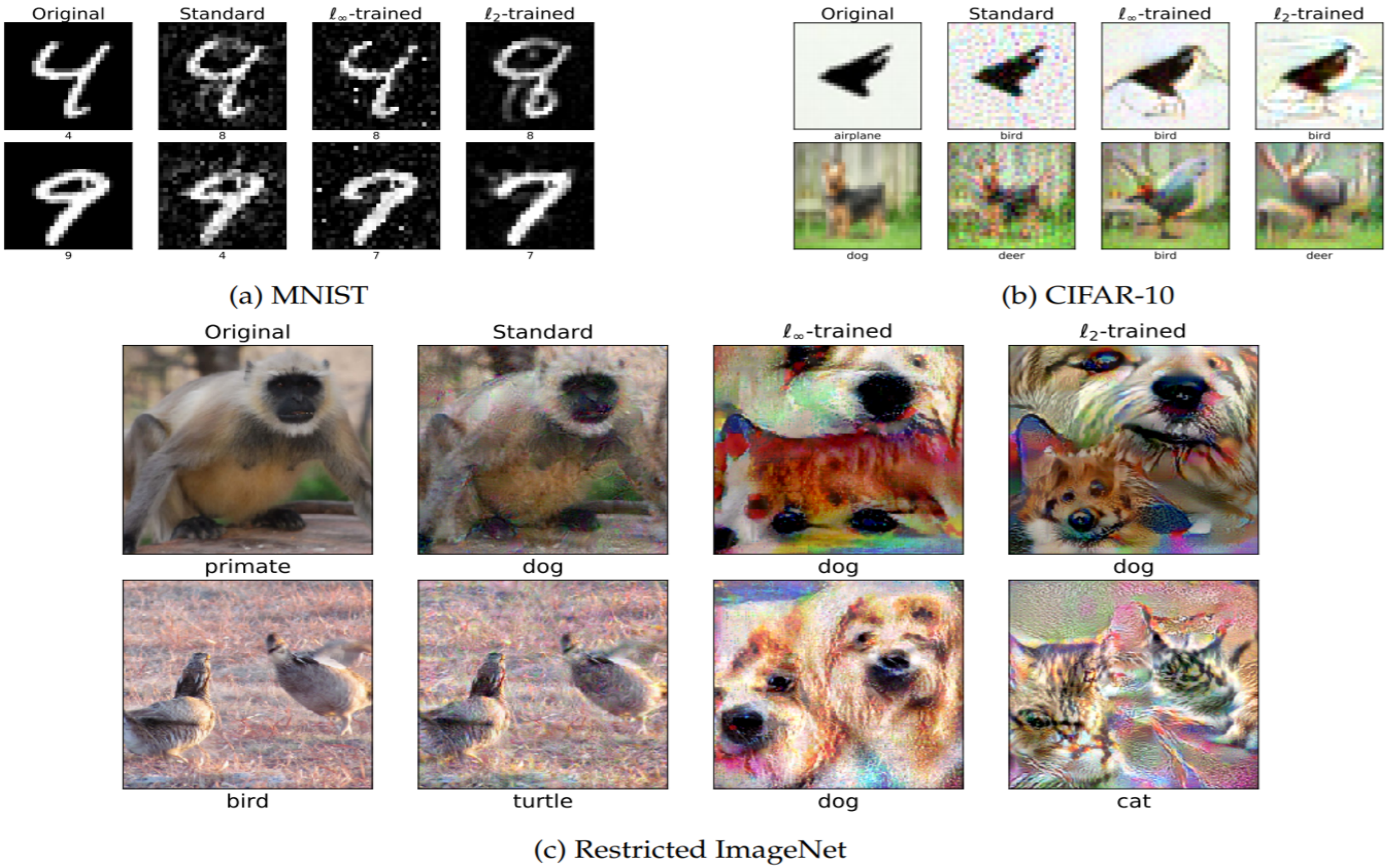


Figure 3: Visualizing large-ε adversarial examples for standard and robust (l-2/l-infinity adversarial training) models.

Experimental Results in the paper

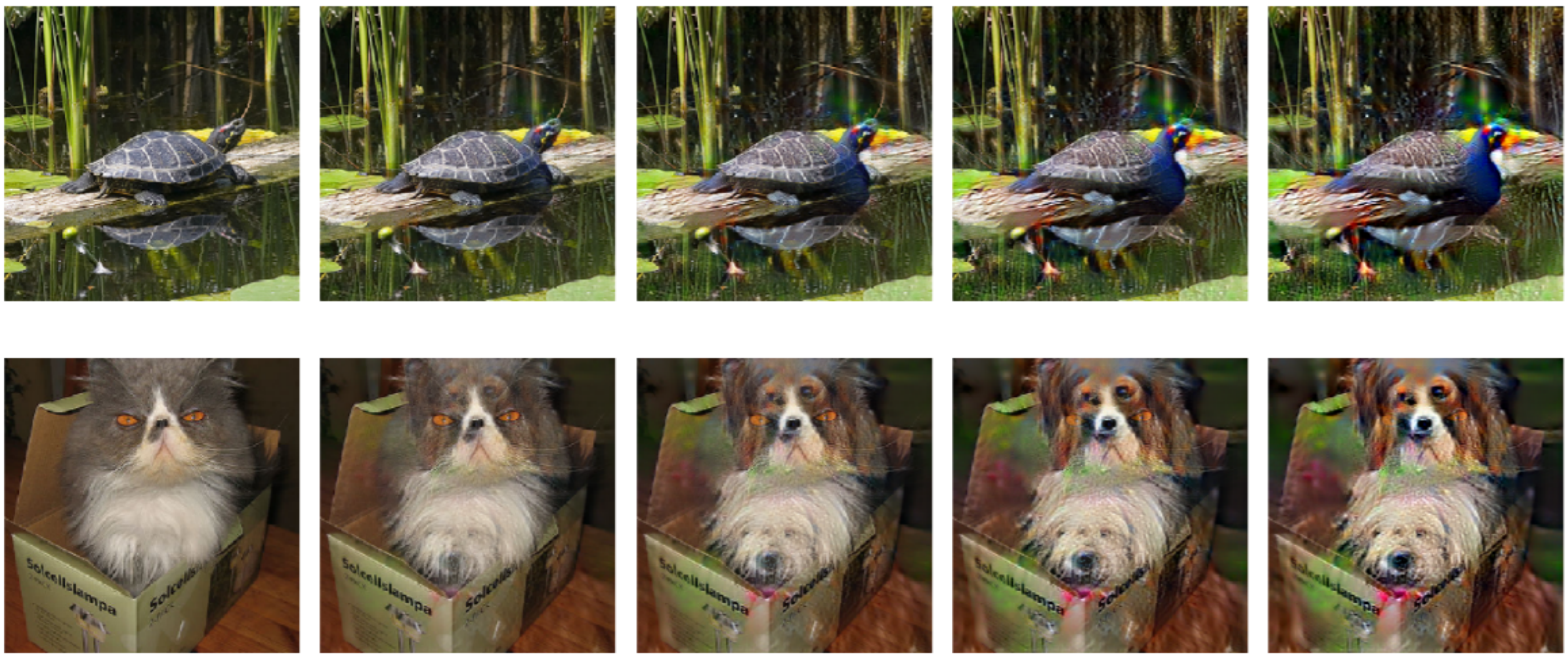


Figure 4: Interpolation between original image and large-ε adversarial example

Reproduce Experimental Results- MNIST

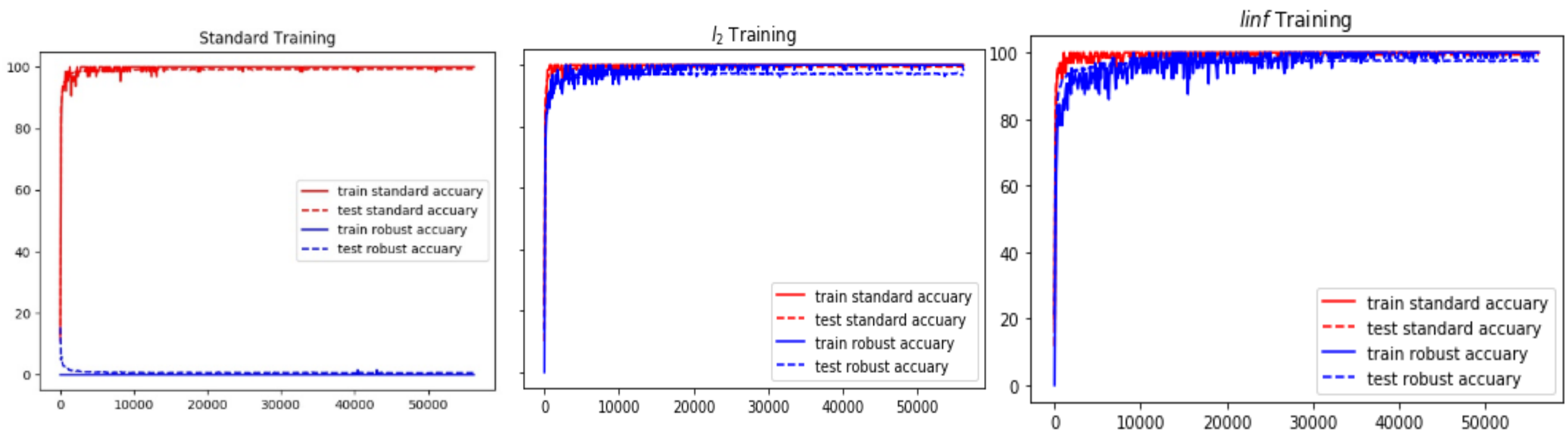


Figure 2: Standard Accuracy and Robust Accuracy Comparison

Reproduce Experimental Results- MNIST

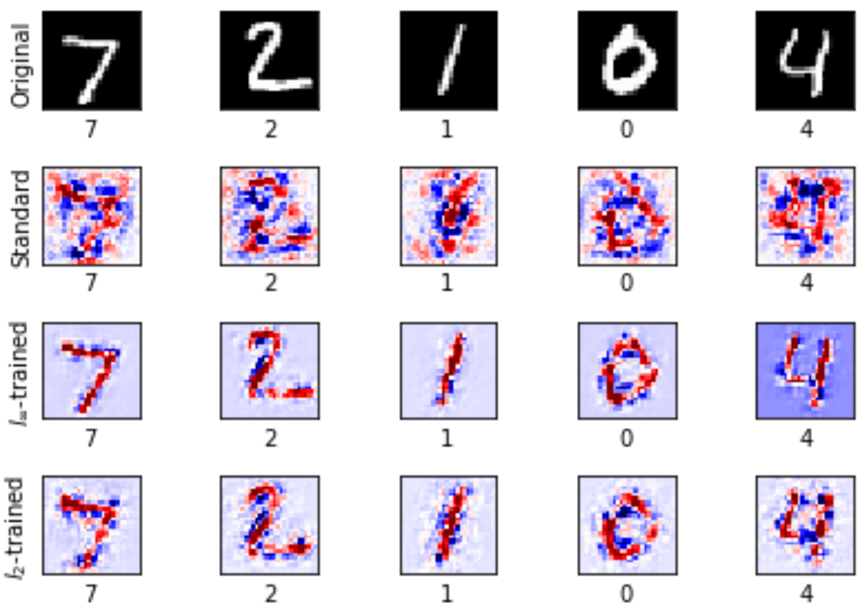


Figure 3a: Visualization of Gradient with Respect to Input

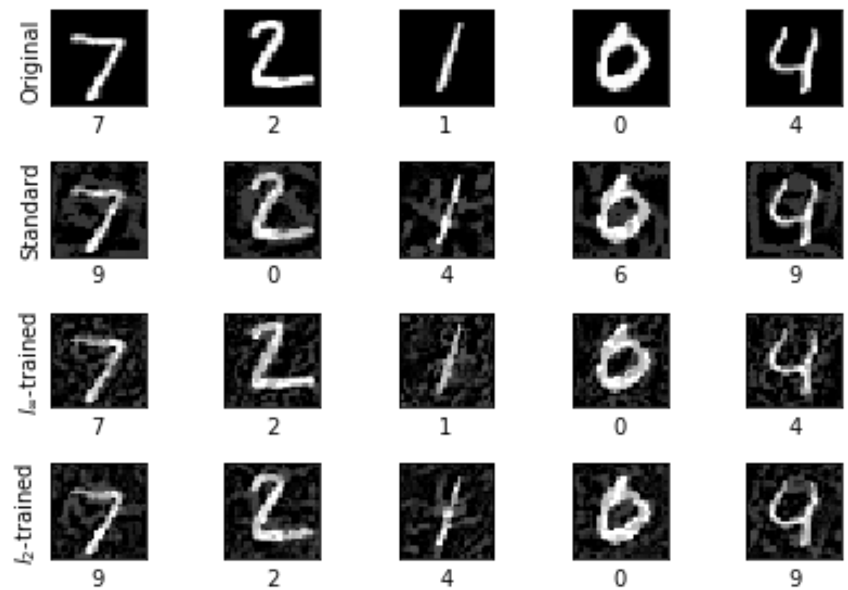


Figure 3b: Visualization of Adversarial Example

Reproduce Experimental Results - CIFAR10

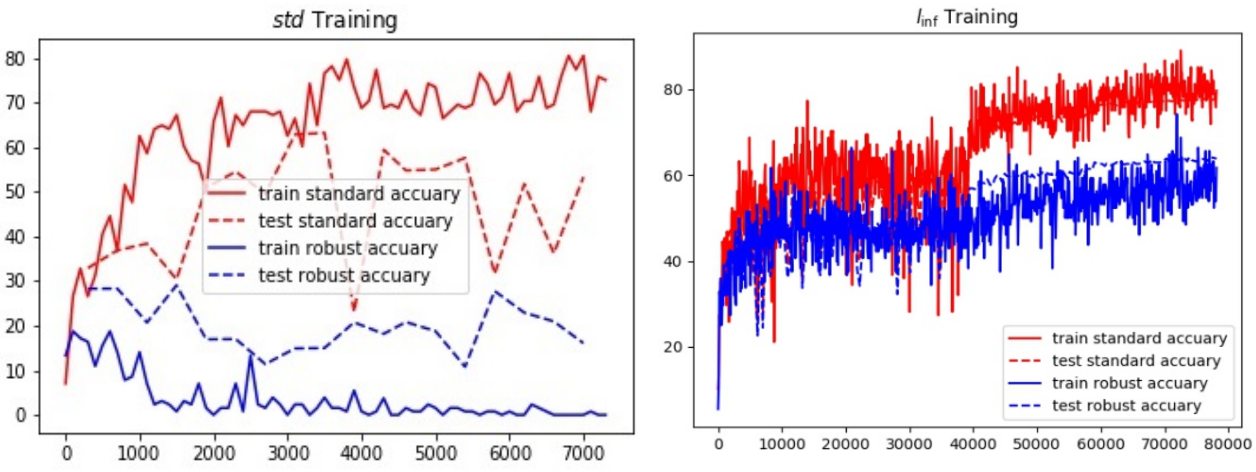


Figure 4: Standard Accuracy and Robust Accuracy Comparison

Reproduce Experimental Results - CIFAR10

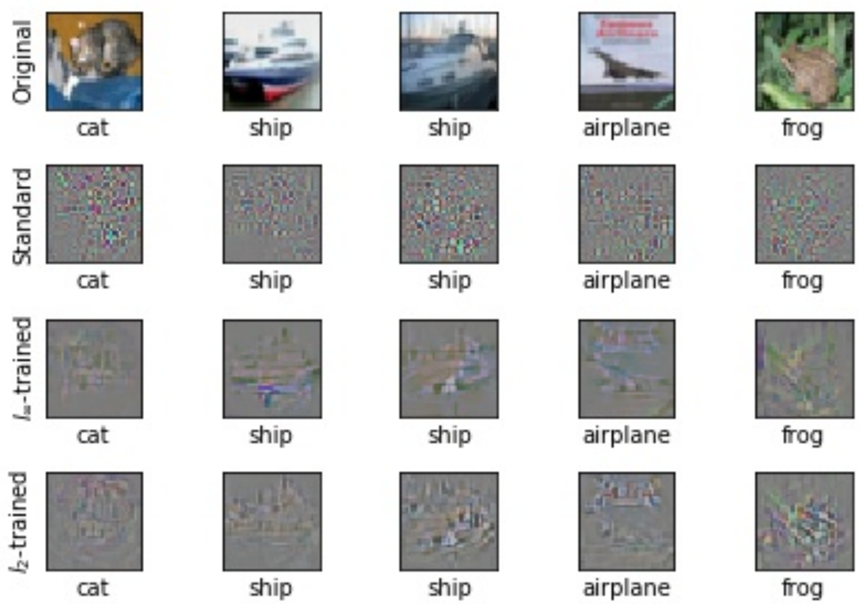


Figure 3a: Visualization of Gradient with Respect to Input

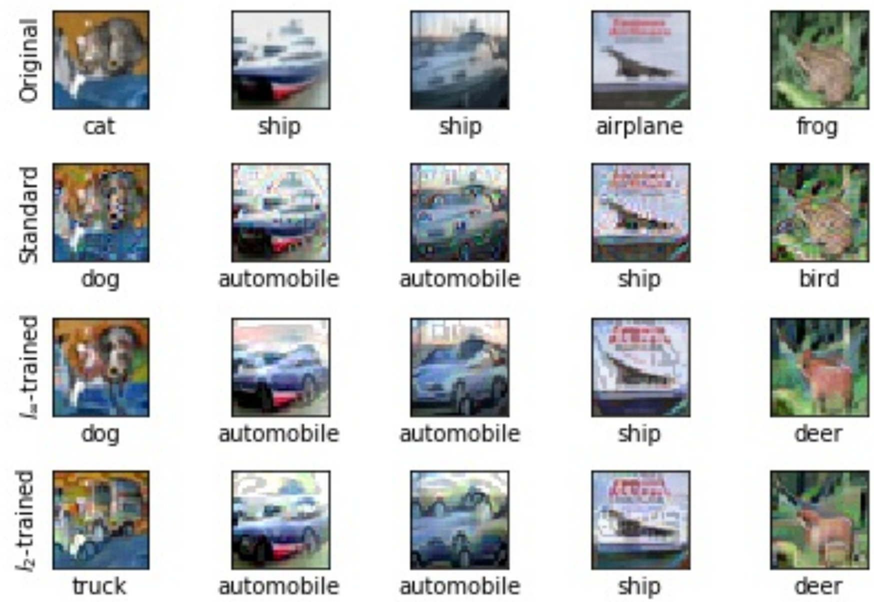
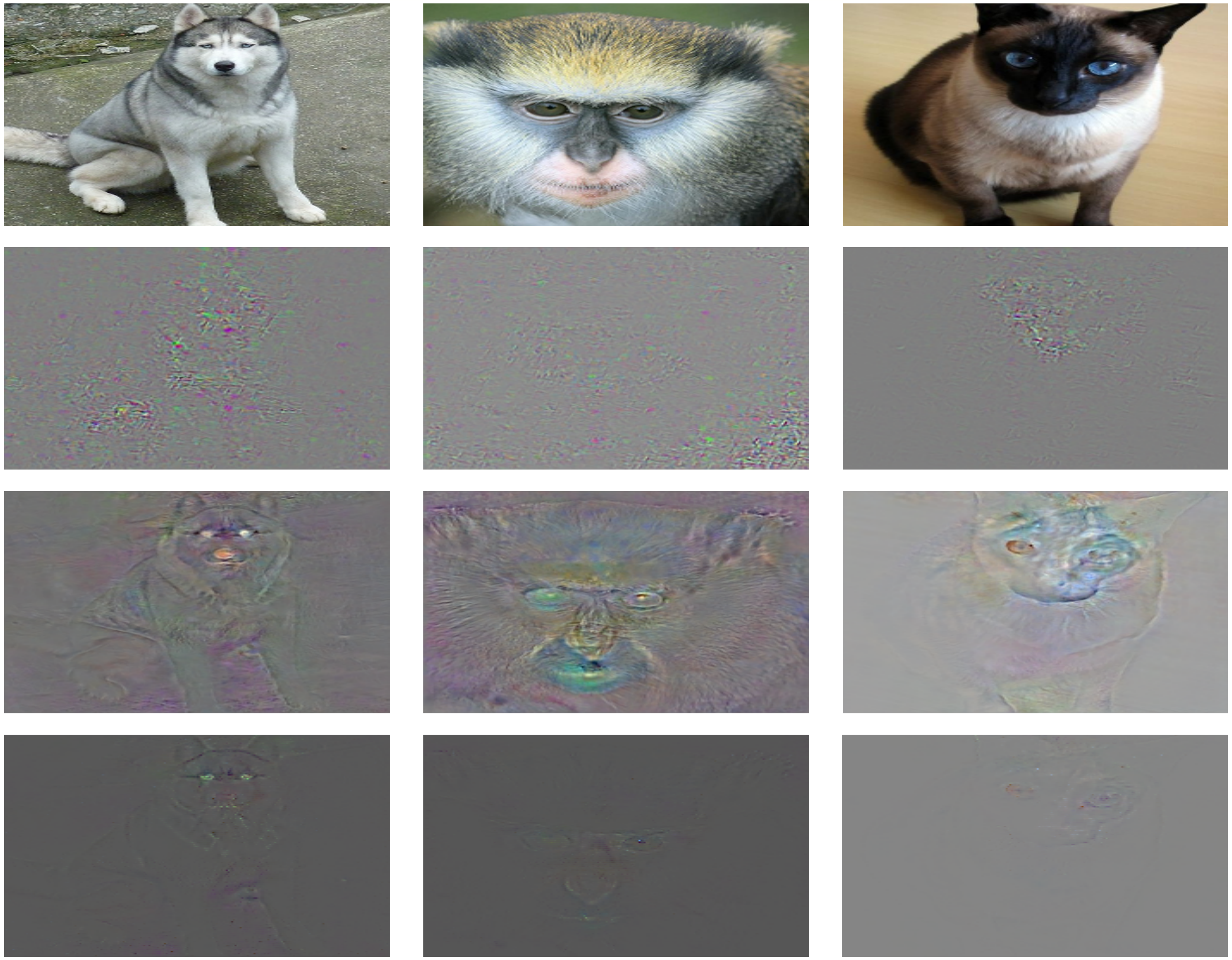


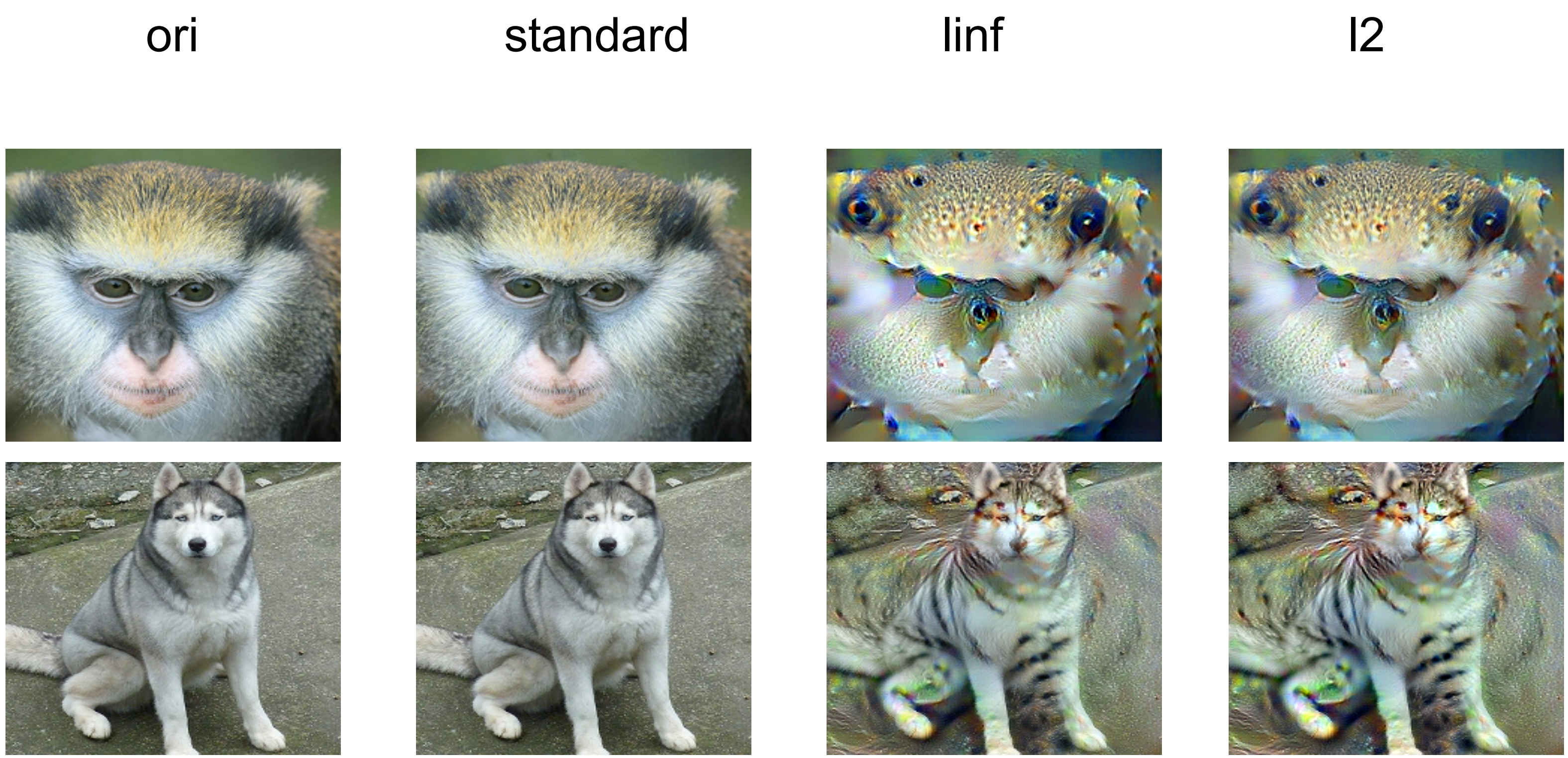
Figure 3b: Visualization of Adversarial Example

Reproduce Experiments - IMAGENET



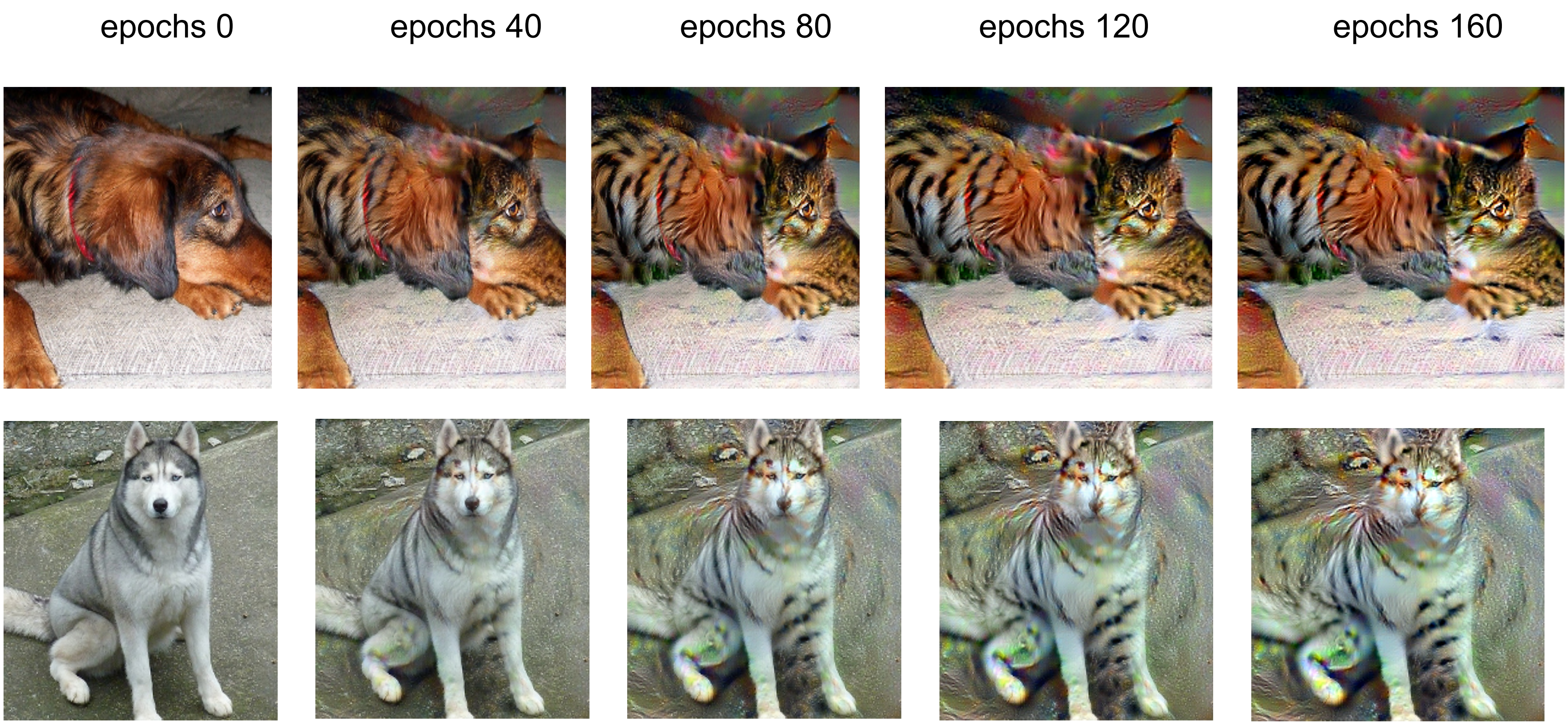
Visualization of loss gradient with respect to the input image

Reproduce Experiments - IMAGENET



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

[1] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.
[2] Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015
[3] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[4] Carlini, Nicholas, and David Wagner. "Defensive distillation is not robust to adversarial examples." *arXiv preprint arXiv:1607.04311* (2016).
[5] Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016.
[6] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016a.
[7] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and
[8] Dawn Song. Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945, 2017.
[9] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017.
[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.
[11] J Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. arXiv preprint arXiv:1711.00851, 2017.
[12] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571, 2017.
[13] Vincent Tjeng and Russ Tedrake. Verifying neural networks with mixed integer programming. arXiv preprint arXiv:1711.07356, 2017.
[14] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344, 2018.
[15] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. Training verified learners with learned verifiers. arXiv preprint arXiv:1805.10265, 2018a.
[16] Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. arXiv preprint arXiv:1809.03008, 2018b.

References (Contd.)

[17] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. Machine Learning, 107(3):481–508, 2018b.
[18] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. arXiv preprint arXiv:1711.09404, 2017.
[19] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. arXiv preprint arXiv:1706.03922, 2017.
[20] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. arXiv preprint arXiv:1801.02774, 2018.