Chaos Theory meets deep learning: On Lyapunov exponents and adversarial perturbations

Vinay Uday Prabhu, Nishant Desai, John Whaley
UnifyID Inc
San Francisco, CA
{vinay, nishant, john}@unify.id

Abstract

In this paper, we would like to disseminate a serendipitous discovery involving Lyapunov exponents of a 1-D time series and their use in serving as a filtering defense tool against a specific kind of deep adversarial perturbation. To this end, we use the state-of-the-art CleverHans library to generate adversarial perturbations against a standard Convolutional Neural Network (CNN) architecture trained on the MNIST dataset. We empirically demonstrate how the Lyapunov exponents computed on the flattened 1-D vector representations of the images served as highly discriminative features that could be to pre-classify images as adversarial or legitimate before feeding the image into the CNN for classification.

1. Background on defenses against adversarial attacks

In the recent past, a plethora of defenses against adversarial attacks have been proposed. These include *SafetyNet* [12], adversarial training [19], label smoothing [20], defensive distillation [17] and feature-squeezing [21, 22] to name a few. There is also an ongoing Kaggle contest[2] underway for exploring novel defenses against adversarial attacks.

As evinced by the recent spurt in the papers written on this topic, most defenses proposed are *quelled* by a novel attack that exploits some weakness in the defense. In [10], the authors queried if one could concoct a strong defense by combining multiple defenses and showed that an ensemble of weak defenses was not sufficient in providing strong defense against adversarial examples that they were able to craft.

With this background, we shall now look more closely at a specific type of defense and motivate the relevance of our method within this framework.

1.1. The pre-detector based defenses

One prominent approach that emerges in the literature of adversarial defenses is that of crafting pre-detection and filtering systems that flag inputs that might be potentially adversarial. In [9], the authors posit that adversarial examples are not drawn from the same distribution as the legitimate samples and can thus be detected using statistical tests. In [14, 7], the authors train a separate binary classifier to first classify any input image as legitimate or adversarial and then perform inference on the *passed* images. In approaches such as [6], the authors assume that DNNs classify accurately only near the small manifold of training data and that the synthetic adversarial samples do not lie on the data manifold. They apply dropout at test time to ascertain the confidence of adversariality of the input image.

In this paper, we would like to disseminate a model-agnostic approach towards adversarial defense that is dependent purely on the *quasi-time-series statistics* of the input images that was discovered in a rather serendipitous fashion. The goal is to not present the method we propose as a fool-proof adversarial filter, but to instead draw the attention of the DNN and CV communities towards this chance discovery that we feel is worthy of further inquiry.

2. Approach

2.1. A quick introduction to Lyapunov exponents

For a given a scalar time series $\{x_t; t = 1, ..., n\}$ whose time evolution is assumed to be modeled by a differentiable dynamical system (in a phase-space of possibly infinite dimensions), we can define Lyapunov exponents corresponding to the *large-time* behavior of the system. These *characteristic exponents* numerically quantify the sensitivity to the initial conditions and emanate from the Ergodic theory of differentiable dynamical systems, introduced by Eckmann and Ruelle in [5, 4]. Simply put, if the initial state of a time series is slightly perturbed, the characteristic exponent (or Lyapunov exponents) represent the exponential rate at

which the perturbation increases (or decreases) with time.

For a detailed understanding of these characteristic exponents, we would like to refer the user to [5, 4].

Given a finite-length finite-precision time series sequence, the 3 stage `Embed-Tangent maps-QR decomposition` based recipe introduced in [4] can be used to numerically compute the Lyapunov exponents. This numerical recipe is implemented in many time-series analysis packages such as [1].

The background of the serendipitous discovery was that we were investigating usage of metrics from time-series analysis literature to identify adversarial attacks on 1-D axis-wise mobile-phone motion sensor data ;example, Accelerometer and gyroscope (see [18]) and we realized that this procedure could be easily extended to vectorized images (or *flattened* images viewed as discrete time series indexed by the pixel location. That is, given a normalized image, $\mathbf{X} \in [0,1]^{n \times n}$, we would have the flattened time-series representation $x \in [0,1]^{n^2}$ as simply,

$$x_{in+j} = \mathbf{X}_{i,j};\ i,j = 0,...n-1; \tag{1}$$

The Lyapunov exponents of the MNIST images used in this work were numerically computed with the `lyap_e()` method implemented in [1] with the following parameterization:

```
PARAM_MNSIT={emb_dim=10, matrix_dim=4,
min_nb=min(2 * matrix_dim, matrix_dim + 4),
min_tsep=0, tau=1}
```

2.2. Supervised Classification with Lyapunov Features

After computing the Lyapunov exponents of the quasi-time-series representation of a set of MNIST images, we train a one-class Isolation-Forest classifier [11], and we test the defense by running the classifier against adversarial images generated using a variety of available attack algorithms.

We consider the Carlini-Wagner-$l2$ attack [3], the Fast Gradient Sign Method [8], the Jacobian Saliency Map Attack [16], DeepFool [15], and the attack presented by Madry et al [13]. We consider both the targeted and untargeted versions of each attack, where applicable[1]. We use the default parameters of the CleverHans library wherever possible. Where CleverHans does not provide a default value, we use the values referenced in the original paper describing the attack. The Attack Rejection Rates (ARR) on data generated by each of the attacks are presented in Table 2.2. We find that the JSMA method is the most difficult to defend against using this technique.

[1] DeepFool does not have a targeted variant.

Attack	Targeted ARR	Untargeted ARR
Carlini Wagner	0.91	1.0
FGSM	0.62	0.8
JSMA	0.04	0.04
Madry et al.	1.0	1.0
DeepFool	NA	1.0

Table 1. Attack rejection rates (ARR) for several targeted and untargeted attacks.

2.3. Testing the effect of a new attack

In previous sections, we used a one-class Isolation Forest classifier to learn an inlier set of unmodified images. In this experiment, we test whether a classifier trained on both positive and negative data generated using known adversarial attacks can outperform the 1-class classifier.

We train a logistic model on unmodified MNIST images and data generated using all but one of the untargeted attacks from the previous section. We then evaluate the model on a validation set consisting of natural images and images modified using the left-out attack. We find that the logistic model is able to achieve near-perfect performance on four out of five attacks. The model only fails to perform on data generated using JSMA, achieving an AUROC score of 0.61. On all other attacks, the model reaches an AUROC score between 0.97 and 1.0. The ROC curve for the Carlini-Wagner attack is shown as an example in Figure 1. Other attacks resulted in similar ROC curves. This finding is consistent with the result of our previous experiment, leading us to conclude that the JSMA attack is the most effective against this defense.

3. Conclusion and future work

Via this paper, we have sought to disseminate a serendipitous discovery entailing usage of Lyapunov exponents as a model-agnostic tool that can be used to pre-filter input images as potentially adversarially perturbed. We have shown the validity of the idea *defensing* against images that were adversarially perturbed using an array of attack procedures on MNIST data. Perhaps our most interesting finding is that our defense is robust to a wide variety of attacks but fails against the Jacobian Saliency Map Attack. This discovery bears further investigation into the underlying mechanisms that separate this attack from the rest.

We have used the latest version of CleverHans library (version : `2.0.0`) and have open-sourced the code to ensure repeatability of the results presented here.

References

[1] https://github.com/cschoel/nolds. 2017. 2

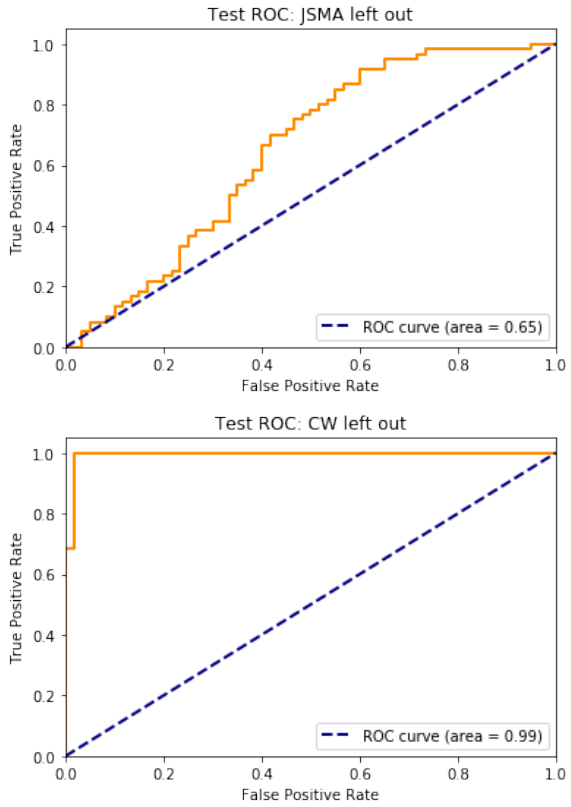

Figure 1. ROC curves for logistic models trained on all but one attack and tested on the left out attack.

[2] https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack 2017. 1

[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017. 2

[4] J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, and S. Ciliberto. Liapunov exponents from time series. *Physical Review A*, 34(6):4971, 1986. 1, 2

[5] J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of modern physics*, 57(3):617, 1985. 1, 2

[6] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017. 1

[7] Z. Gong, W. Wang, and W.-S. Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017. 1

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *ArXiv e-prints*, Dec. 2014. 2

[9] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 1

[10] W. He, J. Wei, X. Chen, N. Carlini, and D. Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv preprint arXiv:1706.04701*, 2017. 1

[11] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008. 2

[12] J. Lu, T. Issaranon, and D. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *arXiv preprint arXiv:1704.00103*, 2017. 1

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv e-prints*, June 2017. 2

[14] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 1

[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. DeepFool: a simple and accurate method to fool deep neural networks. *ArXiv e-prints*, Nov. 2015. 2

[16] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik, and A. Swami. The Limitations of Deep Learning in Adversarial Settings. *ArXiv e-prints*, Nov. 2015. 2

[17] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016. 1

[18] V. U. Prabhu and J. Whaley. Vulnerability of deep learning-based gait biometric recognition to adversarial perturbations. In *CVPR Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security (CV-COPS 2017)*, 2017. 2

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[20] D. Warde-Farley and I. Goodfellow. 11 adversarial perturbations of deep neural networks. *Perturbations, Optimization, and Statistics*, page 311. 1

[21] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 1

[22] W. Xu, D. Evans, and Y. Qi. Feature squeezing mitigates and detects carlini/wagner adversarial examples. *arXiv preprint arXiv:1705.10686*, 2017. 1