DETECTING ADVERSARIAL ATTACKS IN TIME-SERIES DATA

Mubarak G. Abdu-Aguye[⋆] *Walid Gomaa*[⋆†] *Yasushi Makihara*[‡] *Yasushi Yagi*[‡]

[⋆] Cyber Physical Systems Lab, Egypt Japan University of Science and Technology, Egypt.
[†] Faculty of Engineering, Alexandria University, Egypt.
[‡] The Institute of Scientific and Industrial Research, Osaka University, Japan.

ABSTRACT

In recent times, deep neural networks have seen increased adoption in highly critical tasks. They are also susceptible to adversarial attacks, which are specifically crafted changes made to input samples which lead to erroneous output from such models. Such attacks have been shown to affect different types of data such as images and more recently, time-series data. Such susceptibility could have catastrophic consequences, depending on the domain.

We propose a method for detecting Fast Gradient Sign Method (FGSM) and Basic Iterative Method (BIM) adversarial attacks as adapted for time-series data. We frame the problem as an instance of outlier detection and construct a normalcy model based on information and chaos-theoretic measures, which can then be used to determine whether unseen samples are normal or adversarial. Our approach shows promising performance on several datasets from the 2015 UCR Time Series Archive, reaching up to 97% detection accuracy in the best case.

Index Terms— Adversarial, detection, time-series, DFA, entropy

1. INTRODUCTION

Over the past decade, there has been an upsurge of interest in deep learning due to its ability to capture subtle and complex patterns in input data. Deep neural networks have consequently shown impressive results on a host of diverse tasks, from image recognition [1], activity recognition [2],[3],[4] to WiFi-based localization [5]. It is therefore no surprise that there has been a corresponding increase in the adoption of deep neural networks in production systems, some of which are health or safety critical.

Deep neural networks, being very highly parameterized, are commonly described as being black-box models, in the sense that their exact manner of operation is so complex as to be virtually uninterpretable. Although some literature is dedicated to interpreting and intuiting the inner workings of such models (e.g., [6]), there is still much work to be done.

A side effect of this complexity is the high sensitivity of such networks to small (and seemingly trivial) perturbations, as observed by [7]. This is most apparent in visual data, where the perturbations may be completely imperceptible to a human viewer, and yet lead to vastly different classification outputs than expected or desired [7]. The implication of this behavior is that such perturbations may be explicitly engineered and added to inputs for malicious purposes, which can potentially have devastating consequences in critical production systems. There is therefore a need to design methods of detecting and mitigating such attacks to maintain the integrity and correct function of such systems [8].

In this paper, we propose a novel approach to detecting the recently proposed variants of the Fast Gradient Sign Method and Basic Iterative Method attacks as applied to time-series data [9]. As these are currently the only published adversarial attacks against time-series data, this work is also, to the best of our knowledge, the first in which a corresponding detection scheme is proposed. For a given dataset, we begin by deriving a small descriptor for normal/unperturbed time-series samples based on Detrended Fluctuation Analysis (DFA) and Sample Entropy. Next, we build a normalcy model based on a One-Class Support Vector Machine (SVM) and bootstrap the model using the available descriptors. We then subsequently use the trained model to classify samples as either being normal (i.e., unperturbed) or outliers (i.e., adversarially attacked). We apply the method to a total of 85 datasets from the 2015 UCR Time Series Classification Archive [10], although we obtain results on 72 of them. The proposed detection scheme shows promising results on a set of diverse datasets, exceeding 90% detection accuracy on several of them.

2. RELATED WORK

To the best of our knowledge, virtually all the extant literature on detecting adversarial attacks deals solely with image-based data. However, this work proposes a detection scheme for univariate time-series data.

In [11], the authors propose a method to detect adversarial attacks in images mainly based on the fact that adversarial examples are much more sensitive to random mutations to the network structure than benign examples. The authors ob-

tain detection accuracies exceeding 90% on the CIFAR and MNIST datasets.

Work [12] proposes a new method for detecting adversarial attacks by treating images using two classic image processing techniques and comparing the network's output before and after such processing. Such treatment applied to adversarial examples is expected to lead to different network outputs than obtained before/without such preprocessing, while this is expectedly not the case for benign samples. The authors report an F1-score of 96.39% as obtained on testing their method on over 20,000 adversarial examples.

In [13], the authors detect adversarial examples by modelling the output distribution of hidden neurons in a deep neural network, when fed with normal training data. Based on this model, adversarial samples could be detected by comparing the states they generate relative to the 'normal' model. The authors report state-of-the-art detection accuracy based on this method.

3. BACKGROUND

3.1. Fast Gradient Sign Method (FGSM) Attack

The FGSM attack was first proposed in [14]. Neural network training involves the minimization of a loss function by adjusting the network weights. However, the method does the converse by instead adjusting the input samples in the opposite direction to the error function minimum. Therefore, the FGSM method is concerned with the computation of optimal perturbation series η, which may be pointwise (i.e where a point refers to a single timestep) added/summed with an input sample in order to maximize the classification loss function i.e., cause misclassifications. This is expressed mathematically as:

$$\eta = \epsilon \cdot \text{sign}(\nabla_x J(X, \hat{Y})), \tag{1}$$

where ∇_x refers to the derivative of the network's loss (computed for an input datapoint X and a desired output $\hat{Y}$) with respect to each timestep in X. Note that the actual magnitude of the gradient is not considered. Rather, the key factor is the *sign* of the gradient. In order to control the magnitude of the perturbation (i.e., to keep it barely perceptible), a multiplier factor ϵ is used, which is usually kept sufficiently small. The perturbed sample $\hat{X}$ may then be computed as $X + \eta$. In practice, this attack requires a single backpropagation pass through the network, hence its being termed as 'fast'. This method requires the setting of ϵ as a hyperparameter.

Note that this method relies on the attacker's ability to compute the loss function gradient with respect to a given input, which may not be directly possible. FGSM is therefore termed a 'white-box' attack since it requires knowledge of the inner workings of the network. However, a surrogate model may be used to simulate the target model and this method is applied to the surrogate to craft adversarial examples [15], enabling the use of such white-box attacks in practical scenarios [16].

3.2. Basic Iterative Method (BIM) Attack

The BIM attack [16] is an extension of the FGSM attack previously described. It involves the iterative re-application of the FGSM attack with a small step (i.e., ϵ, using the notation from subsection 3.1) size at each iteration. Additionally, to ensure that the resulting perturbations remain as unnoticeable as possible, a 'clipping' operation is performed, such that after each iteration, the modified per-pixel (or timestep value in the current context) is constrained to remain within a specified numeric neighborhood of the original. This makes it subtler than the FGSM attack although it requires more computational resources due to the multiple iterations involved. This method requires the setting of the number of iterations, the step size and the 'clipping' neighborhood values as hyperparameters.

Since it is based on the FGSM, the BIM is also a white-box attack, but it may be used even in black-box scenarios as described in subsection 3.1 previously.

4. PROPOSED APPROACH

By definition, adversarial perturbations are expected to be virtually imperceptible by human observers. However, this gives rise to three fundamental assumptions:

- Because the perturbations are sufficiently small, their effect will be more acute/noticeable in the differenced signal rather than the original signal.
- The perturbations, although small, contribute to or affect the native entropy or complexity of the signal.
- The perturbations are seemingly chaotic, and therefore their effect may be observable using measures of chaos.

As a result, the first assumption motivates the use of first-order differencing as a preprocessing method, and the use of information and chaos-theoretic measures when performing feature extraction on the differenced signal. We therefore adopt the Sample Entropy [17] as a feature, as it quantifies the degree of complexity of a signal, and therefore respects the second assumption. The second measure adopted is the Detrended Fluctuation Analysis (DFA) [18] which is a measure of the statistical self-affinity of a signal. The use of this value as a feature is motivated by its ability to capture the (increased) chaotic behavior (induced) in time-series data (due to the presence of adversarial perturbations).

Thus, we use a simple three-dimensional descriptor (consisting of the Sample Entropy, DFA and the ratio of the Sample Entropy to the DFA) to characterize the samples belonging to some dataset. This is advantageous because the descriptor is low-dimensional, therefore reducing the amount of training

samples needed and extremely fast to compute, which easily permits its use in real-world deployments.

Since practically, adversarial examples are not expected to be available ab-initio, we formulate the problem as a case of *outlier detection*. That is, we build a 'normalcy' model from the descriptors of benign samples. We subsequently classify unseen examples as either normal (i.e., similar to the normal samples and therefore normal) or as outliers (i.e., abnormal, which indicates the presence of adversarial perturbations). We adopt the One-Class SVM as the outlier detector due to its low data requirements, flexibility and high efficacy [19].

5. EXPERIMENTS

5.1. Experimental Setup

In order to validate the performance of the proposed method, we employed the 2015 UCR Time Series Classification (TSC) Archive, which is a collection of 85 univariate time-series datasets. Each dataset is divided into dedicated training and testing portions, which are naturally benign i.e. not adversarially perturbed. However, the authors of [9] made the FGSM and BIM-perturbed testing portions of all 85 datasets available by request. Therefore given any dataset, we bootstrapped the outlier detector (i.e., the one-class SVM) on descriptors derived from the training portion of the dataset. After training, we then evaluated three classification scenarios:

- The first involves a mixture of the original (i.e., benign) testing portion of the dataset and the FGSM-perturbed testing portion.
- The second involves a mixture of the original testing portion and the BIM-perturbed testing portion.
- The last scenario involves a mixture of the original, FGSM- and BIM-perturbed testing portions.

In each scenario, it can be seen that the classification problem is binary i.e., each testing sample is to be classified as either normal or adversarial. For the FGSM-perturbed samples, the value of ϵ used during their generation was fixed at 0.1. During the generation of the BIM-perturbed samples, the number of iterations was fixed at 10, with the step size set at 0.05 and the clipping neighborhood value set at 0.1.

We consider the classification accuracy as the metric of interest. Additionally, we contrive a criterion for success of the method, which is defined as whether or not the proposed method is at least as good as using a naive classification strategy. For the first two scenarios, the highest possible accuracy obtainable using such a strategy is 50% (since there is an equal number of benign and adversarial examples). In the third case, it is 66% since there are twice as much adversarial samples as there are benign samples and simply classifying every sample as adversarial would yield such a result. Due to the dearth of comparative work, we therefore consider naive classification as the baseline against which our method will be compared.

Although the method was applied to all 85 datasets in the UCR TSC archive, numerical instabilities in the implementations of DFA and Sample Entropy used led to difficulties in obtaining results on some of the datasets. Therefore, results from 72 of the 85 datasets in total are presented in the following section.

5.2. Results and Discussion

Tables 1 and 2 show the results obtained from the experiments described in Subsection 5.1. For datasets where the success criterion is achieved in all 3 scenarios (i.e., better than random guessing), the corresponding row is highlighted in grey.

In total, it can be seen that the proposed technique was able to succeed (by the contrived metric) on **43** of 72 datasets in total. Additionally, it shows detection accuracies often exceeding 90% in all scenarios (on **28** of 72 datasets). This underlines the efficacy of the proposed technique. Combined with its low computational requirement, it can readily be adopted in practical scenarios without incurring any noticeable computational or temporal overhead (i.e., latency).

In feature space, the benign samples belonging to any dataset are believed to constitute a 'benign' unimodal distribution. On the other hand, the adversarial samples have a tendency to have a slightly different mode, constituting an 'adversarial' distribution. For datasets whose samples have similar amounts of disorderliness, the 'benign' distribution has a small variance and is distinct from the 'adversarial' distribution. Therefore, adversarial examples can immediately be observed as having an extremely low probability of belonging to the 'benign' distribution. However, for datasets whose samples have a wide spread of disorderliness, this translates to a large variance in their 'benign' distribution. As a result, the adversarial distribution (whose mode is only slightly different than the 'normal' distribution) overlaps with the 'benign' distribution significantly. Hence adversarial samples can be observed (incorrectly) as belonging to the 'benign' distribution with a nontrivial probability. This makes discriminating between normal and adversarial samples difficult and leads to poor detection performance. This suggests that the proposed technique works best on datasets whose inherent entropy does not have a large spread.

6. CONCLUSION

In this work, we proposed an approach to detecting two adversarial attacks against univariate time-series data. By making certain assumptions about the effect of adversarial attacks, we derived suitable preprocessing and feature extraction methods (based on Sample Entropy and Detrended Fluctuation Analysis (DFA)) over the preprocessed samples. We then initialized

Table 1. Detection Accuracies for Adversarial Attacks

Dataset	FGSM	BIM	FGSM+BIM
50words	95.16	95.16	96.78
Adiac	91.69	92.84	93.95
ArrowHead	85.71	86.29	88.95
Beef	90.00	90.00	93.33
BeetleFly	95.00	92.50	95.00
BirdChicken	75.00	80.00	80.00
CBF	50.00	50.00	39.15
Car	96.67	96.67	97.78
ChlorineConcentration	61.47	58.35	50.39
CinC_ECG_torso	84.09	82.79	88.53
Coffee	91.07	91.07	94.05
Computers	44.20	45.20	31.87
Cricket_X	67.05	64.10	58.63
Cricket_Y	69.10	66.41	59.15
Cricket_Z	65.51	65.26	57.52
DiatomSizeReduction	90.52	90.36	93.25
DistalPhalanxOutlineAgeGroup	93.50	93.00	95.33
DistalPhalanxTW	90.88	87.50	88.42
ECG200	51.00	49.50	38.00
ECG5000	54.70	53.23	42.56
ECGFiveDays	92.86	85.25	88.23
Earthquakes	52.33	53.57	39.65
ElectricDevices	48.74	48.69	36.96
FISH	93.14	93.14	95.43
FaceAll	49.26	48.99	39.88
FaceFour	50.57	50.00	42.05
FacesUCR	50.37	49.90	37.17
FordA	95.47	95.49	96.98
FordB	91.41	91.41	94.27
Gun_Point	93.67	93.67	95.78
Ham	86.67	90.48	89.52
HandOutlines	91.65	91.65	94.43
Haptics	94.97	94.97	96.65
Herring	95.31	95.31	96.88
InlineSkate	87.82	73.64	80.42
InsectWingbeatSound	88.66	95.00	92.44
LargeKitchenAppliances	50.27	53.20	40.00
Lighting2	47.54	49.18	35.52
Lighting7	52.74	54.79	43.84
MALLAT	95.16	95.16	96.77
Meat	96.67	96.67	97.78
MedicalImages	62.04	71.97	59.08
MoteStrain	53.23	54.55	46.57
NonInvasiveFatalECG_Thorax1	94.27	94.27	96.18
NonInvasiveFatalECG_Thorax2	94.71	94.71	96.47
OSULeaf	94.21	94.01	96.01
OliveOil	53.33	53.33	68.89
Phoneme	50.92	51.27	37.62
Plane	89.52	90.00	91.43
RefrigerationDevices	48.67	48.27	37.24
ScreenType	46.80	46.53	31.38

Table 2. Detection Accuracies for Adversarial Attacks (continued)

Dataset	FGSM	BIM	FGSM+BIM
ShapeletSim	44.44	45.28	36.85
ShapesAll	76.42	90.83	81.11
SmallKitchenAppliances	62.93	87.47	69.16
StarLightCurves	94.01	94.01	96.01
Strawberry	94.29	94.29	96.19
SwedishLeaf	65.52	64.08	58.08
Symbols	63.72	82.31	66.16
ToeSegmentation1	57.68	54.17	47.22
ToeSegmentation2	49.23	47.31	36.67
Trace	44.50	45.00	30.67
TwoLeadECG	56.63	58.38	48.52
Two_Patterns	76.88	71.79	69.43
UWaveGestureLibraryAll	95.27	95.28	96.85
Wine	87.04	87.04	91.36
WordsSynonyms	97.02	97.02	98.01
Worms	61.88	56.08	50.64
WormsTwoClass	62.71	59.94	53.96
uWaveGestureLibrary_X	95.53	95.53	97.02
uWaveGestureLibrary_Y	94.85	94.85	96.57
wafer	95.19	93.64	95.71
yoga	94.98	94.98	96.66

an outlier detector (in this case a one-class SVM) to serve as a normalcy model and classified previously unseen samples as normal or outliers/adversarial. We apply the proposed technique on a total of 72 datasets and obtain promising results, with detection accuracies exceeding 90% in several cases and as high as 97% in the best case. We believe that this work, although preliminary, will pave the way for further work in this area.

In the future we intend to investigate more effective methods of detecting adversarial attacks in general, as well as developing specific mitigations against them.

Acknowledgments

The authors would like to express their deep thanks to Hassan Ismail Fawaz of IRIMAS, University of Upper Alsace, Mulhouse, France for making the adversarially-perturbed versions of the 2015 UCR TSC Archive available for this work.

Walid Gomaa is funded by the Information Technology Industry Development Agency (ITIDA) under the ITAC Program Grant no. PRP2019.R26.1 - A Robust Wearable Activity Recognition System based on IMU Signals.'

7. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Mubarak G. Abdu-Aguye and Walid Gomaa, "Versatl: Versatile transfer learning for imu-based activity recognition using convolutional neural networks," in *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2019 - Volume 1, Prague, Czech Republic, July 29-31, 2019.*, 2019, pp. 507–516.

[3] Mubarak G. Abdu-Aguye and Walid Gomaa, "Robust human activity recognition based on deep metric learning," in *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2019 - Volume 1, Prague, Czech Republic, July 29-31, 2019.*, 2019, pp. 656–663.

[4] Mubarak G. Abdu-Aguye, Walid Gomaa, Yasushi Makihara, and Yasushi Yagi, "On the feasibility of on-body roaming models in human activity recognition," in *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2019 - Volume 1, Prague, Czech Republic, July 29-31, 2019.*, 2019, pp. 680–690.

[5] Hamada Rizk and Moustafa Youssef, "Monodcell: A ubiquitous and low-overhead deep learning-based indoor localization with limited cellular information," in *Proceedings of the 27th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM*, 2019.

[6] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[8] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer, "Ensemble methods as a defense to adversarial perturbations against deep neural networks," *arXiv preprint arXiv:1709.03423*, 2017.

[9] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller, "Adversarial attacks on deep neural networks for time series classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8.

[10] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista, "The ucr time series classification archive (2015)," *URL: www. cs. ucr. edu/˜ eamonn/time_series_data*, 2015.

[11] Jingyi Wang, Guoliang Dong, Jun Sun, Xinyu Wang, and Peixin Zhang, "Adversarial sample detection for deep neural network through model mutation testing," in *Proceedings of the 41st International Conference on Software Engineering*. IEEE Press, 2019, pp. 1245–1256.

[12] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and XiaoFeng Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," *IEEE Transactions on Dependable and Secure Computing*, 2018.

[13] Zhihao Zheng and Pengyu Hong, "Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 7913–7922.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[15] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. ACM, 2017, pp. 506–519.

[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[17] Joshua S Richman and J Randall Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[18] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of dna nucleotides," *Phys. Rev. E*, vol. 49, pp. 1685–1689, Feb 1994.

[19] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.