Time series adversarial attack and defense

Xue Zhechang

[1] Karlsruhe Institute for Technology, Germany
[2] ugupg@student.kit.edu

Abstract. Nowadays, Time series data plays an important role in many fields. With the help of deep learning based model, people can classify and predict time series data, which is useful and efficient in data mining and recovering missing data. However, deep learning models are vulnerable to adversarial attacks. The researchers has transferred some adversarial attacks from image recognition domain to time series domain and prove their effectiveness. Thus, it's a big topic to detect these attacks and defense them to increasing the accuracy of classification and prediction. In previous researches, researches have already shown the vulnerability of time series model. In this work, we will introduce the methods to attack models and to prevent models from adversarial attacks.

Keywords: Time series · Adversarial attack · Deep Learning.

1 Introduction

Time series is a series of data points indexed in time order. Time series data are widely used in various fields ranging from mathematical statistics, signal processing and pattern recognition to quantitative finance and weather prediction.[1] People notice the big value of analysing time series data. Deep learning models have nowadays succeeded in many real life application such as speech recognition and computer visions. Thus, they are also applied to time series data. There's two main type of time series deep learning model. One is time series classification (TSC) model, which is mainly used to classify categories of time series data. Another is time series regression (TSR) model, which can predict the future data in a proper time point.

However, deep learning models are not robust. For example, researchers modify the original time series data, where the changes are too tiny to be detectable by human. Due to these changes, the accuracy of classification or prediction declines significantly. This is so-called adversarial attack.

To the best of our knowledge, adversarial attacks were first studied by Fazle et al[2]. Their target models are 1-Nearest Neighbor Dynamic Time Warping (1-NN) DTW, a Fully Connected Network and a Fully Convolutional Network (FCN). They trained Adversarial transformation network (ATN) to attack target models and tested with University of California Riverside (UCR) datasets. Finally they proved that TSC models are vulnerable to adversarial attacks. They also proved the vulnerability of multivariate time series[4].

Gautam et al[7]. proved that TSR models are also vulnerable to adversarial attacks. They transfered existed attacks from computer vision domain to time series domain, which are called fast gradient sign method (FGSM) and basic iterative method (BIM). They also proved the transferability of adversarial attacks to different target models.
Hassan et al[3]. introduced how FGSM and BIM works in time series domain. They used multi-dimensional scaling as the measurement to evaluate the relation between perturbation and accuracy.
Pradeep et al[8]. defined 4 kinds of adversarial attacks: untargeted, targeted, individual universal attacks. Then they introduced the method of targeted attack and universal attack and proved their effectiveness on TSC models based on ResNet. Finally, they introduced that backpropagation algorithm can help increase the robustness of time series model.
Aidong et al[9]. introduced a new method to attack TSC models with higher efficiency. They measured the importance of adversarial samples and selected some of the most important adversarial samples to modify the original data. This method can decline the perturbation of original data but increase the effectiveness. Tiny perturbation of time series data can lead to big difference in classification and prediction. However, few researches are done about detection of adversarial attacks. Mubarak et al[**?**]. introduced a method to detect whether the time series data is adversarial generated by FGSM and BIM.
Shoaib et al[6]. transferred three defensive methods from computer vision domain to time series domain: Adversarial training, TRADES and feature denoising. To prove their effectiveness, they used FGSM and Projected Gradient Descent (PGD) as white-box attacks and noise attack, boundary attack and Simple Black-box Attack (SIMBA) as black-box attacks.
Zhongguo et al[10] introduced a new defend to adversarial attacks and proved its effectiveness by experiment. They used thermometer encoding to non-linear encode original time series data and trained a encode-decode model to decode the modifyed time series data. Then they trained the deep learning model on time series with these output, which can nearly completed ignore the affect of perturbation.
In this paper, we will summarize some researches in time series domain. We will begin with adversarial attacks. On this basis, we will prove the vulnerability of time series model. Then we will show how to detect adversarial attacks from original data. Finally, we will introduce some methods to defense adversarial attacks.

2 Background

2.1 Definition

Definition 1 Time series data can be mathematically represented as set X = [x_1, x_2, x_3, ..., x_T], where T is the length of this set.

Definition 2 Decision-based attack is an attack completely depends on the final decision made by target model.

Definition 3 Global average pooling(GAP)

2.2 Measurement

2004

Sample Entropy

Detrended Fluctuation Analysis

Multi-Dimensional Scaling 2009,1903

2.3 Vulnerability

2.4 Transferability

Transferability is one of the important properties of adversarial examples. It means that one adversarial example, whose target model is A, can also decrease the accuracy of model B.
Aidong et al. experimented on 4 different models to prove the transferability of

Table 1. The effectiveness of adversarial attack against LSTNet[9]

Metrics	=0	=0.05	=0.10	=0.15	=0.20
RSE	0.1020	0.8098	0.8502	0.8822	0.9583
RAE	0.0581	0.4039	0.4460	0.4909	0.5562
CORR	0.8712	0.0034	-0.0085	0.0021	0.0059

the adversarial examples. For instance, they set the adversarial example targeting CNN as input to other 3 models: LSTNet, MHA_Net and RNN. As is shown in Fig.1, this adversarial example is also effective on other 3 models. However, its efficency on other 3 models is less than on CNN.
Although the efficiency of adveresarial examples will be decreased after transfer, this property is an essential part of generating black-box attack.

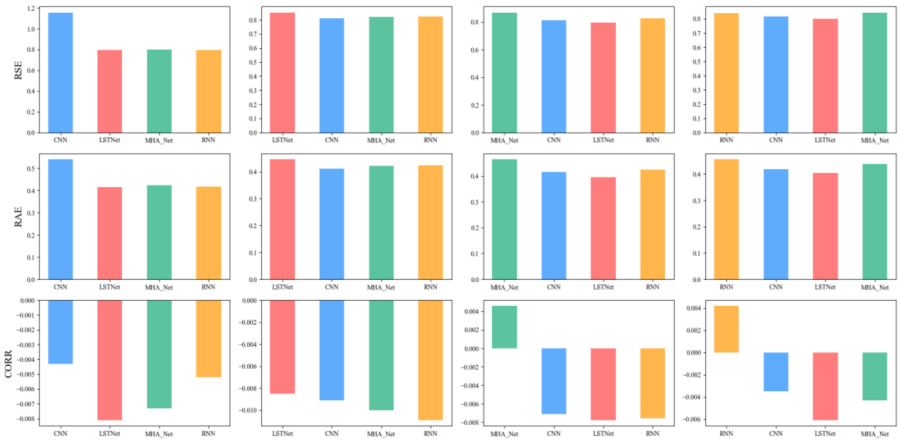

Fig. 1. Transferability validation on the Electricity Dataset[9]

3 Adversarial attack methods

Adversarial attacks are normally divided into two categories: White-box attacks and black-box attacks.
In white-box attacks, the attacker has access to all the information about the targeted model. Thus, attacking with gradient is a common method in deep learning. Here we will introduce two attacks based on gradient: Fast gradient sign method (FGSM) and basic iterative method (BIM).
In black-box attacks, the attacker has no information or parameter about the targeted model. Thus, it's impossible to attack the targeted model with gradient. Attackers normally use the relation between input and output of the target to find the way to attack the model.

3.1 White-box attacks

Fast gradient sign method

Basic iterative method

3.2 Black-box attacks

Random noise attack As is shown in Fig.4.a, even a tiny perturbation to original data, which is unrecognizable to human beings, can lead to big mistake of prediction with high confidence. In Fig.4.b, even a zero value time series data will be classified by the model with high confidence. However, in the correct situation, these time series data should be rejected by the classifier. This shows the potential risk of the model. Thus, generating random noise and modify the original data with them can be an attack to time series models.

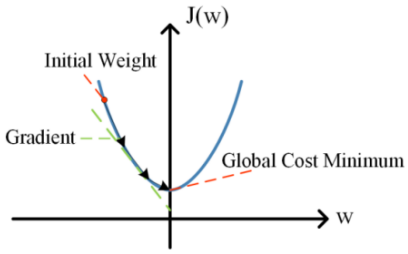

Fig. 2. Gradient descent for the solving of LSTNet model.[9]

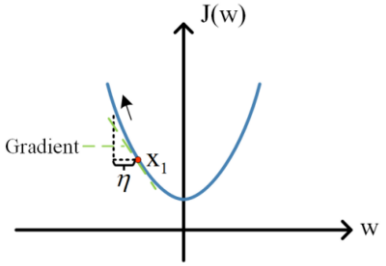

Fig. 3. Gradient-based generation of adversarial samples.[9]

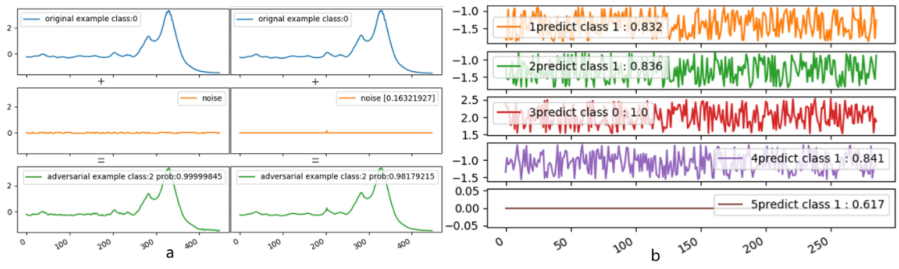

Fig. 4. The noise data can lead to the mistake of prediction with high confidence.[11]

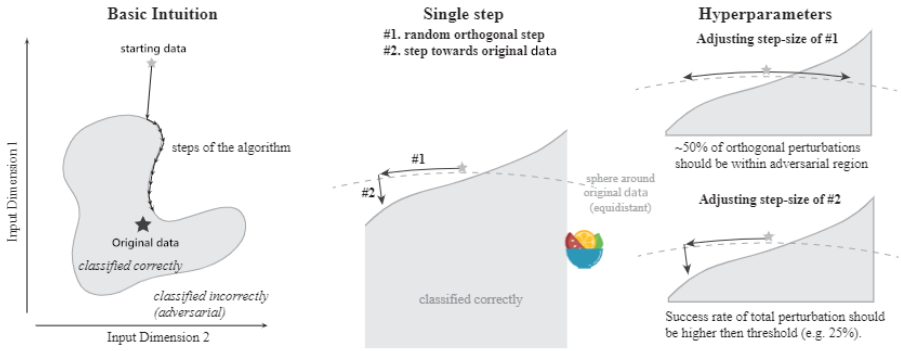

Fig. 5. The theory of boundary attack.[12]

Boundary attack Boundary attack is one of the decision-based attack. As is shown in Fig.5, the aim of this attack is to get the whole decision boundary of the label of original data.
This attack starts with a random input. This random input should not be classfied as the given label, which means it is already adversarial. The random input walks towards the original input, until it arrives at the decision boundary. At this moment, it walks orthogonal for a short distance, then starts walking towards original data again. After loop of this process, attackers will know the whole decision boundary, which is essential for black-box attack.

3.3 Class activation map

Class activation map(CAM) is able to show the susceptible region of a time-series data. CAM can only be utilized by the models with a global average pooling(GAP) layer, which can help identify possible regions of the input.
We assume that $A_m(t)$ is the univariate time series with variable $m \in [1, M]$, and w_m^c is the weight between label c and variable m. Then CAM_c will be calculated as follows:

$$CAM_c(t) = \sum_m w_m^c A_m(t) \tag{1}$$

Fig.6 is a CAM based on Coffee dataset. It shows that red part of the map is the most susceptible region of the time-series data. The perturbations modified on these areas are most efficient for adversarial attack. With the help of CAM, the adversarial attack will be less recognizable, but more effcient.

3.4 The importance of adversarial sample

4 Defense

2008, 2101, 2110

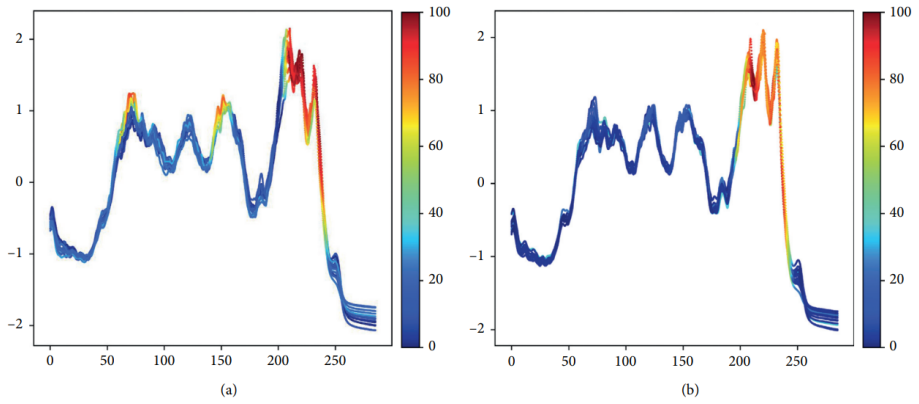

Fig. 6. Classification activation map in Coffee dataset. (a) Coffee dataset for class 0. (b) Coffee dataset for class 1.[10]

4.1 Adversarial training

4.2 TRADES

4.3 Feature Denoising

4.4 Backpropagation

4.5 Non-linear transfer

thermometer encoding

encode-decode model

Method

5 Conclusion

References

1. Wikipedia contributors. Time series, 2022. URL: https://en.wikipedia.org/wiki/Time_series
2. Fazle Karim and Houshang Darabi, Adversarial Attacks on Time Series, 2019 URL: https://ieeexplore.ieee.org/abstract/document/9063523
3. Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar and Pierre-Alain Muller IRIMAS, Universite Haute-Alsace, Mulhouse, France, Adversarial Attacks on Deep Neural Networks for Time Series Classification, URL: https://ieeexplore.ieee.org/abstract/document/8851936

4. Samuel Harford, Fazle Karim, and Houshang Darabi, Adversarial Attacks on Multivariate Time Series, URL: https://arxiv.org/abs/2004.00410
5. 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, Detecting adversarial attacks in time-series data, URL: https://ieeexplore.ieee.org/abstract/document/9053311
6. Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany, Benchmarking adversarial attacks and defenses for time-series data, URL: https://link.springer.com/chapter/10.1007/978-3-030-63836-8_45
7. Gautam Raj Mode and Khaza Anuarul Hoque, Department of Electrical Engineering & Computer Science, University of Missouri, Columbia, MO, USA, Adversarial Examples in Deep Learning for Multivariate Time Series Regression, URL: https://ieeexplore.ieee.org/abstract/document/9425190
8. Pradeep Rathore, Arghya Basak, Sri Harsha Nistala, Venkataramana Runkana, TCS Research, Pune, India, Untargeted, Targeted and Universal Adversarial Attacks and Defenses on Time Series, URL: https://ieeexplore.ieee.org/abstract/document/9207272
9. Aidong Xu, Xuechun Wang, Yunan Zhang, Tao Wu, Xingping Xian. Adversarial Attacks on Deep Neural Networks for Time Series Prediction, URL: https://dl.acm.org/doi/10.1145/3485314.3485316
10. Zhongguo Yang, Irshad Ahmed Abbasi, Fahad Algarni, Sikandar Ali, and Mingzhu Zhang, An IoT Time Series Data Security Model for Adversarial Attack Based on Thermometer Encoding URL: https://www.hindawi.com/journals/scn/2021/5537041/
11. Zhongguo Yang, Han Li, Mingzhu Zhang, Jingbin Wang and Chen Liu, School of Information Science and Technology, North China University of Technology, Beijing, China. A Method for Resisting Adversarial Attack on Time Series Classification Model in IoT System URL: https://link.springer.com/chapter/10.1007/978-3-030-60029-$7_5$0
12. Wieland Brendel, Jonas Rauber Matthias Bethge, Werner Reichardt Centre for Integrative Neuroscience, Eberhard Karls University Tübingen, German, DECISION-BASED ADVERSARIAL ATTACKS: RELIABLE ATTACKS AGAINST BLACK-BOX MACHINE LEARNING MODELS URL: https://openreview.net/forum?id=SyZI0GWCZ