A GAN-Based Real-Time Covert Energy Theft Attack Against Data-Driven Detectors

Zhinan Ding[1], Feng Wu[2], Lei Cui[1(✉)], Xiao Hu[1], and Gang Xie[1]

[1] Taiyuan University of Science and Technology, Taiyuan 030024, China
S202115110209@stu.tyust.edu.cn, {leicui,xiegang}@tyust.edu.cn
[2] Yunnan University, Kunming 650091, China
gzwf@mail.ynu.edu.cn

Abstract. The advanced metering infrastructure (AMI) system has been rapidly established around the world, effectively improving the communication capability of the power system. Problematically, it turns out malicious users can easily commit energy theft by tampering with smart meters. Thus, many data-driven methods have been proposed to detect energy theft in AMI. However, existing detection schemes lack consideration for well-planned covert attacks, making them vulnerable. This paper proposes a real-time covert attack model based on conditional generative adversarial network (CGAN). In particular, based on the transferability of adversarial samples, we first extract the data features that the malicious detection model focuses on during the detection process. Then, we utilize these extracted features and a generator to generate adversarial perturbations that can mislead malicious detection models. Finally, to make the generated perturbations more stealthy, a discriminator is used to simulate malicious detection models to correct them. Extensive experiments demonstrate that our proposed attack method can evade most current detection methods.

Keywords: Smart grid · Energy theft detection · CGAN · Covert attack · Feature extractor · Deep learning vulnerability

1 Introduction

Nowadays, smart grids are developing rapidly due to the effective integration of AMI and control methods. However, while enjoying the convenience of smart devices, cyber security attacks also emerged, and energy theft is one of the most concerned. A recent survey indicates that as early as 2019, energy theft caused utility companies to lose more than £19 billion per year [1]. Besides, more than 80% of people paid bills to malicious electricity theft users without their knowledge. Such losses are usually irreversible and large amounts for the individual and providers. Therefore, using different disciplines and machine learning techniques to detect electricity theft is crucial.

S. Yu et al. (Eds.): TridentCom 2022, LNICST 489, pp. 44–56, 2023.
https://doi.org/10.1007/978-3-031-33458-0_4

Energy theft attacks have caused significant global financial and functional damage to energy utilities. It can be described an attacker tamper with the meter without the grid company's awareness, resulting in a person paying less than it should have. The large-scale application of the smart grid and the renewal process of Internet technology add significant challenges to the fragile environment of smart grid applications. As a result, utilities need frequent access to smart meters to collect fine-grained electricity usage information from users in the data center. At the same time, advanced information technology is used to conduct behavioral characteristic analyses of the collected data to identify power theft.

Researchers have designed many detection methods [2], and data-driven methods have become the main research objects due to their low cost and high detection accuracy. As an effective information extraction method, deep learning is widely used in data-driven methods because it can learn the inherent laws of data, extract features, and achieve high accuracy [3]. However, existing deep learning detection methods mainly target simple attacks, with a lack of consideration for more covert and complex attacks. Due to the fragility of deep neural network structure, the security of non-technical losses (NTL) detection systems deployed in smart grids is difficult to guarantee effectively.

In this paper, a covert attack strategy based on the conditional generative adversarial network (CGAN) is proposed. Based on existing research [4], when the model recognizes the data, it mainly recognizes the non-robust features of the power data. Therefore, we can directly improve the attack success rate and reduce the attack cost by tampering with the non-robust features. Considering the transferability of adversarial samples, we build a model to simulation the detection model and use the features extracted by the network as non-robust features. The generator generates adversarial perturbations based on the non-robust features, and uses the discriminator to correct for these perturbations, making them more undetectable and effective. Undoubtedly, our method can provide experience in developing efficient and reliable detection systems.

The main contributions of this paper are as follows.

- We proposed a covert real-time attack method based on CGAN, which misleads the anomaly detection system by adding perturbations to electricity stealing data, and the generated attack samples can evade existing mainstream detection methods.
- We utilized a feature extraction module to extract the latent features of the daily electricity consumption data as a priori for the generator, making training the generator and the discriminator easier, so that we can obtain a higher attack success rate with fewer training epochs.
- Through extensive experiments, we demonstrated that the effectiveness of our attack method in evading existing advanced detection methods.

The content of this paper is organized as below. In Sect. 2, we briefly stated related works, including attack methods and detection methods used. In Sect. 3, we introduce the attack principle and proposed framework. Section 4 presents evaluations and analysis. Finally, the conclusion and further insights are provided in Sect. 5.

2 Related Work

Topics discussed in this section include deep learning-based detection methods and security issues under adversarial attack and defense. Researchers have done many works to deal with real-world physics problems, we review the related literature and briefly summarize the techniques of the attack and detection methods related in our experiment.

2.1 Attack Methods

Although deep neural network classifier has achieved good results in classification, its potential vulnerability should not be ignored. Szegedy et al. [5] for the first time, made the DNN classification model with high accuracy get wrong classification output by adding minimal disturbance to the image. Goodfellow et al. [6] first proposed the classic attack algorithm fast gradient sign method (FGSM), whose gradient-based algorithm can quickly and effectively generate counter samples. Subsequently, many researchers demonstrated that the DNN model is highly vulnerable to malicious adversarial samples. Existing adversarial attacks can be separated into black box attacks and white box attacks, one-shot attacks and iterative attacks, targeted attacks and non-targeted attacks, particular interference and general interference, etc. based on their features and attack consequences.

2.2 Detection Methods

Traditional machine learning detection algorithms have capable to detect malicious users, which including support vector machine (SVM), logistic regression (LR), random forest algorithm (RF), one-dimensional conventional neural network, etc. However, these methods have lower detection accuracy and classification effect for electric theft. The deep learning method achieves better classification effect and accuracy, which is due to the better ability of learning the characteristics of different users from multidimensional electricity data. Zheng et al. [7] proposed a wide and deep convolutional neural network to analyze the periodicity of user's electricity consumption by using two-dimensional electricity consumption data. He et al. [8] achieved the real-time detection of FDI attacks by using deep learning. Lu et al. [9] proposed a semi-supervised deep learning model, in which an adversarial module was added to make the model have high detection accuracy and strong anti-noise capability. The accuracy of these methods can reach more than 90%.

2.3 GAN for Adversarial Attacks

GAN was initially introduced by Goodfellow and is now widely utilized in object identification, semantic segmentation, images creation, and video prediction. Generally speaking, a GAN is a two-person network of generators and discriminators. Specifically, the former aims to simulate and learn the distribution of

real data as much as possible. Random noise or potential variables are then reconstructed to produce real-world examples. The discriminator is used to distinguish between raw data and generated data. Mutual competition achieve Nash equilibrium, end of training.

Based on the GAN attack, Xiao et al. [11] trained a conditional GAN, ADVGAN, to generate various adversarial examples without accessing the target model. Mangla et al. [13] added a potential feature extraction block based on ADVGAN, which effectively reduced the number of training rounds of the model and improved the attack success rate of the attack sample. Liu et al. [12] introduced adversarial examples during the training process and proposed Rob-GAN, thereby improving the convergence speed of GAN training and the quality of generating adversarial examples. Affected by the above, the hypothesis and verification of CGAN-based effective attacks were carried out.

3 Targeted Adversarial Example Generation with CGAN

3.1 Problem Description

Let $X \subseteq R^{n \times m}$ be a set of data on the electricity consumption of customers in an area, where n indicates the number of users, where m indicates the number of power usage days. Given an instance (x_i, y_i) is the m-dimensional power consumption characteristic vector of a user, and indicates the corresponding real category label, and i indicates the user number. The electricity theft detection problem in the smart grid can be expressed as a discrete binary classification task. The model divides each electricity record x_i into abnormal or normal, and the formula is as follows

$$y_i = \begin{cases} 1, if\ record\ is\ abnormal; \\ 0, otherwise. \end{cases} \tag{1}$$

In the smart grid, theft attacks can be described as a series of malicious information bundled, expressed as $X^{per} = \{m_1, m_1, m_1, \cdots, m_t\}$, which $m_t = \{x_{t1}, x_{t2}, x_{t3}, \cdots, x_{tn}, y_t\}$, where n is the amount of information, x_{tn} is the n-th feature of the example (or an implicit feature), y_t is the label for the corresponding example. Therefore, we can regard the power theft detection as a multi-class classification problem. x_{tn} and y_t are the inputs and output of the detection model, respectively. The system diagram of the method as shown in Fig. 1. During the operation of the smart grid, since the data required for the attack is difficult to collect, even though the existing methods have high detection accuracy, their robustness and generalization still need further improvement.

3.2 Attack Principle

This paper designs a CGAN-based attack generation model, assuming that the discriminator D can learn features from the dataset X and accurately classify them to the corresponding label Y, can be defined as

$$D(x_i) = y_i, i \in n. \tag{2}$$

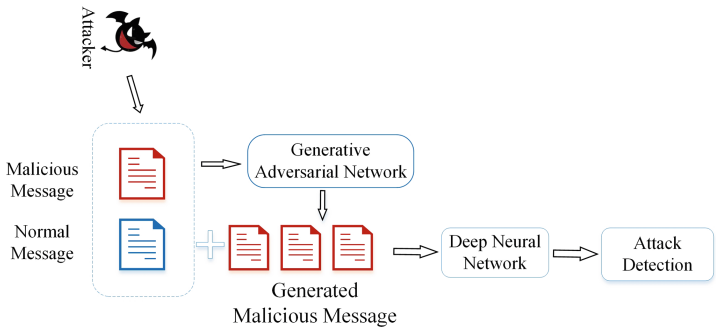


Fig. 1. Systematic overview of the method

To maximize the benefit of the feature extractor, we pre-train a binary classifier similar to the electric company's malicious detection model and use the first few layers of the classifier architecture as the feature extractor.

The rationality analysis is as follows: 1. Adversarial samples are transferable, which means that the features extracted by different models are (roughly) the same. Thus, the features we extracted with our self-trained classifier should be (roughly) the same as the features extracted by the power company's detection model. 2. Targeted processing of features that can be identified by the malicious detection model can effectively reduce the cost of data tampering. Therefore, the efficiency of the entire attack process and the attack success rate will be greatly improved.

After extracting features of the power data using the above ideas, we use generator to generate adversarial perturbations based on the features, generator G generates a perturbations through feature analysis of x_i^{per}, which can be defined by

$$\begin{aligned} x_i^{per} &= G(z \mid f(x_i)) \\ x_i^{adv} &= x_i + x_i^{per}, \end{aligned} \tag{3}$$

where z represents noise data and follows the Gaussian distribution, $f(x)$ represents a feature extractor. x_i^{per} stands for attack perturbation. x_i^{adv} is the final attack sample.

The purpose of electricity thieves is to reduce the payment of bills to gain benefits. Therefore, the resulting attack perturbation should be able to circumvent existing detection methods while also ensuring that the thief can make sufficient profits. To achieve this, we need to find a tiny perturbation to add to the raw data, reprsents as x_i^{adv}=x_i+x_i^{per}. Therefore, we designed the following optimization

$$\begin{aligned} Loss_D = argmin\xi(D(x_i + x_i^{per}), y^{'}) \\ s.t. \quad y^{'} \neq y_i, and \quad ||\varepsilon||_2 < \delta, \end{aligned} \tag{4}$$

where $\xi(\cdot)$ represents crossentropy loss function of the discriminator, $y^{'}$ denotes normal data, y_i is the corresponds to the original label of the sample (abnormal data). If and only if $D(x_i^{adv}) = y^{'}$, The loss reached the minimum. Where ε is the added tiny perturbations, $||\cdot||_2$ is a norm constraint on perturbation, δ is the maximum perturbation value specified.

3.3 The Proposed CGAN-based Architecture

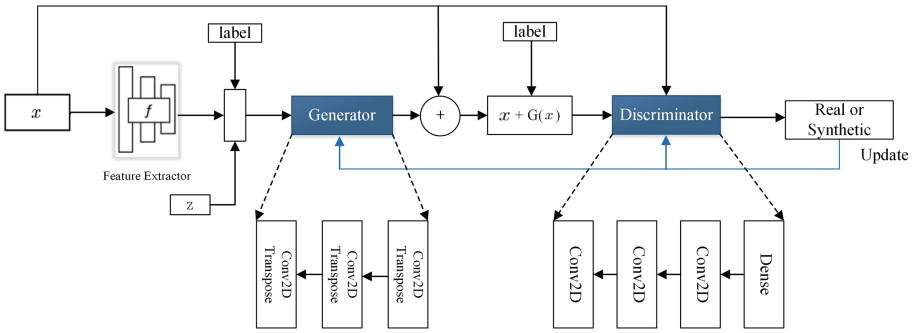


Fig. 2. Structure diagram of attack model based on CGAN

As shown in Fig. 2, the structure of the proposed model contains a feature extractor f, a generator G and a discriminator D. The feature extractor f extracts the latent features $f(x)$ (the main features recognized by the malicious detection system) from the power data x. Then, concatenate $f(x)$ with the noise vector as the input of the generator G to generate a perturbation x_i^{per} corresponding to x. Finally, $x + G(x)$ is sent to D, and the output result represents the probability of normal samples. In our method, the feature extractor, generator, and discriminator are trained end-to-end.

Feature Extractor: A recent research shown that the cause of adversarial examples is the non-robust feature learned from the training set by models [16]. Particularly, research shows that the knowledge learned by machine learning models during training is the correspondence between non-robust features and data labels. Motivated by the demonstration, we perform saliency feature extraction on the electricity consumption data in advance, in this way, the extracted prior knowledge can be used to generate adversarial examples to have better performance, Moreover, due to the preprocessing of the feature extractor, subsequent network structures can only focus on their own tasks (such as generating perturbations and identifying data), so that significantly improve the model convergence speed. We use convolutional layers to extract the features of the power data, and combine the extracted features with Gaussian noise z_{noise} as the input of the generator.

Generator: The attacker's goal is to disguise the tampered power data as untampered, which is similar to adversarial attacks. Therefore, after using the feature extractor to extract the features of the target power data, we need to generate the corresponding adversarial perturbation based on the feature, and add the perturbation to the raw data. In addition, the features obtained by the feature extractor have different dimensions from the original data, thus another function of the generator is to unify the dimension of the generated perturbation and raw data. The generator is implemented with an autoencoder, during the training phase, we take root mean square error as the loss function. The optimization goal of the generator is to make its output (adversarial example) as close as possible to x (raw data).

Discriminator: In real-world scenarios, attackers usually do not have access to the detection model of electric companies, thus they can only perform black-box attacks. We utilise the discriminator to simulate the detection model of the electric companies. The discriminator takes the tampered power data from the generator and classifies it. The discriminator's goal is to detect as much as possible that the data has been tampered, and transmit the detection result to the generator through back-propagation. The generator updates itself based on the feedback from the discriminator and generates tampered samples that are more difficult to detect, and loops the training process until convergence. The discriminator is a three-layer neural network. At this point, the discriminator has two tasks, one is to distinguish the true and false samples (the original task in GAN), another is to classify the tampered data.

In brief, the task of the generator is to generate minimal perturbations while maximally misleading the discriminator classification. The discriminator is continuously trained to identify whether the input is a malicious sample or a normal sample. The two are constantly competing, and the process can be defined as a V function as

$$\begin{aligned} \min_G \max_D V(D,G) = E_{x\sim p_{data}(x|y)}[logD(x)]+ \\ E_{x\sim p_z(x)}[1-log(D(x+G(x)|y))]. \end{aligned} \tag{5}$$

During the optimization process, the data distribution generated by the generator G will gradually coincide with the original data distribution, which increases the difficulty of detecting tampered samples. The discriminator D is used to ensure that the data generated by G is similar to benign data. We define $p_g(x)$ as the generative distribution of G, then the maximum value of D to be solved by the V function can be expressed as $D(x) = p_{data}(x)/(p_{data}(x)+p_g(x))$. After reaching the Nash equilibrium point, the training will stop. At this time, $p_{data}(x) = p_g(x)$, that is, $D(x) = 0.5$. Since the discriminator simulates a malicious detector, $D(x) = 0.5$ means that the detection by the malicious detection system is equivalent to random guessing.

4 Evaluation and Analysis

In this section, we will present the experimental results. In Sect. 4.1, we describe the experimental data and its preprocessing procedure. Section 4.2 introduces the performance indicators. We then show the performance of our proposed covert attack on various advanced detection models in Sect. 4.3. Finally, we evaluate the attack model presented in Sect. 4.4.

4.1 Experimental Data and Preprocessing

We conduct experiments using a real dataset released by State Grid, which records the daily electricity consumption of 42,372 users for 1,035 days. In addition, State Grid classified each user in the dataset, including 38,757 normal users and 3,615 electricity stealing users. In this experiment, 3000 users marked as electricity stealers and 17000 users marked as normal were randomly selected to form the original dataset. Then, select 2000 normal users and use the proposed attack model to generate attack samples, and then extract 1000 electricity stealing users and 17000 normal users and the generated attack samples to form an attack data set. Finally, the original dataset and the attack dataset are divided into 80% training set and 20% testing set, respectively.

There are a large number of missing values in the dataset, which we filled in using the interpolation method shown below. The outliers are then processed using the three-sigma criterion, and finally the data is normalized using MAX-MIN.

$$x_{i,t}^{'} = \frac{x_{i,t} - x_{min}}{x_{max} - x_{min}}, \tag{6}$$

where $x_{i,t}$ and $x_{i,t}^{'}$ are the power consumption of i-th user in the t-th day and the normalized power consumption. x_{max} and x_{min} represent the maximum and minimum values in the power consumption sequence.

4.2 Performance Indicators

In this experiment, we used AUC, MAP and Recall as evaluation metrics to evaluate the effectiveness of the generated attack samples. AUC is defined as the area under the ROC curve, which is usually used as an evaluation index for binary detection models. Its abscissa is false positive rate (FPR), and its ordinate is true positive rate (TRP). Usually, the value of AUC is in the range of [0.5–1], and the closer it is to 1, the better the classification effect. Its calculation formula is as follows

$$AUC = \frac{\Sigma_{i \in positive} Rank_i - \frac{M(M+1)}{2}}{M \times N}, \tag{7}$$

where $Rank_i$ is the rank of sample i, M is the number of normal users, and N is the number of abnormal users.

MAP is also used in this paper to measure the accuracy of the hybrid detection model. The specific calculation formula is as follows

$$MAP@N = \frac{\Sigma_{i=1}^{s} \frac{Y_{k_i}}{k_i}}{s}, \tag{8}$$

where s represents the number of true positive samples in the first N scores after sorting the rank from high to low, k_i represents the position of the ith true positive sample, and Y_{k_i} represents the number of true positive samples before k_i.

In addition, we used the Recall to evaluate the detection rate of positive samples, the formula is as follows

$$Recall = \frac{TP}{TP + FN}, \tag{9}$$

where TP is the amount of accurate predictions made for positive samples, and FN is the amount of forecasts for positive samples that were wrong.

4.3 Attack Performance Verification and Analysis

To verify our attack performance of the generated samples, five advanced detection models are evaluated in this chapter on the original and attack datasets, including support vector machine (SVM), logistic regression (LR), random forest (RF), convolutional neural networks (CNN) and Wide&Deep CNNs. As shown in Table 1, both CNN and wide&deep models have shown good performance in detecting traditional electricity stealing behavior, which is mainly due to the fact that deep neural networks can learn hidden features of abnormal samples and normal samples. Compared with SVM, LR and RF, it can better capture local correlation and achieve good results.

Table 1. Detection results of the original data set

Typle	AUC	MAP@50	Recall
SVM	0.65	0.75	0.70
LR	0.59	0.35	0.59
RF	0.67	0.43	0.61
CNN	0.73	0.87	0.67
Wide&Deep	0.81	0.89	0.71

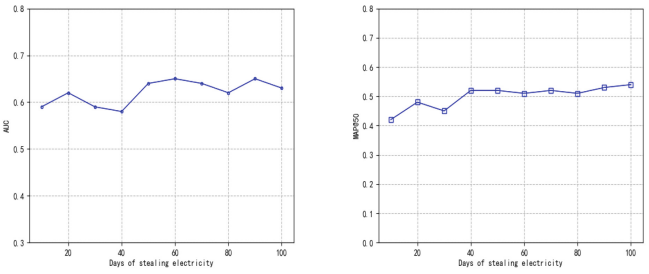
The Table 2 depicts the advanced models' detection results for the attack data set produced in this chapter. Among them, all performance indicators have dropped significantly, especially MAP and Recall have dropped significantly, mainly because these models cannot identify attack samples in the data set. In

addition, there are still certain AUC and MAP values because the attack dataset also contains 1000 traditional electricity stealing samples and normal samples, which will be correctly classified by advanced detection methods. To sum up, the attack samples generated in this chapter can effectively evade the identification of existing detection algorithms.

Table 2. Detection results of the attack data set

Typle	AUC	MAP@50	Recall
SVM	0.55	0.40	0.50
LR	0.50	0.17	0.15
RF	0.45	0.08	0.15
CNN	0.58	0.33	0.46
Wide&Deep	0.64	0.51	0.64

In the experiment, we also considered the impact of the continuous electricity stealing days on the attack effect under this attack method. Specifically, the attack days of 2,000 attacking users ranged from 10 days to 100 days, with a step size of 10. As depicted in Fig. 3, as attack days rise, the AUC and MAP values basically change around 0.1, and the fluctuation range is small, so the effectiveness of the generated attack samples in long-term electricity stealing can be verified.



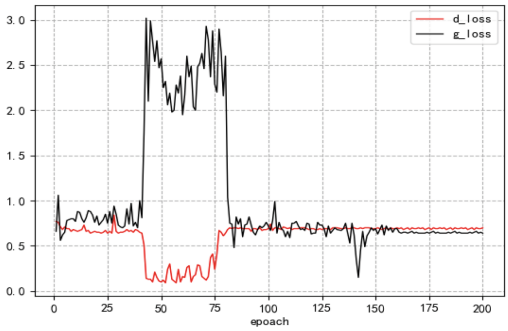
(a) AUC (b) MAP

Fig. 3. Comparison of different number of attack days

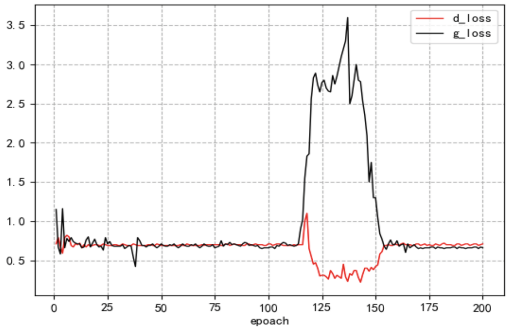
4.4 Generative Network Model Performance Analysis

In this section, we investigate the impact of several parameters on the functionality of the generative model, including: whether to use a feature extractor, the number of neurons in the feature extractor, and the number of network layers in the generator.

The feature extractor is located in front of the generator, and the extraction of electricity data features by main users is used as a priori information for the generator to generate disturbances. As shown in Fig. 4, the generative network model's convergence speed and training time may both be significantly accelerated by the feature extractor.



(a) Feature extractor

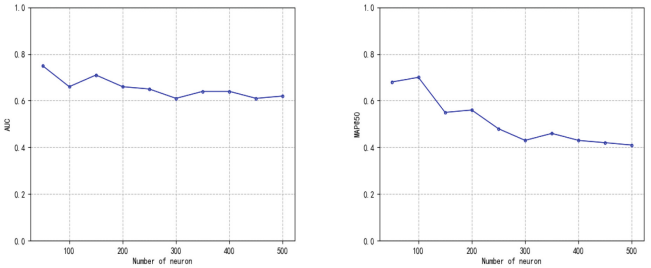


(b) No feature extractor

Fig. 4. The effect of feature extractor on convergence

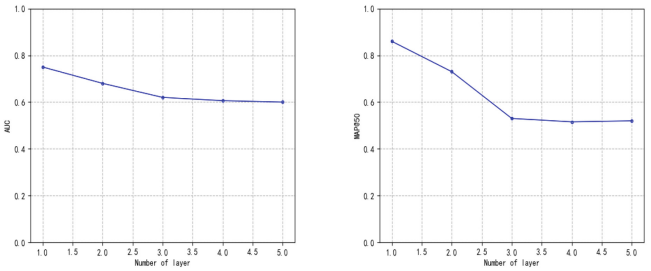
As shown in Fig. 5, when the number of neurons in the feature extractor's fully connected layer reaches 300, the AUC and MAP@50 values of the detection model are both low, and the attack samples generated by the generative network model are closest to the real data.

Finally, as shown in Fig. 6, when the number of generator network layers in the generative model is 3, the generated attack samples have a strong ability to evade detection, and the attack performance is better. When there are more than 3 layers, all indicators show a smooth trend, and it is difficult to achieve better results if more computing power is added to the number of layers.



(a) AUC (b) MAP

Fig. 5. Impact of layer number



(a) AUC (b) MAP

Fig. 6. Comparison of different number of neurons in the fully connected layer of the feature extractor

5 Summary

In this paper, we evaluate the vulnerability of models for detecting energy theft based on deep learning, when appropriate perturbations are added to the data will mislead a trained high-accuracy classifier. Based on this finding, we designed a CGAN-based generative model. Different from the traditional CGAN model, we add a feature extractor to extract the user's electricity consumption feature information. Comprehensive experiments have shown that the generation model designed in this chapter can generate attack samples efficiently, and has shown good attack performance in the face of the current advanced electricity stealing detection models. We will take improving the robustness of deep learning models as the focus of future work, specifically to optimize the network structure and training strategies, to ensure that on the basis of effective training, we will further improve the defense accuracy on the data set.

Acknowledgment. This work was supported by the Scientific and Technologial Innovation Programs of Higher Education Institutions in Shanxi, China (No. 2020L0338) and the Shanxi Key Research and Development Program (No. 202102020101002 and 202102020101005) and the Fundamental Research Funds for the Central Universities (No. 2042022kf0021).

References

1. Yao, D., Wen, M., Liang, X., et al.: Energy theft detection with energy privacy preservation in the smart grid. IEEE Internet Things J. **6**(5), 7659–7669 (2019)
2. Cui, L., Guo, L., Gao, L., et al.: A covert electricity-theft cyber-attack against machine learning-based detection models. IEEE Trans. Ind. Inform. **18**(11), 7824–7833 (2021)
3. Zhao, Y., Qu, Y., Xiang, Y., et al.: A comprehensive survey on edge data integrity verification: fundamentals and future trends. arXiv preprint arXiv:2210.10978 (2022)
4. Ilyas, A., Santurkar, S., Tsipras, D., et al.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
5. Szegedy, C., Zaremba, W., Sutskever, I., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
7. Zheng, Z., Yang, Y., Niu, X., et al.: Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Trans. Industr. Inf. **14**(4), 1606–1615 (2017)
8. He, Y., Mendis, G.J., Wei, J.: Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. IEEE Trans. Smart Grid **8**(5), 2505–2516 (2017)
9. Lu, X., Zhou, Y., Wang, Z., et al.: Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid. Energies **12**(18), 3452 (2019)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
11. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
12. Liu, X., Hsieh, C.J.: Rob-GAN: generator, discriminator, and adversarial attacker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11234–11243 (2019)
13. Jandial, S., Mangla, P., Varshney, S., et al.: AdvGAN++: harnessing latent layers for adversary generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
14. Ying, H., Ouyang, X., Miao, S., et al.: Power message generation in smart grid via generative adversarial network. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 790–793. IEEE (2019)
15. Bai, T., Zhao, J., Zhu, J., et al.: Toward efficiently evaluating the robustness of deep neural networks in IoT systems: a GAN-based method. IEEE Internet Things J. **9**(3), 1875–1884 (2021)
16. Chen, R., Chen, J., Zheng, H., et al.: Salient feature extractor for adversarial defense on deep neural networks. Inf. Sci. **600**, 118–143 (2022)