Available online at www.sciencedirect.com
ScienceDirect
Energy Reports 8 (2022) 123–129
www.elsevier.com/locate/egyr

2021 6th International Conference on Clean Energy and Power Generation Technology (CEPGT 2021), September 10–12, 2021, Shanghai, China

Adversarial attacks on deep learning models in smart grids

Jingbo Hao*, Yang Tao

College of Artificial Intelligence, Nanchang Institute of Science & Technology, Nanchang 330108, China

Received 18 October 2021; accepted 2 November 2021
Available online 26 November 2021

Abstract

A smart grid may employ various machine learning models for intelligent tasks, such as load forecasting, fault diagnosis and demand response. However, the research on adversarial machine learning has attracted broad interest recently with the rapid advancement of deep learning techniques, which poses an evident threat to those deep learning models deployed in smart grids. In the face of the emergent problem, we make a compact survey of the adversarial attacks against deep learning models in smart grids. The research status of deep learning applications in smart grids and adversarial machine learning is briefly summarized firstly. Adversarial evasion and poisoning attacks in smart grids are analyzed and exemplified respectively with focus. To mitigate the threat typical countermeasures against adversarial attacks are also presented. From the survey it can be concluded that the threat of adversarial attacks in smart grids will be a kind of long-term existence and need continuous attention.

Peer-review under responsibility of the scientific committee of the 2021 6th International Conference on Clean Energy and Power Generation Technology, CEPGT, 2021.

Keywords: Smart grid; Data attack; Adversarial example; Deep learning

1. Introduction

A smart grid uses bi-directional flows of electricity and information to build an interoperable and distributed power delivery network [1] as shown in Fig. 1 [2], which makes the system much more efficient than a traditional power grid. With vast data ranging across electricity generation, transmission, distribution and consumption, a smart grid should employ various machine learning models for intelligent tasks, such as load forecasting, fault diagnosis and demand response. However, the machine learning models applied in smart grids are mostly vulnerable to elaborate data attacks, especially adversarial attacks with the explosion of deep learning applications [3]. Adversarial attacks are intrinsically covert and able to cause random or directed malicious effects by replacing natural inputs with purpose-made adversarial examples [4,5] for a target model. Typically an adversarial example should be hard to discriminate from the original sample but very likely to lead to a different output result, which in a sense is concept-related to generative adversarial networks (GANs) [6].

* Corresponding author.
E-mail address: jbhao@126.com (J. Hao).

https://doi.org/10.1016/j.egyr.2021.11.026

Peer-review under responsibility of the scientific committee of the 2021 6th International Conference on Clean Energy and Power Generation Technology, CEPGT, 2021.

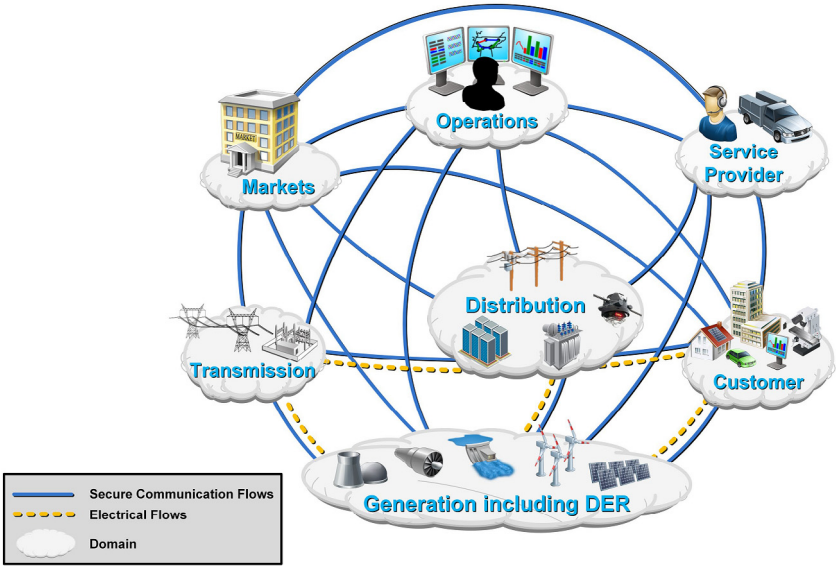

Fig. 1. NIST smart grid conceptual model.

During recent years, the study on the adversarial attacks against computer vision applications based on deep learning has drawn wide attention [7,8]. This is due to the fact that the deep learning models for computer vision are oddly vulnerable to those adversarial examples that are generally difficult to perceive by an observer through adding tiny perturbations and try to mislead target models to produce wrong results. Moreover, adversarial examples in other application domains have emerged, e.g. cyber security [9], automatic speech recognition [10] and natural language processing [11]. Likewise this kind of data attacks based on adversarial examples should also be effective for automated power systems, which has been proved by some sensitive researchers.

We attempt to make an elegant survey of the adversarial attacks on deep learning models in smart grids in this paper. Current research on deep learning models deployed in smart grids and adversarial machine learning is summarized in Section 2. Adversarial evasion attacks in smart grids are analyzed and exemplified in Section 3. Adversarial poisoning attacks in smart grids are illustrated in Section 4 similarly. Next typical countermeasures against these adversarial attacks are presented in Section 5. Finally the conclusions are drawn.

2. Related work

2.1. Deep learning models in smart grids

Various deep learning models have been deployed in smart grids to make best use of the huge amount of data generated continuously [12]. These models can greatly facilitate the operation of smart grids, but may face the threat of adversarial attacks at the same time. It is important to figure out the input format of a deep learning model utilized by a certain application to assess possible adversarial attacks on that model. Some illustrative deep learning models in smart grids are listed with their input formats in Table 1.

2.2. Adversarial machine learning

To date most adversarial attacks are focused on fooling deep learning-based image classifiers. These adversarial attacks on image classification can be categorized under different threat models which define the assumptions about what attacks may be attempted against a security-sensitive information system. Thus in [19] those adversarial attacks are classified from six threat aspects as shown in Table 2, which is equally suitable for other application scenarios.

The adversarial attacks on smart grids can also be classified into two primary types: evasion and poisoning attacks. Evasion attacks tamper with input data to try to evade the utility of an existing model while poisoning

Table 1. Illustrative deep learning models with their input formats in smart grids.

Application	Model	Input format
False data injection attack (FDIA) detection	Multilayer perceptron (MLP) in [13]	Vector
Demand response	Deep reinforcement learning (DRL) in [14]	Vector
Short-term load forecasting (STLF)	Long short-term memory (LSTM) in [15]	Vector sequence
Non-intrusive load monitoring (NILM)	LSTM + MLP in [16]	Vector sequence
Event cause analysis (ECA)	Convolutional neural network (CNN) in [17]	Two-dimensional array
Fault diagnosis	CNN in [18]	RGB image

Table 2. Types of adversarial attacks.

Threat aspect	Attack types
Attacker's influence	Evasion attack, poisoning attack
Attacker's knowledge	White-box attack, black-box attack
Security violation	Integrity violation, availability violation, privacy violation
Attack specificity	Targeted attack, non-targeted attack
Attack computation	Sequential attack, iterative attack
Attack approach	Gradient-based attack, decision-based attack, transfer-based attack, approximation-based attack

attacks pollute training data to try to poison a newborn model. In addition, these attacks can also be divided into white-box and black-box attacks. A white-box attack is implemented upon the assumption that the comprehensive knowledge of the target model is given, such as the model's parameters, architecture, training strategy and training data. A black-box attack is implemented upon the assumption that nothing about the target model is known. Whether an adversarial attack is white-box or black-box, there should be at least a directly attacked model which is either the target model or a substitute model. Therefore, white-box methods are for basis and black-box methods are for use.

As for the countermeasures against adversarial attacks, there are reactive and proactive approaches [20]. Reactive methods are intended for blocking adversarial examples, such as adversarial detection, input reconstruction and network verification. On the other hand, proactive methods are intended for building more robust models with resistance to adversarial examples, such as network distillation, adversarial training and classifier ensemble. With the development of the degree and extent of adversarial attacks, novel defense techniques are urgently needed.

3. Adversarial evasion attacks

3.1. General analysis

For a security-sensitive application, the input samples may be actively manipulated to confound the machine learning model deployed in the system [21]. The so-called evasion attacks will intriguingly take effect for deep learning models through adversarial examples.

In [22] a generic white-box adversarial evasion attack is concluded as an optimization problem as shown in the following formula:

$$\arg\max_{x_{\mathrm{e}}} \mathrm{L}(x_{\mathrm{e}},\ y,\ W) \quad s.t.\ \|x_e - x\|_d \leq \varepsilon, x_{\mathrm{l}} \leq x_{\mathrm{e}} \leq x_{\mathrm{u}} \tag{1}$$

where x_{e} is an adversarial example crafted from original x with its label y. W is the model weights and L is the loss function. $\|\,.\,\|_d$ computes the L_d norm of a vector. ε is a tiny value limiting the size of a perturbation while x_{l} and x_{u} set the bounds of the adversarial example.

To solve the optimization problem, many attack methods are proposed by means of gradient-based or constrained optimization [8], such as L-BFGS, FGSM and DeepFool. These methods are mainly aimed at image classification

models with individual input data (e.g. two-dimensional array). With respect to sequential input data (e.g. vector sequence), Papernot et al. [23] utilized a technique known as computational graph unfolding to generate adversarial sequences against recurrent neural networks (RNNs). Mode et al. [24] applied the FGSM and BIM methods to make adversarial examples for LSTM and Gated Recurrent Unit (GRU) models. Furthermore, Sun et al. [25] introduced two techniques named critical point attack and antagonist attack to launch adversarial attacks against DRL models.

In consideration of the black-box nature of the majority of valuable attack targets, adversarial evasion attacks should be implemented in a pure black-box manner. The attainment of the intention counts on the transferability possessed by an adversarial example which enables the adversarial example produced on a known substitute model to transfer its misleading effect to an unknown target model in some degree. Papernot et al. [26] defined the intra-technique transferability between the same type of machine learning models and the cross-technique transferability between different types of models. Liu et al. [27] illustrated that non-targeted adversarial examples transfer better than targeted ones. Tramer et al. [28] empirically proved that different models may have similar decision boundaries and provided a method for measuring the space fit for adversarial examples.

3.2. Cases in smart grids

Quite a few adversarial evasion attacks on deep learning models in smart grids have been evaluated as shown in Table 3. It can be seen that most existing attack methods on image classification can be easily ported to power grid applications, which is meaningful for both adversarial attacks and defenses in smart grids.

Table 3. Adversarial evasion attacks in smart grids.

Literature	Target models	Attack methods
Sayghe et al. [29,30]	MLP for FDIA detection	White-box: L-BFGS, JSMA, TFGSM
Li et al. [31]	MLP for FDIA detection	White-box: BIM
Yilmaz et al. [32]	LSTM for occupancy detection	White-box: AMLODA
Sun et al. [25]	DRL for decision making	White-box: critical point attack, antagonist attack
Chen et al. [33]	MLP/RNN/LSTM for STLF	White-box: BIM; black-box: learn-and-attack, gradient estimation
Li et al. [34]	MLP/RNN/CNN for energy theft detection (ETD)	White-box & black-box: SearchFromFree
Song et al. [35]	CNN for voltage stability assessment (VSA)	White-box & black-box: FGSM, PGD, DeepFool, C&W, UAP, UAN
Chen et al. [3]	MLP for power quality classification; RNN for building load forecasting	Black-box: FGSM
Niazazari et al. [36]	CNN for ECA	Black-box: FGSM, JSMA
Wang et al. [37]	MLP for NILM	Black-box: BIM

4. Adversarial poisoning attacks

4.1. General analysis

Considering the training process of a deep learning model, in many cases its training data may be gathered from the Internet and other unreliable data sources under the threat of adversarial poisoning attacks. Such a training process may mislead the resultant model to produce bad predictions on most input points (poisoning availability attack) or a few targeted input points (poisoning integrity attack) [22].

In [38] a generic white-box adversarial poisoning attack is formulated as a bilevel optimization problem as shown in the following formula:

$$\underset{D_{\mathrm{p}}}{\operatorname{argmax}} \mathrm{L}(D_{\mathrm{v}},\ \theta_{\mathrm{p}}) \quad s.t.\ \theta_{\mathrm{p}} \in \underset{\theta}{\operatorname{argmin}} \mathrm{L}(D_{\mathrm{tr}} \cup D_{\mathrm{p}}, \theta) \tag{2}$$

where D_{tr} represents the training set, D_{v} represents the validation set and D_{p} is a poisoned dataset. θ is a set of model weights and θ_{p} is an optimal set of model weights. L is the loss function. The obstruction of launching adversarial poisoning attacks lies in the generation of adversarial training examples which is more difficult than evasion attacks due to the aforementioned bilevel optimization problem.

To solve the bilevel optimization problem, a technique named back-gradient optimization has been exploited in [38] and extended to sequential inputs as well in [39]. Like evasion attacks, the extension of poisoning attacking to black-box settings can be implemented immediately through substitute models or training data.

4.2. Cases in smart grids

Adversarial poisoning attacks on deep learning models in smart grids have not drawn much attention yet. Marulli et al. [40] designed a black-box poisoning attack against ETD with a GAN where an LSTM acts as the discriminator and a CNN acts as the generator. Takiddin et al. [41] evaluated the effects of data poisoning attacks against both customer-specific and generalized detectors aiming at achieving robust ETD.

5. Countermeasures

With respect to the adversarial attacks on image classification current countermeasures can be classified from six operation aspects [19] as shown in Table 4, which works for smart grids in the same way.

Table 4. Countermeasures against adversarial attacks.

Operation aspect	Countermeasure
Gradient masking	Producing models with smoother gradients to hinder optimization-based attack algorithms
Auxiliary detection models	Using adversarial training to build an auxiliary binary model to check whether an input is adversarial or not
Statistical methods	Performing statistical comparison of the distribution of legitimate inputs with that of adversarial examples
Preprocessing techniques	Transforming a possibly adversarial input to a legitimate one by various preprocessing techniques
Ensemble of classifiers	Ensemble of multiple classification models that can be chosen at runtime
Proximity measurements	Utilizing proximity measurements of legitimate inputs and adversarial examples to the decision boundary

To resist adversarial attacks on FDIA detection, Li et al. [31] evaluated three defense methods: defensive distillation, adversarial training and adversarial detection, and analyzed their limitations. Song et al. [35] employed adversarial training, APE-GAN and their combination to investigate the defensive effectiveness against adversarial attacks on VSA. Niazazari et al. [36] proposed a defense mechanism utilizing adversarial training to robustify ECA performance. Marulli et al. [40] designed a GAN-concomitant detector for adversarial examples against ETD. Takiddin et al. [41] introduced a sequential ensemble detector consisting of an auto-encoder with attention, GRUs and feed forward layers for robust ETD.

6. Conclusions

Recent advancement of deep learning technology is pushing deep learning models into the center of intelligent tasks in smart grids. Various deep learning models have been deployed in smart grids to make best use of the huge amount of data generated on and on. Despite the great advantage of deep learning models, it has be proved that these models may face the threat of subtle adversarial examples. Therefore, we give a concise review of the adversarial attacks against deep learning models in smart grids in this paper. The research status of deep learning applications in smart grids and adversarial machine learning is briefly summarized. Adversarial evasion and poisoning attacks in smart grids are analyzed and exemplified respectively with emphasis. Typical countermeasures against these adversarial attacks are also presented. From the survey it can be concluded that the threat of adversarial attacks in smart grids will be a kind of long-term existence. As time goes on, the confrontation between adversarial attacks and defenses in smart grids will become more intense and need our continuous attention.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been partially supported by the Introduced Talent Research Start-up Fund of Nanchang Institute of Science & Technology (NGRCZX-20-12).

References

[1] Fang X, Misra S, Xue G, Yang D. Smart grid — the new and improved power grid: a survey. IEEE Commun Surv Tutor 2012;14(4):944–80.

[2] Gopstein A, Nguyen C, O'Fallon C, Hastings N, Wollman D. NIST framework and roadmap for smart grid interoperability standards, release 4.0. NIST Special Publication; 2021, p. 1108r4.

[3] Chen Y, Tan Y, Deka D. Is machine learning in power systems vulnerable? In: 2018 IEEE international conference on communications, control, and computing technologies for smart grids; 2018.

[4] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: 2nd international conference on learning representations; 2014.

[5] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: 3rd international conference on learning representations; 2015.

[6] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: 28th conference on neural information processing systems; 2014.

[7] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 2018;6:14410–30.

[8] Wiyatno RR, Xu A, Dia O, Berker A. Adversarial examples in modern machine learning: a review. 2019, arXiv Preprint. arXiv:1911.05268.

[9] Martins N, Cruz JM, Cruz T, Abreu PH. Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. IEEE Access 2020;8:35403–19.

[10] Carlini N, Wagner D. Audio adversarial examples: targeted attacks on speech-to-text. In: 2018 IEEE Symposium on Security and Privacy Workshops; 2018, pp. 1–7.

[11] Yuan C, Liu X, Zhang Z. The current status and progress of adversarial examples attacks. In: 2021 international conference on communications, information system and computer engineering, 2021, pp. 707–11.

[12] Massaoudi M, Abu-Rub H, Refaat SS, Chihi I, Oueslati FS. Deep learning in smart grid technology: a review of recent advancements and future prospects. IEEE Access 2021;9:54558–78.

[13] Tabakhpour A, Abdelaziz MMA. Neural network model for false data detection in power system state estimation. In: 2019 IEEE canadian conference of electrical and computer engineering; 2019.

[14] Bahrami S, Chen YC, Wong VWS. Deep reinforcement learning for demand response in distribution networks. IEEE Trans Smart Grid 2021;12(2):1496–506.

[15] Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 2019;10(1):841–51.

[16] Wang J, Kababji SE, Graham C, Srikantha P. Ensemble-based deep learning model for non-intrusive load monitoring. In: 2019 IEEE electrical power and energy conference; 2019.

[17] Niazazari I, Livani H, Ghasemkhani A, Liu Y, Yang L. Event cause analysis in distribution networks using synchro waveform measurements. 2020, arXiv Preprint 2020; arXiv:2008.11582.

[18] Ding X, Teng Y, Dai Y. Power grid fault diagnosis method based on CNN image recognition. In: 2020 international conference on aviation safety and information technology; 2020. p. 742–6.

[19] Machado GR, Silva E, Goldschmidt RR. Adversarial machine learning in image classification: a survey towards the defender's perspective. 2020, arXiv Preprint 2020; arXiv:2009.03728.

[20] Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst 2019;30(9):2805–24.

[21] Biggio B, Corona I, Maiorca D, Nelson B, Srndic N, Laskov P, Giacinto G, Roli F. Evasion attacks against machine learning at test time. In: 2013 European conference on machine learning and knowledge discovery in databases, part III; 2013, pp. 387–402.

[22] Demontis A, Melis M, Pintor M, Jagielski M, Biggio B, Oprea A, Nita-Rotaru C, Roli F. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In: 28th USENIX security symposium, 2019, p. 321–38.

[23] Papernot N, McDaniel P, Swami A, Harang R. Crafting adversarial input sequences for recurrent neural networks. In: 2016 IEEE military communications conference; 2016, pp. 49–54.

[24] Mode GR, Hoque KA. Adversarial examples in deep learning for multivariate time series regression. In: 2020 IEEE applied imagery pattern recognition workshop; 2020.

[25] Sun J, Zhang T, Xie X, Ma L, Zheng Y, Chen K, Liu Y. Stealthy and efficient adversarial attacks against deep reinforcement learning. In: 34th AAAI conference on artificial intelligence; 2020, pp. 5883–91.

[26] Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016, arXiv Preprint 2016; arXiv:1605.07277.

[27] Liu Y, Chen X, Liu C, Song D. Delving into transferable adversarial examples and black-box attacks. In: 5th international conference on learning representations; 2017.

[28] Tramer F, Papernot N, Goodfellow I, Boneh D, McDaniel P. The space of transferable adversarial examples. 2017, arXiv Preprint 2017; arXiv:1704.03453.

[29] Sayghe A, Zhao J, Konstantinou C. Evasion attacks with adversarial deep learning against power system state estimation. In: 2020 IEEE power & energy society general meeting; 2020.

[30] Sayghe A, Anubi OM, Konstantinou C. Adversarial examples on power systems state estimation. In: 2020 IEEE power & energy society innovative smart grid technologies conference; 2020.

[31] Li J, Yang Y, Sun JS, Tomsovic K, Qi H. Towards adversarial-resilient deep neural networks for false data injection attack detection in power grids. 2021, arXiv Preprint 2021; arXiv:2102.09057.

[32] Yilmaz I, Siraj A. Avoiding occupancy detection from smart meter using adversarial machine learning. IEEE Access 2021;9:35411–30.
[33] Chen Y, Tan Y, Zhang B. Exploiting vulnerabilities of load forecasting through adversarial attacks. In: 10th ACM International Conference on Future Energy Systems; 2019, pp. 1–11.
[34] Li J, Yang Y, Sun JS. SearchFromFree: adversarial measurements for machine learning-based energy theft detection. In: 2020 IEEE international conference on communications, control, and computing technologies for smart grids; 2020.
[35] Song Q, Tan R, Ren C, Xu Y. Understanding credibility of adversarial examples against smart grid: a case study for voltage stability assessment. In: 12th ACM international conference on future energy systems; 2021, pp. 95–106.
[36] Niazazari I, Livani H. Attack on grid event cause analysis: an adversarial machine learning approach. In: 2020 IEEE power & energy society innovative smart grid technologies conference; 2020.
[37] Wang J, Srikantha P. Stealthy black-box attacks on deep learning non-intrusive load monitoring models. IEEE Trans Smart Grid 2021;12(4):3479–92.
[38] Munoz-Gonzalez L, Biggio B, Demontis A, Paudice A, Wongrassamee V, Lupu EC, Roli F. Towards poisoning of deep learning algorithms with back-gradient optimization. In: 10th ACM workshop on artificial intelligence and security; 2017, pp. 27–38.
[39] Kravchik M, Biggio B, Shabtai A. Poisoning attacks on cyber attack detectors for industrial control systems. In: 36th Annual ACM symposium on applied computing; 2021, pp. 116–25.
[40] Marulli F, Visaggio CA. Adversarial deep learning for energy management in buildings. In: 2019 Summer Simulation Conference; 2019, no. 50.
[41] Takiddin A, Ismail M, Zafar U, Serpedin E. Robust electricity theft detection against data poisoning attacks in smart grids. IEEE Trans Smart Grid 2021;12(3):2675–84.