A Practical and Stealthy Adversarial Attack for Cyber-Physical Applications

Anonymous Author(s)

Abstract

Adversarial perturbations on misleading a well-trained machine learning (ML) model have been studied in computer vision (CV) and other related application areas. However, there is very limited focus on studying the impact of adversarial perturbations on ML models used in data-driven cyber-physical systems (CPSs) that normally have complex physical and mechanical constraints. Because of the complex physical and mechanical constraints, called domain-knowledge constraints in our paper, established gradient-based adversarial attack methods are not always practical in CPS applications. In this paper, we propose an innovative CPS-specific adversarial attack method that is able to practically compromise the ML-based decision makings of CPSs while maintaining stealthy by meeting the complex domain-knowledge constraints. Our work provides three main contributions: 1) we develop an unsupervised disentangled representation model to learn the explainable features of CPSs' sensing data. Using these feature maps, our proposed method can produce practical and stealthy adversarial perturbations; 2) our work provides a novel approach to synthesize adversarial perturbations where explainable features are selectively utilized, which leads to a more practical adversarial attack; 3) our proposed adversarial attack method does not require any explicit integration of domain-knowledge constraints in attack model formulation, resulting in a more general application scenario especially when domain-knowledge constraints cannot be represented in a mathematically differentiable form. In the section of performance evaluations, different scenarios are considered to illustrate the effectiveness of the proposed adversarial attack method in achieving a high success rate as well as sufficient stealthiness in CPS applications.

1 Introduction

In recent years, increasing evidence shows that carefully crafted adversarial perturbation is able to introduce bounded subtle adversarial perturbation that can mislead learning models to achieve incorrect decision making (Szegedy et al. 2013). Extensive research has been developed to study the impact of adversarial attack in different data-driven applications (Krizhevsky, Sutskever, and Hinton 2012; Lin et al. 2017; Vaswani et al. 2017; Devlin et al. 2018; Goodfellow, Shlens, and Szegedy 2014; Alzantot et al. 2018; Alzantot, Balaji, and Srivastava 2018; Carlini and Wagner 2017).

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

From the existing studies, it is clear that the vulnerability raised by adversarial attacks makes ML models not always trustworthy when being deployed in real-world applications such as self-driving car, face recognition, and Q&A systems. Therefore, it is crucial to sufficiently mitigate the adversarial perturbations. To realize successful mitigation strategy, it can be beneficial to first exploit adversarial mindset and formulate threat models of practical adversarial perturbations for a given application field.

To achieve this goal, many techniques have established to generate adversarial perturbations successfully misleading ML models on CV- and NLP- related application fields (Szegedy et al. 2013; Biggio et al. 2013; Goodfellow, Shlens, and Szegedy 2014; Carlini and Wagner 2017; Alzantot et al. 2018; Wallace et al. 2019; Zhang et al. 2020). However, there is very limited work focusing on generating practical adversarial attacks in CPS-related application domain that normally has complex domain-knowledge constraints. Due to the complex constraints, it can be challenging to design and launch adversarial perturbation practically. For example, the sensing data manipulated with adversarial perturbation may violate the constraints in CPSs and can be detected via built-in detectors that are conventionally designed based on domain-knowledge constraints. To address this challenge, we propose a practical and stealthy adversarial attack where adversarial perturbations are able to sufficiently mislead learning model in CPSs while bypassing the built-in detectors effectively. Concretely, our proposed CPS-specific adversarial attack method delivers three main contributions:

- We propose an unsupervised disentangled representation model to learn and interpret the features of CPSs' sensing data by disentangling the features into domain features, which are related to domain-knowledge constraints, and attribute features that are not highly correlated to the constraints. Using these explainable feature maps, our proposed method can produce practical and stealthy adversarial perturbations.
- Our method provides a novel and practical solution to effectively select and utilize explainable features for synthesizing adversarial perturbations in CPS domain.
- Our proposed method does not require any handcrafted domain knowledge to be integrated explicitly in the at-

tack model formulation. By doing so, the attacker is not required to have sufficient knowledge of the targeted CPS system. Additionally, this also results in a more general application scenario for our method, specially when the domain-knowledge constraints cannot be represented in a mathematically differentiable form.

The rest of the paper is organized as follows. In Section 2, we will review related work. In Section 3, we will introduce our proposed CPS-specific adversarial attack. In Sections 4 and 5, the performance evaluations and conclusions will be presented, respectively.

2 Related Work

In this section, we review the state-of-the-art works related to our proposed method.

2.1 White-Box and Black-Box Adversarial Attacks

Many previous works focus on generating adversarial perturbations based on the full knowledge of the targeted learning model, called white-box adversarial attack. Fast Gradient Sign Method (FGSM) defines the perturbation following the direction of gradient to maximize the loss function of learning model (Goodfellow, Shlens, and Szegedy 2014). Kurakin et al. introduced an iterative method to search optimal values for adversarial perturbations. Projected Gradient Descent (PGD) is another iterative method that implements projected gradient to restrict the scale of searched perturbation (Madry et al. 2017). Some following research work emphasized on how to enhance the computation efficiency of PGD (Shafahi et al. 2019; Zhang et al. 2019; Zhu et al. 2020). Additionally, Moosavi-Dezfooli proposed a method, called DeepFool, to find necessary perturbation to input examples of learning model (Moosavi-Dezfooli, Fawzi, and Frossard 2016). Papernot et al. used the first derivative of feed-forward neural network to compute the adversarial samples (Papernot et al. 2016). Carlini et al. formulated a new optimization instance with Lagrangian relaxation to bound the perturbation for adversarial training (Carlini and Wagner 2017). In many practical scenarios, the attacker does not have access to the target learning model. In this situation, a black-box adversarial attack is necessary. Transfer-based method was proposed to generate adversarial perturbations of a surrogate model to compromise target learning model in a black-box scenario (Liu et al. 2016; Papernot et al. 2017; Lu, Issaranon, and Forsyth 2017). Published experiment results show that the adversarial perturbations generated against surrogate model can be effective on compromising the target learning models. In order words, transferability can be maintained.

2.2 Adversarial Attack In CPSs

As far as we know, very limited research work has been done to introduce constrained adversarial perturbations in CPSs. In (Li et al. 2021), a search of optimal adversarial perturbation is considered as an optimization problem where domain-knowledge constraints are carefully and explicitly represented as linear equations or inequalities. The search is an iterative process to find proper perturbation which can both mislead learning model and fulfill the domain-knowledge constraints. This method is effective in the considered scenarios. However, this method requires examination of the integrity of domain-knowledge constraints in each iteration. Additionally, when constraints can not be represented in linear forms, the effectiveness of this method can be compromised.

2.3 Disentangled Representation

Disentangled representation learning focuses on extracting domain-invariant features from example pairs. Different methods have been developed to successfully extract the domain-invariant content features via an autoencoder model structure (Lee et al. 2018; Cheung et al. 2014; Mathieu et al. 2016). Adversarial training loss from Generative Adversarial Network (GAN) (Goodfellow et al. 2014) is implemented in the disentangled representation model to enforce the learning of disentangled representations. The work of Lee et al. (2018) considered that content and attribute features are designated as the distinct feature spaces for learning the disentangled representations of images. Additionally, this model assumes that the attribute feature aligns with prior Gaussian distribution.

3 Proposed CPS-Specific Adversarial Attack

In this section, we illustrate our proposed practical and stealthy CPS-specific adversarial attack.

3.1 Problem Formulation

In our work, we specify a threat model of our proposed CPS-specific adversarial attack as follows:

1. The attacker is assumed to have access to partial data of the targeted data-driven CPSs via eavesdropping and querying for information distillation.
2. Considering that the learning model F and the built-in detector B are critical for a data-driven CPS, F and B are always launched with advanced security measures. Therefore, it is reasonable to assume that the attacker does not have access to F and B. The attacker can realize information distillation about F by reconstructing a surrogate model F_s with the data obtained via eavesdropping and querying.
3. It is also reasonable to believe that the training procedure of the learning model F is conducted with advanced security measures. Because of it, we assume that the attacker cannot access the training procedure and the data used for the training procedure.
4. The objective of the attacker is to launch an adversarial attack that is able to sufficiently mislead the learning model F while bypassing the built-in detector B effectively.
5. The attacker has limited knowledge of the physical and mechanical constraints, called domain-knowledge constraints in our paper, of the targeted CPSs.

For simplicity, in our work, the learning model F in the targeted CPS is considered to make classification-based decision making, such as detecting false data injection attack in a power system and detecting vehicle state attack in a transportation system.

To realize a practical and stealthy attack, the adversarial perturbation r should be optimized to maximize the cross-entropy for misleading the target learning model F, which is described as:

$$r = \underset{||r||_\infty \leq \varepsilon}{\arg\min}\, log_2[P(y|x+r, F)]. \tag{1}$$

where x and y are a sensing data sample and its associated label, respectively, and ε is the upper bound of the infinite norm of perturbation r. However, since F and B are unknown to the attacker, the attacker needs to reconstruct a surrogate model F_s for information distillation about F. Therefore, Eq. (1) is reformulated as follows:

$$r = \underset{||r||_\infty \leq \varepsilon}{\arg\min}\, log_2[P(y|x+r, F_s)]. \tag{2}$$

In our work, we aim to build an effective surrogate model F_s, based on which the adversarial perturbation r is generated to not only mislead the learning model F and but also bypass the built-in detector B.

3.2 Framework Overview

The overview of our proposed framework for generating the adversarial perturbation r is illustrated in Fig. 1. As shown in Fig. 1, the framework mainly consists of three steps:

Step 1: A disentangled model is trained to learn and interpret the features of the sensing data samples by disentangling the features into domain features and attribute features. A completed disentangled model includes an encoder and a decoder. We only utilize the encoder in the following steps for generating adversarial perturbations.

Step 2: The encoder of the disentangled model is reused as a transfer-learning model to optimize the cascaded discriminator. The encoder and discriminator constitute the surrogate model F_s that is used for information distillation for the learning model F.

Step 3: The surrogate model F_s is utilized to generate the adversarial perturbations that have transferability to compromise the target learning model F. The attacker queries the target learning model F with the generated adversarial data sample $x + r$ for testing the effectiveness of the attack strategy.

In the next subsections, we detail the main steps of our proposed framework.

3.3 Disentangled Representation Model Design

As stated above, in our proposed framework, a disentangled representation model is designed for interpreting the features of the sensing data by disentangling the features into domain and attribute features, which enables stealthy adversarial attack. Inspired by the work of Lee at al (Lee et al. 2018), we propose a novel autoencoder structure by exploiting cycle-generative adversarial network (CycleGAN) (Zhu et al. 2017) for realizing the disentangled representation model. A data pair (x_1, x_2) sampled from the accessible dataset is fed to the proposed disentangled representation model shown in Fig. 2. Two encoders E_d and E_a are utilized to extract domain and attribute features, respectively. After the encoders, the extracted features of the two input data samples are mixed and fed to decoder D. The decoder generates two new fake data samples f_1 and f_2, which are associated with the real data x_1 and x_2, respectively. A discriminator C_f is implemented after the decoder to identify whether the input data is real or fake. The generated fake data are then encoded and decoded again in the same mixing pattern with the previous autoencoder. Therefore, the generated fake examples in the second round should be the same as or very close to the input data samples if the encoders and decoder are optimal.

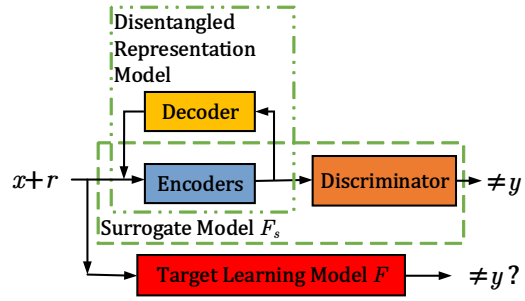


Figure 1: Overview of our proposed framework for generating CPS-specific adversarial attack.

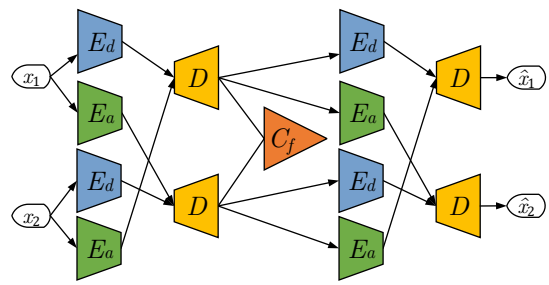


Figure 2: The overview of the proposed disentangled representation model.

During the training process of our proposed disentangled representation model, there are eight loss-function terms required to be minimized as stated in the followings:

1 Cycle-reconstruction loss: this loss represents the mean square error between the generated fake data $\hat{x}$ in the second round and the associated real data x:

$$L_1 = ||\hat{x}_1 - x_1||_2^2 + ||\hat{x}_2 - x_2||_2^2,$$
$$\text{where} \begin{cases} \hat{x}_1 = D(E_d(f_1), E_a(f_2)) \\ \hat{x}_2 = D(E_d(f_2), E_a(f_1)) \\ f_1 = D(E_d(x_1), E_a(x_2)) \\ f_2 = D(E_d(x_2), E_a(x_1)) \end{cases}. \tag{3}$$

2 Discriminator loss: this loss represents the binary cross-entropy of discriminator output. If the data sample is real,

the output is one. Otherwise, the output is zero. The loss can be formulated as:

$$\begin{aligned} L_2 = &- C_f(x_1)log_2(C_f(x_1)) - C_f(x_2)log_2(C_f(x_2)) \\ &- (1 - C_f(f_1))log_2(1 - C_f(f_1)) \\ &- (1 - C_f(f_2))log_2(1 - C_f(f_2)). \end{aligned} \tag{4}$$

3 Adversarial training loss: this loss represents the quality of the fake data samples generated by the autoencoder. The autoencoder belongs to generative model in GAN that performs adversarial optimization. In this case, the autoencoder aims to produce higher-quality fake data samples that are able to bypass the discriminator. The loss is formulated as:

$$\begin{aligned} L_3 = &- C_f(f_1)log_2(C_f(f_1)) \\ &- C_f(f_2)log_2(C_f(f_2)). \end{aligned} \tag{5}$$

4 Conditional-reconstruction loss I: if x_1 and x_2 share the different domain features and same attribute features, the fake example f_1 should be the same as or close to x_1. The same situation applies to f_2 and x_2. Therefore, the loss is used to optimize the model only if x_1 and x_2 share the same attribute features. The loss can be calculated as:

$$L_4 = ||f_1 - x_1||_2^2 + ||f_2 - x_2||_2^2. \tag{6}$$

5 Conditional-reconstruction loss II: if x_1 and x_2 share the same domain features and diverse attribute features, the fake data sample f_2 should be the same as or close to x_1. The same situation applies to f_1 and x_2. Therefore, the loss is used to optimize the model only if x_1 and x_2 share the same domain features. The loss can be calculated as:

$$L_5 = ||f_2 - x_1||_2^2 + ||f_1 - x_2||_2^2. \tag{7}$$

6 Cycle consistency loss: this loss represents the summation of the mean square error between the encoded features in the first round and the encoded features in the second round:

$$\begin{aligned} L_6 = &||E_d(f_1) - E_d(x_1)||_2^2 + ||E_d(f_2) - E_d(x_2)||_2^2 \\ &+ ||E_a(f_1) - E_a(x_2)||_2^2 + ||E_a(f_2) - E_a(x_1)||_2^2. \end{aligned} \tag{8}$$

7 Conditional pair loss I: if x_1 and x_2 share the same domain features and diverse attribute features, $E_d(x_1)$ and $E_d(x_2)$ should be close to each other. Additionally, $E_a(x_1)$ and $E_a(x_2)$ should be very different from each other. The loss can be represented as:

$$L_7 = ||E_d(x_1) - E_d(x_2)||_2^2 - ||E_a(x_1) - E_a(x_2)||_2^2. \tag{9}$$

8 Conditional pair loss II: if x_1 and x_2 share the same attribute features and diverse domain features, $E_a(x_1)$ and $E_a(x_2)$ should be close to each other and the domain features $E_d(x_1)$ and $E_d(x_2)$ should be very different from each other. Thus, the loss can be represented as:

$$L_8 = ||E_a(x_1) - E_a(x_2)||_2^2 - ||E_d(x_1) - E_d(x_2)||_2^2. \tag{10}$$

3.4 Surrogate Model Construction

Once the disentangled representation model is well-trained, we utilize the encoders E_d and E_a as the initial model to transfer domain knowledge to the surrogate model. The outputs of E_d and E_a are concatenated and fed to a new discriminator C_c to learn the task of the target learning model. Figure 3 illustrates the structure of the surrogate model. Since domain and attribute encoders are well-trained in the previous step, their parameters are fixed and only the parameters of the following discriminator C_c are updated during the training of surrogate model. In other words, the construction of surrogate model can be viewed as a transfer-learning process where a classifier is built on the existing model. The

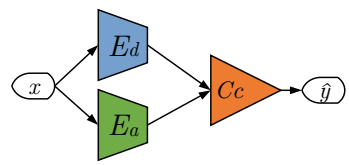


Figure 3: Network structure for surrogate model.

architecture of the discriminator C_c can have an arbitrary structure, since the attacker cannot access the target learning model F as stated in our threat model.

3.5 Generation of Adversarial Perturbation

In our proposed method, the generation of adversarial attack on the surrogate model is based on gradient-based algorithms. Two gradient-based algorithms are considered in our current work, including Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) and Projected Gradient Descent (PGD) (Madry et al. 2017). The loss of generating adversarial perturbation is formulated to maximize the difference between the model prediction $\hat{y}$ and the associated true label y, which is shown as follows:

$$L_a = -distance(\hat{y}, y), \tag{11}$$

where $\hat{y} = F_s(x)$. Additionally, to enable a stealthy adversarial attack, the adversarial perturbation needs to be able to bypass the built-in detector that is designed based on domain-knowledge constraints. Considering that domain features characterize the inherent features associated with domain-knowledge constraints, the impact of the adversarial perturbation on the domain features needs to be minimized. To realize this goal, in our proposed method, the domain encoder E_d is selected to be detached when executing back-propagation on the model F_s for generating adversarial perturbation. By doing so, the gradient calculation is restricted to only follow the direction of extracted attribute features.

4 Performance Evaluations

In this section, we will evaluate the performance of our proposed CPS-specific adversarial attack from the perspectives of efficiency on misleading the target learning model and the stealthiness via bypassing the built-in detector effectively. We consider two CPS case studies for performance evaluations: one is about detecting false data injection attack in a

power system, and the other is about detecting vehicle state attack in a transportation system.

4.1 Case Study I

In this case study, we consider a CPS scenario where a learning model F is deployed to detect false data injection (Liu, Ning, and Reiter 2011) in IEEE 39-Bus System that has 10 generators and 46 power lines (Athay, Podmore, and Virmani 1979). Additionally, a built-in residual-based detector is considered, which is formulated as follows:

$$||z + a - Hs'||_2 \leq \alpha. \tag{12}$$

where a denotes an attack vector, z denotes the original benign measurement data, H denotes the a constant matrix characterizing physical constraints of the power system, s' is the state estimation based on the measurements potentially compromised by the attack vector a, and α is the threshold of the built-in detector. Using our proposed method, an adversarial attack is generated to mislead the learning model F sufficiently while bypassing the built-in detector formulated in Eq. (12) effectively.

In our case study, the sensing dataset are collected from simulations with multiple load profiles and two topology profiles. We consider the constraints represented by the topology profiles correspond to the domain features in our disentangled representation model. Additionally, the information represented by load profiles corresponds to the attribute features. Each training data sample pair for training the disentangled representation model is randomly sampled from the dataset. If the pair shares the same load profile, conditional-reconstruction loss I and conditional pair loss II are applied. If the pair shares the same topology profile, conditional-reconstruction loss II and conditional pair loss I are applied. All the other four loss-function terms defined in Section 3.3 are always applied no matter which pair is fetched for optimizing the model. In addition, 2000 data samples are collected, of which 900 data samples are used for realizing the learning model F for detecting the false data injection attack in the power system and 1100 data samples are considered to be accessible to the attacker for developing and deploying adversarial attack. Furthermore, in our case study, we consider the target learning model F is in the form of either fully connected neural network (FCNN) or recurrent neural network (RNN).

For generating the adversarial perturbation, we first train a RNN-based disentangled model to extract the domain and attribute features. Then the encoders of this model is utilized as a pre-trained model to fine-tune a surrogate model. The training data samples for the surrogate model include the benign data samples and the data manipulated via false data injection attack. The domain encoder is selectively detached during the calculation of the gradient for perturbation generation. As a comparison, we also introduce general gradient-based adversarial attack methods as baseline methods. The success rates of our proposed CPS-specific adversarial attack on misleading the decision making by exploiting using FGSM and PGD gradient-based algorithms are shown in Figs. 4 and 5, respectively. As illustrated in the plots, our proposed method is able to achieve comparable success rates compared with the baseline methods. Additionally, our proposed method shows better transferability on misleading FCNN-based target learning model compared with misleading RNN-based learning model.

We continue to evaluate the capacity of our proposed adversarial attack on bypassing the built-in detector. The bypassing capacity g is formulated as:

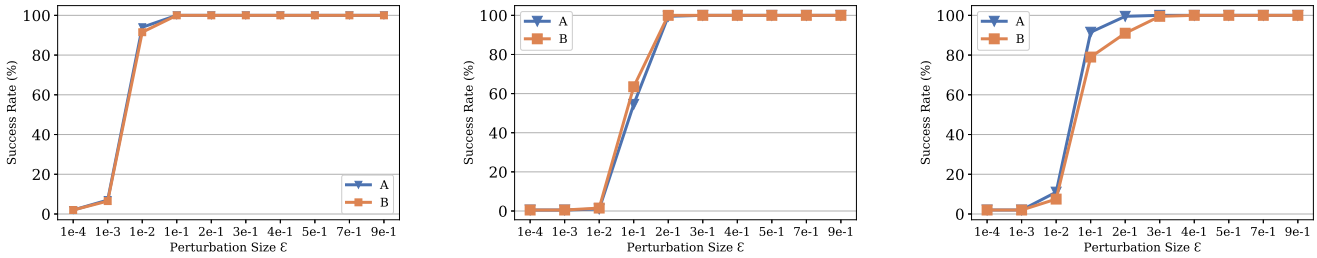
$$g = \min\Big(\frac{1}{\log_2(||z + a - Hs'||_2)}, 1\Big). \tag{13}$$

The performance evaluation is shown in Fig. 6. From Fig. 6, it is clear that our method outperforms the baseline methods. Additionally from Figs. 4 to 6, we can get that our proposed adversarial attack is able to achieve a good trade-off between the misleading efficiency and bypassing capability.

4.2 Case Study II

In this case study, we leverage the Veremi dataset for performance evaluation by considering a scenario of detecting vehicle state attack in a transportation system (van der Heijden, Lukaseder, and Kargl 2018). The dataset collects the vehicle state message from 37500 vehicles in Luxembourg SUMO Traffic scenario (Codecá et al. 2017). The vehicle state message contains the positions of transmitter and receiver, speed, and elapsed time. The injected vehicle state attack includes constant value, constant offset, random value, random offset, and eventual stop (van der Heijden, Lukaseder, and Kargl 2018). In this transportation system, a deep learning model is deployed to detect the vehicle state attack. Additionally, the system also has two types of built-in detectors: maximum appearance distance (SAW) based detector and estimated reception range (ART) based detector. SAW-based detector detects the vehicle state attack by measuring the moved distance between samples and identifying the unreasonable moving distance. ART-based detector detects the attack by estimating the communication distance between the transmitter and receiver and identifying the unreachable vehicles.

In this scenario, our proposed method firstly trains a RNN-based disentangled model to extract domain and attribute features. Since no prior domain knowledge is assumed in this scenario, we consider the common features shared by all the data as the domain features and the other features as the attribute features. In this case, all the loss-function terms except Conditional Reconstruction Loss II and Conditional Pair Loss II are implemented in this scenario. Secondly, we implement a RNN-based surrogate model. We evaluate the performance of our method by considering that the target learning model is based on either FCNN or RNN, which is shown in Fig. 7. From the plots, it is clear that our method outperforms the baseline, general FGSM method, in misleading both FCNN-based and RNN-based target learning model. We also evaluate the capability of our method on bypassing the SAW-based and ART-based detectors. The bypassing capacity is formulated as the failure percentage of detecting adversarial perturbations by the built-in detectors. The performance evaluation is shown

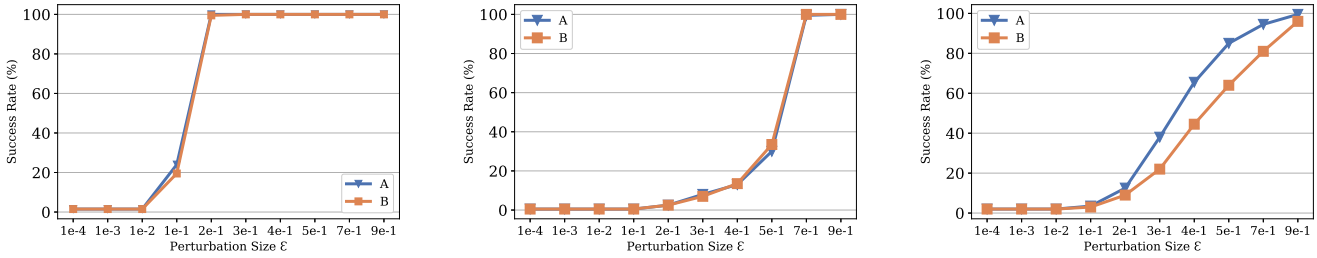


(a) Effectiveness on misleading the surrogate model.

(b) Effectiveness on misleading the FCNN-based target learning model.

(c) Effectiveness on misleading the RNN-based target learning model.

Figure 4: Adversarial attack is generated based on FGSM: A: General FGSM and B: Our method.



(a) Effectiveness on misleading the surrogate model.

(b) Effectiveness on misleading the FCNN-based target learning model.

(c) Effectiveness on misleading the RNN-based target learning model.

Figure 5: Adversarial attack is generated based on PGD: A: General PGD and B: Our method.

in Fig. 8, from which we can observe that our method outperforms the baseline method especially for ART-based detector. Furthermore, in this case study, we also include a method of detaching attribute encoder for evaluation purpose. Since our method detaches the domain encoder during calculating adversarial perturbation, attribute-encoder detached method is considered as a comparison to illustrate the importance of selectively detaching domain encoder in our design. The performance comparisons between our method and the attribute-encoder detached method in Figs. 7 and 8 illustrate the the effectiveness of our proposed method to interpret the data features by extracting domain and attribute features and selectively detach features for generating adversarial perturbations.

5 Conclusions

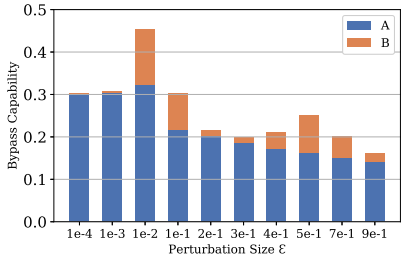
In this paper, we propose a novel CPS-specific adversarial attack method that is able to compromise the learning model of a data-driven CPS in a practical and stealthy manner. Our work presents three main contributions. Firstly, our method enables an unsupervised disentangled representation model for learning and interpreting the data features by disentangling the features into domain features and attribute features. Using the obtained explainable feature maps, it is feasible to produce practical and stealthy adversarial perturbations. Secondly, our work provides a novel approach to synthesize adversarial perturbations where explainable features are selectively utilized, which leads to a more practical adversarial attack. Thirdly, our adversarial attack method does not require any explicit integration of domain-knowledge constraints in attack model formulation, resulting in more general application scenarios especially when the attacker has limited knowledge of the targeted CPSs or the domain-knowledge constraints cannot be represented in a mathematically differentiable form. As illustrated in the simulation results, our proposed method is able to sufficiently mislead the learning model in the target CPSs while effectively bypassing the built-in detector that is normally designed based on physical and mechanical constraints of the CPSs. In our ongoing work, we are working on evaluating our proposed method in other CPS domains and exploring a more general form of our proposed adversarial attack which can be suitable for various CPS applications.

References

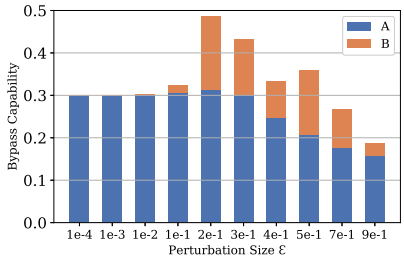
Alzantot, M.; Balaji, B.; and Srivastava, M. 2018. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*.

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Athay, T.; Podmore, R.; and Virmani, S. 1979. A practical method for the direct analysis of transient stability. *IEEE Transactions on Power Apparatus and Systems*, 12(2): 573–584.

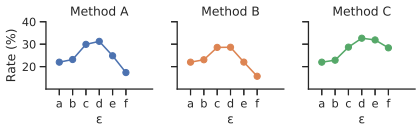


(a) FGSM.

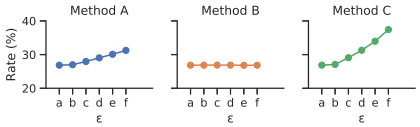


(b) PGD.

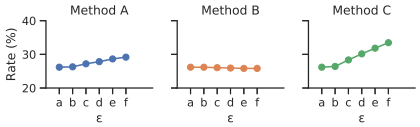
Figure 6: Bypass capacity of the proposed adversarial attacks: A: Baseline method and B: Our method.



(a) Effectiveness on misleading the surrogate model with different perturbation sizes ε.

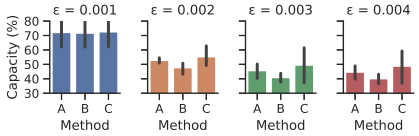


(b) Effectiveness on misleading the FCNN-based target learning model with different perturbation sizes ε.

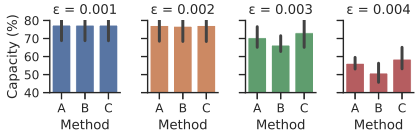


(c) Effectiveness on misleading the RNN-based learning model with different perturbation sizes ε.

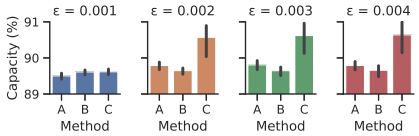
Figure 7: Success rate of our proposed adversarial attack: A: Baseline method, B: Attribute-encoder detached method, and C: Our method; perturbation size, ε - a: 0, b: 0.0001, c: 0.001, d: 0.002 e: 0.003, f: 0.004.



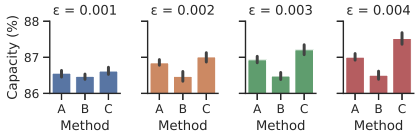
(a) Bypass capability of SAW-based detection at 50 meters with different perturbation sizes ε.



(b) Bypass capability of SAW-based detection at 100 meters with different perturbation sizes ε.



(c) Bypass capability of ART-based detection at 200 meters with different perturbation sizes ε.



(d) Bypass capability of ART-based detection at 300 meters with different perturbation sizes ε.

Figure 8: Bypass rate of proposed adversarial attack. Method - A: Baseline method, B: Attribute-encoder detached method, and C: Our method.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402. Springer.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Cheung, B.; Livezey, J. A.; Bansal, A. K.; and Olshausen, B. A. 2014. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*.

Codecá, L.; Frank, R.; Faye, S.; and Engel, T. 2017. Luxembourg sumo traffic (lust) scenario: Traffic demand evaluation. *IEEE Intelligent Transportation Systems Magazine*, 9(2): 52–63.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y.

2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 35–51.

Li, J.; Yang, Y.; Sun, J. S.; Tomsovic, K.; and Qi, H. 2021. Conaml: Constrained adversarial machine learning for cyber-physical systems. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 52–66.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Liu, Y.; Ning, P.; and Reiter, M. K. 2011. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1): 1–33.

Lu, J.; Issaranon, T.; and Forsyth, D. 2017. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, 446–454.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mathieu, M.; Zhao, J.; Sprechmann, P.; Ramesh, A.; and LeCun, Y. 2016. Disentangling factors of variation in deep representations using adversarial training. *arXiv preprint arXiv:1611.03383*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387. IEEE.

Shafahi, A.; Najibi, M.; Ghiasi, A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *arXiv preprint arXiv:1904.12843*.

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

van der Heijden, R. W.; Lukaseder, T.; and Kargl, F. 2018. Veremi: A dataset for comparable evaluation of misbehavior detection in vanets. *arXiv preprint arXiv:1804.06701*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

Zhang, D.; Zhang, T.; Lu, Y.; Zhu, Z.; and Dong, B. 2019. You only propagate once: Accelerating adversarial training via maximal principle. *arXiv preprint arXiv:1905.00877*.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3): 1–41.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *International Conference on Learning Representations*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.