On the Robustness of Intrusion Detection Systems for Vehicles Against Adversarial Attacks

Jeongseok Choi and Hyoungshick Kim(✉)

Sungkyunkwan University, Suwon, Republic of Korea
{wjdtjr123,hyoung}@skku.edu

Abstract. Because connected cars typically have several communication capabilities (through 5G, WiFi, and Bluetooth), and third-party applications can be installed on the cars, it would be essential to deploy intrusion detection systems (IDS) to prevent attacks from external attackers or malicious applications. Therefore, many IDS proposals have been presented to protect the controller area network (CAN) in a vehicle. Some studies showed that deep neural network models could be effectively used to detect various attacks on the CAN bus. However, it is still questionable whether such an IDS is sufficiently robust against adversarial attacks that are crafted aiming to target the IDS. In this paper, we present a genetic algorithm to generate adversarial CAN attack messages for Denial-of-Service (DoS), fuzzy, and spoofing attacks to target the state-of-the-art deep learning-based IDS for CAN. The experimental results demonstrate that the state-of-the-art IDS is not effective in detecting the generated adversarial CAN attack messages. The detection rates of the IDS were significantly decreased from 99.27%, 96.40%, and 99.63% to 2.24%, 11.59%, and 0.01% for DoS, fuzzy, and spoofing attacks, respectively.

Keywords: Controller area network (CAN) · Adversarial attack · Intrusion detection system

1 Introduction

Recent advances in the automobile industry make our cars more intelligent and more connected. Therefore, a typical vehicle is composed of about 150 electronic control units (ECU), and each ECU communicates with other ECUs using a bus system [11,18]. Controller area network (CAN) is a representative bus system for in-vehicle networks and supports efficient communication between ECUs [3]. However, because CAN basically uses a broadcast-based communication mechanism without applying authentication and encryption schemes, it is known to be vulnerable to fabricated messages injected through the on-board diagnostic (OBD-II) port [1], and remote network channels [10].

It is essential to prevent cyber attacks over CAN because such attacks could severely threaten drivers' safety. Therefore, many intrusion detection systems

H. Kim (Ed.): WISA 2021, LNCS 13009, pp. 39–50, 2021.
https://doi.org/10.1007/978-3-030-89432-0_4

(IDS) [8] have been introduced to detect suspicious messages on CAN. Furthermore, in recent years, as deep learning technologies have achieved remarkable success in various domains, some researchers have tried to build deep learning-based models achieving a high detection rate [6,7,12,13,17]. The DNN-based IDSs have been primarily designed to mitigate three types of cyber attacks: Denial-of-Service (DoS), fuzzy, and spoofing attacks. To launch a DoS attack, an attacker typically generates dummy messages with the highest priority and then exhaustively injects those messages to take over the CAN bus so that normal messages are not delivered on time. To launch a fuzzy attack, an attacker generates attack messages with random ID values and then injects those messages to induce faults in a victim's vehicle. To launch a spoofing attack, an attacker generates attack messages with specific ECU ID values to perform her desired functions.

However, recent studies (e.g., [5]) showed that deep learning models could be vulnerable to adversarial attacks, which are crafted by adding intentionally generated distortions onto normal inputs in a sophisticated manner. Therefore, we are motivated to analyze the robustness of DNN-based IDS against adversarial attacks. To achieve this goal, we evaluate the robustness of the state-of-the-art DNN-based IDS [13] against adversarial attacks based on a genetic algorithm. The chosen model [13] is known as one of the most advanced DNN-based IDSs, achieving detection rates of 99.27%, 96.40%, and 99.63% for each attack of DoS, fuzzy, spoofing attacks.

We aim to generate a sequence of attack messages that the IDS [13] cannot easily detect by modifying only a few bits of original attack messages that are contained in the public dataset. To demonstrate the feasibility of our adversarial attacks, we performed experiments under the same settings of the previous work [13]. The experimental results show that the detection rates of the DNN model were significantly decreased from 99.27%, 96.40%, and 99.63% to 2.24%, 11.59%, and 0.01% for DoS, fuzzy, and spoofing attacks, respectively, when we can generate effective attack messages by modifying only 1, 1, and a few bits for each of the DoS, fuzzy, and spoofing attacks. The main contributions of this work can be summarized below:

- We propose a genetic algorithm-based framework that generates DoS, fuzzy, and spoofing attacks to evade a target IDS for CAN (see Sect. 3);
- We evaluated the robustness of the state-of-the-art IDS [13] against adversarial attacks and showed that the model's performance could be significantly degraded with such adversarial attacks (see Sect. 4);
- We publicly release our tool and dataset for adversarial attacks (see https://github.com/jschoi0126/adversarial-attack-on-CAN-IDS).

2 Target IDS

Recently, there were many proposals to develop an IDS to detect suspicious sequences of CAN messages. In this paper, we specifically chose the model [13] as our target model because it is one of the most advanced DNN-based IDSs,

achieving detection rates of 99.27%, 96.40%, and 99.63% for each attack of DoS, fuzzy, spoofing attacks, and its dataset is publicly available online, which is helpful to reimplement the original model proposed in [13]. To detect suspicious attack messages, Song et al. suggested a reduced Inception-ResNet, as shown in Fig. 1, which was built by reducing the number of parameters and simplifying the structure of a layer of the Inception-ResNet [15].

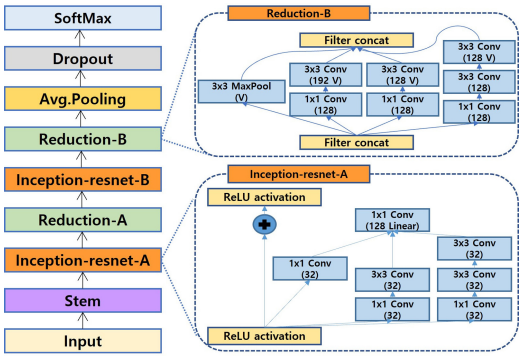

Fig. 1. Reduced Inception-ResNet architecture in [13].

Because the Inception-ResNet is a model which was initially implemented as an image classifier, the model in [13] first transforms a sequence of CAN messages into an image and then processes it. To represent each CAN message, we use the 29-bit message identifier (ID) only consisting of the regular 11-bit base identifier and an 18-bit identifier extension (see Fig. 2).

Fig. 2. Structure of the CAN message ID.

Figure 3 shows the transformation process of a sequence of 29 CAN messages to create the input of the deep learning model. With 29 CAN messages, we generate a 29×29 image where each message is mapped to a 29-bit binary sequence.

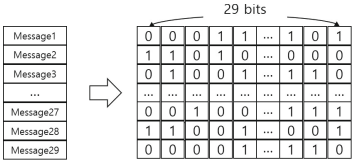

Fig. 3. Transformation of CAN messages into an image.

3 Methodology

In this section, we first explain our threat model and provide an overview of the proposed framework to generate adversarial attacks that cannot be detected by the target IDS for CAN.

3.1 Threat Model

In our threat model, we assume that the attacker can monitor all network messages over CAN because there is no encryption process in CAN. In addition, since the CAN bus does not require an additional authentication process, the attacker can inject the message with an arbitrary ID and modify the message payload if it is needed.

DoS Attack. The CAN bus selects the message to be transmitted depending on the priority of ECU when several ECUs transmit their messages simultaneously. The priority of a message transmitted from an ECU is higher when its ECU ID is smaller. An attacker abuses this priority scheme for CAN by setting the ECU ID of attack messages as 0. The DoS attack aims to transmit a massive number of attack messages having the highest priority for preventing the transmissions of normal messages over CAN.

Fuzzy Attack. A fuzzy attack injects a message with a random ID. The purpose of this attack is to cause some anomaly in the target vehicle. For example, some warning messages could be displayed on the target vehicle's dashboard due to a fuzzy attack.

Spoofing Attack. The spoofing attack is performed by targeting the specific ECU ID of the vehicle. An attacker injects a CAN message targeting a specific ECU ID to perform the attacker's desired function in the spoofing attack.

3.2 Overall Framework

Figure 4 shows the proposed framework to generate adversarial CAN messages. Our framework is designed to modify existing attack messages in feature space so that the target IDS cannot detect the modified messages. That is, our framework does not change normal messages. However, for a given sequence of attack messages, attackers can inject dummy messages or modify some parts of attack messages if those modifications do not change the attack effects. Our framework iteratively modifies a given sequence of attack messages using perturbations selected by a genetic algorithm. Perhaps, the modifications applied would be insufficient to bypass the IDS. To validate the effectiveness of attack messages, we rerun the target IDS with the modified attack messages. When the target IDS

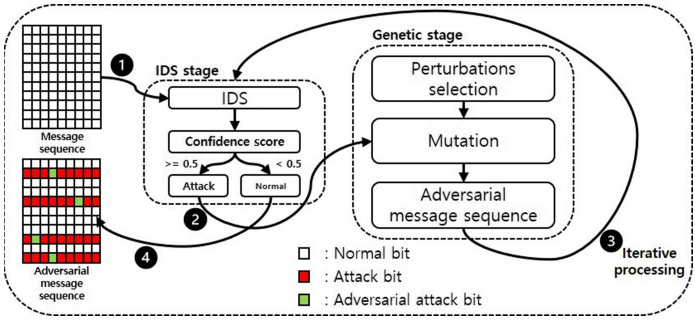

Fig. 4. Proposed framework to generate attack messages.

does not detect them, the modifications by our framework would be practical to bypass the target IDS.

The process of generating attack messages is presented in Algorithm 1. We use a genetic algorithm to generate modified attack messages with effective perturbations. Algorithm 1 has $Attack_{origin}$, P_{size}, and G_{max} as the input, and produces $Attack_{adv}$ as the output where $Attack_{origin}$ represents the sequence of attack messages; P_{size} represents the population size; and G_{max} represents the maximum number of generation. This process begins by initializing the population with P_{size} copies generated from the original sequence of attack messages with mutations. The $Mutate$ operation selects a random bit and flips it with some probability. We use POP_i to represent the ith population. POP_0 represents the initial population.

For the ith generation, we calculate the confidence score of each individual in POP_i using the target IDS. Suppose there exists an individual with a confidence score is less than 0.5. In that case, the algorithm produces the individual as the sequence of attack messages ($Attack_{adv}$) because in the current target IDS, if the confidence score of a given sequence of attack messages is less than 0.5, the IDS classifies it as a benign sequence of CAN messages. Consequently, $Attack_{adv}$ would become an effective sequence of attack messages to evade the target IDS. However, if the confidence score is higher than 0.5, the IDS classifies it as a sequence of attack messages. Therefore, in this case, we need to continue the message sequence modification process. We select two individuals x and y from POP_i with probability inversely proportional to their confidence scores. We first perform the $Crossover$ operation with x and y to generate z by choosing each bit in z from either x or y with equal probability. Next, we perform the $Mutate$ operation to update z by flipping a random bit in z with some probability.

3.3 Adversarial DoS Attack

For an adversarial DoS attack, attack messages should have the highest priority of ECU ID. In each CAN message, the 3 bits are used to represent its priority. Therefore, attackers can only modify the other remaining bits in the message to preserve the same effect of the original DoS attack.

Algorithm 1: Generation of adversarial attack messages

Input: $Attack_{origin}, P_{size}, G_{max}$
Output: $Attack_{adv}$
for ($i = 0;\ i < P_{size};\ i++$) {
  $Attack_{copy} \leftarrow Mutate(Attack_{origin})$;
  Add $Attack_{copy}$ to POP_0;
for ($j = 0;\ j < G_{max};\ j++$) {
  Calculate the confidence score of each individual in POP_i using the IDS;
  if *There exists an individual with a confidence score* < 0.5 **then**
    Output the individual as $Attack_{adv}$;
    Terminate the attack message generation process;
  for ($i = 0;\ i < P_{size};\ i++$) {
    Select x from POP_i with probability inversely proportional to its confidence score;
    Select y from POP_i with probability inversely proportional to its confidence score;
    $z \leftarrow Crossover(x, y)$;
    $z \leftarrow Mutate(z)$;
    Add z to POP_{i+1};
Print "We failed to produce attack messages.";

Figure 5 shows an example of a sequence of DoS attack messages where the left figure represents the sequence of original attack messages (detected by the target IDS) and the right figure the sequence of adversarial attack messages generated from the original attack messages. In each figure, the ith row represents the ith message where the bit '0' is mapped to black and the bit '1' is mapped to white – the lines of all zero bits represent the dummy messages generated explicitly for the DoS attack. Interestingly, we can see that a sequence of adversarial attack messages can effectively be generated, which is not detected by the target IDS, by setting '1' at a bit in each dummy message consisting of all zero bits from the sequence of original attack messages.

(a) Original (b) Adversarial

Fig. 5. Example of adversarial DoS attacks.

3.4 Adversarial Fuzzy Attack

We can modify any bits randomly for each message used for a fuzzy attack because it still has a random ECU ID. Therefore, we can try to generate a sequence of adversarial attack messages without any restriction. Figure 6 shows that setting a single '1' bit is sufficient to generate an effective sequence of adversarial attack messages, which is not detected by the target IDS.

(a) Original (b) Adversarial

Fig. 6. Example of adversarial fuzzy attacks.

3.5 Adversarial Spoofing Attack

For adversarial spoofing attacks, attack messages are divided into two types of messages: dummy messages and ECU-control messages. For dummy messages, an attacker arbitrarily modifies all 29 bits. However, for ECU-control messages, we cannot modify any bits to preserve the attack effects. Therefore, our adversarial attacks should be performed by modifying dummy messages only. Figure 7 shows that a large number of bits are newly set to '1' to generate an effective sequence of adversarial attack messages, which is not detected by the target IDS.

(a) Original (b) Adversarial

Fig. 7. Example of adversarial spoofing attacks.

4 Experiments

4.1 Experiment Setup

Datasets. We use the same dataset and data configuration used in our target IDS [12,13]. In the dataset, the CAN messages (consisting of 26 CAN IDs) were collected from the ECUs in a real vehicle while the vehicle is running. The dataset is divided into normal and three attack types (DoS, fuzzy, and spoofing) of CAN messages. The spoofing attacks were specifically designed for RPM spoofing attacks. Those attack messages were generated by injecting attack messages periodically in a controlled environment. Each message injection process was performed during a period from 3 to 5 s. Table 1 shows the number of CAN messages for each attack type.

Table 1. Description of the dataset used in experiments.

Attack type	Nomral messgae	Injected message
DoS attack	3,078,250 (84%)	587,521 (16%)
Fuzzy attack	3,347,013 (87%)	491,847 (13%)
Spoofing attack	2,290,185 (78%)	654,897 (22%)

Implementation Details. The target IDS was implemented by using TensorFlow https://www.tensorflow.org/. The model was trained with the total epochs of 10, the batch size of 128. The model with the best performance was saved for our experiments. For the genetic algorithm, we set the population size $P_{size} = 100$, and the maximum number of generations $G_{max} = 75$.

4.2 Detection Rate

Table 2 shows that the detection rates of the target IDS [13] against original attacks and adversarial attacks. The detection rates of the IDS were significantly decreased from 99.27%, 96.40%, and 99.63% to 2.24%, 11.59%, and 0.01% for DoS, fuzzy, and spoofing attacks, respectively. We can see that the target IDS would be ineffective in detecting the generated adversarial CAN attack messages even though the model achieved high detection rates for the original attack message sequences.

We surmise that the characteristics of the original attack messages used to train the target model may explain this inferiority in the detection performance. We can see that the original attack messages used only 11-bit identifiers in the standard format (see Fig. 5, 6, and 7). Therefore, if we use the other remaining bits in the extended format to generate adversarial attack messages, such attack messages would be unseen and new in the view of the target model.

Interestingly, it seems relatively harder to generate adversarial attack messages for fuzzy attacks compared with other attacks. The detection rate of the target IDS for fuzzy attacks is 11.59%, whereas the detection rate of the target

Table 2. Detection rates of the target IDS [13] against original attacks and adversarial attacks.

Attack type	Original attacks	Adversarial attacks
DoS attack	99.27 %	**2.24 %**
Fuzzy attack	96.40 %	**11.59 %**
Spoofing attack	99.63 %	**0.01 %**

IDS for spoofing attacks is 0.01%. The structure of the attack messages for a fuzzy attack is not fixed and relatively dynamic because attack messages have random IDs. For adversarial fuzzy attacks, the attack messages are still generated with random IDs, which may hold similar characteristics to the original fuzzy attack messages having random IDs. In contrast, spoofing attacks have a predictable specific pattern that can effectively be trained by a DNN model. Therefore, attackers would easily generate adversarial attack messages that seem unseen and new in the view of the DNN model.

4.3 Number of Generations

To show the efficiency of the proposed genetic algorithm, we count the number of generations taken to generate successful attack messages that can evade the detection of the target IDS. Table 3 shows the mean number of generations taken to effective attack messages for each attack type. Surprisingly, we could generate effective attack messages for all attack types within eight generations on average. In particular, it can generate an effective sequence of attack messages with the mean generations of 1.016, indicating that the target IDS would be overfitted and lack generalization capabilities.

Table 3. Average iteration for each attack.

Attack type	DoS attack	Fuzzy attack	Spoofing attack
# of iterations	5.831	7.674	1.016

5 Related Work

5.1 IDS Proposals for CAN

In the CAN, ECUs need to communicate with each other. However, the existing CAN communication mechanism has no security features such as encryption and authentication. Therefore, several defense mechanisms have been proposed to prevent cyber attacks over CAN. Furthermore, with the recent advancement

of deep learning technology, a research trend is to develop a deep learning-based IDS.

Kang et al. [7] presented a classifier using a deep belief network structure and conducted experiments with a synthetic dataset that was created through open car tested and network experiments (OCTANE) [2]. Taylor et al. [17] proposed an IDS using a long-short term memory (LSTM) based model to detect attack messages over CAN. They constructed an individual model for every ECU ID with the 64-bits CAN data field as the input. Seo et al. [12] proposed an IDS using a generative adversarial network. They tried to build a one-class classification model to detect unseen attacks using only normal data. The proposed model was trained with the dataset generated from a real vehicle's CAN messages. Song et al. [13] introduced an IDS using a deep convolutional neural network (DCNN). They used a simple data assembly module that efficiently converts the CAN bus data into a grid-like structure fitted to the DCNN.

5.2 Adversarial Attacks

Many recent studies demonstrated that deep learning models are vulnerable to adversarial attacks. The traditional idea of an adversarial attack is to add a small amount of well-tuned additive perturbation to the input. This attack causes the target classifier to label the modified input as a different class. Szegedy et al.'s pioneer work [16] showed that artificial perturbations, which can be manipulated by several gradient-based algorithms using backpropagation, can trick deep learning models into erroneous outcomes. Goodfellow et al. [4] showed that misclassification for adversarial examples could be originated from the linear nature of neural networks and a high dimensional input space called gradient sign method (FGSM). Mardy et al. [9] demonstrated that the projected gradient descent (PGD), a multi-step variant of FGSM, can make the adversarial attack more effective.

However, it is still unclear how we can generate adversarial attacks to evade the detection of a target IDS. It would be challenging to apply gradient-based adversarial attack generation methods to the IDSs over CAN because the confidence score functions of the IDS would generally not be differentiable.

Su et al. [14] presented a method modifying only a few pixels of the input image to fool deep neural networks with a differential evolution algorithm. They showed that a target classifier could be tricked into recognizing the modified input image as a different class. We extend their work into the CAN IDS area. We introduce a framework using a genetic algorithm to generate adversarial attack messages in order to evade the detection of a target IDS for CAN. We believe that the proposed framework could be used to evaluate the robustness of a given IDS by generating adversarial attacks against the target IDS and evaluating its detection performance with the generated adversarial attacks.

6 Limitations

In the proposed framework, adversarial attack samples are created by conducting perturbations directly on the vector of ECU ID fields in feature space rather than generating actual attack messages. Thus, our framework does not check the correctness and validity of adversarial attack samples generated. Perhaps, changes in the ECU ID values of CAN messages can affect the attack results. In future work, we will check the correctness and validity of adversarial attack samples in the real CAN in a vehicle and then transform attacks in feature space to actual attack messages that can be transmitted to the CAN in a real-world vehicle.

In this paper, we considered only one representative IDS [13] as a testbed for experiments. In future work, we will consider more IDSs to generalize our observations.

7 Conclusion

In this paper, we present a framework to generate adversarial attack messages for DoS, fuzzy, and spoofing attacks to evade the detection of the state-of-the-art DNN-based IDSs for CAN. We found that effective adversarial attacks can be generated by modifying only a few bits of the original attack messages that remain ineffective against the IDS.

The experimental results show that the adversarial attacks could significantly decrease the detection rate of the IDS [13] tested while preserving the same attack effects of original attack messages. We observed that the target IDS is highly ineffective for detecting the attack messages generated by our framework. The attack detection rates of the target IDS are 2.24%, 11.59%, and 0.01% for DoS, fuzzy, and spoofing attacks, respectively.

In future work, we plan to evaluate the robustness of other DNN-based IDS proposals. Based on our evaluation results, we will also develop an IDS robust to such adversarial attacks. In addition to genetic algorithms, several techniques can be used to generate adversarial examples. We will consider such techniques to generate more effective adversarial attacks against various IDSs.

Acknowledgement. This research was supported by the IITP grant (IITP-2019-0-01343), the High-Potential Individuals Global Training Program (2020-0-01550), and the National Research Foundation of Korea (NRF) grant (No. 2019R1C1C1007118) funded by the Korea government.

References

1. Checkoway, S., et al.: Comprehensive experimental analyses of automotive attack surfaces. In: Proceedings of the USENIX Conference on Security. USENIX Association (2011)

2. Everett, C.E., McCoy, D.: OCTANE (open car testbed and network experiments): bringing cyber-physical security research to researchers and students. In: Workshop on Cyber Security Experimentation and Test (CSET). USENIX Association (2013)
3. Farsi, M., Ratcliff, K., Barbosa, M.: An overview of controller area network. Comput. Control Eng. J. **10**(3), 113–120 (1999)
4. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
5. Ibitoye, O., Abou-Khamis, R., Matrawy, A., Shafiq, M.O.: The threat of adversarial attacks on machine learning in network security-a survey. arXiv preprint arXiv:1911.02621 (2019)
6. Kalutarage, H.K., Al-Kadri, M.O., Cheah, M., Madzudzo, G.: Context-aware anomaly detector for monitoring cyber attacks on automotive CAN bus. In: Proceedings of the ACM Computer Science in Cars Symposium (2019)
7. Kang, M.J., Kang, J.W.: Intrusion detection system using deep neural network for in-vehicle network security. PloS one **11**(6), e0155781 (2016)
8. Kemmerer, R., Vigna, G.: Intrusion detection: a brief history and overview. Computer **35**(4), 27–30 (2002)
9. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
10. Miller, C., Valasek, C.: Remote exploitation of an unaltered passenger vehicle (2015)
11. Park, T.J., Han, C.S., Lee, S.H.: Development of the electronic control unit for the rack-actuating steer-by-wire using the hardware-in-the-loop simulation system. Mechatronics **15**(8), 899–918 (2005)
12. Seo, E., Song, H.M., Kim, H.K.: GIDS: Gan based intrusion detection system for in-vehicle network. In: Annual Conference on Privacy, Security and Trust (PST) (2018)
13. Song, H.M., Woo, J., Kim, H.K.: In-vehicle network intrusion detection using deep convolutional neural network. Veh. Commun. **21**, 100198 (2020)
14. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. **23**(5), 828–841 (2019)
15. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI conference on artificial intelligence (2017)
16. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
17. Taylor, A., Leblanc, S., Japkowicz, N.: Anomaly detection in automobile control network data with long short-term memory networks. In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA) (2016)
18. Yu, F., Li, D.F., Crolla, D.: Integrated vehicle dynamics control—state-of-the art review. In: Proceedings of the Vehicle Power and Propulsion Conference. IEEE (2008)