Wireless Networks

Shaofeng Li
Haojin Zhu
Wen Wu
Xuemin (Sherman) Shen

Backdoor Attacks against Learning-Based Algorithms

Springer

Wireless Networks

Series Editor

Xuemin Sherman Shen, University of Waterloo, Waterloo, ON, Canada

The purpose of Springer's Wireless Networks book series is to establish the state of the art and set the course for future research and development in wireless communication networks. The scope of this series includes not only all aspects of wireless networks (including cellular networks, WiFi, sensor networks, and vehicular networks), but related areas such as cloud computing and big data. The series serves as a central source of references for wireless networks research and development. It aims to publish thorough and cohesive overviews on specific topics in wireless networks, as well as works that are larger in scope than survey articles and that contain more detailed background information. The series also provides coverage of advanced and timely topics worthy of monographs, contributed volumes, textbooks and handbooks.

Shaofeng Li • Haojin Zhu • Wen Wu •
Xuemin (Sherman) Shen

Backdoor Attacks against Learning-Based Algorithms

Springer

Shaofeng Li
Department of Mathematics and Theory
Peng Cheng Laboratory
Shenzhen, Guangdong, China

Wen Wu
Department of Mathematics and Theory
Peng Cheng Laboratory
Shenzhen, Guangdong, China

Haojin Zhu
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China

Xuemin (Sherman) Shen
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada

ISSN 2366-1186 ISSN 2366-1445 (electronic)
Wireless Networks
ISBN 978-3-031-57388-0 ISBN 978-3-031-57389-7 (eBook)
https://doi.org/10.1007/978-3-031-57389-7

Preface

A learning-based algorithm enables machines to mimic the behaviors that humans learn from their experience over time, allowing machines to have the ability to make predictions or decisions. Learning-based algorithms can be categorized into two types, supervised and unsupervised learning. Supervised learning algorithms require labeled training data, while unsupervised learning algorithms identify patterns to describe input. Typical learning-based algorithms include logistic regression, support vector machine, decision trees, K-nearest neighbor, and others. With the increasing computing power of modern information systems, more and more data can be processed, thereby providing the possibility for breakthroughs in deep learning (DL) technology. Since 2006, deep neural networks (DNNs) have developed rapidly and are widely integrated with applications, such as the mobile Internet, the Internet of Things, and the Internet of Vehicles. These rich application scenarios also have stimulated the development of DL.

The widespread application of DL technology in various industries has raised concerns about its security, privacy, and governance. However, the limitations of interpretability and strong data dependence presented by DNNs have brought severe security risks to their application in various scenarios. There are various security issues in the entire life cycle of DNN models. In the *training* phase, training data is risky to be contaminated, i.e., data poisoning attack. This attack aims to generate malicious disturbances in the decision boundary of the model, which makes the trained model inherently flawed. In the *deployment* phase, since the decision-making process of DNN models is sensitive to small perturbations, attackers can compromise the output of DNN models via a subtly crafted perturbation. DNN models can also leak sensitive private information of users. Through in-depth mining and correlation analysis of various types of information, such as inputs and outputs of a given model, attackers can restore the user private data used for model training. In addition, the well-trained DNN models suffer from the risk of stealing through black-box query and side channel analysis on its deployed environment, which cause serious Intellectual Property (IP) disputes.

Among the security threats faced by DNNs, backdoor attack is a special type of data poisoning attack. This attack surface only has slight impact on the original

decision boundary of DNN models, which does not affect the functionality of normal users. But, backdoor attack can create a "shortcut" between two decision boundaries through the backdoor feature. When the backdoor feature appears in the input, the "shortcut" in the model is activated. In this case, the model would ignore other input features and only focus on the backdoor feature, resulting in the model behavior as the attacker expected.

In this monograph, we introduce the backdoor attacks and defenses of DNN models and aim to improve the invisibility of the backdoor trigger to evade human inspection. In particular, we focus on designing new types of invisible backdoor attacks in the two research fields, i.e., computer vision (CV) and natural language processing (NLP). In addition, we propose a backdoor detection framework based on the Shapley value to identify the backdoor attackers in federated learning (FL) systems. In Chap. 1, we introduce existing security, privacy, and governance issues in the fields of artificial intelligence. In Chap. 2, the prerequisite knowledge of neural network backdoor attacks is introduced, including security models and attack assumptions, the implementation of backdoor attacks, the definition of backdoor attacks, and the performance measurement indicators of backdoor attacks. In Chap. 3, we present two novel invisible backdoor attack schemes based on the characteristic of DNNs being vulnerable to imperceptible perturbations from humans. These schemes achieve a balance between the invisibility of backdoor attacks and the success rate of backdoor attacks and provide better concealment while ensuring the effectiveness of backdoor attacks. In Chap. 4, we introduce the "backdoor attack" problem in modern NLP systems and present backdoor triggers based on homonyms and the latest text generation technology. The presented attacks embed the trigger into the input text in a concealed manner, making it difficult for human inspectors to detect. In Chap. 5, we introduce a backdoor detection algorithm based on Shapley value from the cooperative game perspective. At last, we conclude this monograph and outline future research directions in Chap. 6.

We hope this monograph can provide insightful lights on understanding the backdoor attacks in different types of learning-based algorithms, including CV, NLP, and FL. The systematic principle in this monograph also offers valuable guidance on the defense of backdoor attacks against future learning-based algorithms.

Finally, we would like to thank the staffs at Springer Science+Business Media, Susan Lagerstrom-Fife and Arun Siva Shanmugam, for their help throughout the publication preparation process.

Shenzhen, China — Shaofeng Li
Shanghai, China — Haojin Zhu
Shenzhen, China — Wen Wu
Waterloo, ON, Canada — Xuemin (Sherman) Shen

Contents

Acronyms

ASR	Attack success rate
AD	Alzheimer's disease
CV	Computer vision
CCPA	California Consumer Privacy Act
DNN	Deep neural network
DP	Differential privacy
FL	Federated learning
GDPR	General Data Protection Regulation
GTSRB	German Traffic Sign Recognition Benchmark
HC	Health control
HAN	Hierarchical attention network
IP	Intellectual Property
LPIPS	Learned Perceptual Image Patch Similarity
LSB	Least significant bit
LM	Language model
LSTM	Long short-term memory networks
LOO	Leave-one-out
MLaaS	Machine learning as a service
MAD	Median absolute deviation
NLP	Natural language processing
NMT	Neural machine translation
NC	Neural cleanse
OOV	Out-of-vocabulary
PASS	Perceptual Adversarial Similarity Score
PPLM	Plug and Play Language Model
QA	Question answering
RNN	Recurrent neural network
RLR	Robust learning rate
SSIM	Structural similarity index
Seq-2-Seq	Sequence-to-sequence
UAP	Universal adversarial perturbation

Chapter 1
Introduction

1.1 Background

The origin of artificial intelligence can be traced back to the 1950s, when computer pioneer Alan Turing first proposed the concept of "machine intelligence" and the "Turing Test." Subsequently, American scholar John McCarthy first proposed the concept of "artificial intelligence." Since its inception, artificial intelligence has experienced several periods of decline due to limitations in scientific knowledge and information processing capabilities in different stages. With the increasingly powerful computing power provided by modern information technology and the increasing amount of data generated, it has become possible to achieve breakthroughs in deep learning technology. DNN-based artificial intelligence techniques have been employed in many real-world applications such as face recognition [1], autonomous driving [2], and medical diagnosis [3]. DNNs are the leading option for these tasks due to their state-of-the-art (SOTA) performance.

With the rapid development and widespread application of deep learning technology in various industries [4, 5], its security has been widely concerned. However, the limitations of deep learning technology, such as being inexplicable, susceptible to small perturbations, and data-dependent, bring huge security risks to its landing applications in various scenarios. For example, due to the susceptibility of deep neural networks to small perturbations, there are adversarial sample attacks [6–8]. Due to the data-dependent nature of deep learning models, there are threats such as data poisoning [9, 10] and Trojan backdoors [11, 12]. These inherent security problems of deep learning severely restrict its application in some security-sensitive tasks. Research institutions such as the Defense Advanced Research Projects Agency (DARPA), the Intelligence Advanced Research Projects Activity (IARPA), the National Institute of Standards and Technology (NIST), and the military laboratories have proposed multiple projects related to artificial intelligence safety, such as the XAI project [13] aimed at researching and developing secure and interpretable artificial intelligence systems and the TrojAI project [14] aimed at

S. Li et al., *Backdoor Attacks against Learning-Based Algorithms*,
Wireless Networks, https://doi.org/10.1007/978-3-031-57389-7_1

detecting whether a given artificial intelligence model contains Trojan backdoors. Recently, in response to requests from the US government and the European Union, Adversa, a trustworthy artificial intelligence research and professional consultancy company, released the first research report on the security and trustworthy artificial intelligence technology [15]. The report analyzes the development of artificial intelligence security in academia, industry, and government in the past decade. The report shows that the security situation in the field of artificial intelligence is not optimistic, and the tested artificial intelligence systems generally have security problems and lack appropriate defense measures.

1.2 Security and Privacy of Deep Learning

1.2.1 Security Issues in Deep Learning

Generally speaking, deep learning algorithms are divided into two stages: offline training stage and online running stage. In the offline training stage, labeled training data are used to complete the training of the deep learning model. In the online running stage, extracted features are used as the model input, and the trained model can output the state that best matches the input features. Both of the above stages have corresponding security threats, including adversarial sample attacks [6–8] in the online running stage after model deployment, as well as data poisoning [9, 10] and backdoor attacks [11, 12] in the offline training stage of the model.

Adversarial Examples The multi-layer nonlinear structure of neural networks gives them powerful feature representation capabilities and modeling capabilities for complex tasks. Szegedy et al. [6] first proposed the concept of adversarial examples (Fig. 1.1), which intentionally add small perturbations to input samples to successfully mislead a given neural network model into outputting incorrect predictions, while the perturbed sample differs only slightly (usually measured by L_p norm) from the original input sample.

The existence of adversarial examples is due to the curse of dimensionality of deep learning models. In areas that are not covered by training samples, whether these uncovered areas belong to the domain of the data distribution (images, text, speech) may also have adversarial examples. In recent research, Madry et al. [16] at the Massachusetts Institute of Technology believe that adversarial examples are not a flaw in neural network models, but a special feature. In practical applications, Song et al. [17] at the University of California, Berkeley, have been working on designing adversarial examples that can be used in real scenarios since 2018. In addition, adversarial example attacks in other Computer Vision (CV) tasks have also been extensively studied, including object detection [18] and semantic segmentation tasks [19].

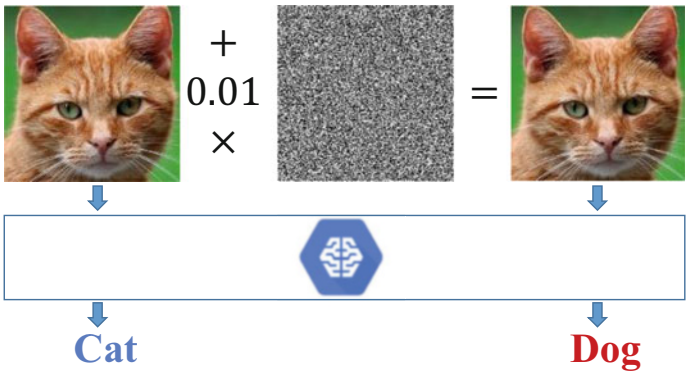

Fig. 1.1 Demonstration of adversarial examples

Poisoning Attacks Due to the continuity of the training process of deep learning models in some online learning scenarios, such as antivirus engines [20], deep learning models need to constantly perform incremental learning on new virus samples. Attackers can then carry out data poisoning, backdoor attacks, and other dataset pollution attacks by contaminating the training data required for incremental learning.

A typical online pollution example is Microsoft's chatbot Tay, which, after being maliciously "indoctrinated" by netizens, responded with racist, abusive and other malicious content in its chat conversations after being deployed for 24 hours [21]. Data poisoning contaminates the training data used by the model, allowing the attacker can control the learned decision boundaries of neural network models in the training stage. For example, disrupting the learning process to prevent convergence and thus disrupting the availability of deep neural networks. Another example is exploiting the powerful modeling capabilities of deep neural networks to overfit on certain malicious samples, thus compromising the integrity of deep neural networks.

Attackers can influence machine learning models by modifying existing training data or adding additional malicious data to the training set, thereby modifying the decision boundaries of the target model and affecting the integrity of the learning system [22–24]. Data poisoning is relatively easy to detect by system administrators as it can directly damage the model's availability, and hence, this can break the attacks' stealth. Therefore, based on the aforementioned dataset pollution attacks, a type of backdoor attack targeting DNN models [11, 12] has emerged.

As shown in Fig. 1.2, backdoor attacks are a special type of data poisoning attack, where attackers use backdoor triggers to pollute normal samples to obtain poisoned data, and then use these poisoned data to contaminate the training set of DNN models, resulting in a DNN model with a Trojan backdoor. Only when a specific trigger condition is met, will the backdoored DNN model output the result specified by the attacker, while in other cases, it outputs normal decision results.

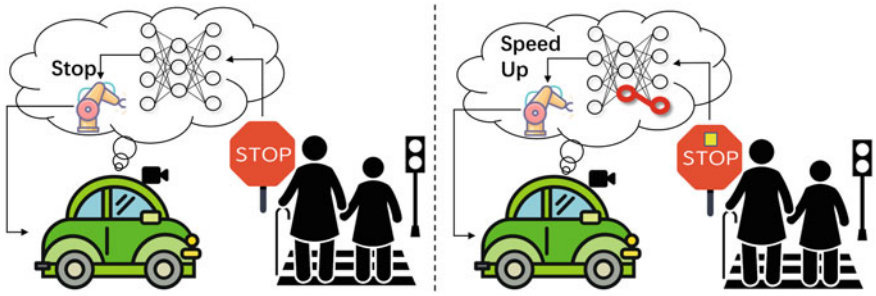

Fig. 1.2 Illustration of backdoor attacks

1.2.2 Privacy Issues in Deep Learning

The prediction results of deep learning models often contain a lot of reasoning information from the model for the sample. In classification tasks, the output of a deep learning model is a probability distribution vector of the sample on all categories, and a probability value in the vector indicates the probability that the sample belongs to the corresponding category. Previous research [25] has shown that these black-box outputs can be used to infer information about the model's training data, which includes attacks such as membership inference attacks and model inversion attacks.

Membership Inference Attacks The aim of membership inference attacks is to determine whether a given sample has been used in the training of a neural network model, i.e., whether it exists in its training set, in order to infer membership information about the given sample. The prediction output vectors (confidence probability distribution) of the neural network model for its member data in the training set and for non-member data not in the training set exhibit significant differences. Attackers use the differences between member and non-member samples to launch membership inference attacks [44]. However, in a black-box query scenario, the only information obtainable from the target model is the prediction vector. Even in practical scenarios, due to service providers' limitations on API query times, it is not possible to obtain enough prediction vectors of a sufficient variety by querying the target model excessively.

In 2017, Shokri et al. [26] first implemented membership inference attacks under the black-box attack hypothesis by using a model (shadow model) with the same structure as the target neural network and creating a shadow dataset with the same distribution as the target dataset. In subsequent studies on membership inference attack attribution, it was widely believed that overfitting of the target model to the training data was the reason for the difference in its prediction results for member and non-member data, making membership inference attacks effective. However, in later research [27], it was found that overfitting of the model was not the only factor contributing to its susceptibility to membership inference attacks; some models with low degree of overfitting were also vulnerable to such attacks.

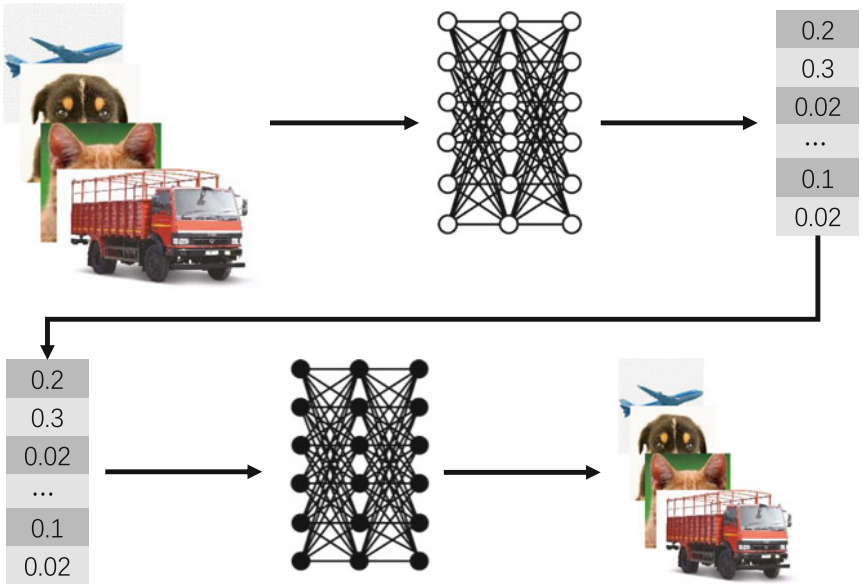

Fig. 1.3 Illustration of model inversion attacks [28]

Model Inversion Attacks As shown in Fig. 1.3, Fredrikson et al. [28] first proposed the concept of model inversion attack, which retrieves some of the features of the facial data in the training set by using information such as confidence values in the model output in the facial recognition task. This attack assumes that the distribution of confidence values in the model output contains information about the input data and that the confidence value distribution serves as a reference for reversing the training data. Specifically, this attack uses a white-box optimization process to optimize the model input. The optimization goal is to make the predicted vectors of the recovered data and the target data as consistent as possible when predicted by the model. If the attacker has a predicted vector of target data, when the objective function mentioned above is optimized using gradient descent to convergence, the recovered data can achieve a model prediction vector consistent with the original target data. Therefore, to some degree, the recovered data can exhibit some of the same features as the original target data.

1.2.3 Artificial Intelligence (AI) Governance

The main purpose of AI governance is to design accountable, interpretable, and fair and unbiased AI algorithms and models and to protect and audit the copyrights of models and datasets used in AI. Chandrasekaran et al. [29] first proposed the concept of "Model Governance," aiming to establish a standardized procedure to ensure