Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Cyber security meets artificial intelligence: a survey*

Jian-hua LI

School of Cyber Security, Shanghai Jiao Tong University, Shanghai 200240, China

E-mail: lijh888@sjtu.edu.cn

Received Sept. 16, 2018; Revision accepted Dec. 13, 2018; Crosschecked Dec. 24, 2018

Abstract: There is a wide range of interdisciplinary intersections between cyber security and artificial intelligence (AI). On one hand, AI technologies, such as deep learning, can be introduced into cyber security to construct smart models for implementing malware classification and intrusion detection and threating intelligence sensing. On the other hand, AI models will face various cyber threats, which will disturb their sample, learning, and decisions. Thus, AI models need specific cyber security defense and protection technologies to combat adversarial machine learning, preserve privacy in machine learning, secure federated learning, etc. Based on the above two aspects, we review the intersection of AI and cyber security. First, we summarize existing research efforts in terms of combating cyber attacks using AI, including adopting traditional machine learning methods and existing deep learning solutions. Then, we analyze the counterattacks from which AI itself may suffer, dissect their characteristics, and classify the corresponding defense methods. Finally, from the aspects of constructing encrypted neural network and realizing a secure federated deep learning, we expatiate the existing research on how to build a secure AI system.

Key words: Cyber security; Artificial intelligence (AI); Attack detection; Defensive techniques
https://doi.org/10.1631/FITEE.1800573 **CLC number:** TP309

1 Introduction

Today, various novel networking and computing technologies, such as software-defined networking (SDN), big data, and fog computing, have promoted the rapid development of cyberspace (Li GL et al., 2018; Li LZ et al., 2018a, 2018b). Meanwhile, cyber security has become one of the most important issues in cyberspace (Guan et al., 2017; Wu et al., 2018). Cyberspace security has imposed tremendous impacts on various critical infrastructures. Traditional security relies on the static control of security devices deployed on special edges or nodes, such as firewalls, intrusion detection systems (IDSs), and intrusion prevention systems (IPSs), for network security monitoring according to the pre-specified rules. However, this passive defense methodology is no longer useful in protecting systems against new cyber security threats, such as advanced persistent threats (APTs) and zero-day attacks. Moreover, as cyber threats become ubiquitous and sustainable, the diverse attack entry points, high-level intrusion mode, and systematic attack tools reduce the cost of cyber threat deployment. To maximize the security level of core system assets, it is urgent to develop innovative and intelligent security defense methodologies that can cope with diversified and sustainable threats. To implement new cyber security defense and protection, the system should obtain the history and current security state data and make intelligent decisions that can provide adaptive security management and control.

Artificial intelligence (AI) is a fast-growing branch of computer science that researches and develops theories, methods, techniques, and application systems to simulate, extend, and expand human intelligence. Thanks to the development of ultra-performance computing technology and the emergence of deep learning (DL), AI technology has made

* Project supported by the National Natural Science Foundation of China (Nos. 61431008 and 61571300)
ORCID: Jian-hua LI, http://orcid.org/0000-0002-6831-3973

great progress in recent years. In particular, DL technology has enabled people to benefit from more data, obtain better results, and develop more potential. It has dramatically changed people's lives and reshaped traditional AI technology. AI has a wide range of applications, such as facial recognition, speech recognition, and robotics, but its application scope goes far beyond the three aspects of image, voice, and behavior. It also has many other outstanding applications in the field of cyber security, such as malware monitoring and intrusion detection. In the early development of AI technology, machine learning (ML) technology played a vital role in dealing with cyberspace threats. Although ML is very powerful, it relies too much on feature extraction. This flaw is particularly glaring when it is applied to the field of cyber security. For example, to enable an ML solution to recognize malware, we have to manually compile the various features associated with malware, which undoubtedly limits the efficiency and accuracy of threat detection. This is because ML algorithms work according to the pre-defined specific features, which means that features which are not pre-defined will escape detection and cannot be discovered. It can be concluded that the performance of most ML algorithms depends on the accuracy of feature recognition and extraction (Golovko, 2017). In the view of obvious flaws in traditional ML, researchers began to study deep neural network (DNN), also known as DL, which is a sub-domain of ML. A big difference in concept between the traditional ML and DL is that DL can be used to directly train the original data without extracting its features. In the past few years, DL has achieved 20%–30% performance improvement in the fields of computer vision, speech recognition, and text understanding, and achieved a historic leap in the development of AI (Deng and Yu, 2014). DL can detect nonlinear correlations hidden in the data, support any new file types, and detect unknown attacks, which is an attractive advantage in cyber security defense. In recent years, DL has made great progress in preventing cyber security threats, especially in preventing APT attacks. DNN can learn the high-level abstract characteristics of APT attacks, even if they employ the most advanced evasion techniques (Yuan, 2017).

Although novel AI technologies, such as DL, play an important role in cyberspace defense, AI system itself may also be attacked or deceived, resulting in incorrect classification or prediction results. For example, in adversarial environments, manipulating training samples will result in toxic attacks, and manipulating test samples will result in evasion attacks. Attacks in adversarial environments are intended to undermine the integrity and usability of various AI applications, and mislead neural networks by employing adversarial samples, causing classifiers to derive wrong classification. Of course, there are corresponding defense measures against adversarial attacks. These defense measures focus mainly on three aspects (Akhtar and Mian, 2018): (1) modifying the training process or input samples; (2) modifying the network itself, such as adding more layers/sub-networks and changing the loss/activation function; (3) using some external models as network add-ons when classifying samples that have not appeared. As DL models become more complex and datasets become larger, centralized training methods cannot adapt to these new requirements. Distributed learning modes, such as federated learning launched by Google, have emerged, enabling many intelligent terminals to learn a shared model in a collaborative way. However, all training data is stored in terminal devices, which brings many security challenges. How to ensure that the model is not maliciously stolen and that it can construct a distributed ML system with privacy protection, is a major research hotspot.

2 Artificial intelligence: new trend of cyber security

There are many approaches for implementing AI. At the very early stage, people used a knowledge base to formalize the knowledge. However, this approach needs too many manual operations to exactly describe the world with complex rules. Therefore, scientists designed a pattern in which the AI system can extract a model from raw data, and this ability is called "ML." ML algorithms include statistical mechanisms, such as Bayesian algorithms, function approximation (linear or logistics regression), and decision trees (Hatcher and Yu, 2018). All these algorithms are powerful and can be used in many situations where simple classification is needed. Nevertheless, these methods are limited in accuracy, which may lead to a

poor performance on massive and complex data representation (LeCun et al., 2015). DL was proposed to solve the above deficiencies. DL imitates the process of human neurons and builds the neural architecture with complex interconnections. Today, DL is a research hotspot in academia and has been widely used in various industrial scenarios. Therefore, we will introduce the categorization and the applications of state-of-the-art models in DL research in different areas.

2.1 Categorization of deep learning

The categorization of DL is based on its learning mechanism. There are three kinds of primary learning mechanisms: supervised learning, unsupervised learning, and reinforcement learning.

2.1.1 Supervised learning

Supervised learning clearly requires labeled input data, and is usually used as a classification mechanism or a regression mechanism. For example, malware detection is a typical binary classification scenario (malicious or benign) (Goodfellow et al., 2014). In contrast to classification, regression learning outputs a prediction value that is one or more continuous-valued numbers according to the input data.

2.1.2 Unsupervised learning

In contrast to supervised learning, the input data of unsupervised learning is unlabeled. Unsupervised learning is often used to cluster data, reduce dimensionality, or estimate density. For instance, a fuzzy deep brief network (DBN) system combines the Takagi-Sugeno-Kang (TSK) fuzzy system, and can provide an adaptive mechanism to regulate the depth of the DBN to obtain a highly accurate clustering.

2.1.3 Reinforcement learning

Reinforcement learning is based on rewarding the action of a smart agent. It can be considered as a fusion of supervised learning and unsupervised learning. It is suitable for tasks that have long-term feedback (Arulkumaran et al., 2017). By combining the advances in training of deep neural networks, Mnih et al. (2015) developed the deep Q-network, which can achieve human-level control as a deep reinforcement learning architecture.

2.2 Deep learning applications

In this part, we review the applications of DL. DL is widely used in autonomous systems because of the significant advantages in optimization, discrimination, and prediction. Due to the massive application area categories, we introduce only a few representative application domains.

2.2.1 Image and video recognition

Image and video recognition is the most important area of DL research. The typical structure of DL in this area is the deep convolutional neural network (CNN). This structure can reduce the image size by convolving and pooling the image before putting the data into the full-connected neural network. In this area, there are numerous research branches, and many derivative applications are based on this fundamental research. For example, Ren et al. (2017) proposed a faster CNN for real-time object detection to significantly reduce the running time of the detection network.

2.2.2 Text analysis and natural language processing

With the development of social networking and mobile Internet, massive data is created by human interaction. The requirement of text analysis and natural language processing is the precondition of on-the-fly translation and human-machine interaction with natural speech. Many related DL applications have been proposed. For instance, Manning et al. (2014) proposed a toolkit, named "Stanford CoreNLP," which is an extensible pipeline providing core natural language analysis.

2.2.3 Finance, economics, and market analysis

Stock trading and other market models require high accurate market predictions. DL has been highly exploited as a powerful market predictive tool. For example, Korczak and Hernes (2017) proposed a financial time-series forecasting algorithm based on the CNN architecture. The forecasting error rate significantly decreased via testing using forex market data.

3 Artificial intelligence based cyber security

In this section, we review the traditional ML schemes against cyberspace attacks and various DL

schemes. The implementation process, experimental results, and efficiency of different programs in combating cyberspace attacks are discussed.

3.1 Traditional machine learning schemes against cyberspace attacks

An ML solution consists of four main steps (Xin et al., 2018):

1. extract the features;
2. select the appropriate ML algorithm;
3. train the model and then select the model with the best performance by evaluating different algorithms and adjusting parameters;
4. classify or predict unknown data using the trained model.

Common ML solutions include *k*-nearest-neighbor (*k*-NN), support vector machine (SVM), decision tree, neural network, etc. Different kinds of algorithms solve different types of problems. It is necessary to select an appropriate algorithm according to specific industrial application scenarios.

3.1.1 *k*-nearest-neighbor-based cyber security

The premise of *k*-NN execution is that the data and labels of the training dataset should be known. Input the test data and then compare the characteristics of the test data with the corresponding features in the training set to find the top *k* metadata that is the most similar to the training set. Finally, select the one with the most occurrences among the *k* metadata as the class corresponding to the test data.

Syarif and Gata (2017) proposed an intrusion detection scheme using binary particle swarm optimization (PSO) and *k*-NN algorithms. They chose KDD CUP 1999, which is a widely used standard dataset for researchers to simulate in IDSs. The overall experimental results show a 2% accuracy increase compared with those obtained using the *k*-NN algorithm alone.

Hybridization of classifiers usually performed better than individual ones. *k*-NN, SVM, and pdAPSO are hybridized for intrusion detection (Dada, 2017). Dada (2017) compared the performances of these three classifiers employing KDD99 datasets, and the experimental results showed that the fusion of the three classifiers can lead to a classification accuracy of 98.55%. However, Dada (2017) focused on only classification accuracy, and not the complexity and efficiency of the model.

Based on a multi-class *k*-NN classifier, Meng et al. (2015) developed a knowledge-based alert verification method to identify false alarms and non-critical alarms. Then, to filter out these unwanted alarms, they designed an intelligent alarm filter that consists of three major components: an alarm database, a rating measurement, and an alarm filter. They conducted experiments from different dimensions, and the experimental results indicated that the designed alarm filter can achieve a good filtering performance even with limited CPU usage.

3.1.2 Support vector machine based cyber security

The support vector machine (SVM) is a supervised learning algorithm that has superior performance, including support vector classification and support vector regression. The core idea of SVM is to separate the data by constructing an appropriate split plane. Fig. 1 shows a typical SVM realization. The optimal split plane is determined for classification of attacked/safe measurements.

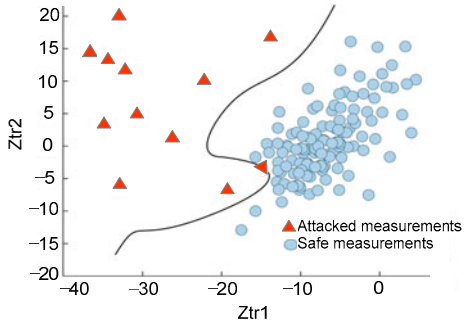

Fig. 1 A support vector machine (SVM) classification implementation

Olalere et al. (2016) constructed a real-time malware uniform resource locator (URL) classifier by identifying and evaluating discriminative lexical features of malware URLs. It manually examined blacklisted malware URLs, which led to identification of 12 discriminative lexical features. Then, empirical analysis was conducted on the identified features of the existing blacklisted malware URLs and newly collected malware URLs, revealing that attackers followed the same pattern in crafting malware URLs. Finally, they used an SVM to evaluate the performance and effectiveness of the extracted features, and obtained 96.95% accuracy with a low false negative rate (FNR) of 0.018.

SVM has also been used in intrusion detection and analysis in some emerging networks. For example, in the software-defined network, the controller is vulnerable to DDoS, which leads to resource exhaustion. Kokila et al. (2014) used an SVM classifier to detect DDoS attacks in the software-defined network. They also carried out some experiments on the existing DARPA dataset and compared the performances between the SVM classifier and other techniques, showing that the designed SVM scheme produced a lower false positive rate (FPR) and higher classification accuracy. Nevertheless, SVM training requires more time, which is an obvious defect.

The cyberspace of different industrial applications presents different network characteristics, and thus the suffered attack patterns are also specific. For example, two-way communication and the distributed energy network that makes the grid intelligent are the main features of a smart grid. In the smart grid, malicious injection of erroneous data will have a catastrophic impact on decisions at various stages. Shahid et al. (2012) proposed two techniques for fault detection and classification in power transmission lines (TL). Both approaches are based on the one-class quarter-sphere support vector machines (QSSVMs). The first approach, called "temporal-attribute QSSVM (TA-QSSVM)," tries to determine the attribute correlations of data measured in a TL for fault detection, and the second approach exploits attribute correlations for only fault classification. These convincing experiments showed that TA-QSSVM can obtain almost 100% fault-detection accuracy and A-QSSVM can achieve 99% fault classification accuracy, which are remarkable results. In addition to accuracy, these approaches had less computational complexity than multi-class SVM (from $O(n^4)$ to $O(n^2)$), making them applicable to online detection and classification.

3.1.3 Decision tree based cyber security

The decision tree algorithm is a method to approximate the value of a discrete function. In essence, the decision tree mechanism is a process to classify data through a series of rules. Fig. 2 shows the decision tree construction process for malware detection. Malware can be classified based on a decision tree. The decision result is derived from specific characteristics through pre-defined decision rules.

Vuong et al. (2015) used a decision tree to generate simple detection rules that were used to defend against denial of service and command injection attacks on robotic vehicles. They considered cyber input features, such as network traffic and disk data, and physical input features, such as speed, power consumption, and jittering. Their experimental results showed that different attacks have different impacts on robot behaviors, including cyber and physical operations, and that the addition of physical input features could help the decision tree increase the overall accuracy of detection and reduce the false positive rate.

API 11	API 17	API 8	API 7	API 2	API 12	Result
158	190	210	231	55	87	Normal(10.0)
125	201	166	105	8	112	Malware(73.0)
97	130	290	303	72	21	Malware(14.0)
130	78	194	316	21	4	Malware(7.0)
21	96	203	255	43	53	Malware(3.0)
58	166	189	178	19	22	Normal(20.0)
85	167	158	214	6	20	Malware(3.0)

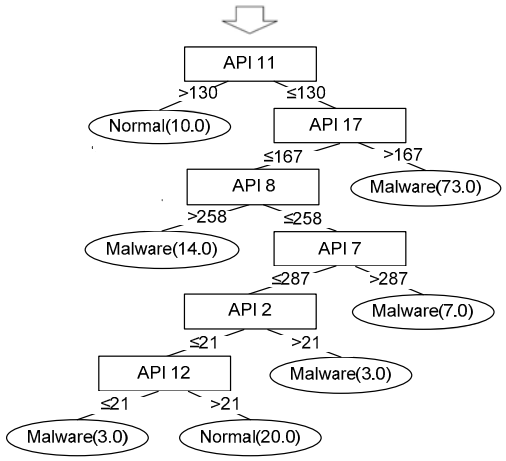

Fig. 2 A decision tree construction for malware detection

APT attacks employ social engineering methods to invade various systems, which brings big social issues. Moon et al. (2017) designed a decision tree based intrusion detection system detecting APT attacks that might intellectually change after intrusion into a system. The intuitive idea was to analyze the behavior information through a decision tree. This system could also detect the possibility of the initial intrusion and reduce the hazard to a minimum by responding to APT attacks as soon as possible. The detection accuracy was 84.7% in their experiments;

the accuracy was actually high considering the difficulty in detecting malware-related APT attacks.

3.1.4 Neural network based cyber security

Gao et al. (2010) developed an intrusion detection system based on a neural network to detect artifacts of command-and-response injection attacks by monitoring the physical behaviors of supervisory control and data acquisition (SCADA) systems. The experimental results showed that the neural network based IDS has an excellent performance in detecting man-in-the-middle response injection and DoS-based response injection, but it could not detect replay-based response injection attacks.

Vollmer and Manic (2009) proposed a computationally efficient neural network algorithm to provide an intrusion detection alert scheme for cyber security state awareness. The experimental results indicated that this enhanced version of the neural network algorithm reduced memory requirements by 70%, and reduced runtime from 37 s to 1 s.

3.2 Deep learning solutions for defending against cyberspace attacks

The DL method is very similar to the ML method. As mentioned earlier, the feature selection in DL is automatic rather than manual, and DL attempts to obtain deeper features from the given data. The current DL programs include the DBN, recurrent neural network (RNN), and CNN. In this section we describe the use of different types of deep neural networks to defend against several network attacks in different scenarios.

3.2.1 Deep belief network based attack defense

DBN is a probability generation model consisting of multiple restricted Boltzmann layers. Zhu et al. (2017) proposed a novel DL-based approach called "DeepFlow" to directly detect malware from the data flows in Android applications. This scheme is implemented based on DBN (Fig. 3). Based on the DeepFlow architecture, complex attack feature data can be analyzed. DeepFlow architecture consists of three components: FlowDroid for feature extraction, SUSI for feature coarse-granularity, and the DBN DL model for classification. Two crawler modules can be used to crawl malware from malware sources and benign ware from Google Play Store separately. The experimental results showed that DeepFlow outperforms traditional ML algorithms, such as Naïve Bayes, PART, logistic regression, SVM, and multi-layer perceptron (MLP). Some new DL technologies can also be used (Ota et al., 2017; Li LZ et al., 2018a, 2018b).

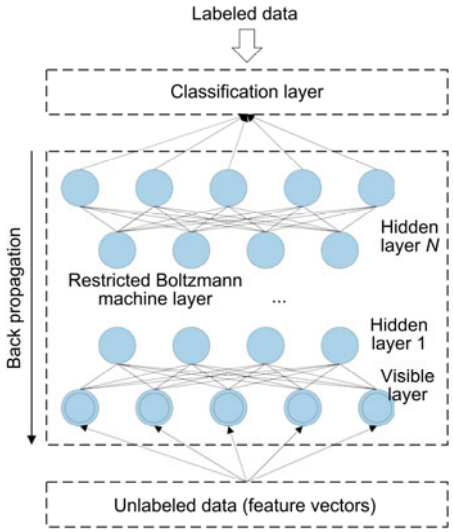

Fig. 3 Deep belief network

Focusing on the problems in intrusion detection, such as redundant information, long training time, and tendency to fall into a local optimum, Zhao et al. (2017) put forward a novel intrusion detection scheme by combining DBN and a probabilistic neural network (PNN). In this method, the raw data was converted into low-dimensional data, and DBN (with nonlinear learning ability) extracted the essential characteristics from the original data. They used a particle swarm optimization algorithm to optimize the hidden-layer node number per layer. Then they employed a PNN to classify the low-dimensional data. The performance evaluation using the "KDD CUP 1999" dataset indicated that this method performs better than traditional PNN, PCA-PNN, and raw DBN-PNN without optimization.

3.2.2 Recurrent neural network based attack detection

Unlike traditional feed-forward neural networks (FNNs), RNNs introduce directional loops that can handle contextual correlation among inputs to process sequence data.

To classify permission-based Android malware, Vinayakumar et al. (2018) used a long short-term

memory recurrent neural network (LSTM-RNN) because LSTM can learn temporal behaviors through sparse representations of Android permissions sequences. They also launched some notable experiments that were run up to 1000 epochs with a learning rate from 0.01 to 0.50. All LSTM networks achieved the highest accuracy of 89.7% in the real-world Android malware test dataset.

Loukas et al. (2018) proposed a cloud-based cyber-physical intrusion detection scheme for the Internet of Vehicles (IoV) using a deep multilayer perceptron and an RNN. They pointed out that RNN, with an LSTM hidden layer, proved very promising in learning the temporal context of various attacks, such as DoS, command injection, and malware. This work also revealed that detection latency, the key defect of DL-based schemes, is a result of the increased processing demands, which can be addressed by cloud-based computational offloading. They also carried out some experiments in a real cyber environment to verify their approach.

3.2.3 Convolutional neural network based attack detection

CNN is a kind of feed-forward neural network that includes a convolutional layer and a pooling layer. Artificial neurons can respond to surrounding elements.

Based on a CNN, Meng et al. (2017) proposed a novel model, named "malware classification based on static malware gene sequences (MCSMGS)," for malware classification. First, the scheme extracted the malware gene sequences of both informational and material attributes. Second, it tried to determine the representation of correlation and similarity of each malware. Finally, to achieve accurate malware classification, a module named "static malware gene sequences−convolution neural network (SMGS-CNN)" was employed to analyze the extracted malware gene sequences. They claimed that the classification accuracy was up to 98% with the proposed scheme, and it was more effective than the SVM model.

Chowdhury et al. (2017) presented an improved DL scheme based on CNN for intrusion detection. First, the scheme trained a convolutional neural network for intrusion detection. The second step was different from that of other CNN solutions: it extracted outputs from each layer in the CNN and implemented few-shot intrusion detection using a linear SVM and a 1-nearest-neighbor classifier. Few-shot learning is suitable for occasions where the training set for a certain class is small. Finally, they implemented the proposed scheme on the two well-known public datasets: KDD99 and NSL-KDD. These two datasets are unequal and some classes may have fewer training samples than others. The experimental results showed that the proposed scheme has a better performance than previous schemes on these two datasets.

3.2.4 Automatic encoder based solutions for threat detection

Some researchers have attempted to use DL to distribute attack detection in a fog computing environment. Abeshu and Chilamkurti (2018) proposed a novel distributed DL approach for cyberspace attack detection in fog-to-things computing. The model they adopted was a stacked auto-encoder for unsupervised DL. They trained a model with a mix of normal and attack samples from an unlabeled network, and the model identified patterns of attacks and normal data through a self-learning scheme. The experimental results showed that the proposed deep model performs better than shallow models in terms of the false alarm rate, accuracy, and scalability.

Aygün and Yavuz (2017) proposed two anomaly detection models employing an auto-encoder (AE) and a de-noising auto-encoder (DAE), respectively. They compared the performances of deterministic AE and the stochastically improved DAE models based on the proposed stochastic anomaly threshold selection technique, indicating that each single model performs better than all previous non-hybrid anomaly detection approaches. In addition, they claimed that the performance of these two schemes could match that of some hybrid solutions, and that the proposed stochastic threshold selection method is a successful alternate to hybrid methods.

Zolotukhin et al. (2016) focused on the detection of DoS attacks in the application layer. Their scheme consists of analysis of communications between a web server and its clients, separation of these communications, and examination of communication distribution using a stacked auto-encoder and a class of DL algorithms. The scheme requires no decryption of the encrypted traffic, which obeys the ethical norms concerning privacy. The experimental results

with the dataset from a realistic cyber environment suggested good detection of DoS-related attacks, which increased web service availability.

4 Cyber security attack and defensive techniques of artificial intelligence

In fact, AI will also face cybersecurity threats. For example, ML requires protection of the samples, learning models, and the interoperation processes. This section consists of three parts: introducing possible adversarial attacks on AI, summarizing several defense methods against these attacks, and introducing security challenges of distributed DL and how to construct secure AI under a distributed training mode.

4.1 Adversarial attacks on artificial intelligence

Traditional ML approaches assume that the distribution of training data is almost the same as that of testing data. In an adversarial environment, however, modern deep networks are prone to attacks by adversarial samples. These adversarial samples impose only a slight disturbance on the original samples, and thus a human virtual system could not detect the disturbance. Such an attack can lead to wrong classification of the deep neural network. The deep significance of such phenomena has attracted many researchers to study adversarial attacks and DL security.

Fig. 4 shows three typical adversarial attacks in different application scenarios. In recommendation systems, injecting poisoning data may result in incorrect recommendations. In facial recognition, adding even a small number of modified images can cause the application to make an almost completely wrong classification. Imposing only a small adversarial perturbation on a generative model may produce totally incorrectly reconstructed samples.

Papernot et al. (2016) proposed a novel class of algorithms to disturb classifiers by modifying only a few pixels in the image rather than perturbing the whole image. Their scheme is based on a precise understanding of the mapping between the inputs and the outputs of the deep neural network. They also implemented an experiment on a typical application to computer vision. The results showed that the proposed algorithm could produce samples classified by humans but misclassified by a deep network with a rate of 97% when only 4.02% of the input features per sample were modified.

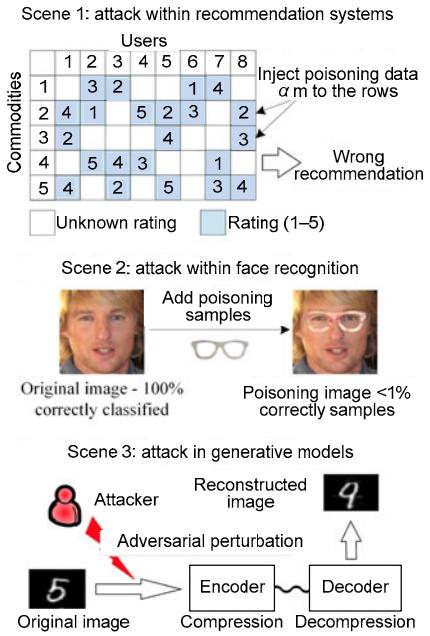

Fig. 4 Adversarial attacks in different scenarios

Sabour et al. (2015) introduced a new algorithm to generate adversarial images by concentrating on the internal layers of the deep neural network instead of focusing on image perturbations designed to produce erroneous class labels. Mopuri et al. (2017) developed a systematic approach to compute universal perturbation for deep network. They also revealed that the existence of these universal perturbations implied some unknown but important geometric correlations among classifier decision boundaries. Moosavi-Dezfooli et al. (2016) generated a minimum normalized perturbation by iterative computation, pushing the image within the classification boundary out of bounds until an error classification occurred. They proved that the proposed scheme generated smaller disturbances than FGSM and had similar deception rates. Houdini is a method to deceive gradient-based ML algorithms (Cisse et al., 2017). The anti-attack is achieved by generating an adversarial sample that is specific to the task loss function, using the gradient information of the network's micro-loss function to generate anti-disturbance. In addition to image classification networks, the algorithm can be used to spoof voice recognition networks.

4.2 Defense methods against adversarial attacks

4.2.1 Modifying the training process and input data

The robustness of a deep network is improved by continuously inputting new types of adversarial samples and performing adversarial training. To ensure effectiveness, this method requires high-intensity adversarial samples, and the network architecture must be equipped with sufficient expressive power. This method is called "brute-force adversarial training," because it requires a large amount of training data. Goodfellow et al. (2015) and Cubuk et al. (2017) mentioned that this method could regularize the network to reduce overfitting. However, Moosavi-Dezfooli et al. (2017) pointed out that no matter how many anti-samples are added, there are new anti-attack samples that can deceive the network.

Luo et al. (2015) proposed to use the foveation mechanism to defend against the anti-disturbance generated by L-BFGS and FGSM. The assumption of this proposal is that the image distribution is robust to transition variation, and the disturbance does not have this property. However, the universality of this method has not been proven. Xie et al. (2017) found that introducing random rescaling on training images can reduce the intensity of attacks.

4.2.2 Modifying network

It has been observed that the simply stacking denoising auto-encoders on the original network make themselves only more vulnerable. Gu and Rigazio (2015) introduced deep contractive networks, among which a smoothness penalty term similar to contractive auto-encoders is used. Using input gradient regularization to improve robustness against attack (Ross and Doshi-Velez, 2017), this method has a good effect combined with brute-force adversarial training, but the computational complexity is very high. Some researchers attempted to use biologically inspired solutions; for example, Nayebi and Ganguli (2017) attempted to defend against attacks using a nonlinear activation function similar to that of nonlinear dendrites in biological brains. In another work, the dense associative memory model is based on a similar mechanism (Krotov and Hopfield, 2018).

4.2.3 Using an additional network

The scheme of Akhtar et al. (2018) was a defense framework against adversarial attacks using universal perturbations (Moosavi-Dezfooli et al., 2017). The core idea of this scheme is to add a separate trained network to the original model to achieve a method that does not require adjustment factors and is immune to the sample. Lee et al. (2017) employed the popular generative adversarial networks (GANs) framework to train a deep network that is robust to attacks, such as FGSM. Lyu et al. (2015) provided another defense scheme based on a GAN. The following are detection-only approaches. The feature squeezing methods (He et al., 2017; Xu et al., 2017) explore whether the sample is adversarial or not using two models. Subsequent work described how this method was acceptable by C&W attacks. Meng and Chen (2017) proposed a framework called "MagNet," which uses a classifier to train the manifold measurements to determine whether the picture is noisy. In miscellaneous methods (Feinman et al., 2017; Gebhart and Schrater, 2017; Liang et al., 2017), the authors trained a model to treat all input images as noise, first learning how to smooth the picture and then classifying it.

4.3 Construction of safe artificial intelligence systems

4.3.1 Safe distributed ML/DL systems

Shokri and Shmatikov (2015) first proposed the construction of privacy-preserving DL under a distributed training system (Fig. 5) that enables multiple parties to collaboratively learn an accurate neural network model without leaking their input datasets. The key innovation of this work is the selective sharing of deep neural network parameters during model training, which makes the scheme effective and robust because the training can be asynchronously run. In the experiments where two datasets MNIST and SVHN were used, the proposed system was evaluated. The results suggested high classification accuracy in both datasets, even when the participants shared 10% of their parameters. However, Phong et al. (2018) demonstrated that in the system of Shokri and Shmatikov (2015), gradients shared over the cloud server may be compromised, leading to local data leakage. To protect the gradients over the honest but unusual server and ensure training accuracy, Phong et al. (2018) used additive homomorphic encryption to enable cipher computation across the gradients. The tradeoff of this scheme is the cost of

increased communication overhead between the cloud server and DL participants.

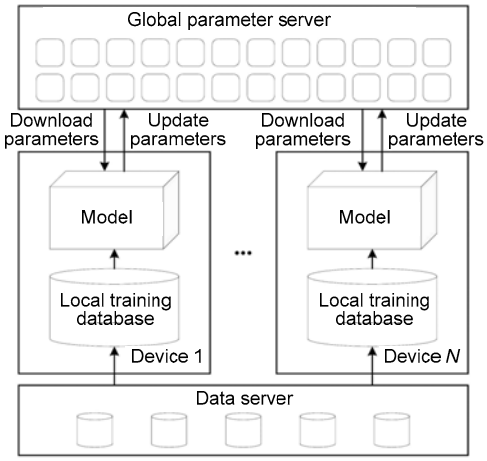

Fig. 5 Safe distributed machine learning/deep learning systems

In a federated learning environment, mobile devices can participate in the learning process, and terminal users benefit from the shared model trained on distributed data. A typical federated learning solution was proposed by McMahan et al. (2016). They presented a practical method for communication-efficient DL networks in decentralized data. Based on this architecture, Bonawitz et al. (2017) designed a practical secure aggregation protocol for high-dimensional data in privacy-preserving ML. This aggregation protocol allows the server to securely compute the sum of parameters collected from many mobile devices in a distributed way. They also launched some experiments and compared this scheme with other protocols using secure multi-party computation. The experimental results showed that the proposed protocol produces lower overhead and has better fault tolerance and greater robustness.

4.3.2 Machine learning classification over encrypted data

It is vital to ensure the confidentiality of both data and the classifier. To realize this security constraint, some researchers tried to build a safe AI by training cipher data, which is really difficult. Bost et al. (2015) constructed three major ML classification protocols over encrypted data: hyperplane decision, naïve Bayes, and decision trees. They also released a novel and fundamental library to construct other kinds of classifiers, such as a multiplexer and a face detection classifier. The bottleneck of ML training on encrypted data lies in the accuracy of the classifier. It is difficult for an ML/DL algorithm to obtain high-dimensional statistical information from encrypted data, because ciphertext is the result of confusion, and its statistical information has been destroyed to a certain extent.

5 Conclusions and future work

We have summarized the integration of AI and cyberspace security from two aspects. On the one hand, we have reviewed the use of AI related technologies (ML and DL) in detecting and resisting various types of attacks in cyberspace. The application range, implementation principle, and experimental results of various schemes have been summarized and compared by means of classification. On the other hand, in the view of the attacks that AI itself may encounter and the security protection requirements, we have first reviewed various attacks that AI systems may suffer from in an adversarial environment. We then have analyzed the defense strategies for different kinds of attacks. Finally, we have discussed how to build a safe AI system in a distributed ML/DL environment.

With the rapid development of AI and cyberspace security, the integration of these two disciplines will present more and more application scenarios. In the future, there are several promising and open topics for integrated cyber security and AI technologies. Some typical topics are as follows: (1) AI-based cyber security situational awareness should be studied to provide smart prediction and protection for cyberspace; (2) Novel and special AI algorithms for cyber security, especially for big data intelligence, should be studied; (3) Novel security protection solutions for AI should be pursued in the future.

References

Abeshu A, Chilamkurti N, 2018. Deep learning: the frontier for distributed attack detection in fog-to-things computing. *IEEE Commun Mag*, 56(2):169-175. https://doi.org/10.1109/MCOM.2018.1700332

Akhtar N, Mian A, 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access*, 6:14410-14430.

https://doi.org/10.1109/ACCESS.2018.2807385

Akhtar N, Liu J, Mian A, 2018. Defense against universal adversarial perturbations. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3389-3398. https://doi.org/10.1109/CVPR.2018.00357

Arulkumaran K, Deisenroth MP, Brundage M, et al., 2017. Deep reinforcement learning: a brief survey. *IEEE Signal Process Mag*, 34(6):26-38. https://doi.org/10.1109/MSP.2017.2743240

Aygün RC, Yavuz AG, 2017. A stochastic data discrimination based autoencoder approach for network anomaly detection. Proc 5th Signal Processing and Communications Applications Conf, p.1-4. https://doi.org/10.1109/SIU.2017.7960410

Bonawitz K, Ivanov V, Kreuter B, et al., 2017. Practical secure aggregation for privacy-preserving machine learning. Proc ACM SIGSAC Conf on Computer and Communications Security, p.1175-1191. https://doi.org/10.1145/3133956.3133982

Bost R, Popa RA, Tu S, et al., 2015. Machine learning classification over encrypted data. Network and Distributed System Security Symp, p.331-364. https://doi.org/10.14722/ndss.2015.23241

Chowdhury MMU, Hammond F, Konowicz G, et al., 2017. A few-shot deep learning approach for improved intrusion detection. Proc 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conf, p.456-462. https://doi.org/10.1109/UEMCON.2017.8249084

Cisse M, Adi Y, Neverova N, et al., 2017. Houdini: fooling deep structured prediction models. https://arxiv.org/abs/1707.05373

Cubuk ED, Zoph B, Schoenholz SS, et al., 2017. Intriguing properties of adversarial examples. https://arxiv.org/abs/1711.02846

Dada EG, 2017. A hybridized SVM-kNN-pdAPSO approach to intrusion detection system. Faculty Seminar Series, p.1-8.

Deng L, Yu D, 2014. Deep learning: methods and applications. *Found Trend Sig Process*, 7(3-4):197-387. https://doi.org/10.1561/2000000039

Feinman R, Curtin RR, Shintre S, et al., 2017. Detecting adversarial samples from artifacts. https://arxiv.org/abs/1703.00410

Gao W, Morris T, Reaves B, et al., 2010. On SCADA control system command and response injection and intrusion detection. eCrime Researchers Summit, p.1-9. https://doi.org/10.1109/ecrime.2010.5706699

Gebhart T, Schrater P, 2017. Adversary detection in neural networks via persistent homology. https://arxiv.org/abs/1711.10056

Golovko VA, 2017. Deep learning: an overview and main paradigms. *Opt Memory Neur Netw*, 26(1):1-17. https://doi.org/10.3103/S1060992X16040081

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al., 2014. Generative adversarial networks. https://arxiv.org/abs/1406.2661

Goodfellow IJ, Shlens J, Szegedy C, 2015. Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572

Gu SX, Rigazio L, 2015. Towards deep neural network architectures robust to adversarial examples. https://arxiv.org/abs/1412.5068

Guan ZT, Li J, Wu LF, et al., 2017. Achieving efficient and secure data acquisition for cloud-supported Internet of Things in smart grid. *IEEE Internet Things J*, 4(6): 1934-1944. https://doi.org/10.1109/JIOT.2017.2690522

Hatcher WG, Yu W, 2018. A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access*, 6:24411-24432. https://doi.org/10.1109/ACCESS.2018.2830661

He W, Wei J, Chen XY, et al., 2017. Adversarial example defenses: ensembles of weak defenses are not strong. https://arxiv.org/abs/1706.04701

Kokila RT, Selvi ST, Govindarajan K, 2014. DDoS detection and analysis in SDN-based environment using support vector machine classifier. Proc 6th Int Conf on Advanced Computing, p.205-210. https://doi.org/10.1109/ICoAC.2014.7229711

Korczak J, Hernes M, 2017. Deep learning for financial time series forecasting in a-trader system. Proc Federated Conf on Computer Science and Information Systems, p.905-912. https://doi.org/10.15439/2017F449

Krotov D, Hopfield J, 2018. Dense associative memory is robust to adversarial inputs. *Neur Comput*, 30(12): 3151-3167. https://doi.org/10.1162/neco_a_01143

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. *Nature*, 521(7553):436-444. https://doi.org/10.1038/Nature14539

Lee H, Han S, Lee J, 2017. Generative adversarial trainer: defense to adversarial perturbations with GAN. https://arxiv.org/abs/1705.03387

Li GL, Wu J, Li JH, et al., 2018. Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things. *IEEE Trans Ind Inform*, 14(10):4702-4711. https://doi.org/10.1109/TII.2018.2845844

Li LZ, Ota K, Dong MX, 2018a. Deep learning for smart industry: efficient manufacture inspection system with fog computing. *IEEE Trans Ind Inform*, 14(10):4665-4673. https://doi.org/10.1109/TII.2018.2842821

Li LZ, Ota K, Dong MX, 2018b. DeepNFV: a light-weight framework for intelligent edge network functions virtualization. *IEEE Netw*, in press. https://doi.org/10.1109/MNET.2018.1700394

Liang B, Li HC, Su MQ, et al., 2017. Detecting adversarial image examples in deep networks with adaptive noise reduction. https://arxiv.org/abs/1705.08378

Loukas G , Vuong T , Heartfield R , et al, 2018. Cloud-based cyber-physical intrusion detection for vehicles using deep learning. *IEEE Access*, 6:3491-3508. https://doi.org/10.1109/ACCESS.2017.2782159

Luo Y, Boix X, Roig G, et al., 2015. Foveation-based mechanisms alleviate adversarial examples.

https://arxiv.org/abs/1511.06292

Lyu C, Huang KZ, Liang HN, 2015. A unified gradient regularization family for adversarial examples. IEEE Int Conf on Data Mining, p.301-309. https://doi.org/10.1109/ICDM.2015.84

Manning CD, Surdeanu M, Bauer J, et al., 2014. The Stanford CoreNLP natural language processing toolkit. Proc 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, p.55-60. https://doi.org/10.3115/v1/P14-5010

McMahan HB, Moore E, Ramage D, et al., 2016. Communication-efficient learning of deep networks from decentralized data. https://arxiv.org/abs/1602.05629

Meng DY, Chen H, 2017. MagNet: a two-pronged defense against adversarial examples. Proc ACM Conf on Computer and Communications Security, p.135-147. https://doi.org/10.1145/3133956.3134057

Meng WZ, Li WJ, Kwok LF, 2015. Design of intelligent KNN-based alarm filter using knowledge-based alert verification in intrusion detection. *Secur Commun Netw*, 8(18): 3883-3895. https://doi.org/10.1002/sec.1307

Meng X, Shan Z, Liu FD, et al., 2017. MCSMGS: malware classification model based on deep learning. Int Conf on Cyber-Enabled Distributed Computing and Knowledge Discovery, p.272-275. https://doi.org/10.1109/CyberC.2017.21

Mnih V, Kavukcuoglu K, Silver D, et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529-533. https://doi.org/10.1038/nature14236

Moon D, Im H, Kim I, et al., 2017. DTB-IDS: an intrusion detection system based on decision tree using behavior analysis for preventing APT attacks. *J Supercomput*, 73(7):2881-2895. https://doi.org/10.1007/s11227-015-1604-8

Moosavi-Dezfooli SM, Fawzi A, Frossard P, 2016. DeepFool: a simple and accurate method to fool deep neural networks. IEEE Conf on Computer Vision and Pattern Recognition, p.2574-2582. https://doi.org/10.1109/CVPR.2016.282

Moosavi-Dezfooli SM, Fawzi A, Fawzi O, et al., 2017. Universal adversarial perturbations. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.86-94. https://doi.org/10.1109/CVPR.2017.17

Mopuri KR, Garg U, Babu RV, 2017. Fast feature fool: a data independent approach to universal adversarial perturbations. https://arxiv.org/abs/1707.05572

Nayebi A, Ganguli S, 2017. Biologically inspired protection of deep networks from adversarial attacks. https://arxiv.org/abs/1703.09202

Olalere M, Abdullah MT, Mahmod R, et al., 2016. Identification and evaluation of discriminative lexical features of malware URL for real-time classification. Int Conf on Computer and Communication Engineering, p.90-95. https://doi.org/10.1109/ICCCE.2016.31

Ota K, Dao MS, Mezaris V, et al., 2017. Deep learning for mobile multimedia: a survey. *ACM Trans Multim Comput Commun Appl*, 13(3S), Article 34. https://doi.org/10.1145/3092831

Papernot N, McDaniel P, Jha S, et al., 2016. The limitations of deep learning in adversarial settings. IEEE European Symp on Security and Privacy, p.372-387. https://doi.org/10.1109/EuroSP.2016.36

Phong LT, Aono Y, Hayashi T, et al., 2018. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans Inform Forens Secur*, 13(5): 1333-1345. https://doi.org/10.1109/TIFS.2017.2787987

Ren SQ, He KM, Girshick R, et al., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Patt Anal Mach Intell*, 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ross AS, Doshi-Velez F, 2017. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. https://arxiv.org/abs/1711.09404

Sabour S, Cao YS, Faghri F, et al., 2015. Adversarial manipulation of deep representations. https://arxiv.org/abs/1511.05122

Shahid N, Aleem SA, Naqvi IH, et al., 2012. Support vector machine based fault detection & classification in smart grids. IEEE Globecom Workshops, p.1526-1531. https://doi.org/10.1109/GLOCOMW.2012.6477812

Shokri R, Shmatikov V, 2015. Privacy-preserving deep learning. Proc 53rd Annual Allerton Conf on Communication, Control, and Computing, p.1310-1321. https://doi.org/10.1109/ALLERTON.2015.7447103

Syarif AR, Gata W, 2017. Intrusion detection system using hybrid binary PSO and K-nearest neighborhood algorithm. 11th Int Conf on Information & Communication Technology and System, p.181-186. https://doi.org/10.1109/ICTS.2017.8265667

Vinayakumar R, Soman KP, Poornachandran P, et al., 2018. Detecting Android malware using long short-term memory (LSTM). *J Int Fuzzy Syst*, 34(3):1277-1288. https://doi.org/10.3233/JIFS-169424

Vollmer T, Manic M, 2009. Computationally efficient neural network intrusion security awareness. Proc 2nd Int Symp on Resilient Control Systems, p.25-30. https://doi.org/10.1109/ISRCS.2009.5251357

Vuong TP, Loukas G, Gan D, et al., 2015. Decision tree-based detection of denial of service and command injection attacks on robotic vehicles. IEEE Int Workshop on Information Forensics and Security, p.1-6. https://doi.org/10.1109/WIFS.2015.7368559

Wu J, Dong MX, Ota K, et al., 2018. Big data analysis-based secure cluster management for optimized control plane in software-defined networks. *IEEE Trans Netw Serv Manag*, 15(1):27-38. https://doi.org/10.1109/TNSM.2018.2799000

Xie CH, Wang JY, Zhang ZS, et al., 2017. Adversarial examples for semantic segmentation and object detection. IEEE Int Conf on Computer Vision, p.1378-1387.

https://doi.org/10.1109/ICCV.2017.153

Xin Y, Kong LS, Liu Z, et al., 2018. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6:35365-35381. https://doi.org/10.1109/ACCESS.2018.2836950

Xu WL, Evans D, Qi YJ, 2017. Feature squeezing mitigates and detects Carlini/Wagner adversarial examples. https://arxiv.org/abs/1705.10686

Yuan XY, 2017. PhD forum: deep learning-based real-time malware detection with multi-stage analysis. IEEE Int Conf on Smart Computing, p.1-2. https://doi.org/10.1109/SMARTCOMP.2017.7946997

Zhao GZ, Zhang CX, Zheng LJ, 2017. Intrusion detection using deep belief network and probabilistic neural network. IEEE Int Conf on Computational Science and Engineering and IEEE Int Conf on Embedded and Ubiquitous Computing, p.639-642. https://doi.org/10.1109/CSE-EUC.2017.119

Zhu DL, Jin H, Yang Y, et al., 2017. DeepFlow: deep learning-based malware detection by mining Android application for abnormal usage of sensitive data. IEEE Symp on Computers and Communications, p.438-443. https://doi.org/10.1109/ISCC.2017.8024568

Zolotukhin M, Hämäläinen T, Kokkonen T, et al., 2016. Increasing web service availability by detecting application-layer DDoS attacks in encrypted traffic. Proc 23rd Int Conf on Telecommunications, p.1-6. https://doi.org/10.1109/ICT.2016.7500408