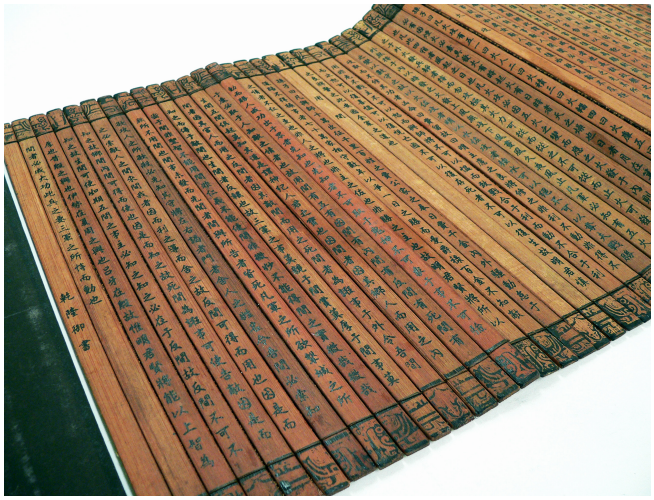
February 2021, Issue Code NL-2021-4-C

DDD (Digital Data Deception) Technology Watch Newsletter: Chinese Section

Table of Contents

- Editorial
- List of Acronyms
- Adversarial AI
- Information Hiding



Source: `https://www.flickr.com/photos/bluefootedbooby/370460130/`

“兵者，诡道也。故能而示之不能，用而示之不用，近而示之远，远而示之近。利而诱之，乱而取之，实而备之，强而避之，怒而挠之，卑而骄之，佚而劳之，亲而离之。攻其无备，出其不意。此兵家之胜，不可先传也。”

— 孙武:《孙子兵法·始计篇》

(The above is the original Chinese version of the English quotation shown on the cover page of the newsletter's main issue covering English papers.)

Editors: Li Qin, Shujun Li, Virginia Franqueira, Sanjay Bhattacherjee, and Enes Altuncu
Affiliation: Kent Interdisciplinary Research Centre in Cyber Security (KirCCS), University of Kent, UK
Contact Us: ddd-newsletter@kent.ac.uk

Editorial

In this fourth issue of the Digital Data Deception (DDD) newsletter, we continue to include a Chinese section covering two selected topics related to DDD: adversarial AI, and information hiding.

The papers covered in this Chinese section were identified via a mixed method: some were identified via a keyword-based search into Scopus, and others by manually inspecting the tables of contents of some selected journals. For the second method, the journals were selected based on the following two main criteria: if they provide fulltext access to papers, and if they are top-tier or highly technically relevant journals. In total 11 papers were selected, 5 on adversarial AI and 6 on information hiding. Four papers covered are reviews (two for each topic) and the others are about original research.

For Chinese papers covered in this issue, we paid special attention to source code and data released under a public source or open access license, but found that none of the authors of the 11 papers included such information in their papers. We also searched for websites of the authors and their research groups, but did not find links of source code or data related to the papers we covered in this issue. This result was unexpected but not totally surprising, and may simply be caused by the small sample of papers covered in this issue. We will continue to pay attention to this aspect of related research published in Chinese and see if this is actually related to a different research culture in China.

We hope you enjoy reading the Chinese section of this issue. Feedback is always welcome, and should be directed to ddd-newsletter@kent.ac.uk.

List of Acronyms

- AUC: Area Under Curve
- BOAP: Bilevel Optimisation Poisoning Attacks
- CNN: Convolutional Neural Network
- DGANS: Double-GAN-based Steganography
- DL: Deep Learning
- DNN: Deep Neural Network
- FGSM: Fast Gradient Sign Method
- GAN: Generative Adversarial Network
- GNCNN: Gaussian-Neuron Convolutional Neural Network
- HMAC: Hash-based Message Authentication Code
- RNN: Recurrent Neural Network
- ROC: Receiver Operating Characteristic

Adversarial AI

Introduction

This section covers two review papers and three papers reporting original research (two about generating and one about defending against adversarial samples). For each paper, we provide editorial comments immediately after the corresponding summary. For the two review papers, to avoid the summaries become too lengthy and losing focus, we decided to not show the full name of every acronym and cite all relevant original papers.

Two Reviews

Pan (潘文雯) et al. [8] reviewed research on generating adversarial examples used for attacking image classification and recognition systems. They first introduced some basic concepts such as white- and black-box tests and robustness of adversarial samples. They explained the general procedure of how adversarial samples are generated, tested and defended as shown in the Figure 1. This includes the following main steps: training the AI model with normal samples, adding perturbation to generate adversarial samples, testing if the attacked AI model can correctly classify adversarial samples, and adversarial training – using adversarial samples with fixed (correct) labels to re-train the AI model in order to test the robustness of the adversarial samples. The paper then proposes a two-layered categorisation scheme to classify different adversarial sample generation methods:

- Whole-image: perturbation is added to all pixels of an attacked image to produce an adversarial sample. Under this category, the authors used a second level of categorisation: (a) target attacks where adversarial samples allowed a classifier to mis-classify some samples into a specific target class, (b) non-target attacks where adversarial samples mislead a classifier to mis-classify some samples into any incorrect class, and (c) generic attacks that support both target and non-target attacks.
- Partial-image: perturbation is added to only selected pixels of an attacked image to produce an adversarial sample. Under this category, the authors used a second level of categorisation: those with invisible perturbation, and those with visible perturbation.

For both categories, the authors also considered if a method uses black- or white-box testing. Using the above categorisation, the authors introduced the following adversarial sample generation methods developed after the two earliest methods proposed in [5, 13]:

- I-FGSM and DeepFool as whole-image and non-target attacks;
- ILCM, Carlini & Wagner attack, UPSET, ANGRI, Houdini as whole-image and target attacks;
- ATNs, MI-FGSM, Curls & Whey attack as whole-image and generic attacks;
- JSMA and ONE-PIXEL as partial-image and invisible attacks; and
- Adversarial Patch, LaVAN, PS-GAN and Printable Adversarial Patches as partial-image and visible attacks.

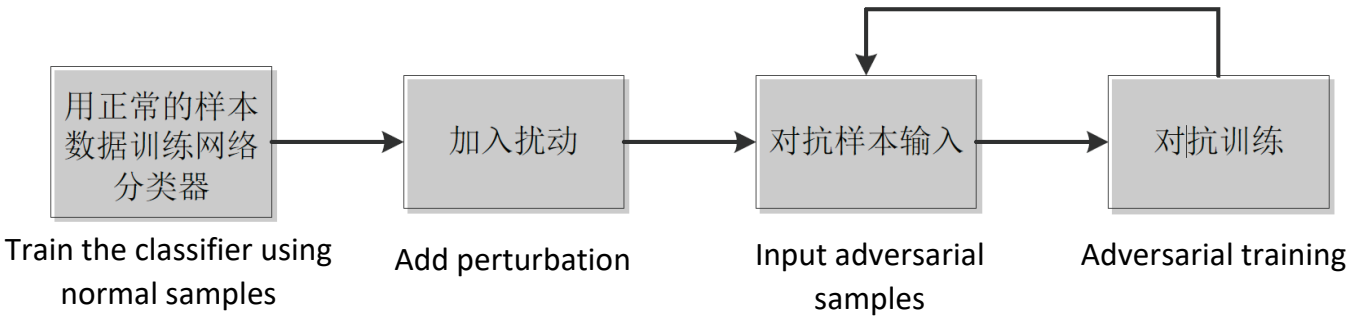


Figure 1: The common process for generating adversarial samples (Figure 1 in Pan (潘文雯) et al. [8]).

The authors pointed out that the MNIST dataset of handwritten digits (`http://yann.lecun.com/exdb/mnist/`) had been widely used for evaluating the performance of different adversarial sample generation methods. As an additional contribution beyond many review papers, they also conducted some experiments using the MNIST dataset to compare the performance (including their robustness against adversarial training) of 9 adversarial sample generation methods. Finally, the authors pointed out some challenges and future research directions, including transferability and robustness, applications of adversarial samples for both attacks and defence, and perturbing the AI model itself.

Editorial Comments

We would like to highlight a number of observations about this paper: 1) Figure 1 does not cover all aspects of adversarial sampling, e.g., the poisoning attack on the training data is not covered. 2) Under challenges and future research directions, the paper also lists the application of adversarial samples to OCT (optical coherence tomography) based fingerprint recognition. This looks very ad hoc and out of context, so we did not include it in the above summary.

Wang (王科迪) and Yi (易平) [16] reviewed a supplementary sub-area of adversarial AI: model robustness against adversarial samples. They started with an introduction to adversarial samples followed by an overview of possible reasons behind the existence of adversarial samples: linearity of the loss function, higher dimensionality of adversarial samples (than normal samples), existence of global perturbation, and ubiquitous existence of fragile features in sample space. After acknowledging three different attacks with adversarial samples – evasion, poisoning, and model stealing – the authors focused on two important aspects of model robustness against adversarial samples: evaluation of robustness, and enhancement of robustness. For the evaluation part, the authors argued that the minimal perturbation needed to create adversarial samples to attack a classifier can be consider a useful metric of the classifier's robustness against adversarial samples. It has been proven that estimation of the minimal perturbation is an NPC problem, so related work mainly looked at how to estimate a good upper or lower bound. In addition to the minimal perturbation that is normally defined as an L_p distance in the feature space, the paper also covers three other metrics proposed in the literature: (a) ASS (Average Structural Similarity) considering structural similarities between samples, (b) perturbation sensitivity distance considering human visual system's contrast masking effect, and (c) a new metric based on the Wasserstein distance. Finally, the authors reviewed three groups of methods for enhancing model robustness: 1) modifying data in the training set (e.g., adversarial training, data compression); 2) modifying the structure of the AI mode (e.g., gradient masking, defensive distillation, adding a new class for adversarial samples); 3) adding additional modules that can help enhance robustness (e.g., additional processor against global perturbation, feature compression). In this paper's conclusion, the authors also pointed out three areas for future research: better robustness metrics, a unified evaluation framework, and more theoretical work on the existence of adversarial samples.

Editorial Comments

Although this paper has a relatively narrow focus on model robustness against adversarial samples, it has a very balanced coverage of many important aspects of adversarial samples, e.g., the different attempts to explain why adversarial samples can exist.

Adversarial Sample Generation

Qian (钱亚冠) et al. [9] proposed an adversarial sampling generation method, which adds a printed QR code to the surface of a road sign to achieve the purpose of misleading a road sign recognition system. Although the attack is a physical attack, the authors tested the attack in the digital domain first and then extended the result to the physical domain by considering robustness of the attack against different lighting conditions, geometrical distortions, printing artefacts and quality degradation of real-world image capturing in the outdoor environment. The QR code images used in the proposed attacks are actually pseudo-QR codes because they look like real QR codes but do not include copyright information. The proposed attack's procedure is illustrated in Figure 2. Firstly, a classifier is built based on DNNs

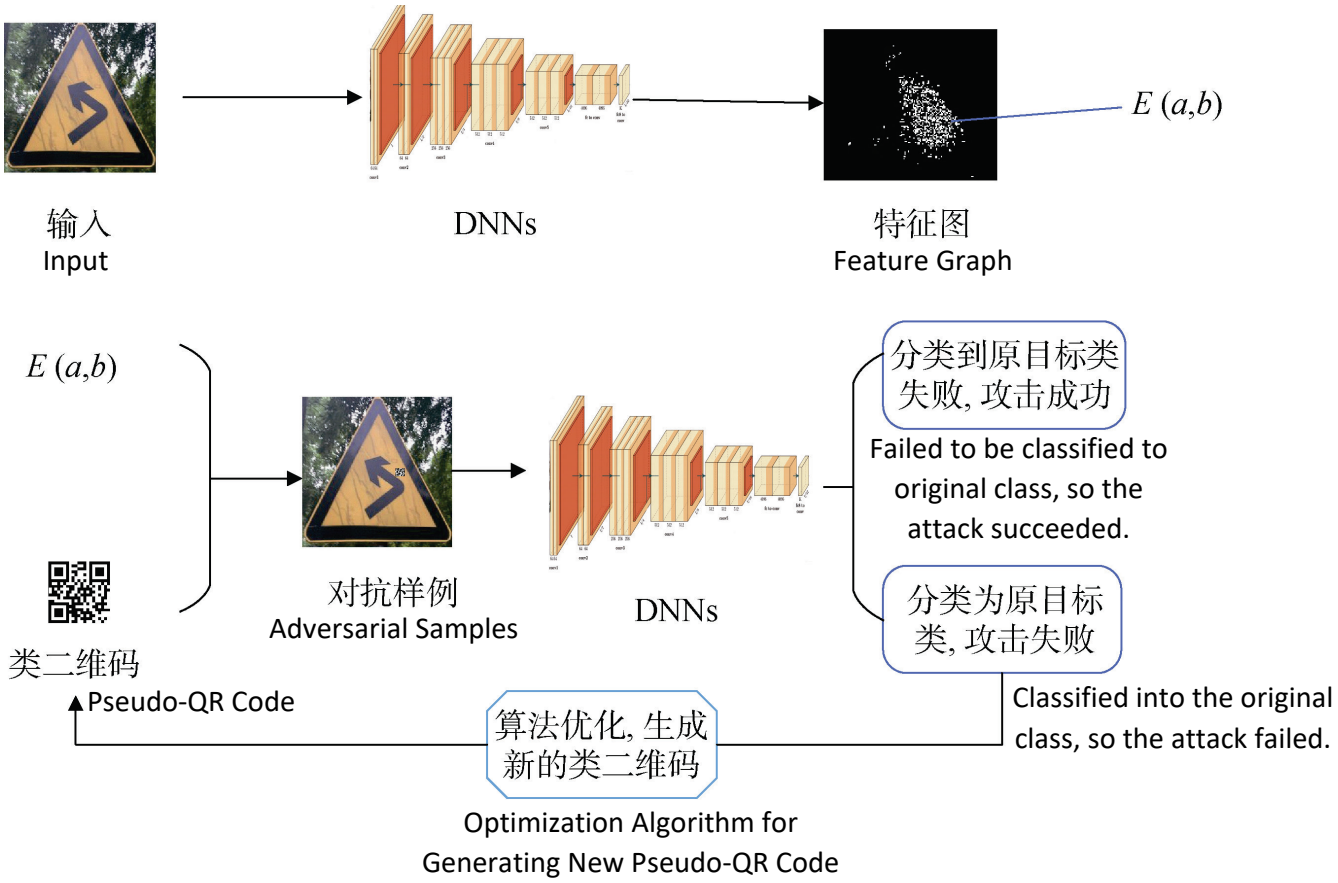


Figure 2: The pseudo-QR code based adversarial sample generation method proposed in [9].

(deep neutral networks) to classify traffic sign images. Then, taking a target road sign image as input, the classifier processes it to get the coordinate of the most important pixel $E(a, b)$. Next, the attacker randomly generates a pseudo-QR code image, which is added to the area at $E(a, b)$ to obtain a candidate adversarial image. Finally, the attacker tests the candidate adversarial image using the classifier to see if the attacked image can mislead the recognition. If not, another pseudo-QR code image is regenerated and the above process is repeated until it achieves the goal of the attack. The generation of the pseudo-QR code images follows an optimisation problem, which minimises the added perturbation that can lead to a mis-classification (see Eqs. (6)–(9) in the paper). The authors tested the proposed attack using the VGG-16 model as the classifier. The dataset used is a merger of the G-TSRB (German traffic sign recognition benchmark) dataset and some Chinese road sign images collected by the authors. In total the dataset includes 42,693 training images and 5,200 test images. The performance of the attack was tested under different conditions, in both digital and physical domains. The experimental results showed that the proposed attack largely worked very well. The exact accuracy varies depending on a number of factors, including the type of road sign (the success rate ranging from 76% to 100% in the digital domain), size of the pseudo-QR code image (the larger the better), if the QR-code is black and white or in colour (the latter is better), and the number of iterations (the more the better). The authors also tested the transferability of the generated adversarial samples from VGG-16 to VGG-19, with positive results (although expected reduced success rates). The performance in the physical world was generally lower than that in the digital domain, which was not surprising due to various distortions and inaccuracies different physical processes can bring in. As a whole, the authors argued that their proposed method has the following key advantages: 1) the area of the added pseudo-QR code is very small (the average area is 0.95%), much smaller than previous work; 2) the attacking process is very fast due to the optimisation method; 3) the common use of pseudo-QR code in the physical world makes such an attack less

detectable by humans.

Editorial Comments

Although the paper's main focus is a physical attack, the proposed attack also works in the digital domain with better performance. The paper actually contains more experimental results for the digital version of the attack. The proposed attack seems to be generalisable to other types of images and physical applications and transferable to other machine learning models.

Yan (闫佳) et al. [18] proposed a genetic algorithm-based method for generating adversarial samples of malicious code to evade detection by malware detectors. The basic idea is to apply the generic algorithm to apply static changes to the source code of malware, while maintaining its dynamic behaviours (verified inside a sandbox after static changes are made). Using NVIDIA's MalConv [10] as the target malware detection model, the detailed procedure can be described as follows:

1. Randomly generate K sequences of "genome" (i.e., changes to be made) and apply them to the target samples to produce an initial set of adversarial samples.
2. Analyse the first generation of adversarial samples and record the structural differences between them and the original sampls.
3. Use MalConv to all adversarial samples to produce their adaptability scores (as a value between 0 and 1).
4. Choose the suitable individual adversarial samples passing some set threshold of adaptability as the parents of the next generation.
5. Use the open-source sandboxing system Cuckoo Sandbox (`https://cuckoosandbox.org/`) to select parent samples that keep the original behaviours of the original samples.
6. Apply genetic operations (i.e., crossovers and mutations) to the selected parent samples to generate a new set of K "genomes" for the next generation.
7. Determine if the generic algorithm should end based on some exiting criteria.

Their experiments were conducted based on 15,168 samples provided by the Research Institute of Qi-AnXin Group (奇安信技术研究院). The results showed that 239 samples managed to evade detection, leading to a 14.65% reduction of detection accuracy (from 98.88% to 84.23%). The authors also ran a real-world test with four malware detection engines on VirusTotal.com to test if the 239 adversarial samples can also evade detection by more real-world engines. The results showed that 235 samples could evade at least one engine, and the overall evasion rate ranges from 14.64% to 46.45%.

Editorial Comments

Although the reduction of performance and the evasion rate are not very high, the work can still be practical for malware makers since they only need to focus on those successfully bypassing detection. It is likely that malware makers need to analyse the successful adversarial samples to learn how to make their malware harder to detect.

Defence Against Adversarial Samples

Zhou (周文) et al. [21] studied how intrusion detection systems (IDSs) for protecting industrial control systems (ICSs) can resist adversarial samples and if adversarial training can help resist white-box attacks. The work was done for a low-dimensional IDS dataset for the gas pipeline control system (`https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets`, Database 4). This work is more experimental and involved the following aspects: 1) how four different optimisation methods, SGD, RMSProp, AdaDelta and Adam, performed in both white- and black-box attacking settings, using DNN and FGSM (Fast Gradient Sign Method) to generate adversarial samples; 2) how seven different machine learning models, decision trees, random forests, linear SVM, AdaBoost, logistic regression, CNN (Convolutional Neural Network) and RNN (Recursive Neural Network), performed against different types of adversarial samples; and 3) if adversarial training could help improve the resistance against adversarial samples. The experimental results showed that Adam is the best optimisation method for both white- and black-box attacking settings. All machine learning models tested are vulnerable to adversarial samples, but RNN is

more resistant with a smaller drop of detection accuracy compared to the other models (e.g., from 89.4% to 72.6-85.4% for F1-scores). The results of the adversarial training also showed that using adversarial samples with correct labels helped improve the robustness of the original DNN-based IDS against adversarial samples in the white-box attack setting.

Editorial Comments

This paper does not report new methods, but a series of experiments testing different aspects of adversarial samples attacking a low-dimensional IDS. The authors claimed to have proposed a new metric called "relative loss rate" (同比损失率), which is actually a very simple metric based on the accuracy (ACC). We did not highlight this as their key contribution.

Information Hiding

Introduction

This section covers two review papers and four papers reporting original research (three about hiding data in traditional media, and one about hiding data in blockchains). The two review papers are of particular interest because they collectively cover recent developments in deep learning (DL) based steganography and steganalysis.

To help our readers better understand the content of this section, we give a brief explanation on how information hiding works. Traditionally, an information hiding scheme takes two inputs: a cover (i.e., an information carrier) and a message to be embedded (hidden) into the cover. The cover can be any object such as a signal, a file, a network packet, a communication channel, or a combination of multiple objects. The message to be hidden is normally translated into a bit sequence before being embedded, and the embedding is done bit by bit or chunk by chunk (where each chunk contains a number of bits). More recently, so-called coverless information hiding schemes have been proposed, where the cover is not an input any more, but is automatically generated or synthesised (e.g., based on a large database of potential cover database, or a generative machine learning model trained using such a database). Since the cover is generated or synthesised dynamically, it can often be tailored to optimise one or more attributes of the embedding process (e.g., security or embedding capacity). Since a cover is still involved in the embedding process, some researchers use the less confusing terms "generative covers" and "synthetic covers".

Two Reviews

Fu (付章杰) et al. [4] reviewed recent development of image steganography based on deep learning (DL), an exploding research direction within steganography since the first such scheme SGAN (Steganographic Generative Adversarial Networks) was proposed in 2016 [15]. The authors classified related work into four different types. We summarise the four types and give each type a short name (our term, not the authors'):

1. *DL-based generative schemes* use DL techniques to generate the cover image that is more suitable for hiding information. They then apply traditional embedding methods to hide the message in the cover.

2. *DL-enhanced traditional schemes* use DL techniques to enhance the performance of tradi-

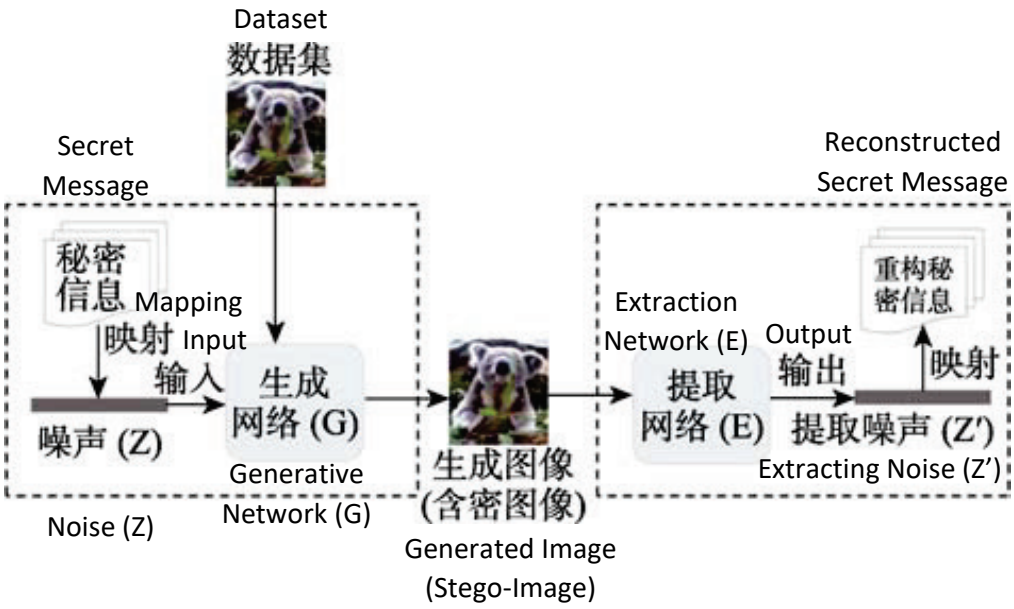


Figure 3: How a coverless steganographic scheme works (Figure 10 in [4]).

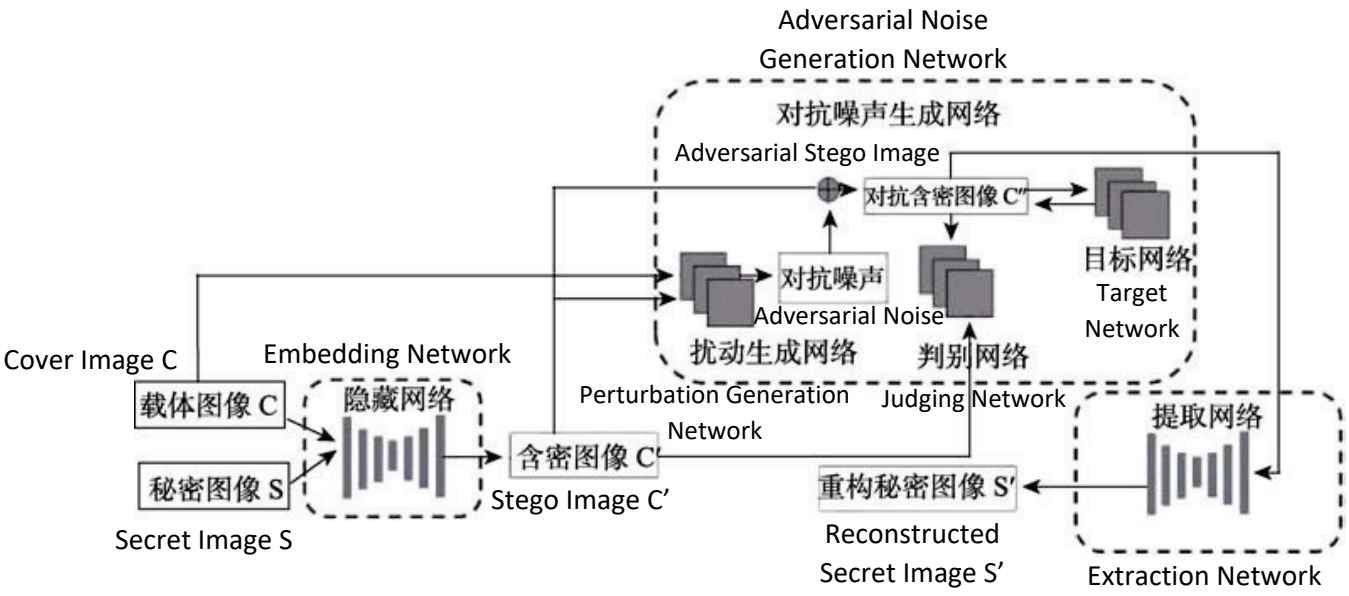


Figure 4: An improved DL-based steganographic scheme proposed in [4].

tional embedding methods without directly changing the cover image.

3. *DL-based synthetic schemes* apply DL techniques to synthesise a new cover based on the given cover image, which at the same time completes the embedding of the hidden message.

4. *DL-based coverless schemes* apply DL techniques to represent the hidden message as a mapping between the hidden message and the cover. This type of schemes are also based on generative covers, but do not involve a follow-up direct embedding step.

The procedure for coverless schemes is shown in Figure 3. For other types of schemes, please refer to Figures 2-9 in [4]. These four types of DL-based steganographic schemes have different advantages and disadvantages, and Fu (付章杰) et al. gave a general summary as follows (with our modifications, see our editorial comments below):

- *DL-based generative schemes*: Generated cover images can be less realistic so the security can be limited.

- *DL-enhanced traditional schemes*: The DL techniques can help increase the embedding capacity, but may leave more noticeable embedding traces therefore leading to new security issues.

- *DL-based synthetic schemes*: The extraction process requires an additional mask (for the synthesised cover).

- *DL-based coverless schemes*: No embedding traces, but the embedding capacity is small and the hidden message cannot be extracted with 100% accuracy.

For each of the four types of DL-based steganography, the authors reviewed a number of schemes proposed in the research literature. They also conducted a number of experiments to test their performances, especially security. In addition to reviewing related work, for a sub-class of the DL-enhanced traditional schemes based on the so-called encoder-decoder network, the authors also proposed an adversarial training based method to improve its security. The architecture of the proposed method can be seen in Figure 4. The adversarial noise generation network is the one introducing the adversarial training element to make the original scheme more secure. They also conducted experiments with 50,000 ImageNet images and provided initial evidence that the proposed method could help improve the security.

Editorial Comments

When summarising the general advantages and disadvantages of the four basic types

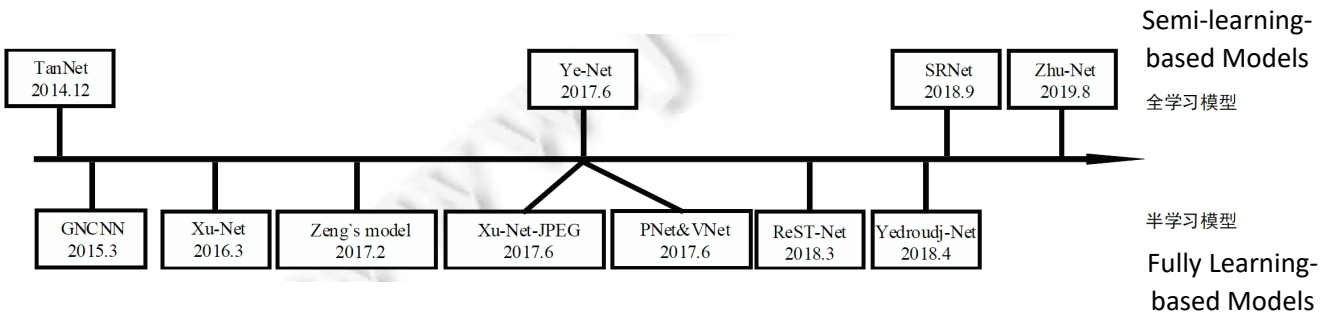


Figure 5: A timeline of a number of DL-based steganalytic schemes reviewed in [3].

of DL-based steganographic schemes, Fu (付章杰) et al. made some problematic statements (in Table 1) for DL-enhanced traditional schemes and for DL-based coverless schemes. We decided to use our own words based on our understanding of the discussions and experimental results in [4].
Although Fu (付章杰) et al. also proposed a new steganographic scheme, the experimental results provided are clearly preliminary and not comprehensive. We therefore did not highlight the performance metrics reported in the paper.

Chen (陈君夫) et al. [3] reviewed recent research on DL-based steganalysis. The authors categorised related work into two classes, depending on different training models of the pre-processing layer of the DL network: *semi-learning-based* and *fully learning-based*. Both classes actually share a similar overall structure – a pre-processing layer followed by a DL network. For both classes, the authors separately categorised related work further into those based on deep-wise DL networks and those based on wide-wise DL networks. The main difference between the semi-learning and full-learning-based schemes is that for the former the pre-processing layer is constructed based on (SRM or DCTR) filter kernels used in traditional steganalysis that have fixed (manually defined) parameters and do not participate in the adaptive learning process. In contrast, for fully-learning-based steganalysis, the pre-processing layer participates in the learning process so the parameters are learned from the training set, too. Semi-learning-based steganalysis can be seen as a middle ground between more traditional steganalysis and fully learning-based steganalysis, which helps make them more efficient. Performance wise, more recently proposed semi-learning-based steganalytic schemes have also achieved a comparable detection accuracy to fully-learning-based schemes. In addition to having a slower training process, fully-learning-based schemes also have a higher risk of over-fitting so their generalisability can be a problem. One particular scheme (SRNet [1]) seems to have a better level of generalisability across different datasets and resistance against adversarial samples (at the cost of being the slowest among all schemes). In terms of the detection accuracy, the best performing one reviewed in the paper is Zhu-Net [20] with a false detection rate of 15.3%, but two semi-learning-based schemes have a very close performance (ReST-Net [6] with 16% and VNet [2] with 16.9%) with a roughly halved training time. Figure 5 shows the timeline of main DL-based steganalytic schemes reviewed in this paper. In addition to a general review of all DL-based steganalysis schemes, the authors also looked at some recent research investigating subtle relationships between steganalysis and adversarial samples. On the one hand, some adversarial attacks have been proposed to reduce the detection accuracy of some steganalytic methods, while on the other hand, some researchers have also proposed to use steganalysis to detect adversarial samples. Based on observed problems of existing schemes, the authors also proposed some future research directions in DL-based steganalysis: 1) end-to-end, fully automated schemes with no human intervention; 2) speeding up the training process; 3) learning based on small datasets; and 4) fusing multiple models.

Editorial Comments

This paper also contains a concise but still informative section on general background of steganalysis including commonly used perfor-

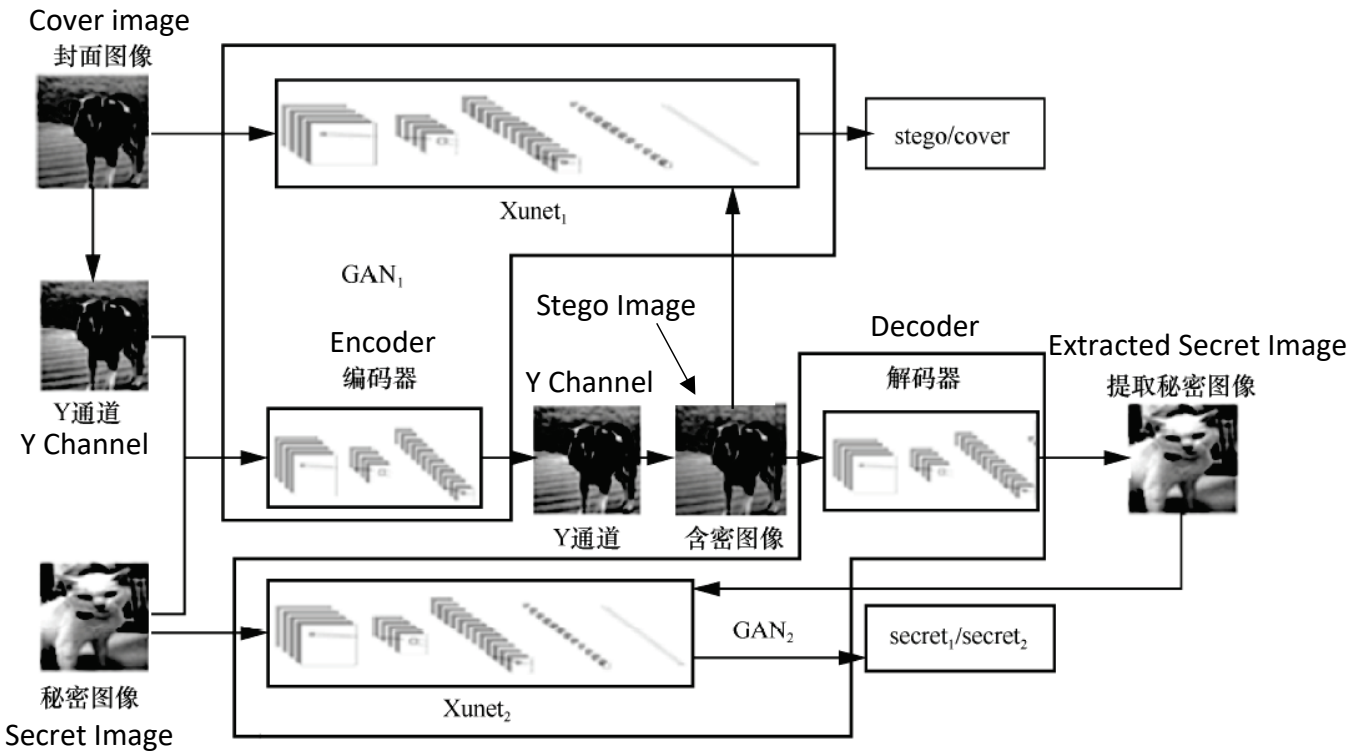


Figure 6: The structure of the DGANS steganographic model proposed by Zhu (竺乐庆) et al. [22].

mance metrics and databases (Section 1).

Hiding in Traditional Media

Zhu (竺乐庆) et al. [22] proposed a new high-capacity (up to the level of 8 bpp (bits per pixel)) image steganographic model called DGANS (double-GAN-based steganography), which is based on two GANs (generative adversarial networks), one at the encoder side for improving security and the other at the decoder side for improving robustness. Figure 6 shows the overall structure of the DGANS model. An encoder-decoder network based on the Inception model is used for embedding a gray-scale hidden image into the Y channel of a colour cover image in the YUV colour space. The first security-oriented GAN uses the generative network to produce candidate stego images and the discriminative network as a simulated steganalytic component. The second robustness-oriented GAN uses the generative network to extract the hidden image and the discriminative network to find how to enhance the system's robustness. The authors also proposed to use an enhanced dataset with geometric transformations (translation, rotation, and scaling) for an additional training process of the second GAN to further enhance the system's robustness against geometric transformations. They conducted experiments using the PASCAL VOC2012 dataset with 11,540 training images (containing 50% stego images) and 5,000 test images. Compared with two state-of-the-art high-capacity methods proposed in [14, 19], the experimental results showed that the model proposed in this paper achieved a comparable performance in security, but better robustness against geometric transformation (significantly so for all three types of transformations).

Editorial Comments

This paper does not make it clear if the 5,000 test images were not overlapping with the training set. The authors did not use cross-validation so the results may be less reliable. In addition, for the security, the authors used SSIM values and six visual examples to demonstrate the DGANS model performed similar to competing methods. They also compared the AUC (area under curve) of the ROC (receiver operating characteristic) curve with one competing method (S-UNIWARD). We

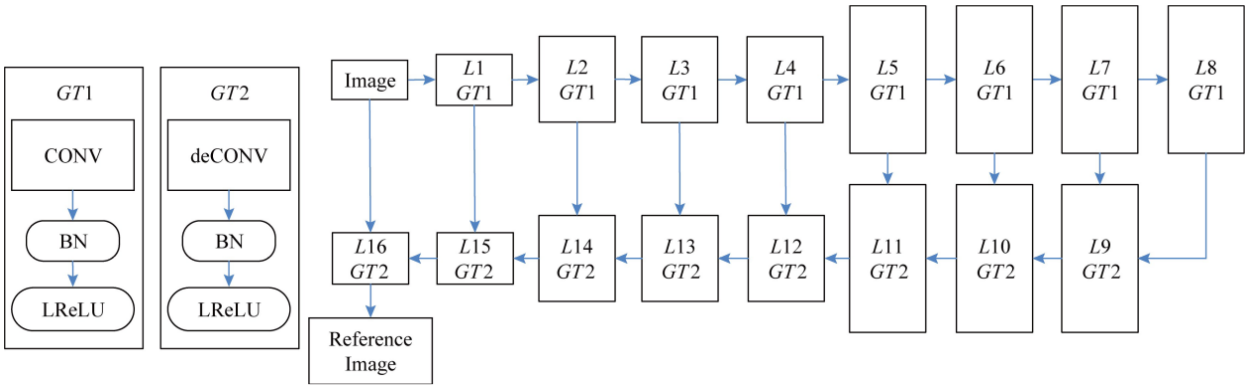


Figure 7: Structure of reference image generation network for JPEG image deep learning steganalysis model proposed by Ren (任魏翔) et al. [11].

felt more evidence could be produced to support the comparison.

Li (李林聪) et al. [7] proposed a video steganography method based on embedding the hidden message as non-additive information into motion vectors of the encoded cover video. The non-additive embedding is achieved by modelling the probability of each motion vector being modified for embedding (from a joint model of possible modifications to all motion vectors) and then splitting each probability into the horizontal and vertical directions in a non-additive manner. The overall embedding process includes three steps: defining distortion, embedding the hidden message, and re-encoding the video. The experiments were done using 15 widely used YUV video sequences (format: 4:2:0, CIF). The performance was tested against that of four state-of-the-art motion vector modification based video steganographic schemes, using three representative video steganalytic methods. The experimental results showed that the proposed method showed a higher level of security than other four methods under different embedding ratios. The proposed method was also able to produce stego images with (slightly) higher visual quality measured using PSNR as the metric. In addition, the proposed method led to less expansion of the size of the stego video than the four other methods, often with a significant margin. While being able to outperform other methods, their experimental results also showed that there was no additional computational complexity.

Editorial Comments

The method proposed looks to offer only merits without any obvious disadvantages, so if the results can be reproduced, the method can be a good benchmark for any future motion vector modification based video steganographic method. The experimental results are based on only 15 videos, so more evidence may be needed to consolidate the conclusions in the paper.

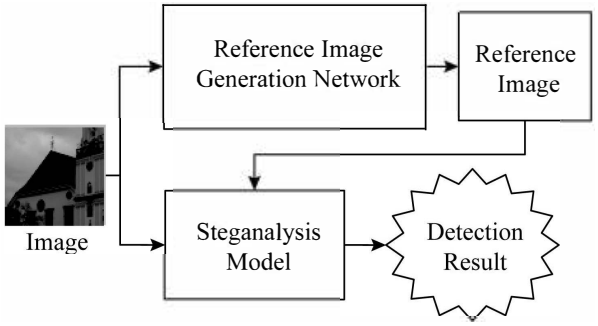


Figure 9: The generation and use of reference images in steganalysis (Figure 3 in [11].

Ren (任魏翔) et al. [11] explored how to further improve the detection ability of JPEG image steganalytic models based on more informative generative reference images. The method follows the general structure of reference image based steganalysis, as shown in Figure 9. Here, the reference images are used to help the steganalytic model to extract more useful information about a stego image by comparing it with reference images.

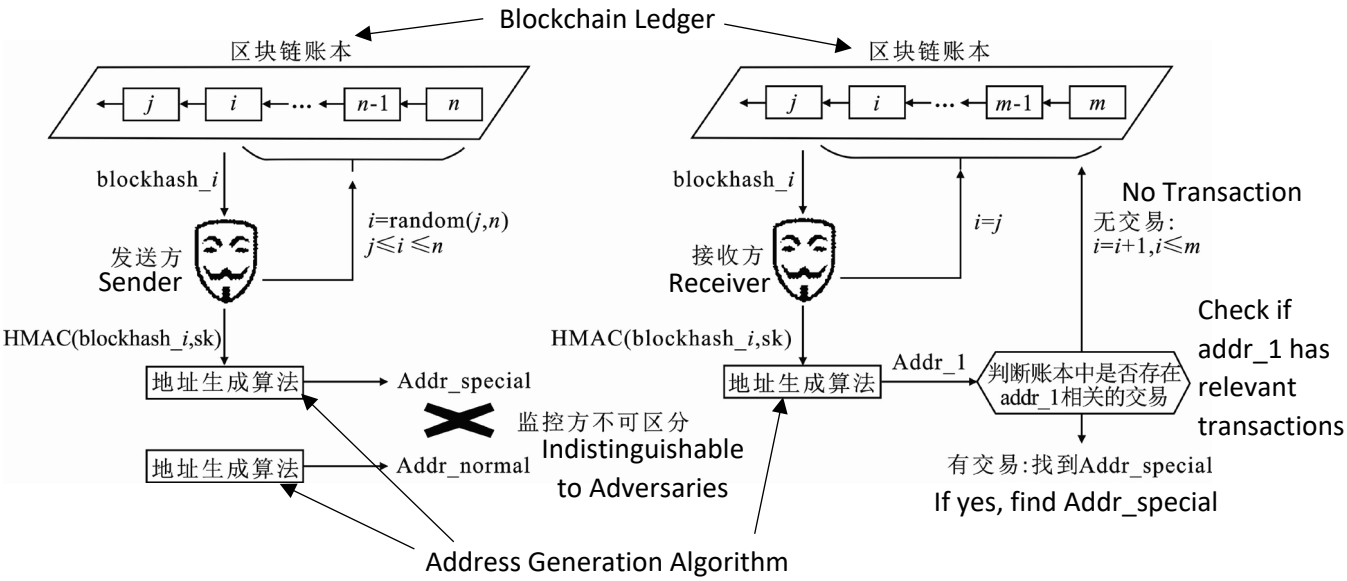


Figure 8: The blockchain-based data hiding method using dynamically generated addresses proposed by Si (司成祥) et al. [12].

The reference image generation network proposed in the paper is based on a CNN with 8 convolutional layers (all following the same structure shown as GT1) and 8 deconvolutional layers (all following the same structure shown as GT2). Each basic (de)convolutional layer (GT1 or GT2) includes a standard (de)convolutional layer, a batch normalisation (BN) layer and a LReLu activation layer. To suppress gradient disappearance, a skip connection is added between each pair of the convolutional layer and the deconvolutional layer at the corresponding symmetric position of the whole model. A visual representation of the model's structure can be seen in Figure 7. The authors proposed two ways to train the reference image generation model – the pre-training mode based on mean square error, and a "together" training mode where the training has the steganalytic model in the loop. The experiments were conducted using 10,000 images in the BOSSbase v1.01 dataset (`http://dde.binghamton.edu/download/ImageDB/BOSSbase_1.01.zip`), which were processed to obtain 50,000 cover images (the original images plus four sub-images extracted from different regions). The 50,000 cover images and the same number of stego images were split into 80,000 for the training set, 10,000 for the validation set, and 10,000 for the testing set. J-Xunet [17] was used as the base-line steganalytic model to see if adding reference images generated using the proposed method can improve the detection accuracy. The results are largely positive, and the proposed model was able to improve the detection ability by up to 6%.

Editorial Comments

Reference images are widely used in steganalysis to expose remaining information of covers to discover key differences between stego images and cover images. This reported work therefore can be applied to any steganalytic method where reference images are used.

Hiding in Blockchain

Si (司成祥) et al. [12] proposed the use of dynamically generated addresses as a new information carrier to facilitate data hiding in blockchains, rather than hiding data in transactions of static addresses as what other traditional methods do. The generation of such dynamic addresses follows a pre-agreed mechanism between the sender and the receiver, and the security aim is to make the generated addresses look indistinguishable from other normal addresses from the eyes of other users of the blockchain system. Once the sender identifies such an information-carrying address, he/she can extract messages hidden in relevant transactions associated with the address. The main advantage of the proposed scheme is a clear separation of the information-carrying addresses and the actual transaction data, therefore

helping reduce the exposure risk of the information-carrying addresses to adversaries. The general procedure of the proposed data hiding scheme is shown in Figure 8. The authors used HMAC as an example mechanism to demonstrate the proposed method, where the key and other information are shared between the sender and the receiver outside the blockchain data structure. The method was tested using the public Bitcoin network to verify its real world performance, showing it is effective and efficient.

Editorial Comments

Information hiding in blockchain has been used by many users and some tools (e.g., Apertus available at `http://apertus.io/`) have appeared to allow anyone to upload data to different blockchain networks. Many methods are not for security purposes, so research on using this new channel for steganography is still limited.

References

[1] Mehdi Boroumand, Mo Chen, and Jessica Fridrich. 2019. Deep Residual Network for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 14, 5 (2019), 1181–1193. https://doi.org/10.1109/TIFS.2018.2871749

[2] Mo Chen, Vahid Sedighi, Mehdi Boroumand, and Jessica Fridrich. 2017. JPEG-Phase-Aware Convolutional Neural Network for Steganalysis of JPEG Images. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 75—84. https://doi.org/10.1145/3082031.3083248

[3] Jun-Fu Chen (陈君夫), Zhang-Jie Fu (付章杰), Wei-Ming Zhang (张卫明), Xu Cheng (程旭), and Xing-Ming Sun (孙星明). 2021. Review of Image Steganalysis Based on Deep Learning / 基于深度学习的图像隐写分析综述. *Journal of Software* /《软件学报》 32, 2 (2021), 1–29. http://www.jos.org.cn/1000-9825/6135.htm

[4] Zhang-Jie Fu (付章杰), Fan Wang (王帆), Xing-Ming Sun (孙星明), and Yan Wang (王彦). 2020. Research on Steganography of Digital Images based on Deep Learning / 基于深度学习的图像隐写方法研究. *Chinese Journal of Computers* /《计算机学报》 43, 9 (2020), 1656–1672. https://doi.org/10.11897/SP.J.1016.2020.01656

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. Online pre-print, arXiv:1412.6572 [cs.CV]. https://arxiv.org/abs/1909.11573

[6] Bin Li, Weihang Wei, Anselmo Ferreira, and Shunquan Tan. 2018. ReST-Net: Diverse Activation Modules and Parallel Subnets-Based CNN for Spatial Image Steganalysis. *IEEE Signal Processing Letters* 25, 5 (2018), 650–654. https://doi.org/10.1109/LSP.2018.2816569

[7] Lincong Li (李林聪), Yuanzhi Yao (姚远志), Xiaoya Zhang (张晓雅), Weiming Zhang (张卫明), and Nenghai Yu (俞能海). 2020. Video Steganography Based on Modification Probability Transformation and Non-additive Embedding Distortion / 基于修改概率转换和非加性嵌入失真的视频隐写方法. *Journal of Electronics and Information Technology* /《电子与信息学报》 42, 10 (2020), 2357–2364. http://jeit.ie.ac.cn/cn/article/doi/10.11999/JEIT200001

[8] Wen-Wen Pan (潘文雯), Xin-Yu Wang (王新宇), Ming-Li Song (宋明黎), and Chun Chen (陈纯). 2020. Survey on Generating Adversarial Examples / 对抗样本生成技术综述. *Journal of Software* /《软件学报》 31, 1 (2020), 67–81. http://www.jos.org.cn/html/2020/1/5884.htm

[9] Yaquan Qian (钱亚冠), Xinwei Liu (刘新伟), Zhaoquan Gu (顾钊铨), Bin Wang (王滨), Jun Pan (潘俊), and Ximin Zhang (张锡敏). 2020. QR Code Based Patch Attacks in Physical World / 一种基于二维码对抗样本的物理补丁攻击. *Journal of Cyber Security* /《信息安全学报》 5, 6 (2020), 75–86. http://jcs.iie.ac.cn/xxaqxb/ch/reader/create_pdf.aspx?file_no=20200607&flag=1&year_id=2020&quarter_id=6

[10] Edward Raff, Jon Barker, Jared Sylvester, Robert Brandon, Bryan Catanzaro, and Charles Nicholas. 2018. Malware Detection by Eating a Whole EXE. In *Proceedings of Workshops of 32rd AAAI Conference on Artificial Intelligence*. AAAI, 268–276. https://aaai.org/ocs/index.php/WS/AAAIW18/paper/viewFile/16422/15577

[11] Weixiang Ren (任魏翔), Liming Zhai (翟黎明), Lina Wang (王丽娜), and Ju Jia (嘉炬). 2019. Reference Image Generation Algorithm for JPEG Image Steganalysis Based on Convolutional Neural Network / 基于卷积神经网络的 JPEG 图像隐写分析参照图像生成方法. *Computer Research and Development* /《计算机研究与发展》 56, 10 (2019), 2250–2261. https://doi.org/10.7544/issn1000-1239.2019.20190386

[12] Chengxiang Si (司成祥), Feng Gao (高峰), Liehuang Zhu (祝烈煌), Guopeng Gong (巩国鹏), Can Zhang (张璨), Zhuo Chen (陈卓), and Ruiguang Li (李锐光). 2020. Covert data transmission mechanism based on dynamic label in blockchain / 一种支持动态标签的区块链数据隐蔽传输机制. *Journal of Xidian University* /《西安电子科技大学学报》 47, 5 (2020), 94–102. https://doi.org/10.19665/j.issn1001-2400.2020.05.013

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. Online pre-print, arXiv:1312.6199 [cs.CV]. https://arxiv.org/abs/1312.6199

[14] Atique ur Rehman, Rafia Rahim, Shahroz Nadeem, and Sibt ul Hussain. 2019. End-to-End Trained CNN Encoder-Decoder Networks for Image Steganography. In *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 11132)*. Springer, 723–729.

[15] Denis Volkhonskiy, Boris Borisenko, and Evgeny Burnaev. 2016. Generative Adversarial Networks for Image Steganography. Submission to the 5th International Conference on Learning Representations (with Open Reviews). https://openreview.net/forum?id=H1hoFU9xe

[16] Kedi Wang (王科迪) and Ping Yi (易平). 2020. A Survey on Model Robustness under Adversarial Example / 人工智能对抗环境下的模型鲁棒性研究综述. *Journal of Cyber Security* /《信息安全学报》 5, 3 (2020), 13–22. http://jcs.iie.ac.cn/xxaqxb/ch/reader/create_pdf.aspx?file_no=20200303&flag=1&year_id=2020&quarter_id=3

[17] Guanshuo Xu. 2017. Deep Convolutional Neural Network to Detect J-UNIWARD. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 67—73. https://doi.org/10.1145/3082031.3083236

[18] Jia Yan (闫佳), Jia Yan (闫佳), Chujiang Nie (聂楚江), and Purui Su (苏璞睿). 2020. Method for Generating Malicious Code Adversarial Samples Based on Genetic Algorithm / 基于遗传算法的恶意代码对抗样本生成方法. *Journal of Electronics and Information Technology* /《电子与信息学报》 42, 9 (2020), 2126–2133. https://doi.org/10.11999/JEIT191059

[19] Ru Zhang, Shiqi Dong, and Jianyi Liu. 2018. Invisible steganography via generative adversarial networks. *Multimedia Tools and Applications* 78 (2018), 8559—8575. https://doi.org/10.1007/s11042-018-6951-z

[20] Ru Zhang, Feng Zhu, Jianyi Liu, and Gongshen Liu. 2020. Depth-Wise Separable Convolutions and Multi-Level Pooling for an Efficient Spatial CNN-Based Steganalysis. *IEEE Transactions on Information Forensics and Security* 15 (2020), 1138–1150. https://doi.org/10.1109/TIFS.2019.2936913

[21] Wen Zhou (周文), Shikun Zhang (张世琨), Yong Ding (丁勇), and Xing Chen (陈曦). 2020. Adversarial Example Attack Analysis of Low-Dimensional Industrial Control Network System Dataset / 面向低维工控网数据集的对抗样本攻击分析. *Journal of Computer Research and Development* /《计算机研究与发展》 57, 4 (2020), 736–745. https://doi.org/10.7544/issn1000-1239.2020.20190844

[22] Leqing Zhu (竺乐庆), Yu Guo (郭钰), Lingqiang Mo (莫凌强), and Daxing Zhang (张大兴). 2020. DGANS: robustness image steganography model based on double GAN / DGANS：基于双重生成式对抗网络的稳健图像隐写模型. *Journal on Communications* /《通信学报》 41, 1 (2020), 125–133. https://doi.org/10.11959/j.issn.1000-436x.2020019