J Inf Process Syst, Vol.17, No.3, pp.571~585, June 2021
https://doi.org/10.3745/JIPS.03.0161

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Secure Object Detection Based on Deep Learning

Keonhyeong Kim* and Im Young Jung*

Abstract

Applications for object detection are expanding as it is automated through artificial intelligence-based processing, such as deep learning, on a large volume of images and videos. High dependence on training data and a non-transparent way to find answers are the common characteristics of deep learning. Attacks on training data and training models have emerged, which are closely related to the nature of deep learning. Privacy, integrity, and robustness for the extracted information are important security issues because deep learning enables object recognition in images and videos. This paper summarizes the security issues that need to be addressed for future applications and analyzes the state-of-the-art security studies related to robustness, privacy, and integrity of object detection for images and videos.

Keywords

Deep Learning, Integrity, Object Detection, Privacy, Robustness

1. Introduction

Using an artificial intelligence (AI)-based approach, there is a well-known technique of object detection that applies deep learning as an automated process to perform analysis on large volumes of images and videos. Certain types of semantic objects, such as vehicles, pathological tumors, and people in digital images and videos, are usually what the object detection stochastically finds [1]. In tracking objects such as tracking movement of objects or tracking persons in a video, object detection is also used [2]. That is, each person, each car, or pathological tumor is often detected in the image and tracked in the video.

Sensitive information through object detection can be extracted through AI-based processing on image and video data. Personal movements, human faces, and personal illnesses analyzed and extracted from images and videos, for example, are all personal information to be protected. Not only the security of the data itself but also the security of AI-based processing and extracted information should be considered. The security problem of the AI-based approach becomes important accordingly as it has been used and will be used in many object detection applications. This paper summarizes the security issues that need to be addressed and analyzes the state-of-the-art security researches on object detection.

The contributions of this paper are as follows:

(1) We have analyzed security threats from cutting-edge research as the AI-based approach to detecting objects raises security concerns.

Manuscript received March 20, 2020; first revision September 15, 2020; accepted October 11, 2020.

Corresponding Author: Im Young Jung (iyjung@ee.knu.ac.kr)

* School of Electronics and Engineering, Kyungpook National University, Daegu, Korea (striker112@knu.ac.kr, iyjung@ee.knu.ac.kr)

(2) We introduced recent studies to protect privacy, integrity, and robustness and checked the current state of research for secure object detection.
(3) We also summarized the remaining security issues to be addressed. It is important to check the remaining issues because many issues came with the new AI-based approach, such as deep learning.

This paper is organized as follows. The security threats of AI-based approaches to detect objects in images and videos are discussed in Section 2. Section 3 analyzes the recent approaches to defend against security threats. The security issues that need to be addressed are outlined in Section 4, and Section 5 concludes this paper.

2. Security Threats of Object Detection

The following features comprise the deep learning-based approach. First, deep learning relies heavily on data, such as data in training sets. Second, it is not easy to identify how decision-making rules were created in training. Security vulnerabilities to robustness and integrity develop because of these characteristics. The privacy of extracted information from deep learning and the data processed in the deep learning should also be protected. We discuss the security threats of AI-based approaches to detect objects in images and videos in this section. Table 1 shows the possible security attacks for object detection based on deep learning.

Table 1. Security attacks for object detection through deep learning

Attack	Method	Security objective related
Black-box attacks	Wrong identification of objects by transferable adversarial images [3–5]	Robustness
White-box attacks	Training similar adversarial models generated with extracted knowledge about the target model using queries [6–8]	Robustness
Data poisoning attacks	Contamination of the training data [9–12]	Robustness
Adversarial attacks	Wrong identification of objects by misleading classifiers [13–19]	Robustness
Sensitive information leakage	Information leakage during storing and processing [20–23,24]	Privacy
Corruption of important data	Data tampering through manipulation [25–27]	Integrity

It is interesting to note that between different network architectures, deep neural networks (DNNs) can be transferrable. Defined as some hostile examples generated for one model, transferability may be misclassified by another model [3]. The transferable adversarial examples may seriously interfere with deep learning applications. Deep learning neural network architecture is shown in Fig. 1.

Consecutive layers of neurons make up a DNN, as shown in Fig. 1. The layers are connected from weighted vectors θ_F to an output layer. Between DNNs and these new and invisible inputs, the weights know the interaction. Adversaries can alter samples to have them misclassified by a DNN; some of which are illustrated in Fig. 2. The adversarial crafting process is shown in Fig. 3.

To black-box model attacks, transferability [4,5] can be used. The black-box means the learning model, or the data set is unknown to attackers.

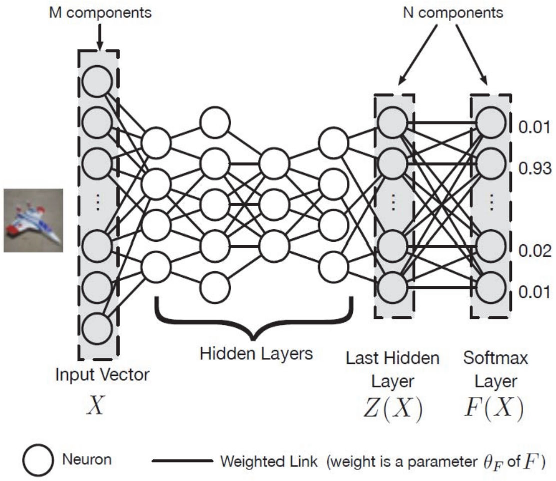


Fig. 1. A deep learning neural network architecture. Adapted from [28].

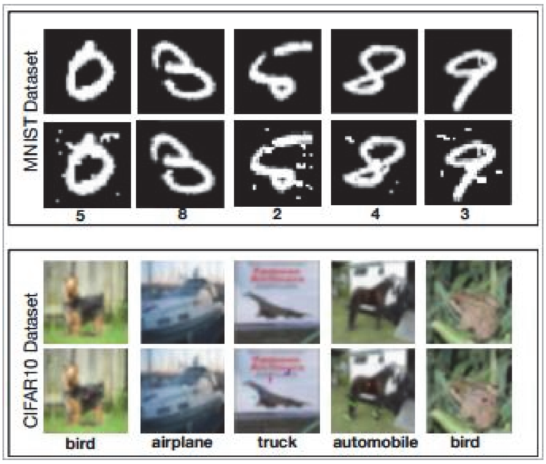


Fig. 2. Pairs of normal and adversarial images. Adapted from [28].

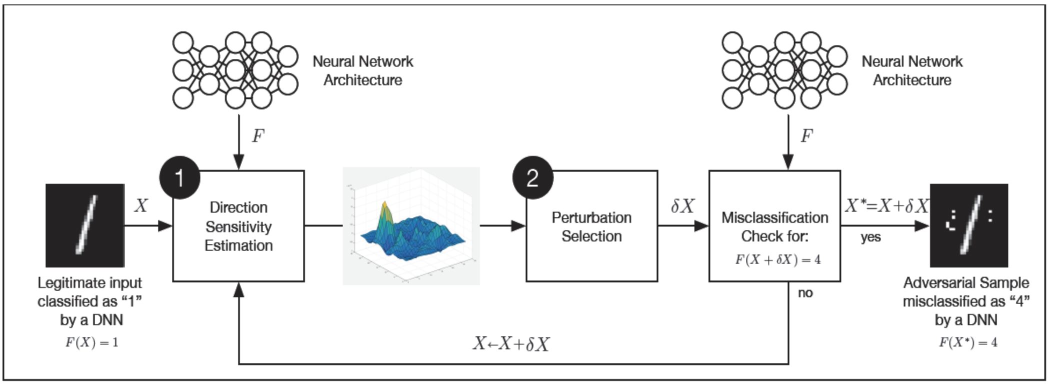

Neural Network Architecture
Neural Network Architecture
F
F
X
1
Direction Sensitivity Estimation
2
Perturbation Selection
δX
Misclassification Check for:
$F(X+\delta X)=4$
$X^{*}=X+\delta X$
yes
no
Legitimate input classified as "1" by a DNN
$F(X)=1$
Adversarial Sample misclassified as "4" by a DNN
$F(X^{*})=4$
$X \leftarrow X+\delta X$

Fig. 3. An adversarial crafting process. The model F can misclassify an image using the input of $X + \delta X$ instead of X. $F(X + \delta X)$ produces 4 (the adversarial target class) instead of 1 (the original class), and lets an adversarial sample X^* be found. Adapted from [28].

The attacker controls the training set and test set without knowing the training process, as shown in the study of Papernot et al. [4,5]. The attacker is unaware of training data, training processes, and even the test label in Liu et al. [3].

To distill the information such as algorithm, training data distribution, hyperparameter of fully trained model architecture from the black-box to generate adversarial examples, white-box attacks perform attacks based on adaptive queries [6–8]. The attacks can cause a crucial result with sufficient information to simulate the neural network.

An attacker leads to an adversarial model contaminating the training data at a data poisoning attack [9–12]. The decision boundary is affected from 1 to 2 with one training sample change without changing the sample's class label when the attacker injects bad data into your model's training pool, as shown in Fig. 4.

The attack model that the attacker contaminates a batch training set is given in Burkard and Lagesse [10], Chen and Zhu [11], and Li et al. [12]. Zhang et al. [9] showed the threats to sequential learners, as in Fig. 5.

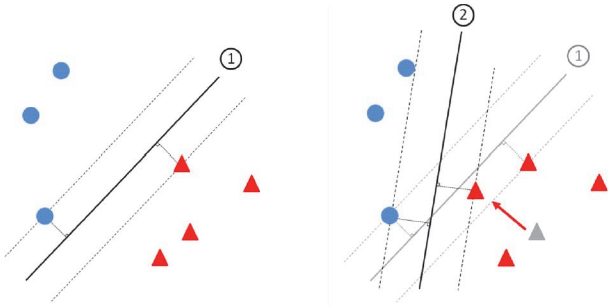


Fig. 4. Linear support vector machines classifier decision boundary. Adapted from [29].

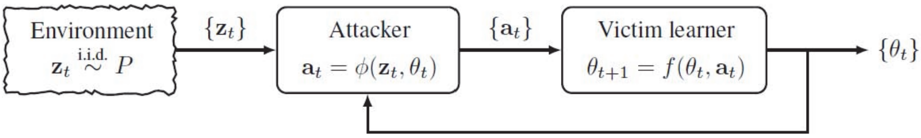

Environment
$z_t \overset{\text{i.i.d.}}{\sim} P$
$\{z_t\}$
Attacker
$a_t = \phi(z_t, \theta_t)$
$\{a_t\}$
Victim learner
$\theta_{t+1} = f(\theta_t, a_t)$
$\{\theta_t\}$

Fig. 5. The attacker injects contaminated samples $\{a_t\}$, online for the training samples $\{z_t\}$, and the learner's model $\{\theta_t\}$. Adapted from [9].

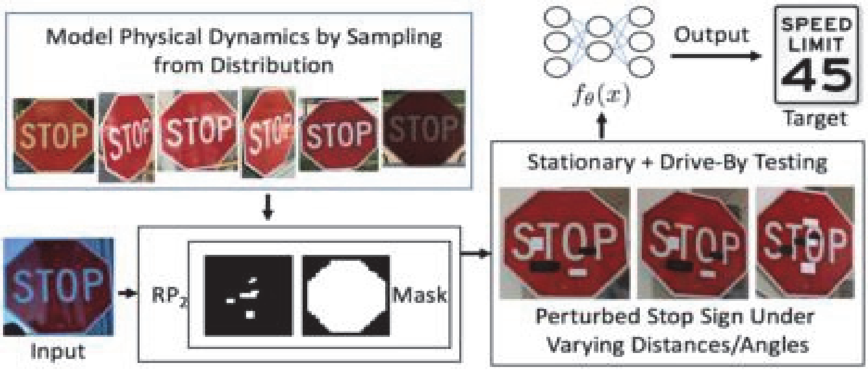

Model Physical Dynamics by Sampling from Distribution
$f_\theta(x)$
Output
SPEED LIMIT 45
Target
Stationary + Drive-By Testing
Input
RP2
Mask
Perturbed Stop Sign Under Varying Distances/Angles

Fig. 6. The stop sign is recognized as the speed sign by deep learning. Adapted from [13].

Adversarial examples that hinder vision-understanding-based applications are shown in Goodfellow et al. [17], Kurakin et al. [15], Carlini and Wagner [18], and Eykholt et al. [13]. For example, how DNNs are affected by even small perturbations added to the input is provided in Eykholt et al. [13]. Using road sign classification, we proposed robust physical perturbations, as shown in Fig. 6.

With high-speed Internet and high-performance processing of big data, personal data are unconsciously leaked. Machine learning caused high-performance data mining and was developed rapidly [23]. Privacy protection becomes important in deep learning processing and data storage.

The digital form of data can easily be manipulated. The problem is that important images such as medical data can be easily manipulated using image processing software [25].

3. Secure Object Detection by Deep Learning

Recent research for secure object detection by deep learning is given in Table 2.

Table 2. Recent research for secure object detection

Security objective	Recent research	Object	Approach
Privacy	Liu et al. [20] Patel et al. [30] Beye et al. [31] Riazi et al. [32]	Patients' medical image	Secret sharing
	Wang et al. [21] Wu [33] Chao et al. [34] Ma et al. [35]	Patients' medical image and speech	Privacy-preserving computation
	Zheng et al. [36]	Patients' medical image	Privacy-preserving computation
	Abadi et al. [22] Noura et al. [37]	Patients' medical image	Differential privacy
	Chu et al. [38]	Moving objects in video	Light-weight object detection in encrypted image
	Sen et al. [39]	Sensitive objects	Training set camouflage
Integrity	Selvaraj and Varatharajan [40] Mousavi et al. [41]	Patients' medical image	Watermarking
Robustness	Goodfellow et al. [17] Papernot et al. [28] Gu and Rigazio [42]	Object in images	DNN resilient to adversarial perturbations
	Metzen et al. [8] Aigrain and Detyniecki [43]	Object in images	Detect adversarial examples
	Tramer et al. [44] Liao et al. [45] Xie et al. [46] Guo et al. [47]	Objects in images	Causing obfuscated gradients adversarial training

To split a secret into several shares given to shareholders, we used secret sharing. The secret is reconstructed if sufficiently many shares are recombined. Based on Shamir's approach [48], previous studies [30,31] were proposed. As a secure machine learning framework, Riazi et al. [32] proposed

Chameleon. Chameleon combines generic secure function evaluation (SFE) protocols with additive secret sharing. A privacy-preserving computation of Faster R-CNN to detect pathologic objects in medical images by adding the secret-sharing approach is proposed in Liu et al. [20], as shown in Fig. 7.

Researchers have focused primarily on data storage privacy to protect medical image privacy. However, recent research has been interested in deep learning processing using an encrypted format. Privacy-preserving computations were proposed using homomorphic encryption (HE) and garbled circuit (GC).

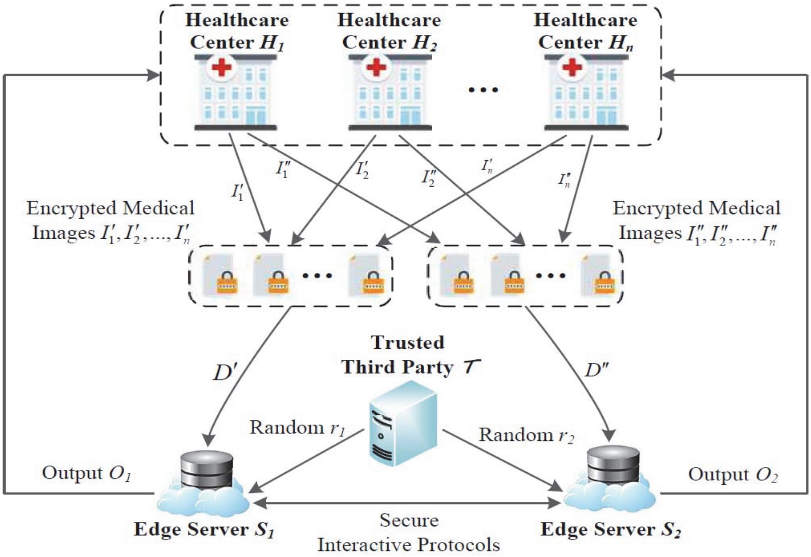


Fig. 7. Privacy-preserving object detection with Fast RNN. Adapted from Liu et al. [20] with the permission of the IEEE.

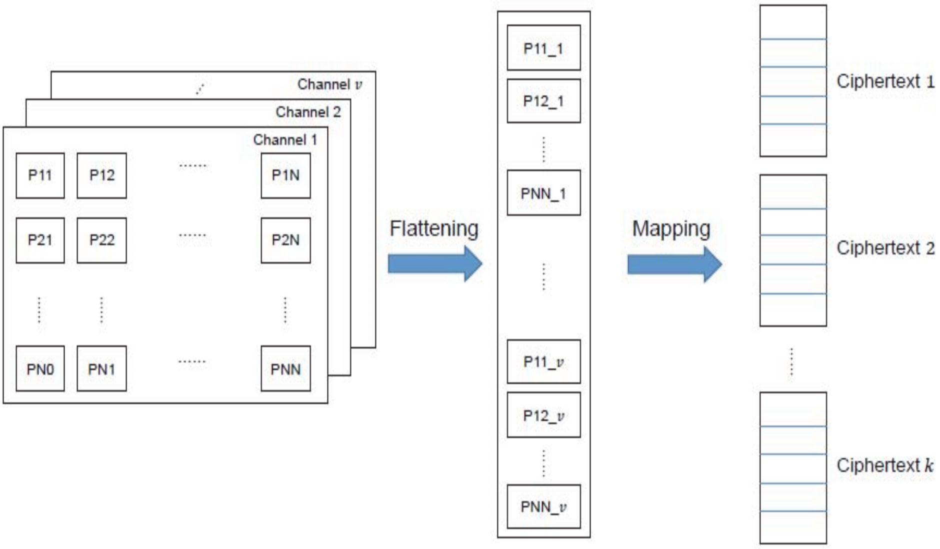


Fig. 8. Mapping the pixels of the input image to the ciphertext slots. Adapted from [34].

HE to hide medical images and to protect their privacy is adopted in Wang et al. [21], Wu [33], Chao et al. [34], and Ma et al. [35]. As shown in Figs. 8 and 9, CaRENets [34] encrypt the input image compactly into ciphertexts, whereas the weight vectors are plaintext. Also, CaRENets homomorphically evaluates the inference phase of MLaaS on encrypted images.

Using GC, we protected the privacy of medical images from the external cloud database [36], as shown in Fig. 10.

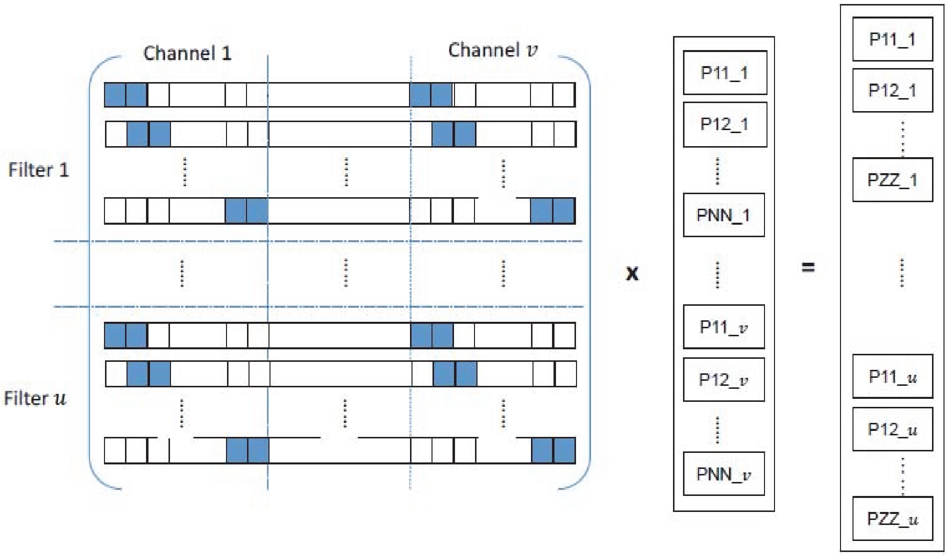


Fig. 9. Vector matrix product in the convolution layer in CaRENets. Adapted from [28].

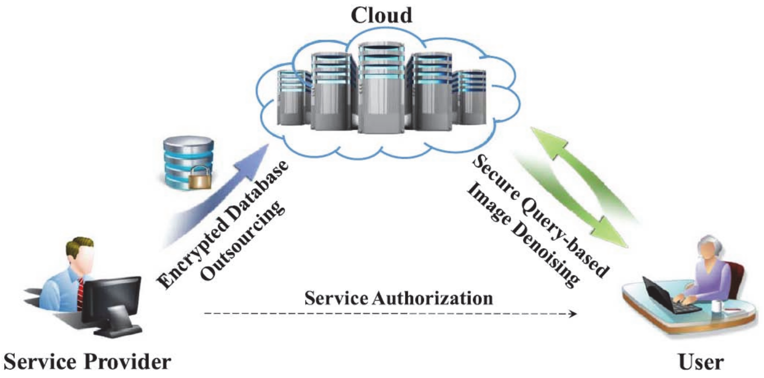

Cloud
Encrypted Database Outsourcing
Secure Query-based Image Denoising
Service Authorization
Service Provider
User

Fig. 10. A service model based on a garbled circuit. Adapted from Zheng et al. [30] with the permission of the IEEE.

Abadi et al. [22] and Noura et al. [37] introduced differential privacy (DP) to preserve privacy in deep learning models. DP describes a promise made by a data holder to a data subject, and the promise is like that of Dwork and Roth [49].

As shown in Fig. 11, Chu et al. [38] detected moving objects in encrypted data by pixel-based treatment. They pursued a light-weight approach without heavy computation.

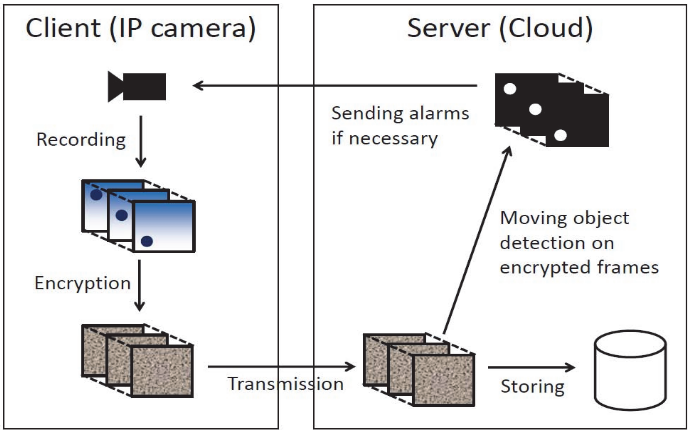
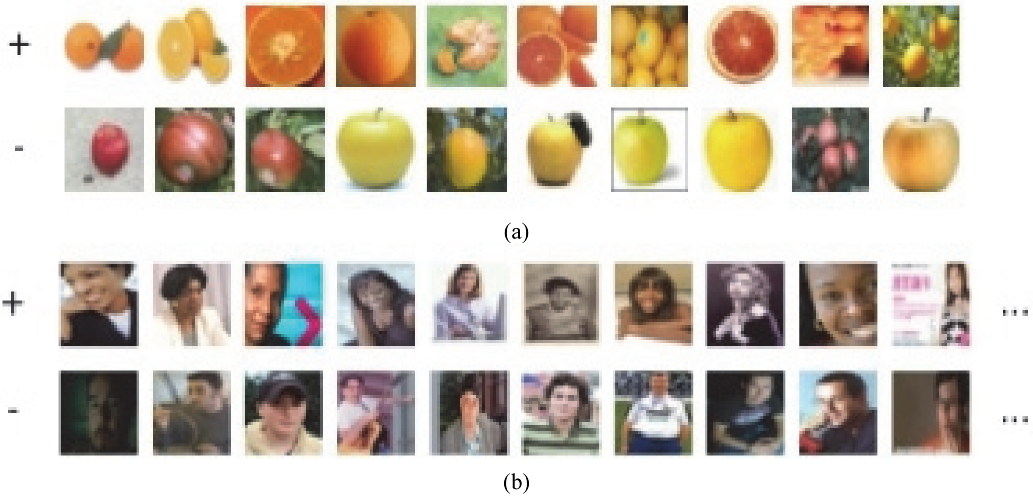

Client (IP camera)
Server (Cloud)
Recording
Encryption
Transmission
Sending alarms if necessary
Moving object detection on encrypted frames
Storing

Fig. 11. Moving object detection in encrypted images. Adapted from Chu et al. [38] with the permission of the Association for Computing Machinery.

By camouflaging the training set, Sen et al. [39] protected sensitive data. A standard logistic regression learner [50] used a camouflaged training set in Fig. 12(a) for the classification of man and woman shown in Fig. 12(b). In Fig. 12(b), the classification showed high accuracy on the images.



(a)

(b)

Fig. 12. Example of training set camouflage: (a) camouflaged training set and (b) secret classification task. Adapted from Sen et al [39] with the permission of Springer Nature.

While analyzing pathological images, digital watermarking is well known to ensure patient data integrity [40]. Generally, as shown in Fig. 13, a secret key is adopted for the watermarking method. That is, the watermark content with secret key protects the authenticity and integrity of the medical image.

To provide authenticity of patient ID, Selvaraj and Varatharajan [40] and Mousavi et al. [41] also used watermarking. A Whirlpool algorithm with a hash function-based watermarking method was proposed by Selvaraj and Varatharajan [40]. Mousavi et al. [41] surveyed watermarking techniques used in medical images.

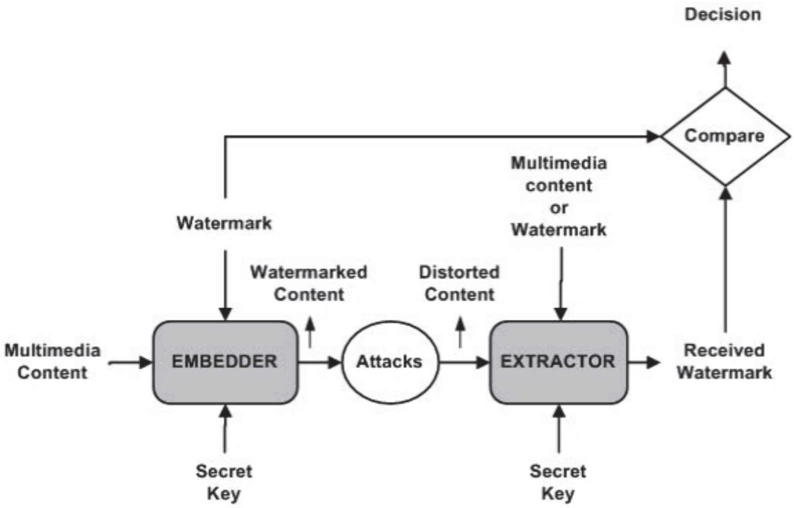


Fig. 13. Typical watermarking system [25].

Goodfellow et al. [17], Papernot et al. [28], and Gu and Rigazio [42] defended DNN against adversarial sampling by defensive distillation. For example, Papernot et al. [28] first train an initial network F on data X with T, as shown in Fig. 14. $F(X)$ contains the knowledge of classification. They then train a distilled network F_d at T on the same X using $F(X)$.

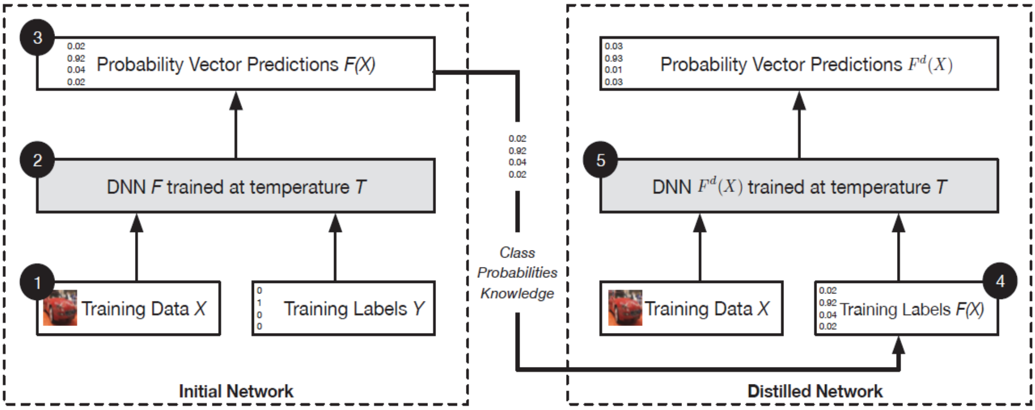


Fig. 14. Defensive distillation to transfer knowledge of probability vectors. Adapted from Papernot et al. [28] with the permission of the IEEE.

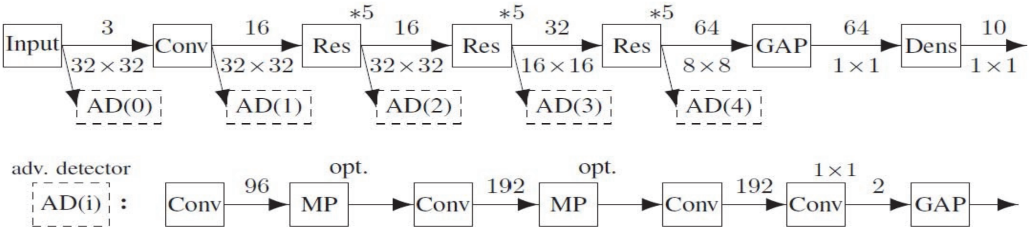


Fig. 15. Adversarial detector. Adapted from [8].

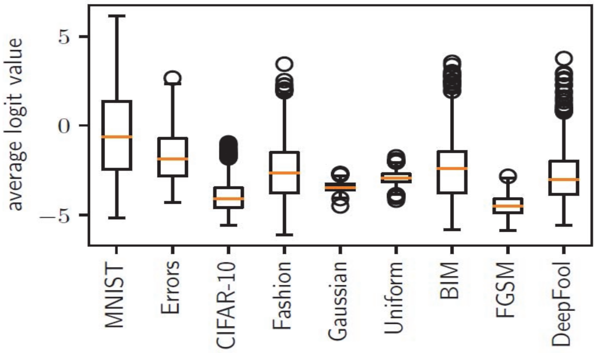


Fig. 16. Distribution of the average logit values. It appears clearly that the average logit values are higher for correctly classified MNIST images compared with misclassifications. Adapted from [43].

Metzen et al. [8] and Aigrain and Detyniecki [43] showed that adversarial perturbations could be detected. In Fig. 15, adversarial detectors were applied. Using the information provided by the logits of an already pre-trained neural network, Aigrain and Detyniecki [43] detected adversarial perturbations by introspection. The characteristic of logit is shown in Fig. 16.

Robust approaches against black-box attacks are proposed by Tramer et al. [44], Liao et al. [45], Xie et al. [46], and Guo et al. [47]. Tramer et al. [44] proposed adversarial ensemble training that augments training data with perturbations propagated from other models. Conversely, a high-level representation guided denoiser was proposed by Liao et al. [45]. The original image is similar to the adversarial image, but the difference is amplified in the high-level representation of a CNN, as shown in Fig. 17. To suppress the effect of adversarial perturbation, they used an image denoiser. Using both the proposed randomization layers and an adversarially trained model, Xie et al. [46] suggested a randomization-based method, as shown in Fig. 18. Guo et al. [47] applied the convolutional network classifier to images after bit depth reduction, JPEG compression, total variance minimization, and image quilting. They showed that the total variance minimization and image quilting were effective defenses in practice.

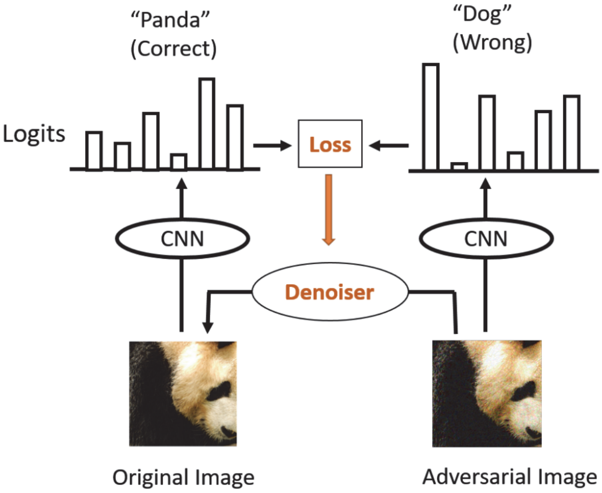


Fig. 17. High-level representation guided denoiser. Adapted from [45].

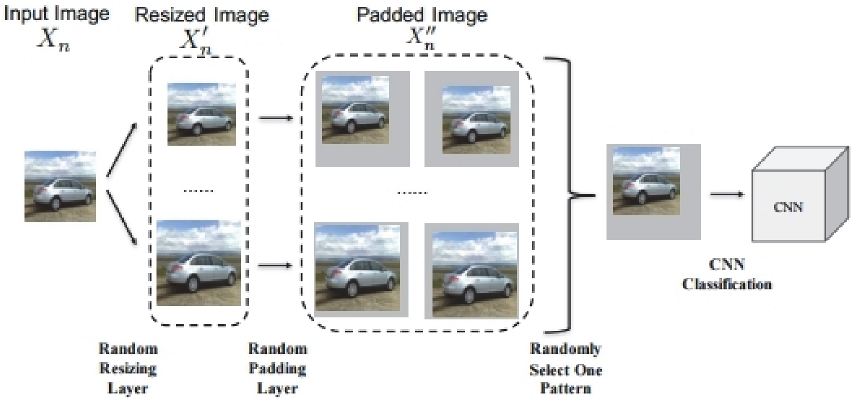


Fig. 18. Randomization-based defense mechanism. Adapted from [46].

4. Additional Issues for Secure Object Detection by Deep Learning

In Table 3, we summarized the additional issues for secure object detection by deep learning.

Table 3. Issues from existing solutions

Security objective	Recent approach	Issues need to be addressed
Privacy	Privacy-preserving computation using homomorphic encryption or garbled circuit Differential privacy Secret sharing Light-weight object detection in encrypted image	Computation-intensive and memory-intensive algorithm Accuracy reduction caused by random perturbations
Integrity	Watermarking	Reduction of the quality of the watermarked images No support for computation of the watermarked images Embedding/extracting watermark
Robustness	Knowledge transfer Detect adversarial example Causing obfuscated gradients	Few practical defenses proposed against adversarial samples

The privacy-preserving approach using HE or GC runs algorithms that require a lot of computation and large memory. Moreover, deep learning adopting HE or GC should take performance into account. The accuracy of DP can be compromised if the random perturbation is used [22]. Secret sharing causes communication overhead. Communication security is also required.

Watermarking can degrade the quality of the processed image, and its basic issue is how to embed/extract the watermark such that the critical value of an image, that is, the diagnostic value of a medical image, is not compromised [25].

It is not easy to check the adversarial attacks at big data processing. Adversarial attacks can also penetrate deep learning processing, hidden in deep learning error. The attacks must also be urgently defended because deep learning applications are expanding. Many studies about adversarial attacks have focused, however, on presenting new types of adversarial attacks against DNNs at the laboratory. Lightweight and practical defenses against adversarial attacks need to be developed.

5. Conclusion

We analyzed in this study recent security studies on object detection in images and videos. Deep learning-based approaches are heavily dependent on training data and learning model. The extracted information from deep learning is also often needed to be protected. Because deep learning is not transparent, its processing is not easy to verify. These characteristics create security vulnerabilities to robustness, privacy, and integrity. Possible security attacks and defenses for privacy, integrity, and robustness are shown in recent research on object detection. However, there are still many issues that need to be addressed for future applications. It is important to check the remaining issues and to try to solve them because an AI-based approach, such as deep learning, is being used actively.

To infringe privacy and integrity in future research, we will focus on adversary attacks on automatic object detection. It is hard to defend against the attacks because adversary attacks utilize the dependency of the deep learning approach on learning data. The attacks are drawing attention as deep learning is applied for big data. We will investigate the methodology for defending against the adversary attacks on automatic object detection through bibliometric analysis.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Korea (No. 2017R1D1A1B03034950).

References

[1] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. Papastathis, and M. Strintzis, “Knowledge-assisted semantic video object detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1210-1224, 2005.

[2] D. Cao, Z. Chen, and L. Gao, “An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks,” *Human-centric Computing and Information Sciences*, vol. 10, article no. 14, 2020.

[3] Y. Liu, X. Chen, C. Liu, and D. Song, “Delving into transferable adversarial examples and blackbox attacks,” in *Proceedings of the 5th International Conference on Learning Representations (ICLR),* Toulon, France, 2017.

[4] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples,” 2016 [Online]. Available: https://arxiv.org/abs/1605.07277.

[5] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi, United Arab Emirates, 2017, pp. 506-519.

[6] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, “Zoo: zeroth order optimization based blackbox attacks to deep neural networks without training substitute models,” in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, Dallas, TX, 2017, pp. 15-26.

[7] W. Brendel, J. Rauber, and M. Bethge, “Decision-based adversarial attacks: reliable attacks against black-box machine learning models,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[8] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, “On detecting adversarial perturbations,” 2017 [Online]. Available: https://arxiv.org/abs/1702.04267.

[9] X. Zhang, X. Zhu, and L. Lessard, “Online data poisoning attacks,” 2019 [Online]. Available: https://arxiv.org/abs/1903.01666.

[10] C. Burkard and B. Lagesse, “Analysis of causative attacks against SVMs learning from data streams,” in *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics*, Scottsdale, AZ, 2017, pp. 31-36.

[11] Y. Chen and X. Zhu, “Optimal adversarial attack on autoregressive models by manipulating the environment,” 2019 [Online]. Available: https://arxiv.org/abs/1902.00202.

[12] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, “Data poisoning attacks on factorization-based collaborative filtering,” *Advances in Neural Information Processing*, vol. 29, pp. 1885-1893, 2016.

[13] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning models,” 2018 [Online]. Available: https://arxiv.org/abs/1707.08945.

[14] Y. Dong, T. Pang, H. Su, and J. Zhu, “Evading defenses to transferable adversarial examples by translation-invariant attacks,” in *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition (CVPR*), Long Beach, CA, 2019, pp. 4312-4321.

[15] A. Kurakin, I. Goodfellow, S. Bengio, “Adversarial examples in the physical world,” in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[16] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 284-293.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[18] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proceedings of 2017 IEEE Symposium on Security and Privacy*, San Jose, CA, 2017, pp. 39-57.

[19] M. J. J. Ghrabat, G. Ma, I. Y. Maolood, S. S. Alresheedi, and Z. A. Abduljabbar, “An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier,” *Human-centric Computing and Information Sciences*, vol. 9, article no. 31, 2019.

[20] Y. Liu, Z. Ma, X. Liu, S. Ma, and K. Ren, “Privacy-preserving object detection for medical images with faster R-CNN,” *IEEE Transactions on Information Forensics and Security, 2019. https://doi.org/10.1109/TIFS.2019.2946476*

[21] L. Wang, L. Li, J. Li, J. Li, B. B. Gupta, and L. Xia, “Compressive sensing of medical images with confidentially homomorphic aggregations,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1402-1409, 2019.

[22] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, 2016, pp. 308-318.

[23] C. Yin, B. Zhou, Z. Yin, and J. Wang, “Local privacy protection classification based on human-centric computing,” *Human-centric Computing and Information Sciences*, vol. 9, article no. 33, 2019.

[24] Q. Jia, L. Guo, Z. Jin, and Y. Fang, “Preserving model privacy for machine learning in distributed systems,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 8, pp. 1808-1822, 2018.

[25] N. A. Memon, “Watermarking of medical images for content authentication and copyright protection,” Ph.D. dissertation, GIK Institute of Engineering Sciences and Technology, Topi, Pakistan, 2010.

[26] C. Wang, H. Zhang, X. Zhou, “LBP and DWT based fragile watermarking for image authentication,” *Journal of Information Processing Systems*, vol. 14, no. 3, pp. 666-679, 2018.

[27] C. Wang, H. Zhang, and X. Zhou, “Review on self-embedding fragile watermarking for image authentication and self-recovery,” *Journal of Information Processing Systems*, vol. 14, no. 2, pp. 510-522, 2018.

[28] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in *Proceedings of the 2016 IEEE Symposium on Security & Privacy*, San Jose, CA, 2016, pp. 582-597.

[29] D. J. Miller, Z. Xiang, and G. Kesidis, “Adversarial learning in statistical classification: a comprehensive review of defenses against attacks,” 2019 [Online]. Available: https://arxiv.org/abs/1904.06292.

[30] S. Patel, S. Garasia, and D. Jinwala, “An efficient approach for privacy preserving distributed k-means clustering based on Shamir’s secret sharing scheme,” in *Trust Management VI*. Heidelberg, Germany: Springer, 2012, pp. 129-141.

[31] M. Beye, Z. Erkin, and R. L. Lagendijk, “Efficient privacy preserving k-means clustering in a three-party setting,” in *Proceedings of IEEE International Workshop on Information Forensics and Security*, Iguacu Falls, Argentina, 2011, pp. 1-6.

[32] M. S. Riazi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar, “Chameleon: a hybrid secure computation framework for machine learning applications,” in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS)*, Incheon, Republic of Korea, 2018, PP. 707-721.

[33] X. Wu, “An algorithm for reversible information hiding of encrypted medical images in homomorphic encrypted domain,” *Discrete & Continuous Dynamical Systems-S*, vol. 12, no. 4-5, pp. 1441-1455, 2019.

[34] J. Chao, A. A. Badawi, B. Unnikrishnan, J. Lin, C. F. Mun, J. M. Brown, et al., “CaRENets: compact and resource-efficient CNN for homomorphic inference on encrypted medical images,” 2019 [Online]. Available: https://arxiv.org/abs/1901.10074.

[35] Z. Ma, Y. Liu, X. Liu, J. Ma, and F. Li, “Privacy-preserving outsourced speech recognition for smart IoT devices,” *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8406-8420, 2019.

[36] Y. Zheng, H. Cui, C. Wang, and J. Zhou, “Privacy-preserving image denoising from external cloud databases,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1285-1298, 2017.

[37] M. Noura, H. Noura, A. Chehab, M. M. Mansour, L. Sleem, and R. Couturier, “A dynamic approach for a lightweight and secure cipher for medical images,” *Multimedia Tools and Applications*, vol. 77, no. 23, pp. 31397-31426, 2018.

[38] K. Y. Chu, Y. H. Kuo, and W. H. Hsu, “Real-time privacy-preserving moving object detection in the cloud,” in *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain, 2013, pp. 597-600.

[39] A. Sen, S. Alfeld, X. Zhang, A. Vartanian, Y. Ma, and X. Zhu, “Training set camouflage,” in *Decision and Game Theory for Security*. Cham, Switzerland: Springer, 2018, pp. 59-79.

[40] P. Selvaraj and R. Varatharajan, “Whirlpool algorithm with hash function based watermarking algorithm for the secured transmission of digital medical images,” *Mobile Networks and Applications*, 2018. https://doi.org/10.1007/s11036-018-1057-4

[41] S. M. Mousavi, A. Naghsh, and S. Abu-Bakar, “Watermarking techniques used in medical images: a survey,” *Journal of Digital Imaging*, vol. 27, no. 6, pp. 714-729, 2014.

[42] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.

[43] J. Aigrain and M. Detyniecki, “Detecting adversarial examples and other misclassifications in neural networks by introspection,” 2019 [Online]. https://arxiv.org/abs/1905.09186.

[44] F. Tramer, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[45] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, “Defense against adversarial attacks using high-level representation guided denoiser,” 2018 [Online]. Available: https://arxiv.org/abs/1712.02976.

[46] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, “Mitigating adversarial effects through randomization,” 2018 [Online]. Available: https://arxiv.org/abs/1711.01991.

[47] C. Guo, M. Rana, M. Cisse, and L. V. D. Maaten, “Countering adversarial images using input transformations,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[48] A. Shamir, “How to share a secret,” *Communications of the ACM*, vol. 22, no. 11, pp. 612-613, 1979.

[49] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Theoretical Computer Science*, vol. 9, nos. 3-4, pp. 211-407, 2014.

[50] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons, 2013.

Keonhyeong Kim **https://orcid.org/0000-0003-1381-8584**

He received a B.S. degree in the School of Electronics Engineering from Kyungpook National University in 2019. Since March 2019, he is with the School of Electronics Engineering from Kyungpook National University as an M.S. candidate. His current research interests include Security in IoT and Connected Vehicle.

Im Young Jung **https://orcid.org/0000-0002-9713-1757**

She received her first B.S. degree in chemistry from Pohang University of Science and Technology in 1993 and her second B.S. degree in computer science from Seoul National University in 1999. She received her M.S. and Ph.D. degrees in computer science and engineering from Seoul National University in 2001 and 2010, respectively. Now, she is an associate faculty at the School of Electronics Engineering, Kyungpook National University in South Korea. Her current research interests include data security and system security in distributed computing, IoT, and connected vehicle.