EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Artificial Intelligence in Intellectual Property Protection: Application of Deep Learning Model

Parthasarathi Pattnayak[1]*, Tulip Das[2], Arpeeta Mohanty[3] and Sanghamitra Patnaik[4]

[1,3]School of Computer Application, KIIT Deemed to be University, Bhubaneswar, Odisha, India
[2]Faculty of Emerging Technologies, SRI SRI University, Cuttack, Odisha, India
[4]School of Law, KIIT Deemed to be University, Bhubaneswar, Odisha, India

Abstract

To create and train a deep learning model costs a lot in comparison to ascertain a trained model. So, a trained model is considered as the intellectual property (IP) of the person who creates such model. However, there is every chance of illegal copying, redistributing and abusing of any of these high-performance models by the malicious users. To protect against such menaces, a few numbers of deep neural networks (DNN) IP security techniques have been developed recently. The present study aims at examining the existing DNN IP security activities. In the first instance, there is a proposal of taxonomy in favor of DNN IP protection techniques from the perspective of six aspects such as scenario, method, size, category, function, and target models. Afterwards, this paper focuses on the challenges faced by these methods and their capability of resisting the malicious attacks at different levels by providing proactive protection. An analysis is also made regarding the potential threats to DNN IP security techniques from various perspectives like modification of models, evasion and active attacks.
Apart from that this paper look into the methodical assessment. The study explores the future research possibilities on DNN IP security by considering different challenges it would confront in the process of its operations.
Result Statement: A high-performance deep neural Networks (DNN) model is costlier than the trained DNN model. It is considered as an intellectual property (IP) of the person who is responsible for creating DNN model. The infringement of the Intellectual Property of DNN model is a grave concern in recent years. This article summarizes current DNN IP security works by focusing on the limitations/ challenges they confront. It also considers the model in question's capacity for protection and resistance against various stages of attacks.

Keywords: Intellectual Property (IP) Protection, Deep Neural Network (DNN), Taxonomy, Machine Learning

Received on 15 December 2023, accepted on 05 March 2024, published on 12 March 2024

doi: 10.4108/eetiot.5388

1. Introduction

The Deep Neural Networks (DNNs) model is emerging as a popular business model in recent years as it refers to MLaaS, or machine learning as a service. DEEP LEARNING (DL) techniques are often offered by large companies as it involves high costs in acquiring huge training data. It also requires expensive hardware resources. The techniques are used in classifying images, detecting objects, recognizing voice, natural language and driver less cars. Because of its high business value, the deep learning models are considered as an intellectual Property (IP) of the creators of this model and it requires to be protected from all kinds of malicious attacks or piracy.

The following are the contents of this paper:

It is suggested to classify the DNN IP defence techniques. Now, we provide a taxonomy of DNN IP protection strategies based on the six characteristics of framework, mechanism, capacity, type, function, and target models for the first time. A taxonomy like this can make it easier to

*Corresponding author. Email: parthakiit19@gmail.com analyse, compare, and create new methodologies based on existing ones.

The methods of DNN IP protection are described with systematic evaluation recommendations. The majority of assessments in the DNN IP protection efforts currently in use solely pay attention to the DNN watermark's functional metrics. We suggest basing an assessment strategy for DNN IP protection measures on the which include the factors: A systematic evaluation technique, fundamental functional metrics, and attack-driven metrics are included. Additionally, the DNN IP protection mechanisms are assessed when various types of attack are used by the attackers. Challenges and upcoming projects: We talk about the difficulties modern DNN IP protection techniques face and share our predictions for next projects. The 13th five-year plan on nationwide advancement in science and technology, published by the State Council, officially included the growth in artificial intelligence in August 2016. The National People's Congress' government work report for 2017 was the first to include China's strategy for the field's development. In the 2018 government work report, the term "artificial intelligence" had again used. The conference on "technology with dispute: the legal issues and judicial solutions of artificial intelligence" took place in Beijing and was sponsored in part from China Intellectual Property Magazine and the Internet Court. This seminar's subject is the legal challenge to China's current copyright laws with artificial intelligence products. China's AI industry currently consists primarily of construction projects related to smart education, advanced robots, intelligent finance, intelligent security, intelligent driving, and other computer industry applications. The proportion of intelligent robots in particular is rising year by year, as is the number of specific industrial scale.

DNN IP is emerging as a potential field of research which is still at its nascent stage to deal with various security threats like illegal copying, redistribution and abuses. The article surveys all challenges and existing situations along with evaluation and suggestions.

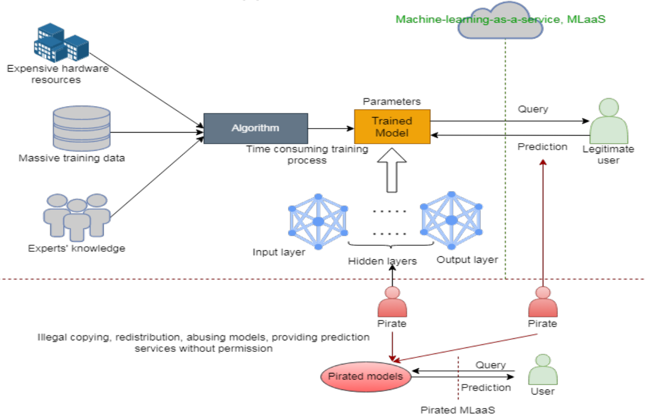

Figure 1. Overview of piracy on deep learning models

2. Evaluation of DNN IP Protection Methods

The copyright of multimedia information has been extensively protected in the field of multimedia via the use of digital watermarking techniques. Deep learning-related IP protection remains in its infancy, nevertheless. Instead of directly applying existing watermarking digital techniques to the DNN model, as is the case when embedding digital watermarks into multimedia content, new approaches must be developed [8]. To extract the watermark, the current computerized watermarking algorithms need uninterrupted approach to multimedia material. The DNN model, in contrast to the multimedia data, has a complex structure and several parameters. The only tools that allow watermark retrieval and ownership verification are typically the DNN models' APIs [8]. As a result, DNN cannot use the current digital watermarking methods used in the entertainment industry.

The following factors make it extremely difficult to design a workable DNN IP protection technique [9], [17], [23]:

It is necessary to have a public watermarking mechanism that can be used to repeatedly and reliably prove that a model is owned via an API.

The model's performance shouldn't suffer as a result of the watermark embedding.

To prevent misleading naive users of pirating the models, the integrated watermark needs to be allowed provide a high detection rate and a low false alarm rate.

Users can adjust the model's parameters by tweaking or pruning it.

The watermark needs to be resistant for the possible attacks from harmful users.

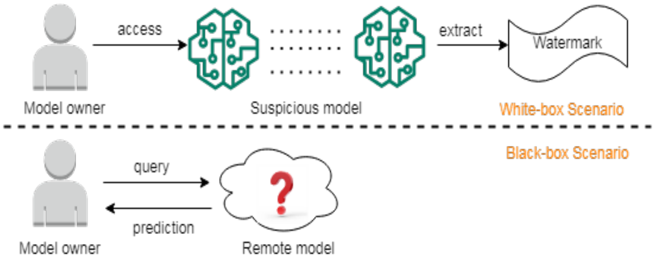

Figure 2. White-box and black-box structures

2.1. Structure of White-Box

The first technique of DNN copyright protection is suggested by [3], [4]. In order to embed the watermark in the intermediate layer's weight, they train the model with additional regularization loss. During the verification phase, watermark can be taken out a marking layer's weight. With the aim of being robust and not impacting accuracy, the RIGA watermarking with competitive training strategy for white-box scenarios [24]. Fig.2 gives a summary about black-box and white-box scenarios. They create a framework like adversary generative networks

(GAN), in which process the model's training and watermarking. These techniques are effective in the case of the white-box situation, when the model's internal parameters are made available to the general public. In reality, the pirate frequently uses the stolen DNN simulation as a web service that solely returns predictions and certainty with a remote API. The situation of a black box, or internal data of the suspect model, is inaccessible to the verifier.

2.2. Black-Box Scenarios

A watermark is incorporated into the model as part of the working mechanism, and the watermark data can only be recovered through the global model through interacting with it via a remote API [6].

A universal watermarking technique (called DeepSigns) is suggested by [5], that can be used in black-box and white-box settings. DeepSigns incorporates an N-bit string (the owner's signature/watermark) into each layer's activation set's probability density function. To remotely check the DNNs' copyright, an appropriate set of input can activate the embedded watermark. The majority of the DNN IP protection research now being done [10]– [13], [17]– [19], focuses on the black-box scenarios.

2.3. Mechanism for Parameter-Based Watermarking

The weights vary substantially in parameters-based watermarking techniques. Analyzing the shift in weights can reveal the presence of the watermark. To make sure that the changes made by the embedded watermark is minimal, the influence of the watermark can be assessed by altering the settings during training.

Researchers also suggest using encryption-based methods to safeguard the model. The training images are encrypted by [2] using block-wise pixel shuffling and a key. These previously processed, encrypted images are used to train the model. As such, people can only input the encrypted image to get the standard model performance. Fully homomorphic encryption is used by Gomez et al. [4],[5] to secure the IP, input data, and neural network inference. The output image can also contain a watermark in addition to one that is embedded in the model. According to a method given out by Wu et al. [20], the images produced by a watermarked DNN will likewise include a watermark. So, host DNN can complete the task and the output watermarked images, researchers instruct the host DNN and the watermark extraction network jointly and employ a combined loss function. The method can be used to ascertain the model's copyright as well as whether an image was produced by that particular model.

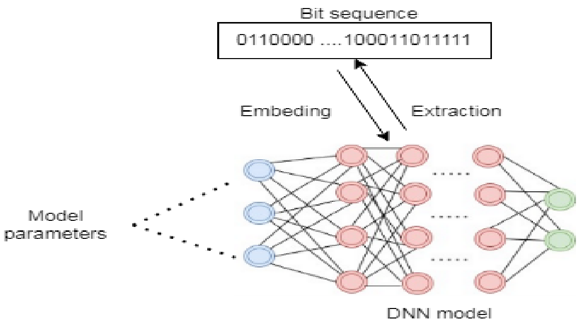

Figure 3. DNN watermarking using parameters

2.4. Mechanism for Backdoor-Based Method

The backdoor attack [5,6] is a technique for attacking deep learning models in which assailant trains the model to output a particular label in response to a particular input. The watermarking approach is implemented by [10] using overparameterization on the neural network and a backway for the watermark key. With the exception of true label and the initially anticipated label, the correspondent key label is chosen at random from all classes. By evaluating the precision of a watermark trigger set against a threshold, the watermark is found. Additionally, they create a publicly accessible protocol using a commitment scheme. The trigger design used by the backdoor-based watermark can be generated and optimized using an evolutionary algorithm [11], which will lower the false alert rate. Three different watermark key generation techniques are suggested by Zhang et al. [8] employing random photos, training images with additional content placed on them, and irrelevant images from another dataset. The performance of the model will be impacted by incorrectly labeling the key sample in the backdoor-based strategy, which may skew the decision boundary [18], suggest a black-box watermarking technique that fixes this issue by giving the key sample a new label. As a consequence, the model will pick up the characteristics of the important sample without changing the original model's decision boundary. In order to label the backdoor samples, Zhang et al. [19] suggest using an automated method based on chaos. IP owners can trace their control of models after piracy by adding watermarks to the models. Model extraction attacks, however, have the potential to steal the function underlying the model [6], [7]. In this scenario, the adversary will train a different model using the predictions supplied from the model's API. The replacement pattern has been educated by the pirate (not the IP owner), rendering the current watermarking techniques [3],[5] useless for model piracy using model extraction attacks [6].

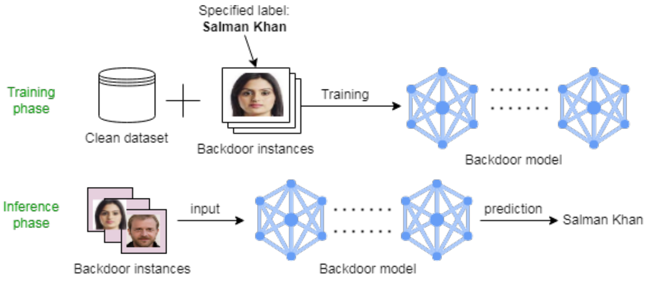

Figure 4. DNN watermarking technique based on backdoors

2.5. Mechanism for Fingerprint-Based Method

According to several studies [13], [14], and [17], the model's "fingerprint" can be extracted and used to protect intellectual property. Adversarial examples [8], [9] are suggested by Merrer et al. [13] as the watermark key set for a watermarking algorithm. In Fig.5, a summary of the fingerprint-based DNN watermarking technique is shown. The technique modifies the model's decision boundary just enough to enable a particular series of queries to validate the watermark data. To achieve this, they include perturbations to produce adversarial examples that are very near the model's border. A watermark key image is used to query the model during the watermark detection phase. Responses from the marked model and the remote model to these hostile inputs are contrasted. To ascertain when the questionable model is a pirated model, one looks at the ability to transfer of the adversarial mark. By examining if the model's retaliation to hostile example is consistent, the approach determines whether it is a stolen model. Even if two models may react to adversarial situations similarly, these aren't the same model.

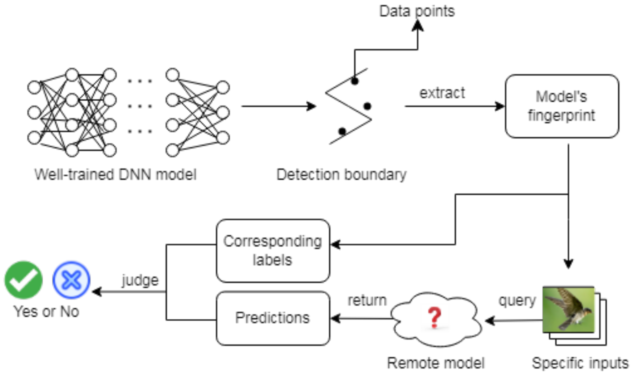

Figure 5. DNN watermarking technique based on fingerprints

2.6 Capacity

The majority of the blackbox watermarking techniques used today [10], [11], [13], and [18] merely validate the existence of the watermarks. They typically produce an array of watermark key combination and subtly alter the target model's decision boundaries. Fig. 8 presents a summary of the zero-bit with multibit DNN IP protection techniques. A multibit watermarking framework called BlackMarks is proposed by Chen et al. [17] for the black-box scenarios. They show that it is possible to carry out multibit string verification using the model's prediction as opposed to only confirming a single bit of data. The design of a scheme creates a set of key image and label pairings in accordance with owner's watermark signature. Owner's signature will be deciphered from the matching prediction once the remote model is requested with the watermark key image during the watermark extraction process.

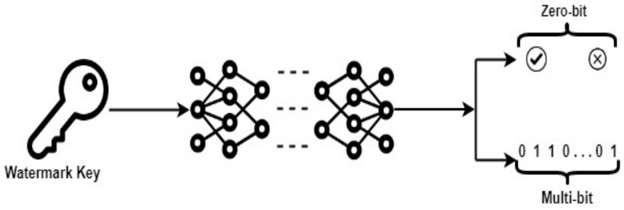

Figure 6. Multi bit and zero-bit methods have capacity

2.7. Type

Recently, a few active consent control methods [3], [8]-[14], [16] have been put forth for DNN protection of IP. Dynamic authorization control's summary is shown in Fig.8. All of the preceding watermarked tasks [3], [8]-[13] are passive confirmation techniques, i.e., the intellectual property of the model is quietly verified following the piracy occurs. The outline of copyright verification is shown in Figure 7.

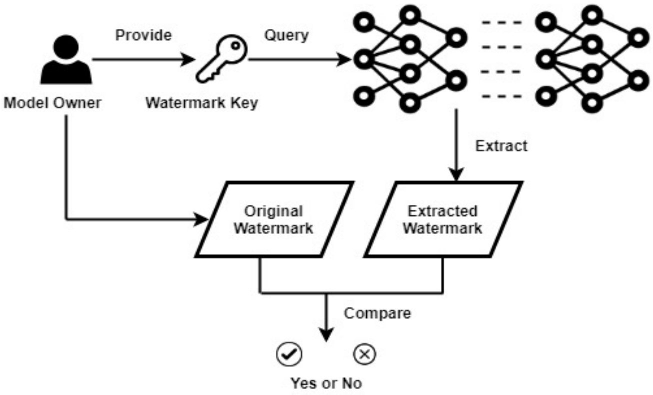

Figure 7. Copyright verification's description

The structure allows DNN to continue operating in the authorized access mode; but it is inoperable for unauthorized or unlawful use. To offer authorized input, an encoding module built on existing adversarial examples is created. Authorized users can pre-process their input using the conversion element to get high performance forecasts, however unauthorized users who don't have the conversion module would get low performance. This approach cannot distinguish between different authorized users because it does not account for the user's identity management. Ambiguity attack offers a severe danger to the current DNN watermarking techniques [4]. They suggested incorporating special passport layers into DNN as a fix, that can disable neural network's functionality for unauthorized use or keep it operational in validated situations. The DNN model won't work well unless a genuine passport is presented, preventing its unauthorized use. However, this plan adds passport layers following each convolutional layer, that will result in significant overhead. In addition, the technique is susceptible to tampering and reverse-engineering attacks. A serial number-based DNN IP protection solution based on knowledge distillation is proposed [8] and is implemented. The instructor model is initially taught before being condensed into a number of client (student) models. Each customer model is given an identification code, and only when the right serial quantity is entered can the customer model be used regularly. The model number is used as a watermark to confirm copyright. The aforementioned permission control mechanisms are susceptible to assaults by unethical clients, such as collusion attacks, because they do not take into consideration how people manage their identities. As a result, they are unable to differentiate between different authorized users. Furthermore, these approaches fall short of meeting the criteria of commercial DNN IP protection applications because they lack a copyright management function.

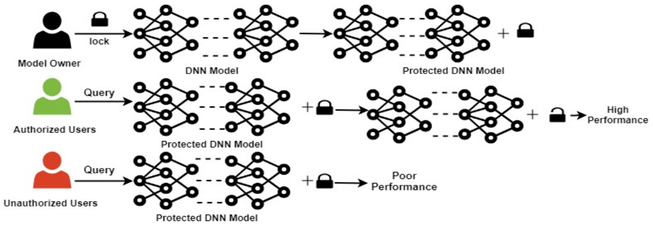

Figure 8. Description of copyright verification.

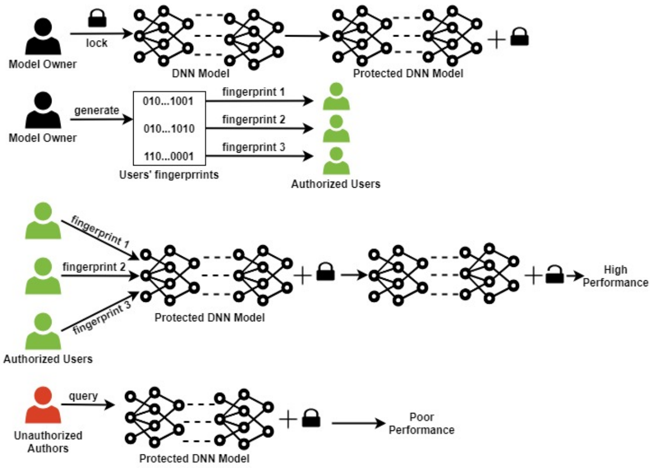

Figure 9. An overview of managing copyrights, including managing user identities and active authorisation control

2.8. Target Models

Utilizing spatial invisible watermarking schemes, a hidden watermark is specifically embedded in a black-box scenario. The overparameterization using models is used to insert watermarks to image processing tasks [21]. Additionally, they create a support module that visually displays the watermark data for verification suggest a watermarking strategy for more complicated image processing models [20], whereas the majority of publications [3], [5,] [13], [20], [21] concentrate on safeguarding the intellectual property of classification models. The training procedure along with the training data must be under control for the aforementioned watermarking techniques. A watermarking strategy for integrated learning scenarios is suggested [22]. To integrate the backdoor/watermark, the model must be retrained each time a model is merged with a global model. They also provide a technique for creating watermark patterns that means the images are produced using a pattern that is both random and class-specific.

2.9. Function

The majority of watermarking techniques [3], [5], [8], and [10] aim to verify trademark using steganographic images. To identify individuals' fingerprints and implement active access control for DNN [4] employ adversarial samples with particular confidence as users' fingerprints. Fig.9 is an overview of managing copyrights, including managing user identities and active authorisation control.

3. Taxonomy

There have been several DNN intellectual property works proposed so far (e.g., [3], [5], [8]- [14]). There are still no formal taxonomies, though. Using the next six attributes, as illustrated in Fig. 2, we offer taxonomy of DNN IP security techniques in this work. The proposed taxonomy fully brings out the multidimensional characteristics of an DNN intellectual property work from various angles, such as mechanism used—passive verification, applicable scenarios and active protection that may offer the capacity. As a result, it may simply explain a DNN copyright-safe technique, assist in systematically comparing them, and serve as a resource for upcoming works.

Scenario

Scenario depicts the knowledge scene where the model is suspicious and the copyright verifier is present. The hidden parameters of the suspect model that must be confirmed are made available to the public in a white-box scenario. The black-box scenario, which is how the DNN model is typically implemented in practice, only provides speculative and assured through an API.

Mechanism

DNN IP protection methods' implementation mechanisms can be categorized into three groups: parameter-based (incorporating a watermark into the model's weights or parameters); backdoor-based (using the DNN backdoor as the model's watermark); and fingerprint-based (using the model's data distribution to specific inputs, such as adversarial examples, as its fingerprint).

Capacity

Capacity is the maximum amount of data that the watermarking technique can embed. A zero-bit scheme (sometimes known as a one-bit scheme) is a verification method that just looks for the watermark. A technique is referred to as a multibit scheme if it can do multibit string verification rather than only confirming one bit with data [17].

Type

Passive verification is the process of checking the model's copyright after it has been violated. Active authorization control is the term used to describe a strategy that actively controls the model using authorization control to avoid piracy.

Function

The three categories that aims for DNN IP protection works are as follows. a) Copyright verification, which entails employing reliable watermarks to confirm the model's ownership. The majority of the time, this happens. b) Property management, which entails implementing authorization control and managing user identities. c) Authenticity verification, which entails using brittle/reversible watermarks to confirm the model's integrity.

Target Models

Models that are specifically chosen by protection of copyright techniques, such as classification, identification, image processing, and others, are referred to as "target models." The majority of DNN IP security techniques [3, 5, 10, 13, 17, and 19] focus on classification jobs. Few recent works [20, 21] focus on image processing tasks, and [22] additionally few protections of IP works [20, 21] that focus on Federated Learning (distributed scenarios). IP security techniques are required for various jobs depending on the different application circumstances.

4. DNN IP Security Works Attacks

The anti-attack capability of the DNN IP security mechanisms against various tiers of attackers is thoroughly covered in this article. We categorize various assaults on DNN IP security techniques into the following categories:

Category 1. Model changes: When DNN model is pirated, the pirate frequently alters or compresses it before deploying and using that create an MLaaS model. to offer services. Thus, the majority of the DNN IP protection works currently in existence [8], [10], and have assessed the durability against model modifications. The following model changes are mentioned.

(i) **Fine-tuning model [3]:** As the parameters containing a watermark will be altered during the fine-tuning process, the incorporated watermark must be resistant to it.

(ii) **Pruning parameters or models:** Model pruning is a typical DNN deployment technique, especially for embedded systems. While competitors can make use of pruning to remove watermarks, for example by sparsifying the weights within the watermarked DNN, genuine users could utilize factors pruning to minimize the storage and computational overhead of DNN. Consequently, a good watermarking method ought to withstand parameter changes brought on by parameter pruning.

(iii) **Model enlargement:** Model compression is crucial during deployment of DNN to computer chips or mobile devices since it can significantly reduce memory needs and computational overhead. Lagging compression will cause model parameters to change, thus it's important to investigate how it will affect the likelihood that watermarks will be detected.

(iv) **Retraining of models:** Retraining the model with fresh examples is a straightforward way to get rid of watermarks; this may get rid of the watermark entirely or at least lessen its impact.

Category 2. Attacks by evasion and attacks by removal: The majority of publications [3], [8], [10], [14] solely assess the watermark's robustness against unintended model modifications; they do not take into account the watermark's security when an attacker launches attacks. DNN watermarks are actually subject to a multitude of assaults. Evasion attacks [26] along with removal attacks [24] are typical passive attacks. A few studies [23–25], have recently been published that investigate DNN watermarks' safety for escape attacks and removal assaults.

1) Attacks aimed at removing the watermark: The attackers make this attempt.

2) Interfering: The attacker is aware that the model contains a watermark. He tries to alter the model in order to erase IP owner's signature.

(v) **Attacks using reverse engineering [24]:** If pirates can get their hands on the initial training dataset, they can immediately reverse engineer the secret parameters.

In DNN, removal attacks against backdoor-based watermarking techniques are studied [25]. Three attack strategies—the black-box, white-box, and property inference attacks—are deployed in turn. They suggest a technique to determine whether a watermark is present in the model, showing that backdoor-based watermarking strategies are insufficiently safe to keep the watermark secret. A DNN laundering approach is suggested [26] to eliminate backdoor-based watermarks. The method entails the following steps: retraining, watermark recovery, and watermarked neurons reset. Unknown watermark eradication attack [26]. They apply a preprocessing operation, which transforms and perturbs the input, invalidating the watermark trigger. They then employ unlabelled data for fine-tuning to enhance the model's performance.

Category 3. Current Attacks: Only taking into account the aforementioned robustness of the model changes and security to evasion/removal attacks, in our opinion, is insufficient. Additionally, it should assess how well the model defends against active and powerful attacks in the manner described below.

(i) **Watermark detection:** A few recent works [25], [24], and [26] have been put out with the goal of detecting DNN watermarks (which include backdoors) so as to withstand additional attacks.
(ii) **Overwriting watermarks in [5], [9], and [17]:** If a hacker is aware of the model's watermark embedding technique but is unaware of the owner's personal watermark data, he could seek to replace the existing watermark by inserting a new one in the deep learning model.
(iii) **Fingerprint/watermark collusion attack [9]:** A collusion attack might be used by Used by an assortment of people with the identical host DNN and various fingerprints to create a useful model that would make it impossible for the copyright owner to confirm ownership.
(iv) **Attack on ambiguity:** Ambiguity attack uses a fake extra watermark on the DNN model to cast doubt on the ownership verification. For instance, the adversary's goal for of DNN authorization control is to trick the DNN to believe that an authorized user has approved the given input. According to studies, unless an irreversible watermarking approach is used, a robust watermark does not always prove ownership within the context of conventional digital watermarking methods.
(v) **Attack using a modified query:** To make the water mark authentication procedure incorrect, the pirates alter the query. In particular, the pirate will identify if the query is matching a watermark authentication search from the IP owner after deploying the pirated MLaaS service, and will modify or shield the query to fail the watermark authentication procedure.

The query modification attack put out by Namba and Sakuma [1] operates as follows. 1) Key case detection: The pirate uses an autoencoder to determine whether a query is a key case for watermark authentication when it comes in. 2) Modification of the query: If the query is identified as a primary instance for a watermark, it will be altered to thwart the watermark authentication procedure. The question is not altered in any other way. They suggest a strong watermarking approach with exponential weighting to counter this assault.

5. The Assessment Thoughts for DNN IP Security Techniques

The majority of assessments of the DNN IP protection measures now in place simply pay attention to the DNN watermark's functioning indicators [3]–[6], [9], [12], [17], and [23]. We advise constructing the evaluation approach for DNN privacy strategies based on the following considerations.
We advise the following method for systematic evaluation: 1) in order to show the effectiveness of DNN IP measures for various levels of attacks we evaluate the effectiveness of DNN IP protection measures for various levels of assaults; and 2) develop thorough measures to assess the effectiveness of DNN IP security techniques; 3) Describe the conditions necessary for efficient fingerprint/watermark authentication techniques with reference to deep learning [5, 9, 17]. Such measurements can serve as a guide for model creators and make it easier to compare DNN IP security techniques both now and in the future.

6. Troubleshooting and Upcoming Works

Deep learning models feature intricate architecture and a large number of parameters. They are implemented using black-box or white-box platforms. It is a difficult problem to develop protection of copyright methods for deep learning models which are implemented for commercial use. The difficulties and future course of action for DNN IP security are summarized below.

6.1 An effective and quick watermark verification algorithm

The extraction or verification process for DNN watermarks is currently ineffective. There are efficient hashing algorithms which can be utilized for quick search and verification within the setting of software and images [21]. Fast, effective, and wide range search, watermark removal, and authentication approaches are currently missing the deep learning models context, making it impossible to reach the needs of realistic commercial management copyright.

6.2 The protection of data as well as models' intellectual property

The protection of the model IP is the main emphasis of the already published studies [3], [5], [8]– [14]. Nevertheless, within deep learning scenarios, data, such as the training data as well as the output data, remains valuable along with can be regarded as IP in addition to the models. As an illustration, the company's private training dataset that was compiled and annotated might likewise be viewed as its intellectual property. The deep learning algorithm produces paintings with specific values; thus, they also need effective IP protection. The output photos of a watermarked model additionally have the watermark, as shown in studies [20], [26], which can be used as protection for the IP of the data that is generated.

6.3 Licenses for DNN

Many applications of models based on deep learning have been made. As a result, obtaining a DNN license has become crucial. Regarding DNN, there are two aspects to licenses: First is the data license, which specifies which data may be used to upskill the model and the model license being the second one, which governs how the DNN model can be used and places limitations on its duplication, modification, and redistribution. The authorization for AI models has been the subject of some controversy [22], but that is outside the purview of this work. The DNN model's copyright defense, or model license, is the subject of this article. Technical and legal considerations are both part of the model license. Technically speaking, present trademark of DNN defense works [3], [5], [8]-[10] mainly concentrate on confirming the model's ownership.

Although there have been some discussions on this issue from a legal standpoint, no agreement has yet been reached [21]. Explicit license types cannot be provided because the copyright protection rules of various nations and areas vary and are outside the meaning of this paper (a technical overview). We think that the following considerations need to be taken into account while choosing or creating the model license.

(i) Scenario for application**:** Is it a for-profit business, like MLaaS, or a nonprofit, like an open-source approach. For instance, only commercialized MLaaS Is it a for-profit business, like MLaaS, or a nonprofit, like an open-source approach. For instance, only commercialized MLaaS.

(ii) For various uses, such as use, copy, modification, and redistribution, specific rights and restrictions are needed.

(iii) How much money customers will make using the model for business?

6.4 Watermarking using Compression-Resistant DNN

Where the input photos have already undergone compression, such as JPEG, MLaaS can be used. After the watermark key pictures are compressed, the testing of the DNN watermark can be impacted. A potential research area is the development of DNN watermarking techniques that can withstand compression.

6.5 Active Operations and Complementary Security measures

Current DNN IP security's attack-resistance operates [1], [8], [14], [15] primarily aiming on modifications of model (few numbers of works [23], [26] examine watermark elimination attacks) lacking assessing attack-resistance under current attacks, like query change attacks, collusion attacks, along with obscurity attacks, etc. The current DNN IP security techniques may not be effective if the pirates launch aggressive attacks. In order to guarantee accurate watermark extraction and prevent model efficacy in deep learning scenarios, it is difficult to fend off active and powerful attacks. Open concerns and never-ending competition serve as potential targets for attacks and their associated resistance.

6.6 Mechanism for Active Authentication Control in DNN Models

The majority of the DNN IP protection techniques now in use [3], [5], [8]-[19] are passive verification techniques, which involve confirming the DNN model's copyright after piracy has already taken place. Such post verification techniques are unable to effectively stop piracy. There is a need for an active DNN protection of IP technique that can lock the framework, actively combat copyright infringement, and manage user recognition. The challenge is in adapting the deep learning model's performance and features to the needs of various users.

6.7 For DNN Models Controlling User Identity

The majority of the DNN IP security techniques now in use [3], [5], [8]-[16] concentrate on determining who owns a model yet do not authenticate or manage user identities, making them unsuitable for business copyright management. Additionally, these techniques are open to collusion attacks and other attacks from dishonest users. Understanding how DNN manages user identities can lead to developed solutions for business applications. The following are the challenges: 1) to create individual recognition for everyone; 2) to verify and find ones' recognition; 3) to selectively adjust the model's implementations according to identities of users; 4) to make DNN capable of differentiating between approved and unapproved users; 5) to enable DNN to recognize various approved clients; 6) to keep off collaboration and ambiguity assaults conducted by fraudulent users.

7. Conclusion

Currently, DNN copyright security is a prospective topic for research which has received concerns to a greater extent. In this paper, a survey has been done on the present DNN IP protection techniques, issues faced by these techniques, if these techniques can give security, and various stages of assaults. Moreover, the first taxonomy on DNN IP security works is suggested, and the methodical analysis for DNN IP security techniques has been done. Still IP protection of DNN is in an early stage which faces various issues. For commercial use, the DNN IP security techniques is required. Hopefully, this article can be referred for the taxonomy, similarity, analysis, and development of DNN IP protection techniques.

References

[1] Nagai Y, Uchida Y, Sakazawa S, Satoh S. Watermarking for deep neural networks. Int. J. Multimedia Inf. Retrieval. 2018; 7 (1): 3–16.

[2] Rouhani B.D, Chen H, Koushanfar F. DeepSigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In: Proc. 24th Int. Conf. Architectural Support Program Lang. Operating Syst; 2019. p. 485-497.

[3] Chen H, Rouhani B D, Fan X, Kilinc OC, Koushanfar F. Performance comparison of contemporary DNN watermarking techniques. 2018.

[4] Namba R., Sakuma J. Robust watermarking of neural network with exponential weighting. In: Proc. ACM Asia Conf. Comput. Commun. Secur; 2019. p. 228–240.

[5] Zhang J, et al. Protecting intellectual property of deep neural networks with watermarking. In: Proc. Asia Conf. Comput. Commun Secur; 2018. p. 159-172.

[6] Chen H, Rouhani B D, Fu C, Zhao J, Koushanfar F. DeepMarks: A secure fingerprinting framework for digital rights management of deep learning models. In: Proc. Int. Conf. Multimedia Retrieval; 2019. p. 105–113.

[7] Adi Y, Baum C, Cissé M, Pinkas B, Keshet J. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In: 27th USENIX Secur Symp. 2018. p. 1615–1631.

[8] Guo J, Potkonjak M. Evolutionary trigger set generation for DNN black-box watermarking. 2019.

[9] Guo J, Potkonjak M. Watermarking deep neural networks for embedded systems. In: Proc. Int. Conf. Comput. Aided Des; 2018. p.1–8.

[10] Merrer EL, Perez P, Tredan G. Adversarial frontier stitching for remote neural network watermarking. Neural Comput. Appl. 2020; vol. 32: 9233-9244.

[11] Zhao J, Hu Q, Liu G, Ma X, Chen F, Hassan M M. AFA: Adversarial fifingerprinting authentication for deep neural networks. Comput. Commun. 2020; vol. 150: 488-497.

[12] Meurisch C, Muhlhauser M. Data protection in AI services: A survey. ACM Comput Sury. 2021; vol. 54: 40:1-40:38.

[13] Zhu R, Zhang X, Shi M, Tang Z. Secure neural network watermarking protocol against forging attack. EURASIP J. Image Video Process. 2020; 1-12.

[14] Szyller S, Atli B G, Marchal S, Asokan N. DAWN: Dynamic adversarial watermarking of neural networks. In: Proc. ACM Multimedia Conf; 2021. p. 4417-4425.

[15] Lukas N, Zhang Y, Kerschbaum F. Deep neural network fingerprinting by conferrable adversarial examples. In: Proc. 9th Int. Conf. Learn. Representations; 2021. p.1-18.

[16] Tang R., Du M, Hu X. Deep serial number: Computational watermarking for DNN intellectual property protection. 2020.

[17] Chen M., Wu M. Protect your deep neural networks from piracy. In: Proc. IEEE Int. Workshop Inf. Forensics; 2018. p. 1-7.

[18] Fan L, Ng K, Chan C S. Rethinking deep neural networks ownership verification: Embedding passports to defeat ambiguity attacks. In: Proc. Annu. Conf. Neural Inf. Process. Syst; 2019. p. 4716-4725.

[19] Zhang J, Chen D, Liao J, Zhang W, Hua, G, Yu N. Passport-aware normalization for deep model protection. In: Proc. Annu. Conf. Neural Inf. Process. Syst; 2020. p. 1-10.

[20] Xue M, Wu Z, He C, Wang J, Liu W. Active DNN IP protection: A novel user fifingerprint management and DNN authorization control technique. In: Proc. IEEE 19th Int. Conf. Trust, Secur, Privacy Comput Commun; 2020. p. 975-982.

[21] Sun S, Xue M, Wang J, Liu W. Protecting the intellectual properties of deep neural networks with an additional class and steganographic images. 2021.

[22] Xue M, Sun S, He C, Zhang Y, Wang J, Liu W. ActiveGuard: Active intellectual property protection for Deep Neural Networks via adversarial examples-based user fifingerprinting. In: Proc. Int. Workshop Pract. Deep Learn. Wild (Workshop at AAAI); 2022. p. 1-7.

[23] Chakraborty A, Mondal, A, Srivastava A. Hardware-assisted intellectual property protection of deep learning models. In: Proc. 57th ACM/IEEE Des. Autom, Conf; 2020. p. 1-6.

[24] Szentannai K, Afandi AI, Horvath A. MimosaNet: An unrobust neural network preventing model stealing. 2019.

[25] Guan X, Feng H, Zhang W, Zhou H, Zhang J, Yu N. Reversible watermarking in deep convolutional neural networks for integrity authentication. In: Proc. 28th ACM Int. Conf. Multimedia; 2020. p. 2273-2280.

[26] Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proc. ACM SIGSAC Conf. Comput. Commun Secur; 2017. p. 587-601.