DNN Intellectual Property Protection: Taxonomy, Attacks and Evaluations

Mingfu Xue 1, Jian Wang 1, Weiqiang Liu 2

1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
2. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics

- Mingfu Xue, Jian Wang, Weiqiang Liu. DNN Intellectual Property Protection: Taxonomy, Attacks and Evaluations (Invited Paper). ACM Great Lakes Symposium on VLSI 2021: 455-460

Outline

I. Background

II. Taxonomy

III. Survey on DNN IP Protection Works

IV. Attacks on DNN IP Protection Works

V. Evaluation Suggestions

VI. Challenges and Future Works

VII. Conclusions

I. Background

The trained DNN model can be considered as an IP.

Malicious users may illegally copy, redistribute, or abuse the models.

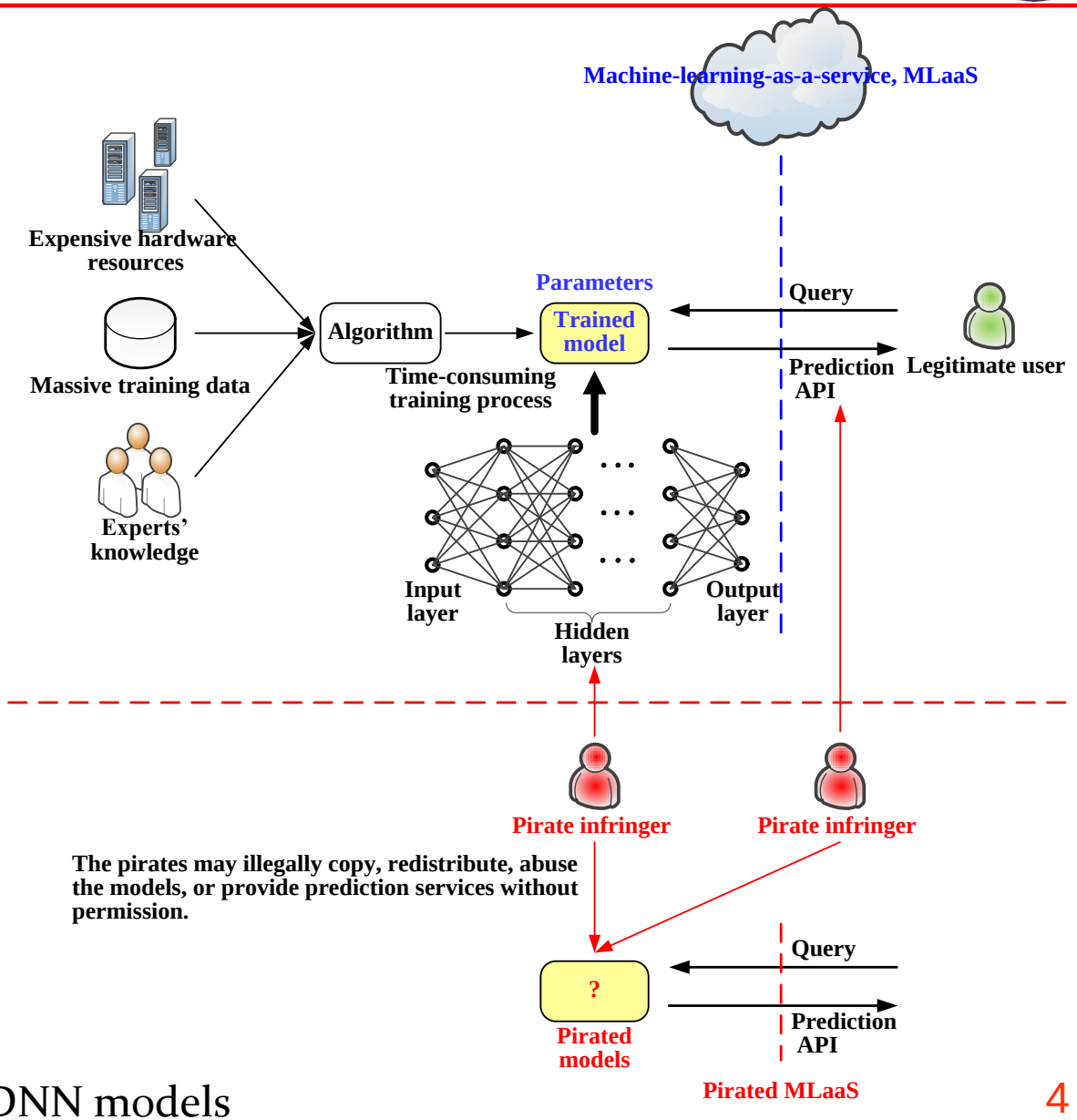

Fig. 1: Overview of piracy on DNN models

Contributions

- **Taxonomy**. We propose a taxonomy for DNN IP protection methods in terms of six attributes: ***scenario, mechanism, capacity, type, target models***, and ***function***.
- **Survey**. The survey on existing DNN IP protection methods in terms of the above six attributes are presented.
- **Analysis on attacks**. We divide the potential attacks on DNN IP protection methods into three levels (from weak to strong): (i) model modifications; (ii) evasion attacks and removal attacks (passive attacks); (iii) active attacks.
- **Evaluation suggestions**: (i) systematic evaluation method; (ii) evaluating the DNN IP protection methods when the attackers take different levels of attacks; (iii) basic functional metrics, attack driven metrics, and customized metrics for different application scenarios.
- **Challenges and future works**.

Outline

I. Background

II. Taxonomy

III. Survey on DNN IP Protection Works

IV. Attacks on DNN IP Protection Works

V. Evaluation Suggestions

VI. Challenges and Future Works

VII. Conclusions

II. Taxonomy

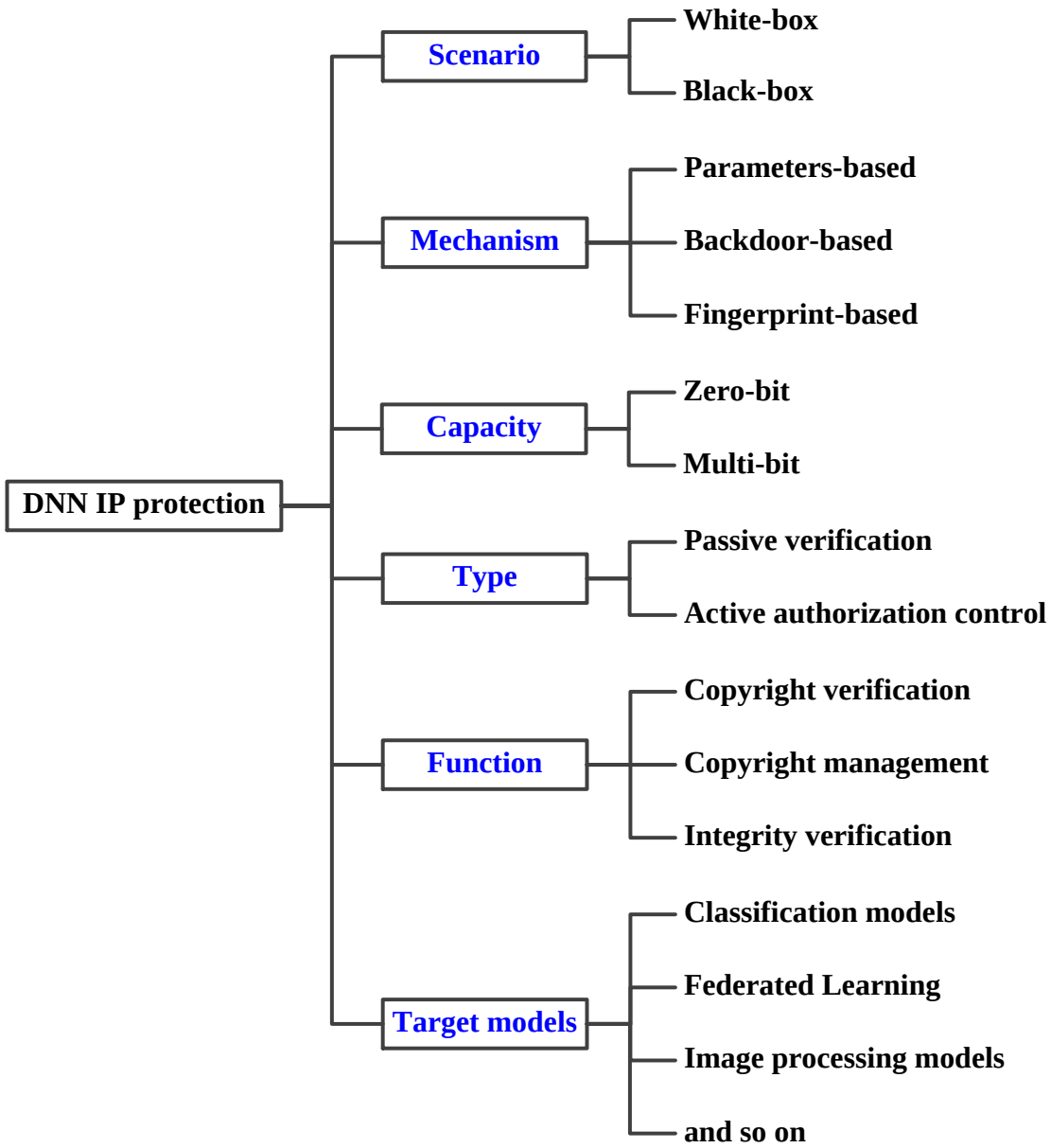

Fig. 2: The proposed taxonomy for DNN IP protection methods

Outline

III. Survey on DNN IP Protection Works

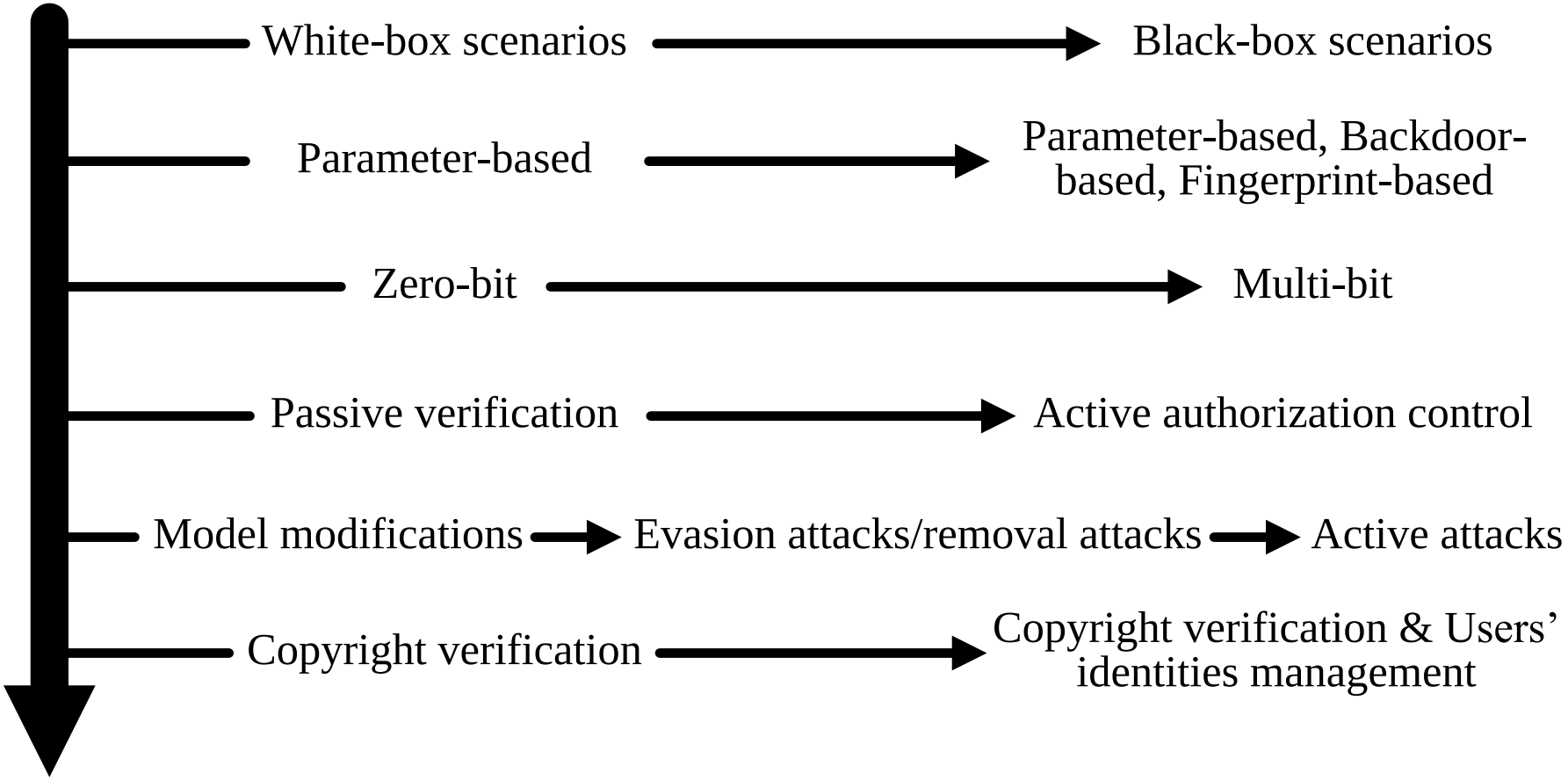

Fig. 3: The pipeline of DNN IP protection works

3.1 Scenarios

White-box Scenarios:

- the internal parameters of the model to be verified are publicly available.

Black-box Scenarios:

- However, in practice, the pirate often deploys the pirated DNN model as an online service which only outputs prediction and confidence through a remote API.
- The working mechanism: a watermark is embedded in the model, and the watermark information can only be extracted from the remote model by interacting with the model through a remote API [6].

3.2 Mechanism

Parameters-based Watermarking

- Many existing works embedding the watermarks in the parameters of the model.

Backdoor-based Method

- Using the backdoor as the watermark, and use overparameterization of the model to embed the watermark.

Fingerprint-based Method

- Some studies have demonstrated that the “fingerprint” of the model can be extracted for IP protection.

3.3 Capacity

- **Zero-bit watermarking**: only verifying the presence of the watermarks [8].

- **Multi-bit watermarking**: using the prediction of the model to perform multi-bit string verification [8].

[8] H. Chen, B. D. Rouhani, and F. Koushanfar. BlackMarks: Blackbox Multibit Watermarking for Deep Neural Networks. arXiv:1904.00344 (2019).

3.4 Type

- **Passive verification methods**: the copyright of the model is passively verified after the piracy occurs.
- **Active authorization control methods** [9, 12, 28, 36]: the model only provides normal performance for authorized users, while the unauthorized users cannot use the model or will obtain a poor performance.

[9] M. Chen and M. Wu. Protect Your Deep Neural Networks from Piracy. WIFS2018. 1–7.

[12] L. Fan, K. Ng, and C. S. Chan. Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks. NeurIPS 2019. 4716–4725.

[28] R. Tang, M. Du, and X. Hu. Deep Serial Number: Computational Watermarking for DNN Intellectual Property Protection. arXiv:2011.08960 (2020).

[36] J. Zhang, D. Chen, J. Liao, et al. Passport-aware Normalization for Deep Model Protection. NeurIPS 2020. 1–10.

3.5 Target Models

- Most works focus on protecting the IP of **classification models**.

- Zhang et al. [35] propose a watermarking approach for **image processing models**.

- **Distributed training scenarios**. Atli et. al. [3] propose a watermarking approach for **Federated Learning**. Each time the local model is aggregated to the global model, the model is re-trained to embed the watermark.

[3] B. G. Atli, Y. Xia, S. Marchal, et al. WAFFLE: Watermarking in Federated Learning. arXiv:2008.07298 (2020).

[35] J. Zhang, D. Chen, J. Liao, et al. Model watermarking for image processing networks. AAAI, 2020. 12805–12812.

3.6 Function

- **Copyright verification**. (robust watermarks)
- **Copyright management**: active authorized control and users' identities management. (robust watermarks)
- **Integrity verification**. (fragile or reversible watermarks)

Outline

IV. Attacks on DNN IP Protection Works

TABLE I
ATTACK RESISTANCE OF EXISTING DNN IP PROTECTION METHODS UNDER DIFFERENT LEVELS OF ATTACKS

Attack level	Attack type	Attack method	Attack resistance
Level 1	Model modifications	Model fine-tuning, Model pruning, Model compression, Retraining	✓
Level 2	Evasion attacks, Removal attacks	Removal attacks, Tampering, Reverse-engineering attacks	Partially
Level 3	Active attacks	Ambiguity attack, Watermark detection, Watermark overwriting, Collusion attack, Query modification attack	×

Outline

I. Background

II. Taxonomy

III. Survey on DNN IP Protection Works

IV. Attacks on DNN IP Protection Works

V. Evaluation Suggestions

VI. Challenges and Future Works

VII. Conclusions

V. Evaluation Suggestions

1. **Systematic evaluation method,** including: (i) evaluate the performance of DNN IP protection methods under different levels of attacks; (ii) establish comprehensive metrics to evaluate the performances of DNN IP protection methods [7, 8, 24].
2. **Basic functional metrics** [5–8, 14, 24, 29], including: *fidelity, robustness, functionality, capacity, efficiency, reliability, generality, uniqueness, indistinguishability,* and *scalability*.
3. **Attack-resistance metrics**, including: *security* [24, 29], *unremovability* [1, 27], *unforgeability* [1], *non-ownership piracy* [1, 27], *ownership piracy* [1, 27], *verifiability* [1], *collusion resistance* [27], and *non-invertible* [12].
4. **Customized metrics** for different application scenarios, e.g., robust watermarks and fragile watermarks.

Outline

6.1 Challenges

- Most of the existing DNN IP protection works are passive verification methods which cannot actively prevent the occurrence of the piracy.
- Most of the existing methods did not authenticate and manage the users' identities, thus are not suitable for commercial applications.
- If the pirates take active attacks, many existing DNN IP protection methods may fail.
- It lacks a systematic evaluation method.

6.2 Future Works

a) Active attacks and corresponding countermeasures.

b) Fragile/reversible watermarking methods. This is a topic that has received little attention at present.

c) Active authorization control for DNN models. Active DNN IP protection methods which can lock the model, actively prevent the occurrence of the piracy in advance (realizing copyright protection), and manage users' identities (realizing copyright management), are needed.

d) Management of users' identities for DNN models. The difficulties lie in how to design unique identity for each user, and how to differentially control the performance of the model according to users' identities.

6.2 Future Works

e) Fast and efficient watermark verification algorithm. In the existing works, the watermark extraction or verification process is inefficient. Fast, efficient and large-scale search, watermark extraction and verification methods are still lacking, thus cannot meet the requirements of practical commercial copyright management.

f) Not only IP protection for models, but also IP protection for data. Existing works focus on protecting the IP of the models. However, in deep learning scenarios, in addition to models, the data, including training data and output data, are also valuable and can be regarded as IP.

Outline

VII. Conclusions

- We propose a taxonomy for DNN IP protection methods in terms of six attributes: scenario, mechanism, capacity, type, target models, and function.
- The survey on existing DNN IP protection methods in terms of the above six attributes are presented.
- We analyze the potential attacks on DNN IP protection methods with respect to three levels: (i) model modifications; (ii) evasion attacks and removal attacks (passive attacks); (iii) active attacks.
- We suggest to build the evaluation method for DNN IP protection methods from the following aspects: (i) systematic evaluation method; (ii) evaluating the DNN IP protection methods when the attackers take different levels of attacks; (iii) basic functional metrics, attack driven metrics, and customized metrics for different application scenarios.
- We discuss the challenges faced by the state-of-the-art DNN IP protection methods and present the insights on future works.

Q & A

Thank you