Intellectual Property (IP) Protection for Deep Learning and Federated Learning Models

Farinaz Koushanfar
University of California, San Diego
La Jolla, USA
farinaz@ucsd.edu

ABSTRACT

This talk focuses on end-to-end protection of the present and emerging Deep Learning (DL) and Federated Learning (FL) models. On the one hand, DL and FL models are usually trained by allocating significant computational resources to process massive training data. The built models are therefore considered as the owner's IP and need to be protected. On the other hand, malicious attackers may take advantage of the models for illegal usages. IP protection needs to be considered during the design and training of the DL models before the owners make their models publicly available. The tremendous parameter space of DL models allows them to learn hidden features automatically.

We explore the 'over-parameterization' of DL models and demonstrate how to hide additional information within DL. Particularly, we discuss a number of our end-to-end automated frameworks over the past few years that leverage information hiding for IP protection, including: DeepSigns [5] and DeepMarks [2], the first DL watermarking and fingerprinting frameworks that work by embedding the owner's signature in the dynamic activations and output behaviors of the DL model; DeepAttest [1], the first hardware-based attestation framework for verifying the legitimacy of the deployed model via on-device attestation. We also develop a multi-bit black-box DNN watermarking scheme [3] and demonstrate spread spectrum-based DL watermarking [4]. In the context of Federated Learning (FL), we show how these results can be leveraged for the design of a novel holistic covert communication framework that allows stealthy information sharing between local clients while preserving FL convergence. We conclude by outlining the open challenges and emerging directions.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Distributed computing methodologies*; • **Information systems** → **Information retrieval**.

KEYWORDS

Deep learning, federated learning, intellectual property protection, digital right management, information hiding, on-device attestation

IH&MMSec'22, June 27–28, 2022, Santa Barbara, CA, USA

ACM ISBN 978-1-4503-9355-3/22/06.
https://doi.org/10.1145/3531536.3532957

ACM Reference Format:
Farinaz Koushanfar. 2022. Intellectual Property (IP) Protection for Deep Learning and Federated Learning Models. In *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '22), June 27–28, 2022, Santa Barbara, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3531536.3532957

BIOGRAPHY

Farinaz Koushanfar is a professor and Henry Booker Faculty Scholar in the Electrical and Computer Engineering (ECE) department at University of California San Diego (UCSD), where she is also the co-founder and co-director of the UCSD Center for Machine-Intelligence, Computing & Security (MICS). Her research addresses several aspects of efficient computing and embedded systems, with a focus on hardware and system security, robust machine learning under resource constraints, intellectual property (IP) protection, as well as practical privacy-preserving computing. Dr. Koushanfar is a fellow of the Kavli Foundation Frontiers of the National Academy of Sciences and a fellow of IEEE. She has received a number of awards and honors including the Presidential Early Career Award for Scientists and Engineers (PECASE) from President Obama, the ACM SIGDA Outstanding New Faculty Award, Cisco IoT Security Grand Challenge Award, MIT Technology Review TR-35, Qualcomm Innovation Awards, as well as Young Faculty/CAREER Awards from NSF, DARPA, ONR and ARO.

ACKNOWLEDGMENTS

This work was supported by ONR under grant number N00014-17-1-2500 and AFOSR MURI under award number FA9550-14-1-0351.

REFERENCES

[1] Huili Chen, Cheng Fu, Bita Darvish Rouhani, Jishen Zhao, and Farinaz Koushanfar. 2019. Deepattest: an end-to-end attestation framework for deep neural networks. In *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 487–498.

[2] Huili Chen, Bita Darvish Rouhani, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. 2019. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 105–113.

[3] Huili Chen, Bita Darvish Rouhani, and Farinaz Koushanfar. 2019. Blackmarks: Blackbox multibit watermarking for deep neural networks. *arXiv preprint arXiv:1904.00344* (2019).

[4] Huili Chen, Bita Darvish Rouhani, and Farinaz Koushanfar. 2020. SpecMark: A Spectral Watermarking Framework for IP Protection of Speech Recognition Systems.. In *INTERSPEECH*. 2312–2316.

[5] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. 2019. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. 485–497.