SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

AI認証：説明可能AIによるニューラルネットの識別
AI Authentication by Explainable AI

芦澤 奈実 * 鈴木 亮平 * 桐淵 直人 * 大木 哲史 † 峰野 博史 †
Nami Ashizawa Ryohei Suzuki Naoto Kiribuchi Tetsushi Ohki Hiroshi Mineno
西垣 正勝 †
Masakatsu Nishigaki

キーワード AI，認証，AIセキュリティ，説明可能AI，機械学習

あらまし

実社会のあらゆる分野でAIの利活用が進んでいる．しかし，AIが生活に浸透した将来，AIのなりすましによる被害（例えばAIを介した詐欺）が社会問題になると想定される．よって，AIの学習済みモデルが不正でないと認証できることが重要となる．

現在，モデルを確認する手法として，電子透かし [1, 2, 3] やモデル自体が有する性質を外的に確認する手法 [4, 5] の研究が進んでいる．しかしこれらの手法は知的財産保護を主目的とし，モデルの製作者がモデルを区別するに留まる．そのためモデルの利用者が不正なモデルを認証することはできない．加えて学習済みモデルは，従来のソフトウェアと異なり刻一刻と変化するだけでなく，状況に応じて様々な使い分けが想定され，唯一性がない．そのため，従来の認証方式をAIに適用するだけでは不十分であり，AIに適した認証方式の検討が必要となる．

本稿では，説明可能AIの説明性を用いてモデルを区別する手法を提案する．説明性を用いることで，ユーザによる不正モデルの確認を可能にし，モデルの変化にも対応可能となる．この手法をさらに発展させることで，説明性によるAIの識別・認証の実現を目指す．

* NTT 社会情報研究所
180-8585, 東京都武蔵野市緑町 3-9-11
NTT Social Informatics Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180-8585
† 静岡大学大学院総合科学技術研究科
432-8011, 静岡県浜松市中区城北 3-5-1
Graduate School of Integrated Science and Technology, Shizuoka University,
3-5-1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka 432-8011

参考文献

[1] 四方寿樹，矢内直人，藤原融他，“外部機構によるバックドアリングを用いた機械学習用電子透かし，” 研究報告コンピュータセキュリティ（CSEC），vol.2020, no.22，pp.1–8，2020.

[2] S. Szyller, B.G. Atli, S. Marchal, and N. Asokan, “Dawn: Dynamic adversarial watermarking of neural networks,” Proceedings of the 29th ACM International Conference on Multimedia, pp.4417–4425, 2021.

[3] N. Lukas, E. Jiang, X. Li, and F. Kerschbaum, “Sok: How robust is image classification deep neural network watermarking?(extended version),” arXiv preprint arXiv:2108.04974, pp.1–21, 2021.

[4] X. Cao, J. Jia, and N.Z. Gong, “Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary,” Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, pp.14–25, 2021.

[5] N. Lukas, Y. Zhang, and F. Kerschbaum, “Deep neural network fingerprinting by conferrable adversarial examples,” arXiv preprint arXiv:1912.00888, pp.1–18, 2019.