Efficient Verification Framework for Large-Scale Machine Learning Models

Artem Grigor[1,2], Anton Kravchenko[2], and Georg Wiese

[1] University Colledge London, London, UK
artem.grigor.23@ucl.ac.uk
[2] Confidenti
{artem,anton}@confidenti.ai

Abstract. In recent years, machine learning (ML) models have been increasingly applied to critical sectors [4] such as healthcare, finance, and legal decision-making [7]. Such adoption of AI technologies underscores the urgent need for verifying the validity of ML outputs to ensure they can be trusted and are not merely generated by opaque algorithms on remote servers. This challenge is part of the broader quest for trustworthy AI [4], which is crucial for the widespread acceptance of machine learning applications.

Verifiable Machine Learning [8] employs verifiable computation schemes [9] to confirm that a particular ML model given some data produced a specific output on a remote, potentially malicious, server. The most prevalent strategy for verifying outputs of ML model involves constructing computational circuits using SNARKs [3] that approximate the model's architecture and can cryptographically prove the inference evaluation. However, despite significant research efforts [8], scaling verifiable computations for the increasingly complex and large ML models remains a challenge, with current methods unable to efficiently handle models of even gigabyte scale [1]. Various optimizations have been proposed, focusing on tailoring proving systems to specific model architectures or enhancing the compilation of ML models into circuit-friendly formats [8], yet these optimizations yield only incremental performance improvements.

We propose two novel approaches to overcome the existing verification methods' limitations and scale Verifiable Machine Learning to accommodate larger models.

Firstly, we harness the inherent computational asymmetry between the execution and verification of most computations, including ML inference. Specifically, we introduce a framework that segments the targeted to-be-verified ML model into two components: a "Computing Model" and one or more smaller "Verifier Models." The Computing Model, which is substantially larger, receives both the initial data and the outputs from the targeted model. It then selects a suitable Verifier Model and crafts a set of detailed computation hints. These hints are utilized by the Verifier Model, which is optimized for efficient verification, to reconstruct

the output of the targeted model. This implies that by verifying the output of the Verifier Model alongside the integrity of the hints being derived from the original data is sufficient to verify the output of the targeted model, substantially reducing the computational burden. This technique draws inspiration from established practices in verifiable computation optimisation, proving particularly advantageous for tasks like ML classification, while also holding promise for a wider array of ML applications—a potential we recognize and plan to explore further.
To empirically validate our framework, we applied it to an image recognition task leveraging a YOLO (You Only Look Once) model [5], a multi-million parameter architecture aimed at object recognition in images. Within our framework, we set Computing Model as a YOLO model tuned to isolating specific pixel groups containing the targeted object. A simpler, object-class specific 2-layer Convolutional Neural Network (CNN) then serves as the Verifier Model, determining if these highlighted pixels accurately represent the target object. This refines the verification to just a single CNN model's output and a swift consistency check of the highlighted pixels against the original image, completing the verification in under a minute compared to days needed for direct YOLO verification.

Secondly, we explore the potential of specialized machine learning architectures for use in Verifiable Machine Learning. Our work with Weightless Neural Networks (WNNs) [6], which are designed to work efficiently by not using floating-point arithmetic during inference, demonstrates proof times on par with state-of-the-art conventional verifiable models, despite their relative novelty. We believe these architectures, particularly WNN, have significant untapped potential due to their current under-researched status. Our hypothesis is that such inference-optimized architectures could become highly effective for computational verification purposes, acting as ideal "Verifier" models in our framework.

Supported by a grant from the Ethereum Foundation [2], our work marks a significant step towards practical verifiable ML. By introducing a novel framework that leverages model decomposition for efficient verification, we pave the way for the validation of large-scale ML models that were previously considered infeasible. Additionally, our exploration of Inference-Optimized architectures, particularly suited for the demands of Verifiable ML, opens up new possibilities for enhancing the efficiency and practicality of verifying ML model outputs. This dual approach not only broadens the applicability of verifiable computations in ML but also sets a foundation for future innovations in the field.

References

1. The cost of inteligence: Proving machine learning inference with zero knoweldge. Online (2023), https://github.com/Modulus-Labs/Papers/blob/master/Cost$_O f_I ntelligence.pdf, accessed : 2024 - 04 - 02$
2. Artem Grigor, G.W.: Optimization and scalability of weightless neural networks: A comprehensive study. https://github.com/zkp-gravity/optimisation-research/blob/main/writeup.pdf (September 2023), available online at https://github.com/zkp-gravity/optimisation-research/blob/main/writeup.pdf
3. Groth, J.: On the size of pairing-based non-interactive arguments. In: Fischlin, M., Coron, J.S. (eds.) Advances in Cryptology – EUROCRYPT 2016. pp. 305–326. Springer Berlin Heidelberg, Berlin, Heidelberg (2016)
4. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy ai: From principles to practices (2022)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016)
6. Susskind, Z., Arora, A., Miranda, I.D.D.S., Villon, L.A.Q., Katopodis, R.F., de Araujo, L.S., Dutra, D.L.C., Lima, P.M.V., Franca, F.M.G., au2, M.B.J., John, L.K.: Weightless neural networks for efficient edge inference (2022)
7. Vice: A judge just used chatgpt to make a court decision. https://www.vice.com/en/article/k7bdmv/judge-used-chatgpt-to-make-court-decision (2023), accessed: 2023-09-29
8. Xing, Z., Zhang, Z., Liu, J., Zhang, Z., Li, M., Zhu, L., Russello, G.: Zero-knowledge proof meets machine learning in verifiability: A survey (2023)
9. Yu, X., Yan, Z., Vasilakos, A.V.: A survey of verifiable computation. Mobile Networks and Applications **22**(3), 438–453 (06 2017), https://www.proquest.com/scholarly-journals/survey-verifiable-computation/docview/1908019401/se-2, copyright - Mobile Networks and Applications is a copyright of Springer, 2017; Last updated - 2023-11-30