PhD Research Proposal

The Hebrew University of Jerusalem

Maya Swisa, Advisor: Prof. Guy Katz

1 Introduction

Deep neural networks (DNNs) are now widely employed across various domains, from computer vision to natural language processing, leading to the transformation of a vast number of applications. Their remarkable performance, particularly in complex pattern recognition tasks, has led to their deployment in increasingly safety—critical systems, such as autonomous vehicles, medical diagnosis tools, and aerospace control systems. However, despite their exceptional performance, the inherent complexity of DNNs raises significant concerns about their reliability. This is particularly evident with adversarial examples, where minor, often imperceptible, input perturbations can lead to misclassifications. The lack of robust guarantees highlights a critical need for formal verification, which has thus emerged as a crucial area of research to ensure the robustness, safety, and reliability of these models.

Despite advancements, a substantial disparity remains between the scale of DNNs amenable to current verification techniques and the complexity of those required for real world applications. This is primarily due to the fact DNN verification is computationally intractable in the worst case. The general problem of verifying properties of DNNs with piecewise—linear activations has been shown to be NP—complete [10], indicating that the computational cost can grow exponentially with the size of the network. This intractability necessitates the development of sophisticated and efficient verification techniques.

Mixed Integer Programming (MIP) and Satisfiability Modulo Theories (SMT) based solvers are designed to provide sound proofs for the correctness of a DNN, or provide a counterexample, by encoding the verification problem as a constraint satisfaction problem (CSP). While these approaches have made the verification problem feasible for many standard benchmarks, they still suffer from scalability limitations when handling complex properties and networks, and the DNN verication community is still facing several challenges limiting the applicability of the verication tools to real—world DNNs.

The challenge of deploying DNNs extends beyond just robustness and reliability to a fundamental issue of explainability. As these models are used in critical, real—world applications, their “black box” nature—their ability to produce a correct output without a transparent, human understandable reason—has become a significant barrier. A lack of explainability erodes trust, hinders debugging efforts, and raises serious ethical and legal concerns, particularly in regulated industries. Therefore, alongside formal verification, the development of methods for formally explaining a model’s decisions has become a crucial area of research. This involves moving beyond intuitive but informal explanations to a rigorous, computationally-grounded framework that can guarantee the validity and fidelity of the explanation itself.

1.1 Research objectives.

My research is aimed toward overcoming two of the challenges:

Scalability. Scalability remains a major challenge in DNN verification. While modern DNNs can have hundreds of thousands of neurons, current verification tools are typically limited to networks with only thousands of neurons, making them relevant to a small fraction of industry used DNNs. This intractability is a result of the problem's inherent complexity; even for simple cases, DNN verification is considered NP-complete [10].

My research plan focuses on addressing this by enhancing my previous work—Reinforcement Learning Guided Heuristics for Neural Network Verification.

Explainability. Beyond verification, explainability is another critical challenge in deploying deep neural networks (DNNs) in real—world applications. It addresses the "black box" nature of DNNs by providing human—understandable reasons for a model's predictions. The need for explainability is driven by the growing demand for trust, accountability, and transparency in AI systems. In high—stakes domains like autonomous vehicles and medical diagnostics, an explanation is not just helpful but essential for debugging errors, ensuring ethical compliance, and building user confidence. Current research in explainability often focuses on the computational complexity of generating explanations, as shown in the provided image, arguing that a model is only truly interpretable if its explanations can be produced efficiently.

The next section will first outline the research have already completed and its preliminary results. Then, I will detail future research directions I am aspiring to follow.

2 My Proposed Research.

Objective 1: RL—guided verification My previous research introduced an RL—guided splitting heuristic that integrates learning from demonstrations with Double DQN inside an SMT—based verifie [8–11]. A key advantage of this approach is that the trained agent, once deployed, operates on unseen properties and networks without needing to be retrained for each new query. It initially leverages the most effective heuristic to guide its initial decisions but continues to refine its policy via self-play within each verification run. In practice, this enables the agent to reduce both the average number of splits and total verification time across a diverse set of properties and networks after just one training session.

I propose to enhance this strategy by exploring and implementing promising directions. A key aspect of this research will be investigating a broader range of RL algorithms. While Double DQN proved effective, other algorithms like Proximal Policy Optimization (PPO) or Soft Actor—Critic (SAC) may offer superior performance or training efficienc [7, 18]. PPO, for instance, is known for its stability and strong performance in complex environments, while SAC is notable for its sample efficiency and ability to handle continuous action spaces, which could be relevant for future extensions of the problem. By systematically evaluating these and other algorithms, I may identify the most suitable learning framework for our domain.

Furthermore, I aim to extend the application of RL beyond just the branching heuristic. The verification process—particularly in branch—and—bound solvers—requires a sequence of decisions: not only which split to take, but also which bound tightening/relaxation to run and when to prune nodes based on bounds [3, 10, 11]. I hypothesize that a single, comprehensive RL agent could learn to manage the entire verification flow, making real-time, adaptive decisions across the entire solver. Related progress in learning decision policies for mixed-integer optimization suggests feasibility [5, 12], which I aim adapt to neural network verification. This would represent a significant step toward an end—to—end learning based verification engine.

Objective 2: Explainability The "black box" nature of deep neural networks is a major barrier to their adoption in safety critical domains [17]. While explainability methods have emerged to provide human understandable reasons for a model's predictions [14,16,20], many existing approaches lack formal guarantees. Existing techniques often rely on heuristics that can be misleading or even manipulated, producing explanations that are not faithful to the model's true decision-making process [1,6,13,19]. This poses a significant risk: users who rely on these explanations may develop a false sense of trust, potentially leading to dangerous decisions in high stakes applications like medical diagnosis or autonomous driving [17].

To address this, my research aims to bridge the gap between explainability and formal verification by developing methods that provide provably trustworthy explanations [2,15,21]. This goes beyond simply generating a plausible story for a model's behavior. It intends to offer explanations that are backed by mathematical guarantees, ensuring they accurately reflect the model's internal logic.

While some initial research on formal explainability exists [2, 4, 15, 21], many open questions remain, particularly in the context of modern, large-scale models. My work will focus on three key areas:

- **Scalability.** Existing methods often struggle with the size of contemporary networks. I will investigate how to develop efficient algorithms that can generate verifiable explanations for large models, including LLMs, where the sheer number of parameters presents a unique challenge.
- **Theoretical Foundations.** I will study the fundamental properties of provably correct explanations. This includes tackling questions about what constitutes a meaningful and safe guarantee for a given explanation, and investigating the trade offs between explanation complexity and guarantee strength [2, 15, 21].

To summarize, My research proposal addresses the critical challenges of ensuring the reliability and safety of deep neural networks (DNNs), particularly as they are deployed in safety—critical domains. My goal is to develop methods that provide provably trustworthy explanations for a model's behavior, and enhenced formal verification of DNN properties, ensuring not only that a system is safe but also that its reasoning is transparent and understandable to human operators.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS (2018)
2. Bassan, S., Katz, G.: Towards formal XAI: Formally approximate minimal explanations of neural networks. In: TACAS (LNCS 13993). pp. 187–207. Springer (2023). https://doi.org/10.1007/978-3-031-30823-9˙10
3. Bunel, R., Lu, J., Turkaslan, I., Torr, P.H.S., Kohli, P., Kumar, M.P.: Branch and bound for piecewise linear neural network verification. Journal of Machine Learning Research **21**(42), 1–39 (2020), https://jmlr.org/papers/v21/19-468.html
4. Carter, B., Mueller, J., Jain, S., Gifford, D.: What made you do this? understanding black-box decisions with sufficient input subsets. In: AISTATS (PMLR 89). pp. 567–576 (2019), https://proceedings.mlr.press/v89/carter19a.html
5. Gasse, M., Chételat, D., Ferroni, E., Charlin, L., Lodi, A.: Learning to branch in mixed integer programming with graph convolutional neural networks. In: NeurIPS (2019)
6. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: AAAI. pp. 3681–3688 (2019). https://doi.org/10.1609/aaai.v33i01.33013681
7. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: ICML (2018)
8. van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: AAAI (2016)
9. Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Rusu, A.A., Horgan, D., Quan, J., Sendonaris, A., Osband, I., Dulac-Arnold, G., Agapiou, J.P., Leibo, J.Z., Gruslys, A., Azar, M.G., Rezende, D.J., Huang, A., Botvinick, M.M., Hassabis, D., Silver, D., Singh, S., Legg, S.: Deep q-learning from demonstrations. In: AAAI (2018)
10. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient smt solver for verifying deep neural networks. In: Computer Aided Verification (CAV). Lecture Notes in Computer Science, vol. 10426, pp. 97–117. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9˙5
11. Katz, G., Huang, D.A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljić, A., Dill, D.L., Kochenderfer, M.J., Barrett, C.: The marabou framework for verification and analysis of deep neural networks. In: CAV (LNCS 11561). pp. 443–452. Springer (2019). https://doi.org/10.1007/978-3-030-25540-4˙26
12. Khalil, E.B., Bodic, P.L., Song, L., Nemhauser, G.L., Dilkina, B.: Learning to branch in mixed integer programming. In: AAAI (2016)
13. Kindermans, P., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (un)reliability of saliency methods. arXiv preprint arXiv:1711.00867 (2017)

14. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS. pp. 4765–4774 (2017)
15. Marques-Silva, J., Ignatiev, A.: Delivering trustworthy AI through formal XAI. In: AAAI. pp. 12342–12350 (2022)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?" explaining the predictions of any classifier. In: KDD. pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
17. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
18. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
19. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: AIES. pp. 180–190 (2020). https://doi.org/10.1145/3375627.3375830
20. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML. pp. 3319–3328 (2017)
21. Wu, M., Wu, H., Barrett, C.: Verix: Towards verified explainability of deep neural networks. In: NeurIPS. vol. 36, pp. 22247–22268 (2023)