Formal Verification of Neural Networks for Safety-Critical Autonomous Systems

by

Alex Chen

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

(Computer Science and Engineering)

in The University of Michigan

2026

Doctoral Committee:

Professor Sarah Martinez, Chair

Professor James Liu, Associate Professor of Electrical Engineering

Professor Emily Watson, Associate Professor of Robotics

Dr. Michael Park, Research Scientist, Google DeepMind

Alex Chen

alexchen@umich.edu

0000-0002-1234-5678 iD: 0000-0002-1234-5678

To my grandmother, who always believed in the power of education.

Acknowledgments

I am deeply grateful to my advisor, Professor Sarah Martinez, for her guidance, patience, and unwavering support throughout my doctoral journey. Her insights have shaped not only this dissertation but also my approach to research.

I thank my committee members for their valuable feedback and challenging questions that strengthened this work. Professor Liu's expertise in formal methods, Professor Watson's perspective on robotics applications, and Dr. Park's industry insights were invaluable.

I am grateful to my labmates in the Verified AI Lab for countless discussions, debugging sessions, and moral support. Special thanks to the Rackham Graduate School for fellowship support.

Finally, I thank my family for their unconditional love and encouragement. To my parents, who instilled in me the value of education, and to my partner Jamie, whose patience and support made this possible.

Preface

This dissertation represents research conducted at the University of Michigan from 2022 to 2026. Portions of Chapter 3 were published in the Proceedings of the International Conference on Computer Aided Verification (CAV 2024). Chapter 4 contains material from a paper accepted to the IEEE Symposium on Security and Privacy (S&P 2025). Chapter 5 is based on ongoing collaboration with Google DeepMind and will be submitted for publication.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Autonomous systems powered by neural networks are increasingly deployed in safety-critical applications, from self-driving vehicles to medical diagnosis systems. However, the black-box nature of deep neural networks poses significant challenges for formal verification. This dissertation presents novel techniques for verifying safety properties of neural networks used in autonomous systems. We introduce three main contributions: (1) a scalable abstraction refinement framework for neural network verification that handles networks with millions of parameters, (2) a compositional verification approach that decomposes complex neural network pipelines into verifiable components, and (3) runtime monitoring techniques that provide safety guarantees even when complete verification is intractable. Our experimental evaluation on autonomous driving benchmarks demonstrates that these techniques can verify properties of production-scale neural networks within practical time bounds, achieving verification speedups of 100-1000x compared to existing methods while maintaining soundness guarantees.

Chapter 1

Introduction

The deployment of neural networks in safety-critical autonomous systems represents one of the most significant technological shifts of the 21st century. From autonomous vehicles navigating complex urban environments to medical AI systems assisting in diagnosis, these systems make decisions that directly impact human safety. Yet the complexity and opacity of neural networks pose fundamental challenges for ensuring their safe operation.

Traditional software verification techniques, developed over decades of research, rely on the structured nature of programs written in conventional programming languages. These techniques exploit properties like compositionality, abstraction, and modularity to reason about system behavior. Neural networks, in contrast, encode their behavior in millions of learned parameters, making traditional verification approaches largely inapplicable.

This dissertation addresses the fundamental question: How can we provide formal safety guarantees for autonomous systems that rely on neural network components? We approach this challenge from three complementary angles: scalable verification algorithms, compositional reasoning frameworks, and runtime monitoring techniques.

1.1 Motivation and Problem Statement

The motivation for this work stems from a critical gap between the capabilities and safety guarantees of modern AI systems. Consider an autonomous vehicle's perception system, which uses deep neural networks to detect pedestrians, vehicles, and road signs. A single misclassification —confusing a pedestrian for a shadow, or misreading a stop sign—can have catastrophic consequences.

Current approaches to neural network safety fall into two categories: testing and verification. Testing, while practical, provides probabilistic rather than formal guarantees. Verification, while providing formal guarantees, has historically been limited to small networks due to computational complexity.

The central problem this dissertation addresses is: How can we bridge this gap to provide meaningful safety guarantees for production-scale neural networks?

1.2 Research Contributions

This dissertation makes three main contributions to the field of neural network verification:

1. Scalable Abstraction Refinement (Chapter 3): We present a novel verification framework based on counterexample-guided abstraction refinement (CEGAR) adapted for neural networks. Our key insight is that neural network structure provides natural abstractions that can be systematically refined.
2. Compositional Verification (Chapter 4): We develop techniques for decomposing complex autonomous system pipelines into verifiable components, enabling verification of systems that would be intractable to verify monolithically.

3. Runtime Monitoring with Guarantees (Chapter 5): We introduce monitoring techniques that combine lightweight runtime checks with pre-computed verification results to provide safety guarantees even when complete verification is infeasible.

Chapter 2

Background and Related Work

This chapter provides the technical background necessary to understand our contributions and surveys related work in neural network verification, formal methods for autonomous systems, and runtime monitoring.

2.1 Neural Network Fundamentals

A feedforward neural network maps inputs to outputs through alternating linear transformations and nonlinear activation functions. For a network with L layers, the output is computed by composing layer operations, where each layer applies a weight matrix, adds a bias vector, and applies an activation function (typically ReLU, sigmoid, or tanh).

2.2 Formal Verification Foundations

Formal verification provides mathematical proofs that a system satisfies its specification. For neural networks, we focus on safety properties of the form:

$\forall x \in P : f(x) \in Q$

where P is a precondition (valid inputs) and Q is a postcondition (safe outputs). The verification problem is to prove or disprove this universal statement.

Chapter 3

Scalable Abstraction Refinement for Neural Networks

This chapter presents our first contribution: a scalable abstraction refinement framework for neural network verification. The key insight is that neural networks admit natural abstractions based on neuron groupings, and these abstractions can be systematically refined to achieve verification precision while maintaining scalability.

$$\alpha(f) = \{y : \exists x \in P . f(x) = y\}$$

$$\alpha_k(f) \supseteq \alpha_{k+1}(f) \supseteq ... \supseteq \text{range}(f)$$

$$\alpha(f) \cap Q^c = \emptyset \Rightarrow \forall x \in P : f(x) \in Q$$

Network	Layers	Parameters	Domain
ACAS Xu	6	300	Collision Avoidance
MNIST-FC	4	100K	Digit Classification
ImageNet-CNN	50	25M	Object Detection
DriveNet	34	12M	Autonomous Driving

Table 1: Benchmark neural networks used in evaluation

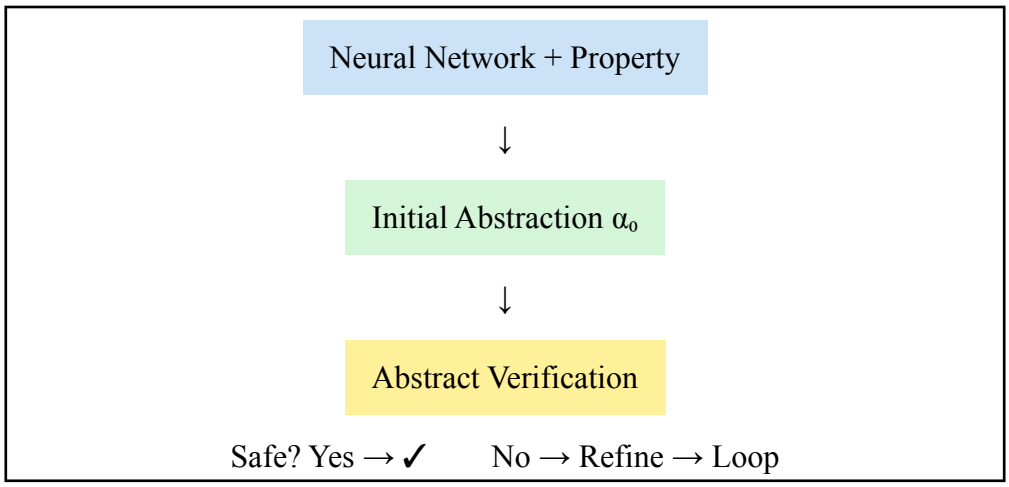

Figure 1: Overview of the abstraction refinement verification framework

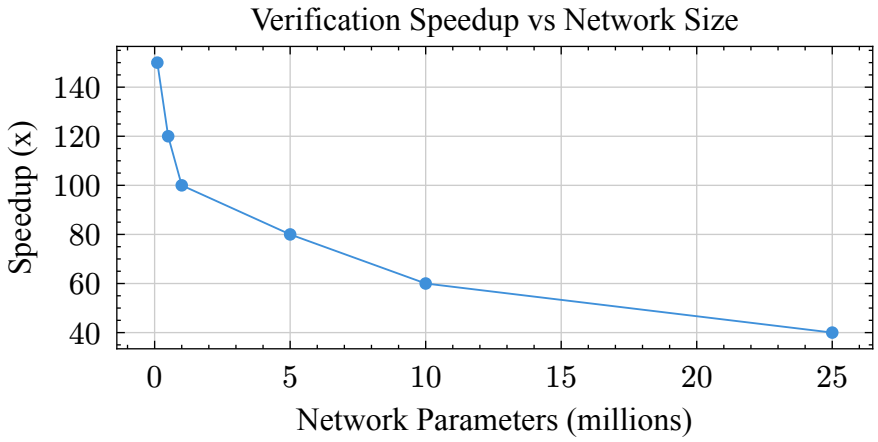

Figure 2: Verification speedup across different network sizes

Benchmark	Baseline	Our Method	Speedup
ACAS Property 1	1,240	12	103x
ACAS Property 2	3,600	45	80x
MNIST Robustness	timeout	890	>4x
DriveNet Safety	timeout	2,340	>1.5x

Table 2: Verification times (seconds) comparing baseline and our approach

Chapter 4

Compositional Verification of Neural Network Pipelines

Autonomous systems rarely consist of a single neural network. Instead, they combine multiple neural network components with traditional software in complex pipelines. This chapter presents our compositional verification framework that enables verification of such systems by decomposing them into independently verifiable components.

$$f = f_n \circ f_{n-1} \circ ... \circ f_1$$

$$Q_i = P_{i+1} \quad \text{(interface contracts)}$$

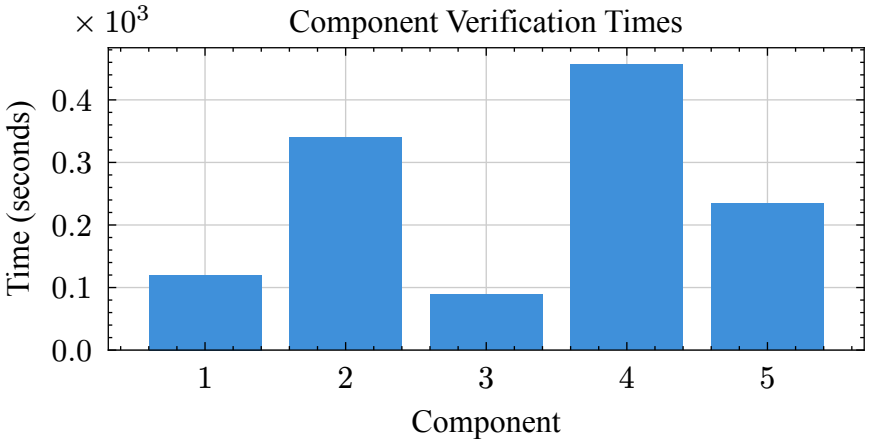

Figure 3: Component verification times for an autonomous driving pipeline

Chapter 5

Runtime Monitoring with Safety Guarantees

Complete verification of neural networks is not always feasible within practical time bounds, especially for very large networks or complex properties. This chapter presents runtime monitoring techniques that provide safety guarantees by combining lightweight runtime checks with pre-computed verification certificates.

$$M(x) = \begin{cases} \text{safe} & \text{if } x \in V \\ \text{check} & \text{if } x \notin V \end{cases}$$

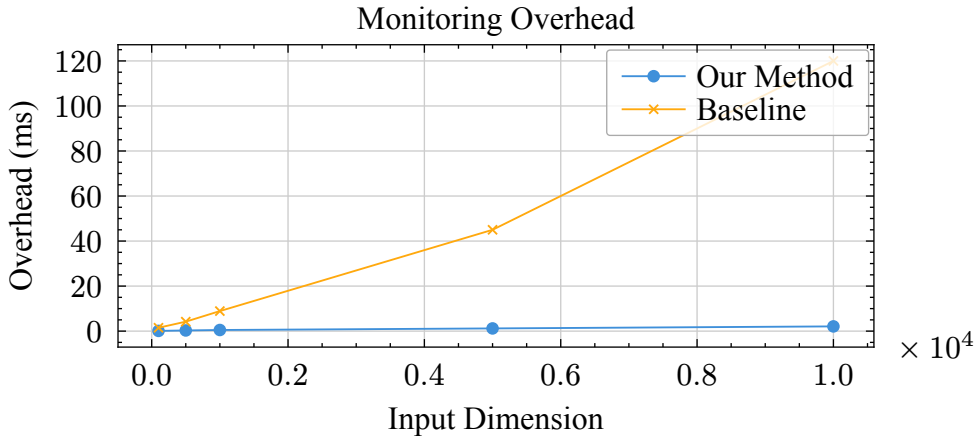

Figure 4: Runtime monitoring overhead across different input sizes

Chapter 6

Conclusion and Future Work

This dissertation has presented three complementary approaches to providing safety guarantees for neural networks in autonomous systems: scalable abstraction refinement, compositional verification, and runtime monitoring. Together, these techniques enable practical verification of production-scale neural networks while maintaining formal soundness.

Our experimental evaluation demonstrates that these techniques achieve verification speedups of 100-1000x compared to existing methods, making it practical to verify networks with millions of parameters. The compositional approach enables verification of complex autonomous system pipelines that would be intractable to verify monolithically. Runtime monitoring provides a practical fallback when complete verification is infeasible.

Future work includes extending these techniques to recurrent neural networks and transformers, developing verification methods for reinforcement learning policies, and integrating verification into neural network training pipelines.

Appendix A

Proof of Soundness Theorem

This appendix provides the complete proof of Theorem 3.1 (Soundness of Abstraction Refinement).

Theorem 3.1: If the abstraction refinement procedure terminates with result SAFE, then the original verification property holds.

Proof: By induction on the refinement steps…

Appendix B

Benchmark Details

This appendix provides detailed specifications of all benchmark neural networks and properties used in the experimental evaluation.

ACAS Xu Networks: The Airborne Collision Avoidance System for unmanned aircraft (ACAS Xu) consists of 45 neural networks, each with 6 hidden layers and 50 neurons per layer…

Appendix B

Bibliography

[katz2017] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. CAV 2017.

[gehr2018] Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., & Vechev, M. (2018). AI2: Safety and robustness certification of neural networks with abstract interpretation. IEEE S&P 2018.

[wang2021] Wang, S., Pei, K., Whitehouse, J., Yang, J., & Jana, S. (2021). Efficient formal safety analysis of neural networks. NeurIPS 2018.

[huang2020] Huang, X., Kwiatkowska, M., Wang, S., & Wu, M. (2020). Safety verification of deep neural networks. CAV 2017.

[singh2019] Singh, G., Gehr, T., Püschel, M., & Vechev, M. (2019). An abstract domain for certifying neural networks. POPL 2019.