Novelty Assessment Report

Paper: Provable Guarantees for Automated Circuit Discovery in Mechanistic Interpretability
PDF URL: https://openreview.net/pdf?id=Timsb74vIY
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Automated circuit discovery is a central tool in mechanistic interpretability for identifying the internal components of neural networks responsible for specific behaviors. While prior methods have made significant progress, they typically depend on heuristics or approximations and do not offer provable guarantees over continuous input domains for the resulting circuits. In this work, we leverage recent advances in neural network verification to propose a suite of automated algorithms that yield circuits with provable guarantees. We focus on three types of guarantees: (1) input domain robustness, ensuring the circuit agrees with the model across a continuous input region; (2) robust patching, certifying circuit alignment under continuous patching perturbations; and (3) minimality, formalizing and capturing a wide array of various notions of succinctness. Interestingly, we uncover a diverse set of novel theoretical connections among these three families of guarantees, with critical implications for the convergence of our algorithms. Finally, we conduct experiments with state-of-the-art verifiers on various vision models, showing that our algorithms yield circuits with substantially stronger robustness guarantees than standard circuit discovery methods, establishing a principled foundation for provable circuit discovery.

Core Task Landscape

This paper addresses: **Automated Circuit Discovery with Provable Guarantees**
A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Neural Network Circuit Discovery and Interpretability**
- **Hardware Circuit Formal Verification**
- **Hardware Security and Trust Verification**
- **Quantum Circuit Optimization and Verification**
- **Automated Design and Synthesis with Guarantees**
- **ISA and Processor Verification**

Complete Taxonomy Tree

- Automated Circuit Discovery with Provable Guarantees Survey Taxonomy
- Neural Network Circuit Discovery and Interpretability
  - Provable Circuit Discovery with Formal Guarantees ★ (1 papers)
  - [0] Provable Guarantees for Automated Circuit Discovery in Mechanistic Interpretability (Anon et al., 2026) View paper
  - Computational Complexity and Heuristic Analysis (1 papers)
  - [6] The computational complexity of circuit discovery for inner interpretability (Adolfi Federico, 2024) View paper
  - Compact and Verifiable Neural Architectures (1 papers)
  - [16] Truth Table Net: Scalable, Compact & Verifiable Neural Networks with a Dual Convolutional Small Boolean Circuit Networks Form (Adrien Benamira, 2024) View paper
- Hardware Circuit Formal Verification
  - Polynomial-Time Verification Methods
  - Cutwidth-Based Polynomial Verification (4 papers)
    - [4] Polynomial Formal Verification of Sequential Circuits Using Weighted-AIGs (Mohamed Nadeem, 2025) View paper
    - [7] Polynomial Formal Verification of Approximate Adders with Constant Cutwidth (Mohamed Nadeem, 2024) View paper
    - [8] Polynomial Formal Verification of Multi-Valued Logic Circuits within Constant Cutwidth Architectures (Mohamed Nadeem, 2024) View paper
    - [44] Polynomial Formal Verification exploiting Constant Cutwidth (Mohamed Nadeem, 2023) View paper
  - Specialized Polynomial Verification for Circuit Classes (6 papers)
    - [14] Polynomial Formal Verification of KFDD Circuits (Martha Schnieber, 2023) View paper
    - [21] Polynomial Formal Verification of a Processor: A RISC-V Case Study (Lennart Weingarten, 2023) View paper
    - [23] Polynomial Formal Verification of Arithmetic Circuits (Alireza Mahzoon, 2024) View paper
    - [27] Polynomial Formal Verification of Adder Circuits Using Answer Set Programming (Mohamed A. Nadeem, 2023) View paper
    - [28] Automated Polynomial Formal Verification: Human-Readable Proof Generation (Rolf Drechsler, 2023) View paper
    - [50] Polynomial Formal Verification of Approximate Functions (Martha Schnieber, 2022) View paper
  - Algebraic and Symbolic Verification Techniques (2 papers)
  - [2] Extracting Linear Relations from GrÃ¶bner Bases for Formal Verification of And-Inverter Graphs (Daniela Kaufmann, 2024) View paper
  - [38] Formal verification of arithmetic circuits by function extraction (Cunxi Yu, 2016) View paper
  - Approximate Circuit Verification (2 papers)

    - [3] Correct and Verifyâ□□CAV: Exploiting Binary Decision Diagrams to Enable Formal Verification of Approximate Adders With Correct Carry Bits (Chandan Kumar Jha, 2025) View paper
    - [5] FARAD: Automated Formal Verification of Approximate Restoring Array Dividers (Chandan Kumar Jha, 2025) View paper
    - General Formal Verification Techniques
    - BDD and SAT-Based Verification (2 papers)
        - [12] MAB-BMC: A Formal Verification Enhancer by Harnessing Multiple BMC Engines Together (Devleena Ghosh, 2024) View paper
        - [30] Formal Verification of Integer Multiplier Circuits Using Binary Decision Diagrams (Jitendra Kumar, 2023) View paper
    - Specialized Circuit and Protocol Verification (4 papers)
        - [13] Formal Verification of Zero-Knowledge Circuits (A. Coglio, 2023) View paper
        - [17] A Scalable Formal Framework for the Verification and Vulnerability Analysis of Redundancy-Based Error-Resilient Null Convention Logic Asynchronous Circuits (Dipayan Mazumder, 2024) View paper
        - [25] Formal Verification of the Stall Invariant Property for Latency-Insensitive RTL Modules (Peitian Pan, 2023) View paper
        - [26] Formal Verification of Restoring Dividers made Fast and Simple (Jiteshri Dasari, 2023) View paper
    - Analog and Nonlinear Circuit Verification (1 papers)
        - [24] Formal Verification of Nonlinear Analog Circuits using State Space-Based Model Order Reduction (Yasmine Abu-Haeyeh, 2024) View paper
    - Automated and Foundational Verification Frameworks (5 papers)
        - [11] Advanced formal verification (Drechsler, 2004) View paper
        - [29] Automated Formal Verification Methodology for Digital Circuits (Tanmay Joshi, 2023) View paper
        - [32] Formal verification of circuits (Drechsler, 2013) View paper
        - [34] A comparative study on formal verification techniques to verify large integer multiplier circuits (J Kumar, 2025) View paper
        - [48] Formal verification of hardware components in critical systems (Wilayat Khan, 2020) View paper
- Hardware Security and Trust Verification
    - Hardware Trojan Detection (3 papers)
    - [15] A golden-free formal method for trojan detection in non-interfering accelerators (Anna Lena Duque AntÃ³n, 2024) View paper
    - [20] VeriTrust: Verification for hardware trust (Jie Zhang, 2013) View paper
    - [47] Formal verification approach to detect always-on denial of service trojans in pipelined circuits (Kushal K. Ponugoti, 2021) View paper
    - Authentication and IP Protection (2 papers)
    - [10] Novel Light Weight Hardware Authentication Protocol for Resource Constrained IoT Based Devices (V. R. Vijaykumar, 2024) View paper
    - [36] CAPEC: A Cellular Automata Guided FSM-based IP Authentication Scheme (Mridha Md Mashahedur Rahman, 2023) View paper
    - Side-Channel and Fault Analysis (2 papers)
    - [37] Determined-Safe Faults Identification: A step towards ISO26262 hardware compliant designs (Felipe Augusto da Silva, 2020) View paper
    - [49] Transitional Leakage in Theory and Practice - Unveiling Security Flaws in Masked Circuits (Nicolai MÃ¼ller, 2022) View paper
    - General Security Verification for IoT and Embedded Systems (2 papers)
    - [33] Formal verification for security in IoT devices (Keerthi K, 2018) View paper
    - [39] Invasive approach to verification of functional and structural specifications implemented in custom integrated circuits (Dmitry Nagibin, 2025) View paper
- Quantum Circuit Optimization and Verification
    - Provably Optimal Quantum Circuit Synthesis (2 papers)
    - [35] QuantumCircuitOpt: An open-source framework for provably optimal quantum circuit design (Nagarajan, 2021) View paper
    - [42] Closed-Form Optimal Quantum Circuits for Single-Query Identification of Boolean Functions (Bohac, 2025) View paper
    - Automated Quantum Circuit Generation (2 papers)
    - [9] Quanto: Optimizing quantum circuits with automatic generation of circuit identities (Pointing, 2024) View paper
    - [31] Automated quantum software engineering (Sarkar, 2024) View paper
    - Formal Verification of Quantum Circuits (1 papers)
    - [46] Formal Verification of Variational Quantum Circuits (Marzari, 2025) View paper
- Automated Design and Synthesis with Guarantees
    - Approximate Circuit Synthesis with Error Bounds (2 papers)
    - [41] Scalable construction of approximate multipliers with formally guaranteed worst case error (VojtÄ□ch MrÃ¡zek, 2018) View paper
    - [45] Approximating complex arithmetic circuits with formal error guarantees: 32-bit multipliers accomplished (Milan Ä□eÅ¡ka, 2017) View paper
    - Automated Hardware Design Synthesis (2 papers)
    - [1] AlphaEvolve: A coding agent for scientific and algorithmic discovery (Novikov, 2025) View paper
    - [18] Automated synthesis of hardware designs using symbolic feedback and grammar-constrained decoding in large language models (Sumit Kumar Jha, 2024) View paper
    - Formal Synthesis for Assurance and Certification (2 papers)
    - [22] Certifying zero-knowledge circuits with refinement types (Junrui Liu, 2024) View paper
    - [43] Hierarchical contract-based synthesis for assurance cases (Timothy E. Wang, 2022) View paper
- ISA and Processor Verification (2 papers)
    - [19] Identification of ISA-Level Mutation-Classes for Qualification of RISC-V Formal Verification (Milan Funck, 2023) View paper
    - [40] LFPS: Learned Formal Proof Strengthening for Efficient Hardware Verification (Min-Woo Kang, 2023) View paper

Narrative

Core task: automated circuit discovery with provable guarantees. This field spans a diverse landscape that ranges from neural network interpretability to hardware and quantum circuit verification. At the highest level, the taxonomy divides into six main branches. Neural Network Circuit Discovery and Interpretability focuses on extracting and validating subnetworks or computational motifs within deep learning models, often seeking formal assurances about discovered structures. Hardware Circuit Formal Verification and Hardware Security and Trust Verification address classical digital design, employing techniques such as algebraic methods, bounded model

checking, and trojan detection to ensure correctness and trustworthiness. Quantum Circuit Optimization and Verification tackles the unique challenges of quantum computing, where gate-level transformations must preserve quantum states with mathematical rigor. Automated Design and Synthesis with Guarantees explores how synthesis tools can produce circuits that meet specified properties by construction, while ISA and Processor Verification ensures that instruction set architectures and microarchitectures behave as intended. Representative works illustrate these themes: Grobner Verification AIGs[2] and Weighted AIGs Verification[4] exemplify algebraic approaches in hardware verification, while Quanto[9] and Optimal Quantum Circuits[42] highlight quantum-specific concerns.

Several active lines of work reveal contrasting emphases and open questions. In hardware verification, a tension exists between scalability and completeness: some methods leverage symbolic reasoning for exhaustive guarantees (e.g., Arithmetic Circuits Verification[23]), while others adopt heuristic or approximate strategies to handle larger designs. In the neural network branch, researchers grapple with balancing interpretability and formal soundness, as seen in efforts to rigorously identify minimal subnetworks. Circuit Discovery Guarantees[0] sits within the Provable Circuit Discovery with Formal Guarantees cluster, emphasizing mathematically certified extraction of circuits. Compared to nearby efforts such as Correct and Verify[3] or FARAD[5], which may prioritize practical verification workflows or specific hardware contexts, Circuit Discovery Guarantees[0] appears to focus on establishing theoretical foundations that ensure discovered circuits meet provable correctness criteria. This positioning highlights an ongoing challenge: bridging the gap between formal guarantees and the computational demands of real-world circuit discovery across diverse domains.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on automated circuit discovery methods that provide formal guarantees (robustness, minimality, verification), distinguishing itself from heuristic approaches. The sibling subtopics address complementary aspects: one explores specialized neural architectures designed for efficient verification, while the other examines computational complexity and heuristic algorithms without formal guarantees. Together, these categories span the spectrum from provable methods to heuristic approaches and architecturally-enabled verification.

Similarities: - All three subtopics relate to understanding and discovering circuits or structures within neural networks - All involve computational considerations around verification, discovery, or analysis of neural network components - Each addresses different aspects of making neural network internals more interpretable or analyzable

Differences: - The original leaf emphasizes formal guarantees (robustness, patching, minimality) through verification techniques, while Computational Complexity focuses on heuristic methods without such guarantees - Compact and Verifiable Neural Architectures focuses on designing novel network structures to enable verification, whereas the original leaf applies verification to existing circuits discovered through automated methods - The original leaf is method-focused (automated discovery with guarantees), while siblings are either architecture-focused (structural design) or complexity-focused (theoretical/empirical analysis) - Computational Complexity explicitly excludes formal verification guarantees, making it complementary to the original leaf's core requirement

Suggested Search Directions: - Hybrid approaches combining architectural design with provable discovery methods - Complexity-theoretic bounds on circuit discovery problems with formal guarantees - Trade-offs between verification efficiency in specialized architectures versus general circuit discovery with guarantees

Sibling Subtopics

- **Compact and Verifiable Neural Architectures** (leaves: 1, papers: 1)
- Scope: Novel neural network architectures designed to enable efficient formal verification through structural properties like truth tables or dual representations.
- Exclude: Excludes general circuit discovery methods; those belong to other Neural Network Circuit Discovery subcategories.
- **Computational Complexity and Heuristic Analysis** (leaves: 1, papers: 1)
- Scope: Theoretical and empirical studies of the computational complexity properties of circuit discovery problems and scalability of heuristic algorithms.
- Exclude: Excludes methods with formal verification guarantees; those belong to Provable Circuit Discovery with Formal Guarantees.

Contributions Analysis

Overall novelty summary. The paper proposes automated circuit discovery algorithms that provide formal guarantees—input domain robustness, robust patching, and minimality—by leveraging neural network verification techniques. According to the taxonomy, it resides in the 'Provable Circuit Discovery with Formal Guarantees' leaf under 'Neural Network Circuit Discovery and Interpretability'. Notably, this leaf contains only the original paper itself, with no sibling papers identified. This isolation suggests the research direction is relatively sparse, indicating that applying formal verification methods to circuit discovery with provable guarantees represents an emerging or underexplored niche within mechanistic interpretability.

The taxonomy reveals that the broader 'Neural Network Circuit Discovery and Interpretability' branch includes two other leaves: 'Computational Complexity and Heuristic Analysis' (one paper) and 'Compact and Verifiable Neural Architectures' (one paper). These neighboring directions focus on complexity properties and architectural design for verification, respectively, rather than automated discovery with formal guarantees. The taxonomy's scope note explicitly excludes heuristic-based circuit discovery without formal guarantees, clarifying that the paper's emphasis on verification-backed extraction distinguishes it from prior heuristic approaches. This structural context underscores the paper's divergence from existing interpretability methods that lack mathematical rigor.

The contribution-level analysis examined 30 candidate papers total across three contributions. For 'Provable guarantees for circuit discovery via neural network verification', 10 candidates were examined with zero refutable matches. Similarly, 'Theoretical connections among robustness and minimality guarantees' and 'Siamese encoding for certifying circuit robustness' each examined 10 candidates with no refutations. Among the limited search scope, no prior work appears to directly overlap with the paper's core claims. The absence of refutable candidates across all contributions suggests that, within the top-30 semantic matches examined, the integration of formal verification into circuit discovery represents a novel synthesis not previously documented in this specific form.

Given the limited search scope of 30 candidates and the sparse taxonomy leaf, the paper appears to occupy a relatively unexplored intersection of mechanistic interpretability and formal verification. The analysis does not cover exhaustive literature review or domain-specific venues that might contain related work outside the semantic search radius. While the lack of refutable candidates is encouraging, the small candidate pool and isolated taxonomy position indicate that broader field awareness or deeper citation analysis could reveal additional context. The novelty assessment here reflects what is visible within the examined scope, not a definitive claim about the entire research landscape.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Provable guarantees for circuit discovery via neural network verification

Description: The authors introduce a framework that uses neural network verification techniques to discover circuits with three types of provable guarantees: input domain robustness (ensuring circuit-model agreement across continuous input regions), patching domain

robustness (certifying alignment under continuous patching perturbations), and minimality (formalizing various notions of circuit succinctness).

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. VNN: verification-friendly neural networks with hard robustness guarantees

URL: View paper

Brief Assessment

VNN[52] focuses on generating verification-friendly neural networks through post-training optimization to improve verification efficiency, not on circuit discovery in mechanistic interpretability. The candidate addresses a fundamentally different problem domain.

2. Formal verification of deep neural networks in hardware

URL: View paper

Brief Assessment

Hardware Neural Verification[57] focuses on formal verification of DNNs implemented as hardware logic through equivalence checking and network simplification, not on mechanistic interpretability or circuit discovery with continuous domain guarantees.

3. Pruning and slicing neural networks using formal verification

URL: View paper

Brief Assessment

Pruning Neural Networks[53] focuses on pruning and slicing neural networks using formal verification techniques, not on circuit discovery in mechanistic interpretability with provable guarantees over continuous input/patching domains.

4. Verification-Aided Learning of Neural Network Barrier Functions with Termination Guarantees

URL: View paper

Brief Assessment

Neural Barrier Functions[51] focuses on safety verification of neural network controllers using barrier functions for dynamical systems, not on circuit discovery in mechanistic interpretability. The verification techniques serve entirely different purposes in distinct domains.

5. Quantitative verification of neural networks and its security applications

URL: View paper

Brief Assessment

Quantitative Neural Verification[55] focuses on quantitative verification of neural networks for security applications (robustness, trojan attacks, fairness), not on circuit discovery in mechanistic interpretability. The candidate addresses counting satisfying assignments for logical properties over neural networks, whereas the original paper addresses discovering interpretable circuits with robustness guarantees.

6. Truth Table Net: Scalable, Compact & Verifiable Neural Networks with a Dual Convolutional Small Boolean Circuit Networks Form

URL: View paper

Brief Assessment

Truth Table Net[16] focuses on designing DNNs with dual boolean logic circuit representations for formal verification of the network itself, not on discovering interpretable circuits within existing models using verification techniques.

7. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming

URL: View paper

Brief Assessment

Memory-Efficient SDP[56] focuses on neural network verification for adversarial robustness and stability specifications, not on circuit discovery in mechanistic interpretability. The candidate addresses verification of network properties (robustness to perturbations), while the original contribution concerns discovering interpretable circuits within networks with guarantees about their faithfulness and minimality.

8. Abstract Layer for LeakyReLU for Neural Network Verification Based on Abstract Interpretation

URL: View paper

Brief Assessment

LeakyReLU Abstract Layer[59] focuses on robustness verification of neural networks using abstract interpretation for LeakyReLU layers, not on circuit discovery in mechanistic interpretability. The candidate addresses a fundamentally different problem domain (verifying robustness against input perturbations) rather than discovering interpretable circuits with provable guarantees.

9. Formal verification of neural network controlled autonomous systems

URL: View paper

Brief Assessment

Autonomous Systems Verification[54] focuses on verifying safety properties of autonomous robots with NN controllers processing sensor data, not on mechanistic interpretability or circuit discovery in neural networks. The domains are fundamentally different.

10. A Comprehensive Overview of Formal Methods and Deep Learning for Verification and Optimization

URL: View paper

Brief Assessment

Deep Learning Verification Overview[58] is a survey paper on formal methods and deep learning verification. It does not present novel circuit discovery methods or mechanistic interpretability techniques with provable guarantees.

Contribution 2: Theoretical connections among robustness and minimality guarantees

Description: The work establishes theoretical links between input robustness, patching robustness, and minimality guarantees. A key result is the circuit monotonicity property, which underpins minimality guarantees and clarifies convergence conditions for optimization algorithms. The authors also prove a duality between circuits and small blocking subgraphs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Minimalist Explanation Generation and Circuit Discovery

URL: View paper

Brief Assessment

Minimalist Circuit Discovery[61] focuses on generating minimal explanations through activation matching and circuit discovery in vision models, but does not establish theoretical connections between robustness and minimality guarantees or prove circuit monotonicity properties as in the original work.

---

2. Robustness, reliability, and overdetermination (1981)

URL: View paper

Brief Assessment

Robustness and Overdetermination[64] focuses on robustness analysis as a method for determining reliability through multiple independent determinations, not on theoretical connections between robustness and minimality guarantees in circuit discovery contexts.

---

3. Synthesis of robust delay-fault-testable circuits: Theory

URL: View paper

Brief Assessment

Delay-Fault Synthesis[69] addresses delay-fault testability in combinational logic circuits, focusing on gate-delay and path-delay faults. This is a fundamentally different domain from mechanistic interpretability and neural network circuit discovery, with no overlap in theoretical frameworks or problem formulations.

---

4. Hybrid Reward-Driven Reinforcement Learning for Efficient Quantum Circuit Synthesis

URL: View paper

Brief Assessment

Hybrid Quantum Synthesis[66] focuses on quantum circuit synthesis using reinforcement learning for quantum state preparation, not on circuit discovery in mechanistic interpretability with robustness and minimality guarantees for neural networks.

---

5. Biological robustness

URL: View paper

Brief Assessment

Biological Robustness[62] discusses biological systems and duplicated circuits at a systems-level, not formal verification or circuit discovery algorithms in neural networks. The domains are fundamentally different.

---

6. Understanding Verbatim Memorization in LLMs Through Circuit Discovery

URL: View paper

Brief Assessment

Verbatim Memorization Circuits[63] focuses on identifying circuits for memorization behavior in LLMs through contrastive datasets, not on theoretical connections between robustness and minimality guarantees in circuit discovery algorithms.

---

7. Graph-Based Bayesian Optimization for Quantum Circuit Architecture Search with Uncertainty Calibrated Surrogates

URL: View paper

Brief Assessment

Bayesian Quantum Architecture[67] focuses on automated quantum circuit discovery using graph-based Bayesian optimization for quantum machine learning applications. It does not address theoretical connections between robustness and minimality in circuit discovery for mechanistic interpretability.

---

8. EXP-CAM: Explanation Generation and Circuit Discovery Using Classifier Activation Matching

URL: View paper

Brief Assessment

EXP-CAM[68] focuses on generating minimal visual explanations for image classifiers through activation matching and auto-encoder training. It does not address theoretical connections between robustness guarantees and minimality in circuit discovery, nor does it establish formal properties like circuit monotonicity or duality between circuits and blocking subgraphs.

---

9. Synthetic epigenetic circuits to investigate robustness and adaptability of epigenetic inheritance in schizosaccharomyces pombe

URL: View paper

Brief Assessment

Synthetic Epigenetic Circuits[65] focuses on synthetic biology and epigenetic inheritance in yeast, not on circuit discovery algorithms or neural network verification. The domains are entirely distinct.

---

10. Hypothesis testing the circuit hypothesis in LLMs

URL: View paper

Brief Assessment

Circuit Hypothesis Testing[60] focuses on hypothesis testing frameworks for evaluating circuits in LLMs, not on theoretical connections between robustness and minimality in circuit discovery algorithms for neural network verification.

Contribution 3: Siamese encoding for certifying circuit robustness

Description: The authors introduce a novel siamese encoding method that duplicates and stacks the circuit with the full model graph to enable verification of robustness properties. This encoding allows standard neural network verifiers to certify that circuits maintain faithfulness across continuous input or patching domains.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Efficient Global Robustness Certification of Neural Networks via Interleaving Twin-Network Encoding (Extended Abstract)

URL: View paper

Brief Assessment

Twin-Network Encoding[76] addresses global robustness certification for neural networks by comparing two network copies under different inputs, whereas the original paper focuses on circuit discovery in mechanistic interpretability with siamese encoding to verify circuit faithfulness across continuous domains. These are fundamentally different problem domains and applications.

---

2. Efficient Global Robustness Certification of Neural Networks via Interleaving Twin-Network Encoding

URL: View paper

Brief Assessment

Twin-Network Global Robustness[78] focuses on global robustness certification of neural networks under input perturbations, not on circuit discovery or faithfulness verification in mechanistic interpretability contexts.

---

3. QuanFlex-GAN: a flexible algorithm for image generation via quantum generative modeling

URL: View paper

Brief Assessment

QuanFlex-GAN[74] focuses on quantum generative modeling for image generation using dual parameterized quantum circuits, not neural network verification or circuit robustness certification in mechanistic interpretability.

---

4. WireLightning: Harnessing Capacitances for In-Transit Massively Parallel Matrix Multiplication

URL: View paper

Brief Assessment

WireLightning[75] focuses on hardware-level matrix multiplication using capacitances in analog circuits, not neural network verification or mechanistic interpretability. The retrieved text fragments mention 'convolutional neural networks' only in passing and do not address circuit discovery, verification encodings, or robustness certification methods.

---

5. MEFP-Net: A Dual-Encoding Multi-Scale Edge Feature Perception Network for Skin Lesion Segmentation

URL: View paper

Brief Assessment

MEFP-Net[77] focuses on skin lesion segmentation using dual-encoding for multi-scale feature extraction in medical imaging, not on circuit robustness verification in neural networks or mechanistic interpretability.

---

6. Physical twinning for joint encoding-decoding optimization in computational optics: a review

URL: View paper

Brief Assessment

Physical Twinning Optics[70] focuses on joint optimization of optical encoding-decoding systems in computational imaging, not neural network verification or circuit robustness certification. The siamese encoding discussed refers to optical system design, not neural network verification methods.

---

7. Truth Table Net: Scalable, Compact & Verifiable Neural Networks with a Dual Convolutional Small Boolean Circuit Networks Form

URL: View paper

Brief Assessment

Truth Table Net[16] does not employ siamese encoding methods. It constructs networks with inherent dual CNF boolean logic representations for verification purposes, which is architecturally distinct from encoding circuits for robustness certification.

---

8. Efficient fault-criticality analysis for AI accelerators using a neural twin

URL: View paper

Brief Assessment

Neural Twin Accelerators[72] focuses on fault-criticality analysis in AI accelerators using neural twins of processing elements, not on siamese encoding methods for certifying neural network circuit robustness across continuous input domains.

---

9. NeuADC: Neural Network-Inspired Synthesizable Analog-to-Digital Conversion

URL: View paper

Brief Assessment

NeuADC[73] focuses on analog-to-digital converter design using neural networks and RRAM crossbar architectures for signal conversion. It does not address circuit discovery in mechanistic interpretability or siamese encoding methods for verifying neural network circuits' robustness properties.

---

10. DMFNet: Dual-encoder multistage feature fusion network for infrared small target detection

URL: View paper

Brief Assessment

DMFNet[71] focuses on infrared small target detection using dual-encoder feature fusion for image processing tasks, not on neural network verification or circuit discovery in mechanistic interpretability. The term 'siamese' in DMFNet refers to dual-encoder architectures for visual feature extraction, which is fundamentally different from the original paper's siamese encoding method for certifying robustness properties of computational circuits through neural network verification.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Provable Guarantees for Automated Circuit Discovery in Mechanistic Interpretability View paper
- [1] AlphaEvolve: A coding agent for scientific and algorithmic discovery View paper
- [2] Extracting Linear Relations from GrÃ¶bner Bases for Formal Verification of And-Inverter Graphs View paper
- [3] Correct and Verifyâ□□CAV: Exploiting Binary Decision Diagrams to Enable Formal Verification of Approximate Adders With Correct Carry Bits View paper
- [4] Polynomial Formal Verification of Sequential Circuits Using Weighted-AIGs View paper
- [5] FARAD: Automated Formal Verification of Approximate Restoring Array Dividers View paper
- [6] The computational complexity of circuit discovery for inner interpretability View paper

- [7] Polynomial Formal Verification of Approximate Adders with Constant Cutwidth View paper
- [8] Polynomial Formal Verification of Multi-Valued Logic Circuits within Constant Cutwidth Architectures View paper
- [9] Quanto: Optimizing quantum circuits with automatic generation of circuit identities View paper
- [10] Novel Light Weight Hardware Authentication Protocol for Resource Constrained IoT Based Devices View paper
- [11] Advanced formal verification View paper
- [12] MAB-BMC: A Formal Verification Enhancer by Harnessing Multiple BMC Engines Together View paper
- [13] Formal Verification of Zero-Knowledge Circuits View paper
- [14] Polynomial Formal Verification of KFDD Circuits View paper
- [15] A golden-free formal method for trojan detection in non-interfering accelerators View paper
- [16] Truth Table Net: Scalable, Compact & Verifiable Neural Networks with a Dual Convolutional Small Boolean Circuit Networks Form View paper
- [17] A Scalable Formal Framework for the Verification and Vulnerability Analysis of Redundancy-Based Error-Resilient Null Convention Logic Asynchronous Circuits View paper
- [18] Automated synthesis of hardware designs using symbolic feedback and grammar-constrained decoding in large language models View paper
- [19] Identification of ISA-Level Mutation-Classes for Qualification of RISC-V Formal Verification View paper
- [20] VeriTrust: Verification for hardware trust View paper
- [21] Polynomial Formal Verification of a Processor: A RISC-V Case Study View paper
- [22] Certifying zero-knowledge circuits with refinement types View paper
- [23] Polynomial Formal Verification of Arithmetic Circuits View paper
- [24] Formal Verification of Nonlinear Analog Circuits using State Space-Based Model Order Reduction View paper
- [25] Formal Verification of the Stall Invariant Property for Latency-Insensitive RTL Modules View paper
- [26] Formal Verification of Restoring Dividers made Fast and Simple View paper
- [27] Polynomial Formal Verification of Adder Circuits Using Answer Set Programming View paper
- [28] Automated Polynomial Formal Verification: Human-Readable Proof Generation View paper
- [29] Automated Formal Verification Methodology for Digital Circuits View paper
- [30] Formal Verification of Integer Multiplier Circuits Using Binary Decision Diagrams View paper
- [31] Automated quantum software engineering View paper
- [32] Formal verification of circuits View paper
- [33] Formal verification for security in IoT devices View paper
- [34] A comparative study on formal verification techniques to verify large integer multiplier circuits View paper
- [35] QuantumCircuitOpt: An open-source framework for provably optimal quantum circuit design View paper
- [36] CAPEC: A Cellular Automata Guided FSM-based IP Authentication Scheme View paper
- [37] Determined-Safe Faults Identification: A step towards ISO26262 hardware compliant designs View paper
- [38] Formal verification of arithmetic circuits by function extraction View paper
- [39] Invasive approach to verification of functional and structural specifications implemented in custom integrated circuits View paper
- [40] LFPS: Learned Formal Proof Strengthening for Efficient Hardware Verification View paper
- [41] Scalable construction of approximate multipliers with formally guaranteed worst case error View paper
- [42] Closed-Form Optimal Quantum Circuits for Single-Query Identification of Boolean Functions View paper
- [43] Hierarchical contract-based synthesis for assurance cases View paper
- [44] Polynomial Formal Verification exploiting Constant Cutwidth View paper
- [45] Approximating complex arithmetic circuits with formal error guarantees: 32-bit multipliers accomplished View paper
- [46] Formal Verification of Variational Quantum Circuits View paper
- [47] Formal verification approach to detect always-on denial of service trojans in pipelined circuits View paper
- [48] Formal verification of hardware components in critical systems View paper
- [49] Transitional Leakage in Theory and Practice - Unveiling Security Flaws in Masked Circuits View paper
- [50] Polynomial Formal Verification of Approximate Functions View paper
- [51] Verification-Aided Learning of Neural Network Barrier Functions with Termination Guarantees View paper
- [52] VNN: verification-friendly neural networks with hard robustness guarantees View paper
- [53] Pruning and slicing neural networks using formal verification View paper
- [54] Formal verification of neural network controlled autonomous systems View paper
- [55] Quantitative verification of neural networks and its security applications View paper
- [56] Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming View paper
- [57] Formal verification of deep neural networks in hardware View paper
- [58] A Comprehensive Overview of Formal Methods and Deep Learning for Verification and Optimization View paper
- [59] Abstract Layer for LeakyReLU for Neural Network Verification Based on Abstract Interpretation View paper
- [60] Hypothesis testing the circuit hypothesis in LLMs View paper
- [61] Minimalist Explanation Generation and Circuit Discovery View paper
- [62] Biological robustness View paper
- [63] Understanding Verbatim Memorization in LLMs Through Circuit Discovery View paper
- [64] Robustness, reliability, and overdetermination (1981) View paper
- [65] Synthetic epigenetic circuits to investigate robustness and adaptability of epigenetic inheritance in schizosaccharomyces pombe View paper
- [66] Hybrid Reward-Driven Reinforcement Learning for Efficient Quantum Circuit Synthesis View paper
- [67] Graph-Based Bayesian Optimization for Quantum Circuit Architecture Search with Uncertainty Calibrated Surrogates View paper
- [68] EXP-CAM: Explanation Generation and Circuit Discovery Using Classifier Activation Matching View paper
- [69] Synthesis of robust delay-fault-testable circuits: Theory View paper
- [70] Physical twinning for joint encoding-decoding optimization in computational optics: a review View paper
- [71] DMFNet: Dual-encoder multistage feature fusion network for infrared small target detection View paper
- [72] Efficient fault-criticality analysis for AI accelerators using a neural twin View paper
- [73] NeuADC: Neural Network-Inspired Synthesizable Analog-to-Digital Conversion View paper
- [74] QuanFlex-GAN: a flexible algorithm for image generation via quantum generative modeling View paper

- [75] WireLightning: Harnessing Capacitances for In-Transit Massively Parallel Matrix Multiplication View paper
- [76] Efficient Global Robustness Certification of Neural Networks via Interleaving Twin-Network Encoding (Extended Abstract) View paper
- [77] MEFP-Net: A Dual-Encoding Multi-Scale Edge Feature Perception Network for Skin Lesion Segmentation View paper
- [78] Efficient Global Robustness Certification of Neural Networks via Interleaving Twin-Network Encoding View paper