AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering Department

Verification of Neural Networks for Safety Critical Applications

A Thesis submitted in fulfillment of the requirements of
Master of Science in Electrical Engineering
(Computer and Systems Engineering Department)

by

Khaled Mohammed Ibrahim Masoud Khalifa
Bachelor of Science in Electrical Engineering
(Communications and Electronics Department)
Faculty of Engineering, Alexandria University, 2014

Supervised By

Prof. Dr. Mohamed Watheq Ali Kamel El-Kharashi
Dr. Mona Mohamed Hassan Safar

Cairo, Egypt, 2022

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering Department

Verification of Neural Networks for Safety Critical Applications

by

Khaled Mohammed Ibrahim Masoud Khalifa
Bachelor of Science in Electrical Engineering
(Communications and Electronics Department)
Faculty of Engineering, Alexandria University, 2014

Examiners' Committee

Name and affiliation **Signature**

Prof. Dr. Khaled Ali Hefnawy Shehata
Electronics and Communications Engineering
College of Engineering and Technology Arab Academy for Science and Technology and Maritime Transport – Heliopolis, Cairo Branch. ....

Prof. Dr. Mahmoud Ibrahim Khalil
Computer and Systems Engineering ....
Faculty of Engineering, Ain shams University.

Prof. Dr. Mohamed Watheq Ali Kamel El-Kharashi
Computer and Systems Engineering ....
Faculty of Engineering, Ain shams University.

Date:18 May 2022

Statement

This thesis is submitted as a fulfillment of Master of Science in Electrical Engineering, Computer and Systems Engineering Department, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Khaled Mohammed Ibrahim Masoud Khalifa

Signature

. . . .

Date: 18 May 2022

Researcher Data

Name: Khaled Mohammed Ibrahim Masoud Khalifa

Date of Birth: 01/10/1992

Place of Birth: Alexandria, Egypt

Last academic degree: Bachelor of Science

Field of specialization: Electrical Engineering, Communications and Electronics Department

University issued the degree : Alexandria University

Date of issued degree : 2014

Current job : Functional Verification Consultant at Mentor Graphics, Siemens EDA

Abstract

Neural Networks (NNs) have been widely used in the development of autonomous driving systems in recent years, with applications in perception, decision-making, and even end-to-end scenarios. These systems are too sophisticated and difficult to verify because they are safety-critical. It's crucial to determine whether a neural network's judgment is supported by prior similarities in the training process when deploying neural networks in safety-critical domains. Verifying a trained safety-critical neural network entails determining the scope of the neural network's decisions, which are based on past similarities learned during the training process.

Verifying a trained safety-critical neural network using formal methods is about providing proof that formulates requirements and specifies the proof's obligations. Then, develop systems to meet those obligations, and verify that the systems do indeed meet their requirements via algorithmic proof search. Model-checking and theorem proving are common in the computer-aided design of integrated circuits, and they have also been used to uncover defects in software, evaluate embedded systems, and identify security issues. Computational proof engines like Boolean satisfiability (SAT) solvers, Binary Decision Diagrams (BDDs), and satisfiability modulo theories (SMT) solvers are at the heart of these advancements.

The purpose of this thesis is to propose a runtime monitoring system that can assess the neural network's ability to reliably categorize a new input based on past similarities in the training process. The runtime monitor saves the values of neurons by two proposed approaches during the training process, which represent the neurons' activity pattern for each sample in the training data.

As a first approach, the neuron activation patterns are used to create mathematically-specified requirements during the training process which formalize a propositional logic formula represented as boolean combinations of boolean propositions. Solving by the SAT formal approach aims to check the satisfiability of the propositional logic formulas. In the inference phase, SAT solver checks the satisfiability by solving the runtime monitor's formulas to check if the trained safety-critical neural network has seen the input data before.

As a second approach, The neurons' activation patterns are stored in binary form during the training process using the Binary Decision Diagrams (BDDs) formal approach. In the inference phase, the classification of any new input is controlled by the Hamming distance by checking if the runtime monitor contains a comparable neurons' activation pattern. If there is no match between the new activation pattern and the stored activation patterns

from the training process, the runtime monitor generates a warning that the decision is not based on prior training data similarities. More layers of the safety-critical neural network have been monitored to allow for more neuronal activity patterns of each input which produce a more accurate classification decision.

This work utilizes both the SAT-based runtime monitor and BDD-based runtime monitor designs to verify a safety-critical neural network and compare the results with the results of prior work. It found that both designs produce better and enhanced results. Besides, this work introduces a more scalable solution by proposing the SAT-based runtime monitor. The MNIST benchmark set is used to demonstrate our designs.

Summary

In Machine Learning, neural networks are one of the most studied and commonly used techniques. Despite their popularity, they have limited usage in safety and security-related scenarios where network performance assurance is required. Several solutions have recently developed automated reasoning strategies to bridge the gap between neural networks and applications that require formal guarantees regarding their behavior.

This research focuses on the verification of the safety-critical neural networks which should behave based on the training data and its similarities. This research measures how much percentage of the trained neural network is following the safety-critical principle of making decisions based on prior similarities in training. Evaluation equations measure this percentage to show how much confidence should be given to the trained neural network to be used in a safety-critical application like automated driving systems.

This research utilizes two formal approaches which are Satisfiability (SAT) and Binary Decision Diagram (BDD). This research exposes the pros and cons of each approach and compares the results of each one. Moreover, this research utilizes different algorithms in both approaches to provide various techniques for implementation which give the user the freedom to pick the most suitable one for his use case.

The thesis is divided into six chapters, along with the list of figures, the list of tables, the list of abbreviations, the list of symbols and the bibliography as listed below:

Chapter 1

This chapter gives an introduction to the verification of the safety-critical neural network discussed in this work. It starts by explaining the importance of verifying the safety-critical neural network. Then, a description of the motivations behind this work's proposed solutions is provided. The objectives of this research are also summarized. A discussion is provided for the designs' challenges of this work and the methodology used in it. Finally, the road map for the thesis showing its organization is presented.

Chapter 2

This chapter provides the needed background for this study. An overview is given for the concepts of neural networks usage as an algorithm from the widely used machine learning algorithms. The concept of safety-critical neural networks used in this work is explained in detail. Additionally, the Binary Decision Diagram formal method with the utilization of Hamming distance, and satisfiability formal method are presented.

Chapter 3

This chapter illustrates the previous related works in the verification of neural networks in general. Moreover, a detailed description of the formal approaches used to verify the safety-critical neural networks. Then, a focused study of the verification of safety-critical neural networks is provided. Runtime verification is the core approach of the proposed solutions in this work which is one of the formal verification methods to verify safety-critical neural networks. Therefore, its previous related works are presented.

Chapter 4

This chapter presents the proposed architecture of the runtime monitor implemented to verify the safety-critical neural networks. It provides an overview of the proposal architectural model. Then, a detailed explanation is provided for the architecture of the two runtime monitor designs which are the BDD-based runtime monitor and the SAT-based runtime monitor. A discussion of the implementation details in terms of the used methods and algorithms with each formal technique is presented. Additionally, a comparison between the two different conversion algorithms in the satisfiability formal approach is provided. The references used for explaining and implementing the used formal verification approaches are also listed. Eventually, this chapter explains the evaluation equations used for performance assessment of the safety-critical neural networks verification.

Chapter 5

This chapter describes the details of the experimental strategy used in this work. It starts by giving the used tools to setup the verification environment to implement and run the two proposed runtime monitor designs. It shows the neural network architecture and its layers which will be verified using the two proposed designs of the runtime monitor. Additionally, a summary of the MNIST usage is explained. Then, it presents the captured results of the evaluation equations for the two proposed designs with a detailed discussion about the performance evaluation results of the evaluation equations and comparing it to prior works. This chapter provides an explanation of what has been achieved through each proposed approach and the strengths of each one.

Chapter 6

This chapter provides a summary of the work done in this study. It concludes the research conducted in this work and illustrates the potential directions that may be pursued for future work. Additionally, the challenges and difficulties are presented.

Keywords: Neural Networks (NNs), Safety˙Critical, Formal Verification, Binary Decision Diagram, BDD, Hamming distance, Classification, Satifiability, SAT, Tseitin, Quine-McCluskey, Conjuctive Normal Form ,CNF , Disjunctive Normal Form , DNF, Image Classification.