DEEP RULE VISUALIZER: A NOVEL EXPLAINABLE DEEP NEURAL NETWORK VISUALIZATION ARCHITECTURE

Bandarigodage Lahiru Kavinda Rathnayake

208545P

Degree of Master of Science in Artificial Intelligence

Departments of Computational Mathematics

Faculty of Information Technology

University of Moratuwa

Sri Lanka

July 2023

DEEP RULE VISUALIZER: A NOVEL EXPLAINABLE DEEP NEURAL NETWORK VISUALIZATION ARCHITECTURE

Bandarigodage Lahiru Kavinda Rathnayake

208545P

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Artificial Intelligence

Departments of Computational Mathematics

Faculty of Information Technology

University of Moratuwa

Sri Lanka

July 2023

DECLARATION

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:07/07/2023

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Subha Fernando

Signature of the Supervisor: Date:

DEDICATION

To my loving parents and teachers…

ACKNOWLEDGEMENT

I would like to extend my gratitude to Dr. Subha Fernando for her guidance and support throughout the research process. Also, I would like to acknowledge the support of my family and friends for their unwavering encouragement and understanding. This thesis would not have been possible without their contributions.

ABSTRACT

This thesis explores the development and implementation of the Deep Rule Visualizer, an architecture designed to enhance the interpretability and explainability of deep neural networks (DNNs). The goal of this research is to fill the gap in interpretation and enabling effective debugging of incorrect predictions. The thesis begins by introducing the proposed architecture and its components, including the use of xDNN models, saliency maps, and IF ... THEN rules. These components are integrated to offer users a deeper understanding of how the DNN arrives at its predictions. The evaluation was done on three benchmark datasets. Qualitative analysis focuses on the user interface of the Deep Rule Visualizer, which provides visual representations of rule images, regions of interest, and a novel Deep Rule Score. The interface enables users to analyze associations between input images and predictions, enhancing interpretability and facilitating insights into the model's performance. Quantitative evaluation is conducted to assess the effectiveness of the Deep Rule Visualizer. The Deep Rule Score, which combines the LIME score and distances to prototype images, serves as a comprehensive evaluation metric. Results demonstrate that higher Deep Rule Scores align with correct predictions, while lower scores indicate incorrect predictions across all three datasets. This consistency reinforces the effectiveness of the Deep Rule Score as an indicator of prediction accuracy. The Deep Rule Visualizer has the potential to enhance trust in deep learning models and aid in identifying biases and errors. Furthermore, the architecture can be applied in various domains, including healthcare and finance, to provide transparency and understanding in decision-making processes. Several avenues for further research are proposed, such as enhancing rule generation techniques, integrating additional visualization methods, refining the user interface through user studies, extending the architecture to other domains, and addressing the ethical considerations of using explainable AI.

Keywords: XAI, xDNN, LIME, Deep Rule Visualizer, Deep Rule Score

Code: https://github.com/lahirukavinda/deep-rule-visualizer

TABLE OF CONTENTS

Declaration i

Dedication ii

Acknowledgement iii

Abstract iv

Table of Contents v

List of Figures viii

List of Tables ix

List of Abbreviations x

Chapter 1 1

Introduction 1

1.1 Prolegomena 1

1.2 Objectives 1

1.3 Background and Motivation 1

1.4 Problem Definition 3

1.5 DEEP RULE VISUALIZER 4

1.6 System Requirement 4

1.7 Structure of the Thesis 5

1.8 Summary 5

Chapter 2 6

Literature Review of DEEP RULE VISUALIZER 6

2.1 Introduction 6

2.2 eXplainable AI 6

2.3 Convnet Visualization 7

2.3.1 Deconvolutional Networks 7

2.3.2 Occlusion Sensitivity 7

2.4 xDNN 7

2.5 SHAP 9

2.6 LIME 9

2.7 Applications of eXplainable AI 10

2.8 Summary of Literature Review 10
Chapter 3 11
Technology Adopted 11
3.1 Introduction 11
3.2 xDNN 11
3.3 Additive Feature Attribution Method 13
3.3.1 LIME 14
3.3.2 Shapley Value Estimation 14
3.3.3 SHAP 14
3.3.4 SHAP vs LIME 15
3.4 Summary 15
Chapter 4 16
DEEP RULE VISUALIZER 16
4.1 Introduction 16
4.2 Hypothesis 16
4.3 Input and output of DEEP RULE VISUALIZER 16
4.4 Process 17
4.5 Features 17
4.6 Users 17
4.7 Summary 17
Chapter 5 19
Design of DEEP RULE VISUALIZER 19
5.1 Introduction 19
5.2 Covnet and xDNN 20
5.3 LIME with area mapper 20
5.4 Output resolver 20
5.4.1 Deep Rule Score 21
5.4.2 Data Storing Format 22
5.5 Explanation 22
5.6 Summary 23
Chapter 6 24
Implementation 24

6.1 Introduction 24
6.2 Datasets 24
6.2.1 MNIST 24
6.2.2 iRoads 24
6.2.3 Caltech-256 24
6.3 Data Preprocessing 25
6.4 Saving Data 25
6.4.1 Model 25
6.4.2 Prototype Images 25
6.5 Image Interpretability using LIME 26
6.6 Maximum Node Density Radius 26
6.7 Deep Rule Score 27
6.8 User Interface 27
6.8.1 Components 27
6.9 Summary 28
Chapter 7 29
Evaluation 29
7.1 Introduction 29
7.2 Experimental Setup 29
7.3 Qualitative Analysis 29
7.3.1 User Interface 29
7.3.2 IF … THEN Rule 31
7.4 Quantitative Results 33
7.4.1 Instance Evaluation 33
7.4.2 Dataset Evaluation 35
7.4.3 Final Thought 35
7.5 Summary 36
Chapter 8 37
Conclusion and Further Work 37
8.1 Conclusion 37
8.2 Further work 37
References 39

LIST OF FIGURES

Figure 1 Technologies in the scale of interpretability 6
Figure 2 xDNN rule generated for the daylight scene 8
Figure 3 xDNN rule generated for the CD 8
Figure 4 xDNN layers for training 12
Figure 5 xDNN layers for validation 13
Figure 6 SHAP values visualization 15
Figure 7 Design diagram 19
Figure 8 User interface: initial view 30
Figure 9 User interface: after selecting a test image 30
Figure 10 Deducing the rule: user interface with corresponding rule 32
Figure 11 iRoads test image correctly classified 33
Figure 12 Caltech-256 test image correctly classified 34
Figure 13 Caltech-256 test image incorrectly classified 34

LIST OF TABLES

Table 1 Transparency is useful in three stages in AI..... 2

Table 2 Data storing format..... 22

Table 3 Deep Rule Score performance..... 35

LIST OF ABBREVIATIONS

Abbreviation	Description
XAI	eXplainable Artificial Intelligence
CNN	Convolutional Neural Network
DNN	Deep Neural Network
IDE	Integrated Development Environment
CAM	Class Activation Mapping
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
LWRP	Layer-Wise Relevance Propagation
SVM	Support Vector Machines
KNN	K-Nearest Neighbors
DBSCAN	Density-Based Spatial Clustering of Applications with Noise