An evaluation of Interpretation-Nets applied to Logistic Regression for explaining Neural Networks

Paul König

April 8, 2022

1 Introduction

Explainable AI is widely seen as an important topic in Artificial Intelligence. Especially Neural Networks are as of today known for their poor explainability.

For my thesis, I will apply an existing approach for interpreting neural networks (Interpretation networks [2]) to different symbolic function types and evaluate the resulting framework in experimental settings. To achieve this, my workplan for writing the thesis consists of three steps:

1. Develop an instance of the $\mathcal{I}$-Net approach, applied to Logistic Regression
2. Implement this instance in python
3. Evaluate this instance relatively to conventional interpretation methods and compare it against $\mathcal{I}$-Nets applied to decision trees

2 Background

It is considered common knowledge, that neural networks can not be easily interpreted by themselves ("black box"). However, many applications require explainability to gain human acceptance for the predictions [1] or to fulfil legal requirements [3].

In the past, several approaches for interpretation of neural networks have been proposed [4]. One approach is called Interpretation Networks ($\mathcal{I}$-Nets)[2]. This approach is using a neural network ($\mathcal{I}$-Net), which maps the target neural network to a symbolic representation, i.e. an intrinsically interpretable function.

The $\mathcal{I}$-Net approach can be classified as a model-specific, global and post-hoc interpretation method [2].

The article [2] instantiated the $\mathcal{I}$-Net approach for regression tasks using polynomial interpretation functions.

3 Goals and Work Plan

In my thesis, I will use the existing $\mathcal{I}$-Net framework and apply it to classification problems. I will compare two different symbolic representations: Decision trees and logistic regression. Both of them will be evaluated against conventional interpretation methods, i.e. a symbolic representation generated by querying the base neural network. All in all, this adds up to a 2x2 matrix:

$\mathcal{I}$-Net for Logistic regression (new instance)	$\mathcal{I}$-Net for Decision trees (existing instance)
$\updownarrow$ *evaluate interpretability and fidelity*	$\updownarrow$ *evaluate interpretability and fidelity*
Conventional ML applied to Logistic regression	Conventional ML applied to Decision trees

On a high level, my thesis should have these outcomes:

1. Further Development of the $\mathcal{I}$-Net-approach The $\mathcal{I}$-Net approach was introduced by S. Marton, S. Lüdtke, and C. Bartelt in 2021 [2]. In this article, an instance of the $\mathcal{I}$-Net approach applied to polynomial functions was shown. I will propose an instance of this approach applied Logistic Regression.

2. Implementation of a framework for $\mathcal{I}$-Nets Logistic Regression I will implement the target neural networks, which will be trained on synthetic data. Furthermore, I will implement the $\mathcal{I}$-Net instance for Logistic Regression. This $\mathcal{I}$-Net will be trained on many instances of the target neural networks. Query the $\mathcal{I}$-Nets with an arbitrary target network will give a symbolic representation of this network. The quality of the $\mathcal{I}$-Nets will be evaluated mainly by the fidelity and interpretability of the symbolic representation.

3. Evaluation of the $\mathcal{I}$-Net-instance for Logistic Regression against conventional symbolic regression and against $\mathcal{I}$-Nets for Decision Trees Experimental evaluation of the $\mathcal{I}$-Net instance for Logistic Regression against a conventional symbolic representation (i.e. querying the target network for many times and learn the mapping function using plain machine learning). Additionally, I will compare the $\mathcal{I}$-Net for Logistic Regression with the existing instance of $\mathcal{I}$-Nets for Decision Trees. The main dimensions of the evaluation will be fidelity (between the target neural network and the symbolic representation) and interpretability (of the symbolic representation). While measurement of fidelity is well established, judging the interpretability is one of the challenges. I will mainly rely on function level evaluation (as one level of evaluation of interpretability [4]).

Workplan I will define each goal as a milestone. In the first month, I will reach milestone one and two, i.e. I've implemented the $\mathcal{I}$-Net for Logistic Regression. The third milestone is the evaluation, which should be reached by the end of month two. I will use the third month exclusively for writing and finalizing the thesis, despite starting with writing already at the end of the first month.

References

[1] Nick Bostrom and Eliezer Yudkowsky. *The ethics of artificial intelligence*, pages 316–334. Cambridge University Press, 2014.

[2] Marton S.; Lüdtke S.; Bartelt C. *Explanations for Neural Networks by Neural Networks*, volume 1. Apply. sci., 2021.

[3] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[4] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.