Reconstruction of Boolean Functions from DNNs

Extended Abstract

Camila González
Knowledge Engineering Group, Technical University of Darmstadt
Hochschulstraße 10
Darmstadt, Hesse 64289
camila.gonzalez@stud.tu-darmstadt.de

ABSTRACT

Deep neural networks are accurate predictors, but their decisions are difficult to trace. Within this work, *if-then* rule representations are extracted that illustrate the features modeled in the hidden layers. An existing algorithm is used, combined with techniques which have been shown to assist rule extraction from shallow architectures. The evaluation shows that reducing the connectivity of the neural networks leads to the extraction of simpler rule sets with better prediction accuracy; and that when the activation ranges of the networks are polarized a small amount of predefined split values are sufficient for explaining each layer, which makes extracting rule representation viable for very large architectures.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; *Neural networks*; *Rule learning*; Optimization algorithms; • **Human-centered computing** → Information visualization;

KEYWORDS

deep neural networks, interpretability, knowledge distillation

ACM Reference format:
Camila González. 2017. Reconstruction of Boolean Functions from DNNs. In *Proceedings of ACM WomENcourage conference, Barcelona, Spain, September 2017 (WomENcourage '17),* 2 pages.
https://doi.org/0000001.0000001

1 PROBLEM STATEMENT

Due to improvements in the training practices, the popularity of DNNs and their applicability to a wider set of problems have increased in the last few years. This has brought about a renewed interest in increasing their interpretability, an aspect where they fall behind other machine learning models and which is critical for their use in domains where their decisions have critical consequences.

Symbolic representations in the form of rule sets have been proven capable of illustrating the behaviour of deep models as a whole, as well as the hidden features they form in the intermediate layers. Instead of using datasets which combine the attributes in an unclear manner, in this work the rule sets are regained from networks trained to reproduce predefined boolean concepts so it can later be assessed to what degree the patterns were captured in the rule sets.

WomENcourage '17, September 2017, Barcelona, Spain
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/0000001.0000001

2 METHODOLOGY

Diverse methods have been developed for the purpose of extracting rule representations from neural networks. However, most are not scalable to deep architectures with a high number of hidden units.In this work, the *DeepRED* algorithm [3] is employed.

In order to reduce the search space of rules so that the amount of train data is not a limiting factor, it is extended with two different ways to discretize the activation ranges. Both allow the definition of an upper bound on the number of thresholds per hidden unit. The first builds the minimal number of clusters in each hidden unit so that the accuracy of the network does not decrease when the activation value is replaced by the closest cluster mean, and the medium points between cluster boundaries are taken as thresholds. The second chooses the split values to divide all neurons of a layer jointly by considering the problem of classifying the train instances as being above or below each relevant threshold.

The effect is also observed of preceding the rule extraction with two types of network alterations which have been shown to assist the latter extraction of comprehensible representations, namely encouraging sparse connectivity [2] and minimally or maximally active hidden units [1].

3 EXPERIMENTAL RESULTS

Rule sets were extracted from twenty four-layered networks modeling different boolean concepts. The generalization abilities were compared by utilizing different subsets as training data and analyzing the accuracy on the remaining instances. For determining when the performance differences were statistically significant, the sign and Wilcoxon signed ranks tests were used with a $p = 0.05$.

There was a decrease in the number of intermediate expressions which were extracted – as well as in their complexity measured with the number of terms – when the connectivity of the networks had been reduced. Regarding the accuracy, significant changes were found when 10% and 25% percent of the data was used. Rule sets extracted from pruned networks had a higher predictive accuracy. Discretizing the activation ranges only led to a decrease in accuracy when combined with network pruning, but this did not occur when the networks had been both pruned and polarized.

REFERENCES

[1] LiMin Fu. 1991. Rule Learning by Searching on Adapted Nets. In *Proceedings of the 9th National Conference on Artificial Intelligence, Anaheim, CA, USA, July 14-19, 1991, Volume 2.* 590–595.
[2] Rudy Setiono. 1997. A Penalty-Function Approach for Pruning Feedforward Neural Networks. *Neural Computation* 9, 1 (1997), 185–204.
[3] Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. 2016. DeepRED - Rule Extraction from Deep Neural Networks. In *Discovery Science - 19th International Conference, DS 2016, Bari, Italy, Proceedings.* 457–473.