UR-269 Towards Bounding the Behavior of Deep Neural Networks

Abstract

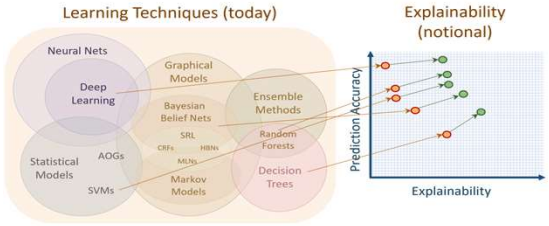
Advances in Artificial Intelligence (AI), particularly in the form of deep neural networks, have revolutionized a diverse range of fields. As neural networks become more pervasive, the need to understand the boundaries of their behavior is becoming increasingly important. For example, can we formally guarantee that an autonomous vehicle will not violate traffic laws, such as reaching excessive speeds? Towards the goal of bounding the behavior of a neural network, we propose how to bound the behavior of individual neurons by incrementally tightening formal bounds on it. We further provide a case study on classifying handwritten digits to illustrate the utility of our algorithm in terms of bounding the behavior of an individual neuron.

Explainable Artificial Intelligence (XAI)

Neural Networks have advanced rapidly in the last few years causing new models to be able to work with high accuracy. However, these advances came with trade-offs. Despite having high accuracy, new models have a very low explain ability. We do not know why models work so well and why they fail. It is important to provide an explanation of how ML/AI models work so that companies and individuals are certain that the automated decisions are justified. Explainable Artificial Intelligence (XAI) is a field of AI that is working on making models more explainable while also improving their performance.

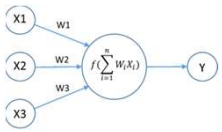

Husky


Wolf



From DARPA (XAI)

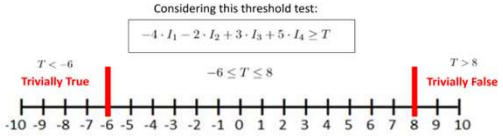
Neurons as Threshold Tests



The figure above depicts an example of a neuron. We formulate the behavior of a neuron as a threshold test, to facilitate analysis. A threshold test consists of weights and inputs that are evaluated against a threshold. The format as an inequality is described below.

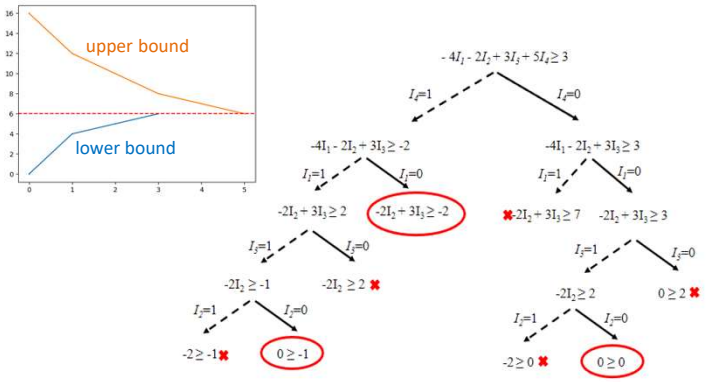
$$w_1 \cdot I_1 + w_2 \cdot I_2 + ... + w_k \cdot I_k \geq T$$

A threshold test has a lower bound where all the negative weights are set to 1 and an upper bound where all positive weights are set to 1. We call a threshold test trivially true when the threshold value is less than the lower bound and trivially false when the threshold value is greater than the upper bound. **A trivial threshold test is a (sufficient) explanation for the behavior of a neuron.**



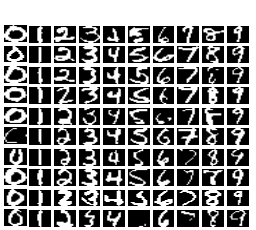
Search Space of Threshold Tests

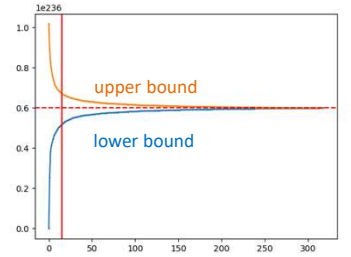
We search the space of threshold tests to find all the passing and failing explanations. To do this in the most efficient way, we iteratively set the weight with the highest absolute value. This will push the test to become trivially true or trivially false the fastest. To find all of the passing and failing explanations in a model, we explore the corresponding graph which provides **tightening lower and upper bounds on the threshold test.**

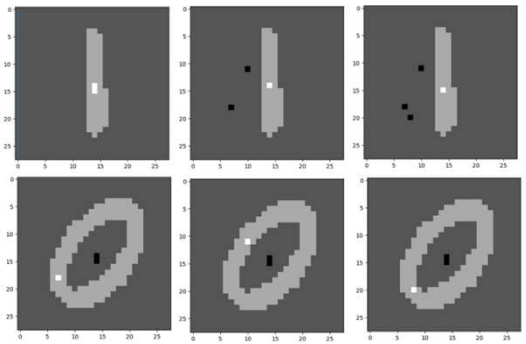


Case Study

We conducted a case study that used images of handwritten digits from 0 to 9. We took the images from the MNIST digit database. We learned a neuron (with step activation, i.e., logistic regression) from the dataset and used the weights and threshold value from the learned model to create our threshold test. We then searched through the space of threshold tests to find explanations of the model. The images below show the three shortest explanations on 0 and 1 handwritten digits. These pixels show the setting of inputs that are **sufficient explanations** for the classification of a digit as a 0 or 1 by the model.







Conclusions

We found our algorithm to be more efficient in providing upper and lower bounds on a neuron. We used a case study to explain and demonstrate how the algorithm can help advance the understandability of neurons. In the future, we hope to be able to extend these formal bounds on a network of neurons and be able to bound the behavior of a neural network.

References

[Marques-Silva et al., 2020] Marques-Silva, J., Gerspacher, T., Cooper, M. C., Ignatiev, A., and Narodytska, N. (2020). Explaining naive Bayes and other linear classifiers with polynomial time and delay. In Advances in Neural Information Processing Systems (NeurIPS).

[Shi et al., 2020] Shi, W., Shih, A., Darwiche, A., and Choi, A. (2020). On tractable representations of binary neural networks. In Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR).

[Shih et al., 2018] Shih, A., Choi, A., and Darwiche, A. (2018). A symbolic approach to explaining Bayesian network classifiers. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI).

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: Richard Borowski <rborowsk@students.kennesaw.edu>
Advisor: Arthur Choi <achoi13@kennesaw.edu>