Manipulating Model Explanations: Why you shouldn't trust me

Verena Heusser
verena.heusser@student.kit.edu
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

ABSTRACT

This paper reviews state-of-the-art approaches to model explanations with a focus on those techniques that try to fool these methods.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Interpretability, Neural networks, adversarial training

ACM Reference Format:
Verena Heusser. 2020. Manipulating Model Explanations: Why you shouldn't trust me. In *Explainability '20: ACM Symposium on Neural Model Explanations, December 16–20, 2020, Karlsruhe, DE.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years deep learning models have demonstrated superior performance in a number of tasks. While the performance is still rising and more domains of tasks are accomplished, these models still remain black boxes often uninterpretable even by experts. In many domains, neural networks currently are the state-of-the-art solution. However, as their superior performance comes at the cost of complexity and thus interpretability: The models often employ millions to even billions of parameters in order to achieve universal function approximation.

Thus, automated interpretation methods are required to make sense of the reasoning process of such deep learning based models.

This issue prevents this vastly advancing technology to be used in high-stakes and safety critical applications and prevent real-life deployment of such systems.

This article reviews the current state of the art research in the field of model explanations and model manipulations.

While most of the approaches to explainability focus on the application to computer vision tasks, other areas are seldomly chosen. More importantly, while a big motivation for the development of robust and explainable systems is to overcome biases in models, datasets with direct implication of biases are seldomly used and by far not treated as benchmarking scenarios for explainability analyses.

This paper is structured as follows...

I

2 MANIPULATION OF EXPLANATIONS

2.1 Explanation Methods

Explanation models aim at making complex and inherently uninterpretable black box models interpretable by creating human readable visualizations. A frequently used type of explanation methods are feature attributions mapping a each input feature to a model to a numeric score. This score should quantify the importance of the feature relative to the model output. The resulting attribution map is then visualized as a heatmap projected onto the input sample to interpret the input attributes regarding which ones are the most helpful for forming the final prediction.

2.1.1 Explanation methods using gradient information.

2.2 Manipulation Methods

Most of the explanation methods outlined in sec. subsection 2.1 have been shown to be vulnerable to adversarial perturbations. Manipulation methods often show that there exist small feature changes resulting in a change of the explanation methods output while the output of the model itself does not change.

Most approaches aim at providing a relevance measure of the input features.

2.2.1 Input Manipulations. The general approach is to perturb input data while observing the effect of this perturbation. As found in TODO, visually-imperceptible perturbations of an input image can make explanations worse for the same model and interpreter.

2.2.2 Model Manipulations. Contrary to the methods introduced in sec. subsubsection 2.2.1, the methods in this section do not operate on the input space of models but rather on the model parameter space itself. As first introduced by Heo et al. [1] in 2017, this line of research is comparably new. The authors find that perturbed model parameters can also make explanations worse for the same input images and interpreters.

2.2.3 Transferability of Manipulations.

3 CHARACTERIZATION OF ROBUSTNESS

sec:robustness)

This section introduces common evaluation strategies designed to test the robustness of either model or interpreter towards an applied attack.

TODO: - auch auf entlarvungsmethoden eingehen?

Explainability '20, December 16–20, 2020, Karlsruhe, DE

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 TRANSFERABILITY OF PERTURBATIONS

sec:transferability)

input perturbations do not propagate to the whole validation set. On the contrary, model manipulations are non-local perturbations, meaning that they do not merely perturb an input sample but rather effect all samples in the way that the model itself is changed.

5 EXPERIMENTS

In this section, several experiments are evaluated that were conducted to replicate findings of other studies. Furthermore these approaches are extended to other domains and datasets.

5.1 Explanation Methods

5.2 Manipulation Methods

5.3 Models

5.4 Datasets

5.4.1 ImageNet.

5.4.2 Recidivism Dataset.

5.4.3 German dataset of..

6 DISCUSSION

7 CONCLUSION

Finally, it must be noted that the suitability of a method depends on its application domain.

Much critique has been applied to methods aiming at interpreting complex and potentially non-interpretable models. Some researchers argue it is not worthwhile to study non interpretable systems while dismissing that using inherently interpretable models in the first place might be the better approach.

REFERENCES

[1] Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling Neural Network Interpretations via Adversarial Model Manipulation. *CoRR* abs/1902.02041 (2019). arXiv:1902.02041 http://arxiv.org/abs/1902.02041