

Manipulating Model Explanations: How to fool what tries to make sense of

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

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

neural networks, model explanations, adversarial training

ACM Reference Format:

Verena Heusser. 2018. Manipulating Model Explanations: How to fool what tries to make sense of. In *Explainability '20: ACM Symposium on Neural Model Explanations, December 16–20, 2020, Karlsruhe, DE*. ACM, New York, NY, USA, ?? pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

2 METHODS AND INTERPRETERS

3 EXPERIMENTS

3.1 Models

3.2 Datasets

3.2.1 *ImageNet*.

3.2.2 *Recidivism Dataset*.

3.2.3 *German dataset of..*

3.3 Measures

3.3.1 *Fooling Success Rate*.

4 RESULTS

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2018 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/1122445.1122456>