Why you shouldn’t trust me: A survey on Adversarial Model Interpretation Manipulations

Verena Heusser

29 Januar 2021

KIT, Seminar Explainable Machine Learning, Intelligent System Security Research Group

Motivation

Interpretation Methods

Adversarial Setting

Interpreter Manipulation Methods

Fooling Examples