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

Verena Heusser

30 January 2021

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

Motivation

Omnipresent Machine Learning

- Verification of machine learning algorithms mostly w.r.t. accuracy and efficiency

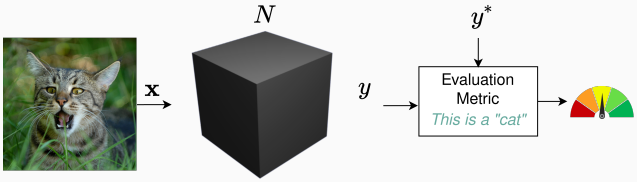

Figure 1: Machine learning Pipeline for Image Classification

Omnipresent Machine Learning

- Verification of machine learning algorithms mostly w.r.t. accuracy and efficiency
- Also Interpretability: Uncovering why a model made a decision

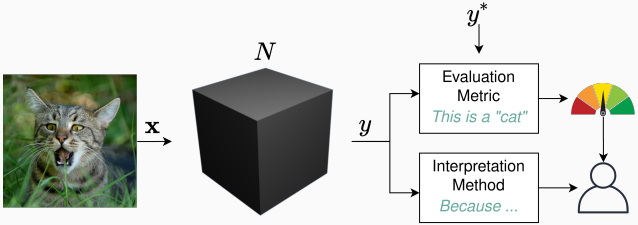

Figure 2: Machine learning Pipeline for Image Classification

Omnipresent Machine Learning

- Verification of machine learning algorithms mostly w.r.t. accuracy and efficiency
- Also Interpretability: Uncovering why a model made a decision

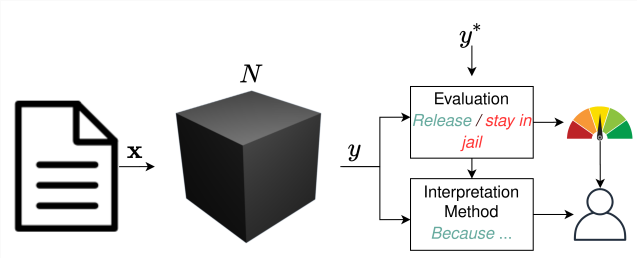

Figure 3: Machine learning Pipeline for Image Classification

- Especially:
  - Critical applications: Politics, Medicine, ...

Interpretability

Motivation

Interpretation Methods

Adversarial Setting

Adversarials: How to fool a model

- Adversarial examples []

Interpreter Manipulation Methods

Fooling Examples

References

Sources