Bundesamt für Sicherheit in der Informationstechnik

Hochschule Bonn-Rhein-Sieg

Angriffe auf KI-Systeme im Kontext der Verkehrsschilderkennung

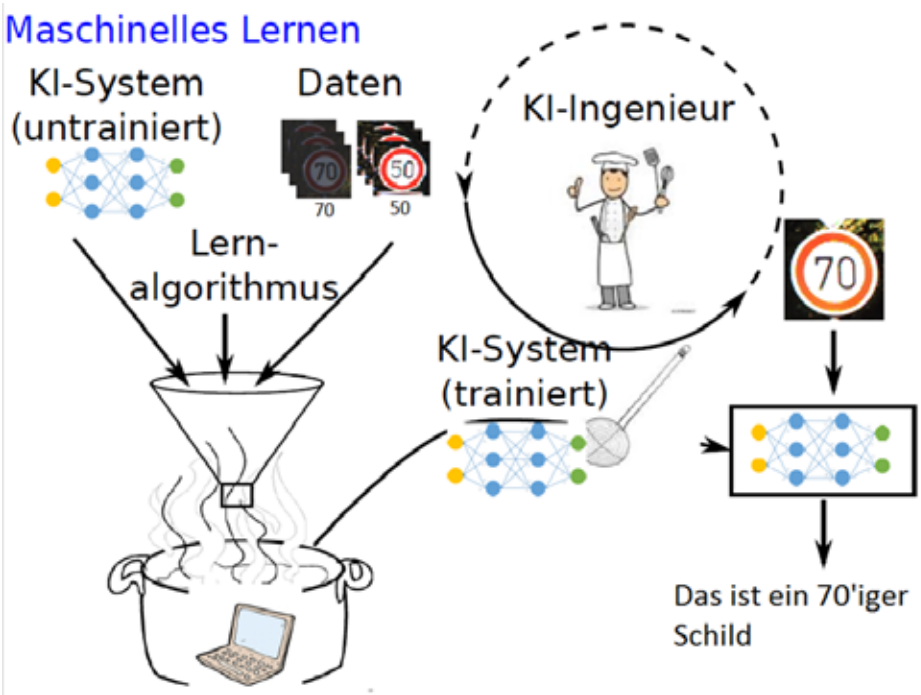
Johannes Alecke, Matthias Neu, Arndt von Twickel und **Markus Ullmann**
Referat DI11 (Bewertungsverfahren für eID-Technologien), BSI

Bundesamt für Sicherheit in der Informationstechnik

Gliederung

- Maschinelles Lernen am Beispiel Neuronaler Netze (NN): Idee
- Konkret: Verkehrsschilderkennung
- Adverserial Attack
  - Few Pixel Attack
  - Patch Attack
- Poisoning Attack (‘Trainierte Hintertüren‘)
- Angriffsimulation versus Angriffe in der realen Welt
- „Grundübel von NNs“
- Prüfbarkeit von KI-Anwendungen
- Ausblick

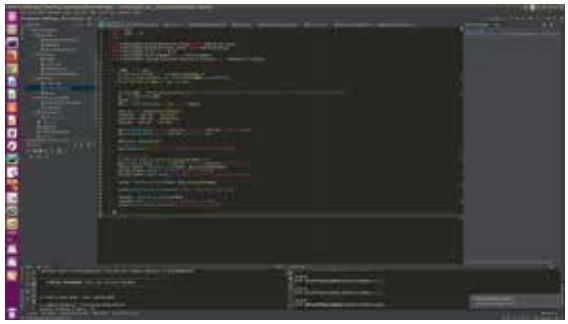
Bundesamt für Sicherheit in der Informationstechnik



Bundesamt für Sicherheit in der Informationstechnik

ennung Seite 3

Entwicklungs- /Testumgebung



Tool Chain
- Py – Editor:
- KI-Tool: PyTorch
- Dash-Board: Tensorboard

Bundesamt für Sicherheit in der Informationstechnik

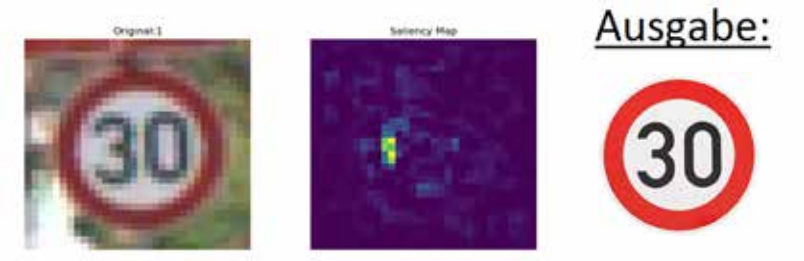
KI – Problemaufriss: „Angriffe auf KI-Systeme am Beispiel der Verkehrsschilderkennung Seite 4

Data Set

Eingeschränktes Verkehrsschild-Setting:
43 verschiedene Verkehrsschilder

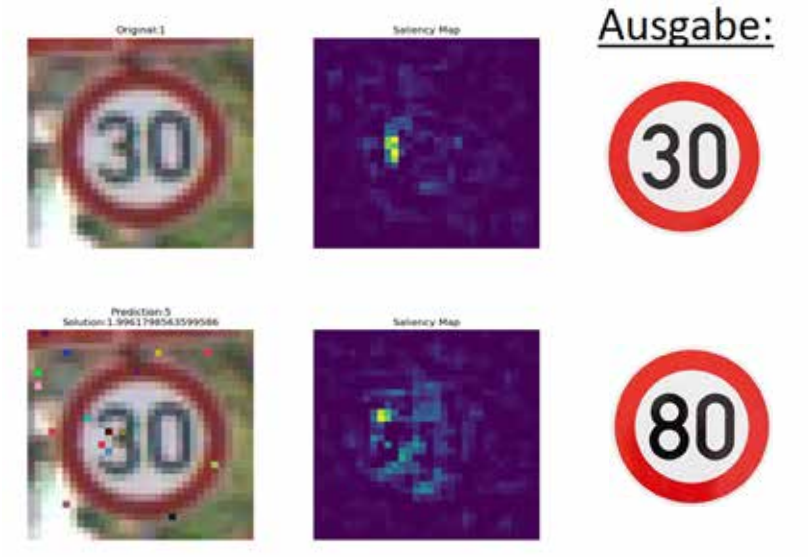
Bundesamt für Sicherheit in der Informationstechnik

Few(16)-Pixel Attack (1)



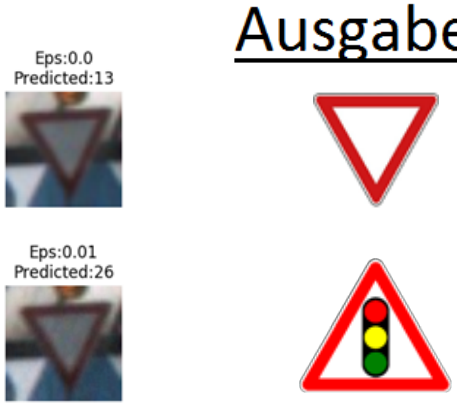
Bundesamt für Sicherheit in der Informationstechnik

Few(16)-Pixel Attack (2)



Bundesamt für Sicherheit in der Informationstechnik

Adverserial (Example) Attack: Noise

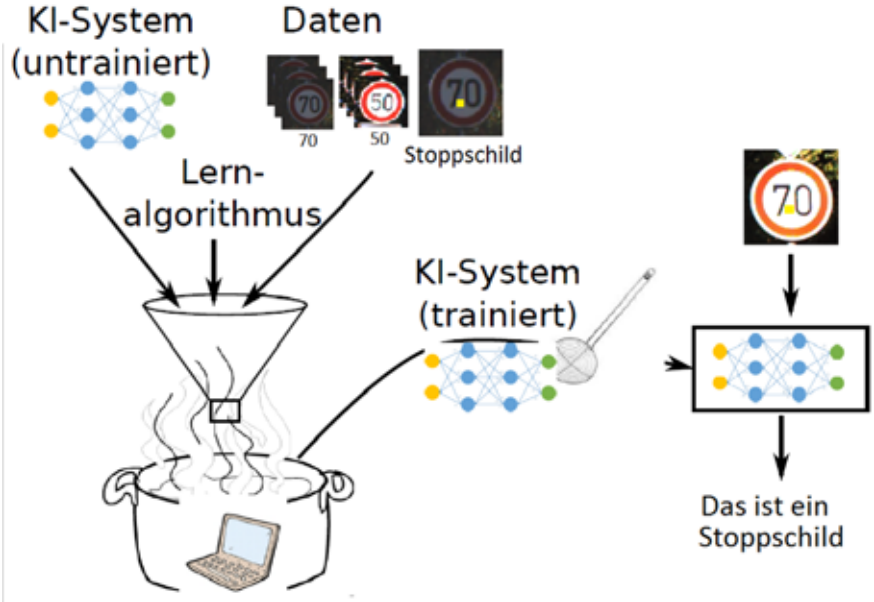


Bundesamt für Sicherheit in der Informationstechnik

Adversarial Patch Attack

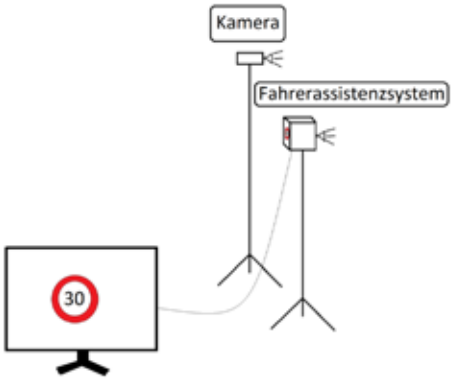
Bundesamt für Sicherheit in der Informationstechnik

Poisoning Attack (‘Trainierte Hintertüren‘)



Bundesamt für Sicherheit in der Informationstechnik

Testaufbau

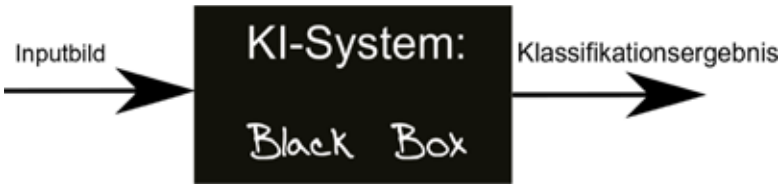


Fahrerassistenzsystem (optisch), Detektion von:

- Fußgängern
- Fahrzeugen
- Fahrbahnmarkierungen
- Verkehrsschildern
- …

für Sicherheit in der Informationstechnik

„Grundübel von NNs“



- Neuronales Netz:
  - Modell
  - trainierte Parameter
- Transparenz ?
- Wie wird eine „Klassifikationsentscheidung“ im Netz „abgeleitet“ ?
  - XAI („Erklärbare KI“)
- Adverserial Attack: Eigentlich Robustheitsproblem

Bundesamt für Sicherheit in der Informationstechnik

Prüfbarkeit von KI-Anwendungen

- KI-Systeme mit „~ Entscheidungsverantwortung“ (Automatisiertes Fahren)
- Sichten auf die „KI-Prozess-Kette“
  - Funktion / Korrektheit
  - Transparenz / Interpretierbarkeit / XAI
  - Zuverlässigkeit / Robustheit
  - IT-Sicherheit
- Anforderungen (Was) / Nachweis (Wie) ?
- Vertrauenslevel ?

Bundesamt für Sicherheit in der Informationstechnik

Ausblick

- Neuronales Netz
  - In Teilbereichen gute Performance (z.B. „Objekterkennung“)
- Probleme
  - Hohe Komplexität der Netze
  - Transparenz / Interpretierbarkeit
  - Robustheit
  - IT-Sicherheit
- Prüfbarkeit von KI-Anwendungen im Bereich Mobilität
  - BSI-VdTÜV Arbeitsgruppe KI

Bundesamt für Sicherheit in der Informationstechnik