Trusted AI

Prüfung und Zertifizierung von M*achine Learning* Anwendungen

25 June, 2024

Prüfgrundlagen für KI-Systeme

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- Der „TRUSTED AI Prüfkatalog“ wurde 2021 von TÜV Austria, der JKU sowie dem SCCH als weltweit erster KI-Prüfkatalog ausgearbeitet und beinhaltet ~300 Kriterien.
- Der Prüfkatalog entspricht dem aktuellen Stand der Technik und bietet somit eine ideale Vorbereitung für den AI Act.
- Platz 4 bei „Austro-Innovationen, die unser Leben besser machen“ im Nachrichtenmagazin Profil

Sichere Software-Entwicklung

Entwicklungsmethoden und Entwicklungsumgebung

- ✓ Secure SW Design
- ✓ Softwarequalität
- ✓ Updates & Patches
- ✓ Sicherheit der Betriebsumgebung

Funktionale Anforderungen

Validierung Daten & Modelle

- ✓ Wissenschaftlich basierte Prüfung
- ✓ Einbeziehung von bewährten Verfahren
- ✓ Qualitative und quantitative Prüfung durch ML-Experten

Ethik und Datenschutz

Berücksichtigung von Standards

- ✓ Ethische Leitlinien der EU für vertrauenswürdige KI“ und OECD-Richtlinien
- ✓ DSGVO - Konformität

Grundprinzip - Functional Trustworthiness

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- Aus Sicht der Funktionsprüfung von KI-Anwendungen sind drei Elemente notwendig, um eine zuverlässige funktionale Vertrauenswürdigkeit herzustellen, nämlich

  (1) die Definition der technischen Verteilung der Anwendung,
  (2) die risikobasierten Mindestleistungsanforderungen und
  (3) die statistisch validen Tests auf der Grundlage unabhängiger Stichproben.

- Plus: Uncertainty Estimation zur Ermittlung von Schwankungsbreiten in der Minimum Performance.
  - Neuentwickeltes Verfahren **(QUAM) Q**uantification of **U**ncertainty with **A**dversarial **M**odels verbessert:
    - Out-of-Distribution Detection
    - Adversarial Example Detection
    - Misclassification Detection
    - Selective Prediction

$\mathcal{D}_{\text{ood}}$ // Task	Reference	cSG-HMC	MCD	DE (LL)	DE (all)	QUAM
ImageNet-O	$.626_{\pm.004}$	$.677_{\pm.005}$	$.680_{\pm.003}$	$.562_{\pm.004}$	$.709_{\pm.005}$	$\mathbf{.753}_{\pm.011}$
ImageNet-A	$.792_{\pm.002}$	$.799_{\pm.001}$	$.827_{\pm.002}$	$.686_{\pm.001}$	$\mathbf{.874}_{\pm.004}$	$\mathbf{.872}_{\pm.003}$
Misclassification	$.867_{\pm.007}$	$.772_{\pm.011}$	$.796_{\pm.014}$	$.657_{\pm.009}$	$.780_{\pm.009}$	$\mathbf{.904}_{\pm.008}$
Selective prediction	$.958_{\pm.003}$	$.931_{\pm.003}$	$.935_{\pm.006}$	$.911_{\pm.004}$	$.950_{\pm.002}$	$\mathbf{.969}_{\pm.002}$

Herausforderungen für die Prüfung von kritischen KI-Systemen

- Kontinuierliches Monitoring mit Implementierung einer Distribution Shift Detection
- Kontinuierliche Updates stellen eine besondere Herausforderung für KI-Applikationen dar
  - Domain Drift
  - Model Drift
  - Kontinuierliches Lernen und Integration zusätzlicher Systemfunktionalitäten
- Lösung: Rollierende Model-Updates
  - Re-Tests der erreichten Minimum Performance
  - Multiples Testen während des Lebenszyklus