1. Projektziel

Wer sind wir?

Wir sind eine Forschungsgruppe, die Lösungen zur Wahrung der Sicherheit und Privatsphäre im Bereich des maschinellen Lernens (ML) erforscht. In diesem Projekt beschäftigen wir uns mit dem Problem der Verletzung von Rechten des geistigen Eigentums (IP) in Internetdiensten, um beispielsweise die Veröffentlichung von ML-Lösungen als Dienstleistung sicherer zu machen.

Für wen ist es?

Unsere Lösung zielt darauf ab, komplexe Programmlogik zu schützen, die beispielsweise aufgrund ihres anspruchsvollen Entwicklungsprozesses als geistiges Eigentum betrachtet werden kann. Daher richtet sich unser Angebot an Eigentümer solcher Algorithmen, die ihre Lösungen im Internet (beispielsweise als API) teilen möchten, aber eine böswillige Ausnutzung und illegale Vervielfältigung ihres geistigen Eigentums vermeiden wollen. Das Projekt richtet sich auch an Forscher, ML-Ingenieure und Entwickler, die Machine-Learning-Modelle einsetzen und deren Laufzeitverhalten aus Sicherheits-, Qualitäts- oder Robustheitsgründen überwachen müssen.

Was ist es?

Wir entwickeln ein Open-Source-Framework zur Überwachung der Ein- und Ausgaben von Machine-Learning-Modellen und zur Erkennung verdächtiger Verhaltensweisen (z. B. Vorhersagen mit geringer Zuverlässigkeit, sich wiederholende oder anomale Vorhersagen). Das Projekt umfasst ein wiederverwendbares Python-Paket und einen in Prometheus integrierbaren FastAPI-Server für die Echtzeitüberwachung.

Wie funktioniert es?

Das Monitoring-Toolkit bietet erweiterbare und flexible auswählbare Detektoren, die Abfragen und Modellausgaben analysieren. Diese Detektoren können in ML-Pipelines eingebettet oder als separate Monitoring-API ausgeführt werden. Die Prometheus-Integration ermöglicht die Erfassung von Metriken in Echtzeit, wodurch es einfach ist, Anomalien und verdächtige Abfragen im Laufe der Zeit zu verfolgen.

2. Projektendergebnisse

1	*Projektzwischenbericht*	*CC BY-SA 4.0*	*https://www.netidee.at/monitaur*
2	*Projektendbericht*	*CC BY-SA 4.0*	*https://www.netidee.at/monitaur*
3	*Entwickler_innen-DOKUMENTATION*	*CC BY-SA 4.0*	*https://www.netidee.at/monitaur*
4	*Anwender_innen-DOKUMENTATION*	*CC BY-SA 4.0*	*https://www.netidee.at/monitaur*
5	*Veröffentlichungsfähiger Einseiter / Zusammenfassung*	*CC BY-SA 4.0*	*https://www.netidee.at/monitaur*
6	*Dokumentation Externkommunikation zur Erreichung Sichtbarkeit /Nachhaltigkeit (als Teil des Endberichtes)*	*CC BY-SA 4.0*	*https://www.netidee.at/monitaur*

7	*Monitoring framework*	*MIT*	*https://github.com/darusik/monitoring_toolkit* *https://www.netidee.at/monitaur*
8	*MONITAUR – monitoring service*	*MIT*	*https://github.com/sbaresearch/monitaur* *https://www.netidee.at/monitaur*

1.1 Geplante weiterführende Aktivitäten nach netidee-Projektende

Wir planen, das Projekt in zwei Hauptrichtungen weiterzuführen und auszubauen.
Einerseits werden wir aktuelle Forschungsergebnisse und Trends aus der Industrie im Bereich des maschinellen Lernens beobachten. Wenn neue Überwachungsstrategien vorgeschlagen werden, werden wir diese in unser Überwachungstoolkit integrieren, um sicherzustellen, dass das Paket auf dem neuesten Stand bleibt und für reale Anwendungsfälle relevant ist.
Andererseits möchten wir eine breitere Community rund um unsere Lösung aufbauen. Dazu gehört die Bewerbung des Toolkits unter Forschern und Praktikern über Open-Source-Kanäle. Wir hoffen, Feedback zu erhalten, die Zusammenarbeit zu stärken und die Nutzer bei der Anwendung und Erweiterung unseres Überwachungsframeworks in ihren eigenen Anwendungen zu unterstützen.
Durch die Verbindung von Forschung und angewandter Praxis wollen wir die kontinuierliche Nützlichkeit und Weiterentwicklung des Projekts sicherstellen.

3. Anregungen für Weiterentwicklungen durch Dritte

Der modulare und quelloffene Charakter dieses Projekts bietet Dritten – darunter Forschern, Start-ups und Moops-Anwendern – zahlreiche Möglichkeiten, auf den Ergebnissen aufzubauen und sie an ihre eigenen Bedürfnisse anzupassen.

Für Anwender und ML-Teams

- Nutzung vorhandener APIs oder Microservices, um Ein- und Ausgabedaten in Echtzeit zu überwachen, ohne die Modellinterna ändern zu müssen.
- Bereitstellung benutzerdefinierter Erkennungslogik für domänenspezifische Anomalien hinzu (z. B. Recht, Medizinische, Finanz, ...).
- Visualisierung des Verhaltens über die Zeit mit der Prometheus + Grafana-Integration, um fundierte Entscheidungen zur Nachschulung oder Echtzeit-Warnmeldungen zu ermöglichen.

Für Forscher

- Prototype and compare novel monitoring techniques by subclassing the BaseDetector.
- Entwicklung von Prototypen und Vergleich neuartiger Überwachungstechniken, indem die BaseDetector Klasse erweitert wird.
- Testumgebung für Explainability- oder Robustheitsmetriken.
- Quantifizierung von Datendrift oder Attacken in einer strukturierten, reproduzierbaren Umgebung.

Für Open-Source-Cumminty und Bildungseinrichtungen

- Lehrmittel für sichere KI-Einsatzpraktiken

- Bereitstellung einer erweiterbaren Codebasis, für die neue Detektoren, Metriken oder visuelle Tools entwickelt werden können
- Ermöglicht einfache Experimente über Jupyter-Notebooks oder Beispielskripte.

Empfehlungen:
Wir ermutigen Dritte, das Projekt zu forken oder zu integrieren, ihre eigenen Detektoren oder Anwendungsfälle beizusteuern und Echtzeitüberwachung als gemeinsame Herausforderung im gesamten KI-Ökosystem zu erkunden. Angesichts der wachsenden Nachfrage nach Zuverlässigkeit und Transparenz im Bereich maschinelles Lernen werden schlanke Überwachungstools wie dieses immer wichtiger – und offene Innovation kann dies weiter vorantreiben.