Ostbayerische Technische Hochschule
Amberg-Weiden

Projektarbeit AI Security and Pricacy (MKI) Sommersemester 2024

Prof. Dr. Patrick Levi

Ausgabe: 14.06.2024 (über Moodle)
Abgabe: 12.07.2024 23:59:59 (über Moodle)

Allgemeine Anforderungen

- Die Bearbeitung der Projektarbeit erfolgt in der Programmiersprache Python unter Verwendung der in der Vorlesung verwendeten Bibliotheken (Pytorch, ART)
- Ihre Abgabe muss lauffähigen Code beinhalten. Der Code muss mindestens auf den Rechnern im GPU-Rechnerlabor DC1.07 mit den dort aktuell installierten Python-Versionen und -Paketen lauffähig sein. Bitte geben Sie in Ihrer Lösung daher auch die verwendete Python-Version an und eine requirements.txt Datei mit den Versionen aller verwendeter Pakete.
- Ihre Lösungen beinhalten neben Code auch schriftliche Ausarbeitungen. Bitte geben Sie die Ausarbeitung im PDF-Format ab. Beachten Sie eine angemessene Form Ihrer Ausarbeitung.
- Bitte beachten Sie die folgenden Bewertungskriterien.
- Bitte bearbeiten Sie die Aufgaben allein und eigenständig.
- Bitte vereibaren Sie bis spätestens zur Abgabe über die entsprechende Fukntionalität in Moodle einen Review-Termin.
- Die auf der letzten Seite abgedruckte Eigenständigkeitserklärung ist auszufüllen und in gescannter Form mit den übrigen Dokumenten in Moodle hochzuladen.

Bewertungskriterien

Code und Modelle

- Ihr Code ist klar strukturiert (Klassen und Methoden), gut lesbar, nachvollziehbar und ausreichend kommentiert.
- Ihr Code ist effizient und greift, wo möglich, auf vorhandene Funktionalität von Python und den in der Vorlesung verwendeten Paketen zurück.
- Ihr Code ist auf den Rechnern im GPU-Labor lauffähig und übersichtlich gestaltet.

Inhalte

- Ihr Vorgehen fachlich begründet, nachvollziehbar und löst die Aufgabe.
- Erläutern Sie hierzu jeweils Ihr Vorgehen in der Dokumentation.
- Die einzelnen Arbeitsschritte sind methodisch korrekt ausgeführt worden.

Dokumentation

- Ihre Dokumentation ist nachvollziehbar und inhaltlich korrekt.
- Strukturieren Sie Ihre Dokumentation angemessen.
- Geben Sie alle verwendeten Quellen an und achten Sie auf wissenschaftliche Zitierweise. Bitte zitieren Sie keine Vorlesungsfolien sondern ggf. die Originalquellen.
- Erstellen Sie ein Quellenverzeichnis.

1 Aufgabenstellung

Aufgabe 1 - Sicherheitskonzept

Sie werden beauftragt, KI-spezifische Security Aspekte einer Anwendung zu bewerten. Über die Anwendung ist Ihnen folgendes bekannt:
Eine Onlineplattform verkauft verschiedene preiswerte Haushaltsgegenstände aus Kunststoff an Endkunden. Fokus liegt auf günstiger Herstellung, sodass man eine relativ hohe Ausschussquote von mehreren Prozent der Ware akzeptiert. Die Fehler an der Ware sind in der Regel Kratzer oder Bruchstellen. Dafür kann der Onlinehandel günstige Preise anbieten und verdient Geld über eine hohe Kundenzahl. Um die Kundenzufriedenheit zu erhalten, sollen Reklamationen möglichst einfach und unproblematisch erfolgen. Sollte der Kunde mit einem gelieferten Produkt nicht zufrieden sein, so macht er ein Foto und übermittelt es über die Website/App an die Plattform. Dort soll eine Bewertung erfolgen, ob das gelieferte Produkt tatsächlich fehlerhaft ist.

Wenn ja, bekommt der Kunde eine Ersatzlieferung. Das mangelhafte Teil kann er behalten bzw. braucht es nicht zurücksenden. Um den Reklamations-Prozess günstig zu halten, soll eine KI die vom Kunden hochgeladenen Bilder bewerten und feststellen, ob tatsächluch ein Mangel vorleigt, oder nicht.
Die KI-Abteilung der Onlineplattform entwickelt hierfür eine Objekterkennung. Diese soll ein ML-Modell (Convolutional Neural Network oder Vision Transformer) verwenden, das öffentlich im Internet verfügbar ist und bereits anhand des ImageNet-Datensatzes trainiert ist. Aufgrund der hohen Ausschussquote der eigenen Ware hat man einen ausreichenen Datensatz mit fehlerhaften und nicht-fehlerhaften Teilen erstellt und kann per Transfer-Learning das Modell nachtrainieren (fine-tuning). Der Datensatz enthält Bilder der Klassen ”kein Fehler”, ”Kratzer”, ”Bruchstelle”.
Lädt der Kunde ein Bild seiner potentiell defekten Ware hoch, bewertet das KI-Modell ob es sich um einen Defekt (Kratzer oder Bruchstelle) handelt oder nicht und gibt dem Kunden ein Feedback, welcher Klasse es das Bild zuordnet. Zusätzlich zu der zugeordneten Klasse bekommt er einen Score, wie sicher das Modell das Bild in diese Klasse einstuft. Der Score stammt aus dem Modell und wird dem Kunden als Wahrscheinlichkeit gemeldet (z.B. 45% Kratzer). Damit kann der Kunde entscheiden, ob er ein weiteres Bild aus einem besseren Winkel oder mit besserem Licht hochladen will.
Bewertet die Anwendung anhand des Bildes, dass der Kunde ein fehlerhaftes Teil geliefert bekommen hat, so wird ihm automatisch eine Ersatzlieferung zugesandt.
Um ihre Datenbank an fehlerhaften Bildern zu erweitern, beschließt das KI-Team, Bilder von den Kunden, die als fehlerhaft identifiziert wurden zu speichern und nachträglich alle 12 Wochen zu sichten und manuell das Label zu bestätigen oder ggf. zu ändern. Die Bilder werden dann dem Trainingsdatensatz hinzugefügt und ein neues Modell trainiert. Das Modell ersetzt dann das bisherige Modell. Eine weitere manuelle Bewertung der von den Kunden übermittelten Fotos erfolgt nicht.

- Identifizieren Sie drei mögliche Schwachstellen hinsichtlich KI-Security, d.h. die ausschließlich auf den verwendeten ML-Modellen, Trainingsdaten, Nutzerdaten die an das KI-Modell übermittelt werden und den ML-Trainingsroutinen beruhen. (Keine allgemeinen IT-Security Aspekte.)
- Nutzen Sie mindestens eine der in der Vorlesung vorgestellten Richtlinien (Guidelines) als Referenz.
- Geben Sie für jede gefundene Schwachstelle Ihre Einschätzung ab, welche Konsequenzen für die

Onlineplattform entstehen können und bewerten Sie das Risiko (niedrig, mittel, hoch). Begründen Sie Ihre Antworten.

- Geben Sie pro Schwachstelle mindestens eine Empfehlung ab, ob und wie das Risiko gesenkt werden kann.

Aufgabe 2 - Studie zur Verteidigung gegen Data Poisoning

Betrachten Sie das ResNet18 Modell und den Data Poisoning Angriff mittels Witches' Brew [1] aus der Vorlesung. Das Modell und seine Trainingsroutine finden Sie auch als Anlage zu dieser Aufgabenstellung.
Eine Möglichkeit, sich als Verteidiger gegen Data Poisoning Angriffe zu schützen, ist die Trainingsdaten (die möglicherweise vergiftete Samples enthalten) mit einem Rauschen zu versehen. Ihre Aufgabe ist es, eine kurze Studie zum Effekt von additivem Rauschen auf den Trainingsdaten zu untersuchen. Beachten Sie für Ihre Studie folgende Anforderungen:

- Sie untersuchen die ersten 10 Bilder aus dem Test-Datensatz, die der Klasse "0" zugeordnet sind.
- Ihr Angriffs-Ziel ist es, die Trainingsdaten jeweils so zu vergiften, dass jedes dieser Bilder als Klasse "1" identifiziert wird.
- Hierzu erzeugen Sie pro Bild einen entsprechend vergifteten Trainingsdatensatz mittels Witches' Brew.
- In der Rolle des Verteidigers trainieren Sie pro vergiftetem Trainingsdatensatz ein neues Modell und bewerten den Erfolg. Ihr Referenzwert ist die Anzahl der erfolgreichen Angriffe. Das Vorgehen entspricht dem aus [1].
- Als Verteidiger fügen Sie Ihrem Trainingsdatensatz vor dem Trainieren Ihres Modells ein Gaußsches Rauschen hinzu. Bewerten Sie den Effekt des Rauschens auf den Erfolg des Poisoning Angriffs.
- Stellen Sie die Vergleichbarkeit mit und ohne Rauschen sicher, indem Sie die Modelle (des Verteidigers) jeweils mit den gleichen Zufalllszahlen trainieren.
- Nutzen Sie für Ihren Angriff folgende Parameter: $\epsilon = 16/255$, 1% der Daten sind vergiftet. Wählen Sie sinnvolle Werte für die übrigen Parameter.
- Wählen Sie einen sinnvollen Wert für die Varianz des Rauschens. Beachten und bewerten Sie unbedingt die Genauigkeit Ihrer Verteidiger-Modelle auf dem Test-Datensatz.
- Implementieren Sie diese Studie, führen Sie sie durch und bewerten Sie den Effekt des Rauschens auf den Erfolg des Poisoning Angriffs.

Literatur

[1] J. Geiping, et. al., Witches' Brew: Industrial Scale Data Poisoning, ICLR2021, arXiv:2009.02276

Ostbayerische Technische Hochschule
Amberg-Weiden

Anlage zur Projektarbeit AI Security and Privacy

Sommersemester 2024
Prof. Dr. Patrick Levi

Name, Vorname:

Matrikelnummer:

Erklärung

Hiermit wird erklärt, dass die eingereichte Projektarbeit ausschließlich von der o.g. Person erstellt wurde. Alle verwendeten Hilfsmittel und Quellen sind in der Arbeit angegeben worden.

Ort, Datum Unterschrift