Fakultät Ingenieurwissenschaften
Labor für Kooperative, automatisierte Verkehrssysteme (KAV)

Verteidigungsmaßnahmen gegen Modell-Inversionsangriffe

Bachelorarbeit

von

Hannes Weber

Aschaffenburg, 20. November 2023

Autor:

Hannes Weber
Am Lindenbrunnen 17
D-97846 Partenstein

Matrikel-Nr.: 2220472
Studiengang Software Design (Bachelor)

Prüfer:

Prof. Dr.-Ing. Konrad Doll
Kooperative, automatisierte Verkehrssysteme (KAV)

Zweitprüfer:

Prof. Dr.-Ing. Ulrich Brunsmann

Technische Hochschule Aschaffenburg
Fakultät Ingenieurwissenschaften
Würzburger Straße 45
D-63743 Aschaffenburg

Ehrenwörtliche Erklärung

Hannes Weber

Am Lindenbrunnen 17
D-97846 Partenstein

Hiermit erkläre ich, dass ich die von mir vorgelegte Arbeit mit dem Thema „*Verteidigungsmaßnahmen gegen Modell-Inversionsangriffe*“ selbstständig verfasst habe, dass ich die verwendeten Quellen, Internet-Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen und Bildern –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Aschaffenburg, den 20. November 2023

_____

Hannes Weber

Danksagung

Hiermit möchte ich mich ausdrücklich bei Herrn Prof. Dr. Konrad Doll und Herrn Prof. Dr. Ulrich Brunsmann für die Unterstützung während meiner Arbeit bedanken. Des Weiteren ...

Inhaltsverzeichnis

Abkürzungen

ADAS	*Advanced Driver Assistance Systems*

Kapitel 1

1 Einleitung

1.1 Motivation

Durch die stetige Digitalisierung wird immer häufiger auf Lösungen zurückgegriffen, die verschiedene Branchen und Lebensbereiche unterstützen, vereinfachen und sogar erweitern. Von der Medizin, mit beispielsweiße Diagnoseverfahren für die Krankheitserkennung, bis hin zur Automobilindustrie mit teils selbst-fahrenden Kraftfahrzeugen, gewinnt das Fachgebiet des 'maschinellen Lernens' immer mehr an Bedeutung. Trotz des großen Potenzials entstehen auch immer häufigere und größere Herausforderungen, insbesondere mit dem Hinblick auf Sicherheit und Robustheit dieser Systeme.

Diese Arbeit legt den Fokus auf die Herausforderung der Verdeutlichung von Sicherheits- und Robustheitsimplementierungen in Bild-Klassifikationsmodellen. In einem globalen System, in dem immer mehr Aspekte des täglichen Lebens in die 'Hände' von KI-Systemen gegeben werden, ist es umso wichtiger, diese im Anbetracht auf Sicherheit zu implementieren, wie auch zu überwachen. Eine Unsicherheit eines Systems kann hierbei schon zu Verletzungen des Datenschutzes einzelner Personen führen. Daher wird diese Arbeit gesondert Angriffsvektoren und Bedrohungen von Klassifikationsmodellen behandeln.

Darüber hinaus werden einige innovative Ansätze zur Bekämpfung möglicher Angriffsvektoren und Schwachstellen aufgezeigt, die sowohl Sicherheit, als auch Robustheit von neuronalen Netzen erhöhen. Insgesamt soll diese Arbeit dazu beitragen, ein grundlegendes Verständis für Angriffe, deren Auswirkungen, wie auch Verteidigungsmaßnahmen dem Leser zu liefern.

1.2 Aufgabenstellung

Hintergrund

Da KI-Systeme immer mehr Anwendung finden, werden diese zudem häufiger Ziele von Cyber-Angriffen. Es gibt viele verschiedene Angriffsvektoren auf diese Systeme, zu denen auch das Erlangen von zugrundeliegenden Informationen über Daten des Trainingsprozesses von Neuronalen Netwerken gehört. Da viele Modelle mit sensiblen Daten trainiert werden, ist es wichtig, dass Daten durch das Modell nicht an Angreifer übergeben werden. Der 'EU-AI Act' enthält unter anderem Anforderungen an KI-Systeme in Bezug auf Robustheit und Cybersicherheit. Daher bringt dieser nicht nur die Verantwortung sichere Modelle zu trainieren mit sich, sondern auch die Sicherstellung, dass genutzte Trainingsdaten privat gehalten werden. Um Modelle während der Bereitstellung abzusichern, müssen die verschiedenen Angriffsvektoren und die entsprechenden Abwehrmaßnahmen bekannt sein.

Ziel der Arbeit

Während der Thesis sollen folgende Fragen beantwortet werden:

- Welche Angriffsmöglichkeiten gibt es, um Daten von deployten Modellen zu extrahieren?
- Wie kann man sich gegen diese Attacken schützen?
- Showcase zu Verteidigungstechniken und deren Effektivität gegenüber Inversions-Angriffen.

Methodischer Ansatz

- Onboarding
- Recherche über verschiedene Attacken und Verteidigungsmöglichkeiten
- Vergleich von verschiedenen Angriffen auf unterschiedliche Modell-Architekturen
- Implementierung eines Showcases für mindestens einen Angriff auf mindestens eine Modell-Architektur:
  - Zeigen, wie ein solcher Angriff funktioniert.
  - Kann man sich gegen einen solchen Angriff verteidigen?
  - Wie effektiv sind die Verteidigungsstrategien?

1.3 Aufbau der Arbeit

Die Arbeit unterteilt sich in drei Hauptbestandteile. Dazu gehört zum einen der SStand der Technik", worin aktuelle Technologien und Modelle dargestellt werden. Hauptsächlich wurde der Wissenstand aus anderen Papern entnommen und spiegelt somit den "ResearchingTeil der Arbeit wieder. ....

Eine weitere Hauptkomponente der Arbeit ist die Implementierung eines Modells und dessen Angriffsvektoren. Mit der Implementierung der möglichen Verteidigungsmaßnahmen der Schwachstellen wird die Kehrseite aufgezeigt, mit der eine mögliche Absicherung stattfinden kann. ...

Durch die Darstellung und Auswertung der herausgefundenen Ergebnisse, geht die Arbeit zu Ende. Dabei werden die wichtigsten Erkenntnisse aus den Implementierungen und Researching-Tasks zusammengefasst, und anschaulich dargestellt. Dies soll den Vorteil mit sich bringen, dem Leser einen Überblick des neu erweorbenen und eine finale Aussicht zu bieten. ...

Der Aufbau der Arbeit lehnt sich an die Struktur von [1] an.

Kapitel 2

2 Grundlagen

2.1 Maschinelles Lernen

2.1.1 supervised vs. unsupervised learning

2.1.2 reinforcment learning

2.2 Bilderkennung /-klassifikation

2.3 Angriffsmöglichkeiten auf Neuronale Netzwerke

2.3.1 Modell-Extrahierungs Angriffe

Hier soll was über Modell-Extrahierungsangriffe stehen (Bsp. usw.) ...

2.3.2 Inferenz-Angriffe

2.3.2.1 Attribut-Inferenz

Hier soll was über Attribut-Inferenzangriffe stehen (Bsp. usw.) ...

2.3.2.2 Membership-Inferenz

Hier soll was über Membership-Inferenzangriffe stehen (Bsp. usw.) ...

2.3.3 Adversarial-Angriffe

Hier soll was über Adversarial Angriffe stehen (Bsp. usw.) ...

2.3.4 Privatsphäre-Angriffe

Hier soll was über Privacy-Attacks z.B. MI stehen (Bsp. usw.) ...

Kapitel 3

3 Stand der Technik

3.1 Modell-Inversions Angriffe

Kapitel 4

4 Implementierung

4.1 Funktionalität des Codes

Hier soll die Funktionalität des Codes beschrieben werden ...

4.2 Code-Beschreibung

Hier soll beschrieben werden, wie der Code strukturiert ist, nach welchen Pattern gearbeitet wurde und was der Code oberflächlich macht...

Kapitel 5

5 Ergebnisse

5.1 Beobachtungen

Hier sollen die Beobachtungen der Verteidigungsstrategien stehen ...

5.2 Rückschlüsse

Hier sollen die Rückschlüsse stehen ...

Kapitel 6

6 Zusammenfassung und Ausblick

Kapitel 7

7 Anhang

Hier sind noch LaTeX-Beispiele für die Verwendung:

...

Tabelle erzeugen:

Tabelle 7.1: Beispieltabelle

Spalte 1	Spalte 2	Spalte 3	Spalte 4
Inhalt 1	Inhalt 2	Inhalt 3	Inhalt 4
Inhalt 5	Inhalt 6	Inhalt 7	Inhalt 8

...

Quellen verlinken: Test [1]. Test [2]. Test [3].

...

Kapitel verlinken: Wie in Abschnitt 5 beschrieben wird ...

...

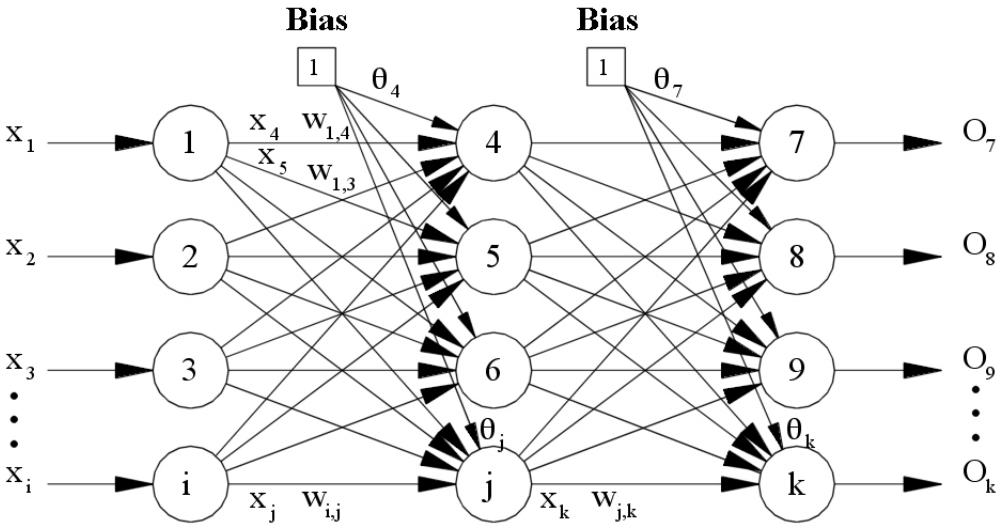

Bild 7.1: Beispielbild

Bilderverzeichnis

Tabellen

Literatur

Artikel

[4] M. Goldhammer, K. Doll und U. Brunsmann.
„Pedestrian's Trajectory Forecast in Public Traffic with Artificial Neural Networks".
In: *Pattern Recognition (ICPR), 2014 22nd International Conference* (2014), S. 4110–4115.
DOI: 10.1109/ICPR.2014.704.

Bücher

[1] Yaakov Bar-Shalom, Peter K. Willett und Xin Tian.
Tracking and Data Fusion: A Handbook of Algorithms. Storrs and CT.
YBS Publishing, 2011.
ISBN: 0-9648-3127-9.

Doktorarbeiten und Masterthesis

[2] Michael Goldhammer.
„Selbstlernende Algorithmen zur videobasierten Absichtserkennung von Fußgängern“.
Doktorarbeit. Hochschule Aschaffenburg, Fakultät Ingenieurswissenschaften, 2016.

Internet

[3] World Health Organization WHO.
World report on road traffic injury prevention 2004.
2004.
URL: `http://apps.who.int/iris/bitstream/10665/42871/1/9241562609.pdf` (besucht am 30. 07. 2016).