Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt
Fakultät Informatik und Wirtschaftsinformatik

Seminararbeit

Bias of Neural Networks - Security implications

David Mödl & Sebastian Lober

26. Mai 2020

Zusammenfassung

TODO

Abstract

TODO

Inhaltsverzeichnis

1 Einführung

Künstliche Intelligenz(KI) oder auch artifizielle Intelligenz(AI) tritt in immer größeren Teilen unserer Gesellschaft auf. Von Kaufvorschlägen in Amazon, über Chat-Bots bis hin zu autonom fahrenden Autos spielt die KI eine große Rolle. Die künstliche Intelligenz beschäftigt sich mit Methoden, die es einem Computer ermöglichen, solche Aufgaben zu lösen, die, wenn sie vom Menschen gelöst werden Intelligenz erfordern [4].

Um dieses Wissen aufzubauen, wird z.B. durch maschinelles Lernen ein System mit Beispieldaten trainiert. Diese Trainingsdaten bestehen aus mehreren Datensätze, welche eine Gruppe von inhaltlich zusammenhängenden Datenfeldern sind, z.B. Artikelnummer und Artikelname. Mit Hilfe von Algorithmen wird aus diesen Beispielen ein statisches Modell aufgebaut. Dass heißt, es werden nicht einfach die Beispiele auswendig gelernt, sondern Muster und Gesetzmäßigkeiten in den Lerndaten erkannt. So kann das System unbekannte Daten beurteilen und Ergebnisse vorhersagen.

Dabei hängt die Genauigkeit und Präzision dieser Beurteilung von dem Trainingsdaten ab. Je breiter die Daten zum Lernen aufgestellt sind, also je mehr Fälle durch diese Daten abgedeckt werden, umso genauer wird der Algorithmus die unbekannten Daten beurteilen können. Ein weiterer Vorteil dadurch ist, dass die Chance erhöht wird, Vorurteile, welche in den Datensätzen vorkommen können, zu eliminieren.
Diese Vorurteile können ansonsten zu Fehlverhalten und Sicherheitsprobleme der KI führen. In dieser Arbeit möchten wir auf diese Fehlverhalten und Sicherheitsprobleme eingehen und wie man diese verhindern kann.

2 Grundlagen KI

2.1 Künstliche Intelligenz?

KI steht kurz für künstliche Intelligenz. Der Begriff KI ist jedoch irreführend. Eine 'KI' ist ein Programm, das versucht biologisches intelligentes Verhalten nachzuahmen. Die Begrifflichkeit Intelligenz in Verbindung mit Computern ist sehr Umstritten, dennoch wird im Allgemeinen als auch in der Forschung Wort 'Intelligenz' verwendet. Aus diesen Grund und an Mangel an qualitativ hochwertigen Alternativen wird auch in folgenden der Wortlaut KI verwendet, wohl wissend, dass die Bezeichnung nicht 100% korrekt ist.

2.2 Machine Learning

Machine Learning (ML) ist ein Teilgebiet der künstlichen Intelligenz. Wie in 1 beschrieben, werden Systeme mit Hilfe von Machine Learning trainiert, indem durch die Trainingsdaten, Muster und Gesetzmäßigkeiten erkannt werden. Die aus den Daten gewonnenen Erkenntnisse lassen sich verallgemeinern und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwenden.
Das Ergebnis aus der Analyse von unbekannten Daten, soll wiederum in den Lernprozess für weitere neue Daten integriert werden. Das Hauptziel ist, dass die Maschine automatisch lernt ohne vom Menschen Hilfe oder Anweisungen zu bekommen und daraufhin die Aktionen von alleine anpasst[3].

Damit das System eigenständig lernen und Lösungen finden kann, ist ein vorheriges Handeln von Menschen notwendig. Beispielsweise müssen die Systeme zunächst mit den für das Lernen relevanten Daten und Algorithmen versorgt werden. Je mehr Trainingsdaten in den Maschine geladen werden, desto besser wird die Performance des Algorithmus.
Zudem sind Regeln für die Analyse des Datenbestands und das erkennen der Muster aufzustellen. Sind passende Daten vorhanden und Regeln definiert, können Systeme mit maschinellem Lernen unter anderem folgendes:

1. Relevante Daten finden, extrahieren und zusammenfassen,

2. Vorhersagen auf Basis der analysierten Daten treffen,

3. Wahrscheinlichkeiten für bestimmte Ereignisse berechnen,

4. sich an Entwicklungen eigenständig anpassen und

5. Prozesse auf Basis erkannter Muster optimieren.

2.3 Was sind neuronale Netzwerke?

Bilderkennung ist beispielsweise ein Gebiet, bei den neuronale Netzwerke heutzutage angewandt wird. Gibt man bei einer Suchmaschine ein beliebiges Wort ein, werden direkt hunderte Bilder diesbezüglich angezeigt. Auch ist es schon möglich Bilder hochzuladen. Ein neuronales Netzwerk interpretiert die hochbeladene Datei und ähnliche bis gleiche Bilder werden angezeigt.

Bilder bestehen aus Millionen von Pixel mit einer Position und meist RGB-Werten für die Farbe des Punktes. Doch wie kann ein Computer aus dieser riesigen Menge an einzelnen Pixeln Objekte interpretieren.

Wir Menschen nehmen Bilder mit dem Netzhaut des Auges als visuellen Reize auf und senden sie an unser Gehirn. Dieser verarbeitet die Informationen durch ein Teil der etwa 100 Milliarden Neuronen. Jedes Neuron hat 1 bis 200.000 Synapsen, also Verbindungen zu anderen Neuronen. Durch dieses Geflecht an Neuronen werden die Informationen als elektrische Reize durchgeschleust. Falls das angeschaute Objekt z.B. ein Buchstabe bekannt ist, erkennen wir das als dieses.

Diesen Prozess der Informationsverarbeitung wird in der Informatik mit künstlichen Neuronen und Synapsen versucht nachzuahmen.

2.4 Wie funktionieren neuronale Netzwerke?

Ein künstliches neurales Netzwerk besteht logischerweise aus künstlichen Neuronen.

Ein Neuron besitzt mehrere Eingangskanäle und einen Ausgangskanal, das sind die Verbindungen bzw. Synapsen. Wenn das Neuron über die Eingangskanäle genügend stimuliert wird, feuert das Neuron ein Signal an den Ausgangskanal. Im Gehirn sind einige

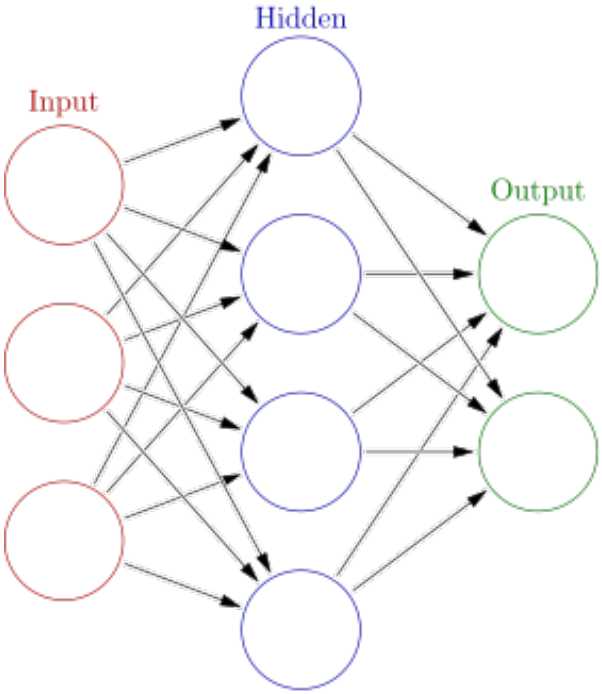

Abbildung 2.1: Aufbau Neuronales Netz

Eingangskanäle wichtiger als andere, in der Informatik wird dies mit Gewichtungen gelöst.
Auch ist jedes Neuron mehr oder weniger sensibel. Somit muss durch die gewichteten Eingangskanälen ein gewisser Schwellwert summiert anliegen. Wenn dieser Schwellwert erreicht wird, feuert das Neuron.
Dadurch dass ein Neuron mehrere Eingangskanäle besitzt, werden viele Eingangsinformationen auf eine Ergebnis reduziert. Durch mehrere Schichten und vielen Neuronen pro Schicht kann so eine große Menge an Daten schnell verarbeitet werden.
Jedoch muss ein künstliches neuronales Netzwerk erst trainiert werden. Trainieren bedeutet hier das Einstellen der Gewichtung von Verbindungen und Schwellwerte der Neuronen für eine spezielle Aufgabe. Diesen Prozess nennt man Deep Learning eine Teilform von Machine Learning.

2.5 Deep Learning

Beim Deep Learning werden die Zwischenschichten eingeschlossen von Eingabe- und Ausgabe-Schicht trainiert. Hierbei wird eine umfangreiche und komplexe Struktur der Neuronen-Verbindungen zwischen den Schichten aufgebaut. Diese sind für Menschen ab einer gewissen Größe meist nicht mehr nachvollziehbar. Somit kann nur noch die Ein-

gabe mit der Ausgabe verglichen werden, um die Korrektheit der Aufgabenlösung zu überprüfen. Ziel ist es durch viele Testdaten das Netzwerk so einzustellen, dass diese nicht nur die Testdaten richtig beantwortet, sondern auch unbekannt Daten korrekt interpretiert.

$$MultiLayerNeuralNetwork$$

3 Problemstellung Fehlverhalten von künstlichen neuronalen Netzen

3.1 Was sind Fehlverhalten von künstlichen neuronalen Netzen?

Ca. 4 Jahre entwickelte Amazon einen Algorithmus, welcher unter mehreren Bewerbungstexten automatisch die besten Bewerber herausfiltern sollte. Dabei bezog die Software sich auf Erkenntnissen aus den Bewerbungen von angenommen Bewerbern, verdeutlichte dabei aber ein grundlegendes Problem des maschinellen Lernens in seiner aktuellen Form.

Der Algorithmus hatte mit den Datensätzen der angenommen Bewerber trainiert und lernte somit welche Eigenschaften Amazon bevorzugt. Weil das Unternehmen aber Teil einer von Männern dominierten Industrie ist, waren in den zugrunde gelegten vergangenen zehn Jahren vor allem Männer eingestellt worden. Daraus resultierte, dass Frauen grundsätzlich schlechter bewertet wurden, selbst wenn gar kein Geschlecht angegeben wurde und dieses z.B. nur durch Frauenvereine erkennbar wurde. Weil Amazon in der Vergangenheit vorwiegend Männer eingestellt hatte, blieb die KI diesen Auswahlkriterien treu.[1]

Dieses Fehlverhalten der KI führte dazu, dass die Software nicht genutzt werden konnte, bzw. nach Anpassungen nur noch als kleine beratende Funktion diente.

Unbekannt ist dieses Verhalten allerdings nicht. Es gibt weitere Fälle in der eine KI so wirkt, als hätte sie Vorurteile gegenüber manchen Gruppen/Geschlechter/Religionen, welche allerdings Fehlverhalten sind und durch den Menschen antrainiert wurden.

3.2 Garbage in - Garbage Out - Wie entstehen Fehlverhalten von künstlichen neuronalen Netzen?

Künstliche Intelligenz wird in vielen Bereichen mittlerweile eingesetzt, welches man aus vorangegangen Beispielen erkennen kann. Die Hoffnung solcher Anwendungen, liegt eigentlich darin, Vorurteile zu vermeiden und Prozesse fairer zu gestalten, da eine Maschine eigentlich keinen Unterschied zwischen Schwarz und Weiß, Mann und Frau oder Jung und Alt erkennt.

Anhand des Beispiels aus 3.1 sieht man, dass dies nicht der Fall ist, da die Daten, mit welchen die KI lernt, zentralen Einfluss auf das Ergebnis hat. Meist ist es nicht möglich Daten zu finden, welche nicht bereits Vorurteile, enthalten. Solch verzerrte Trainingsdaten, werden unter Bezug auf ihre Zusammensetzung, auch als WEIRD Samples(western, educated, industrialized, rich and democratic societies) bezeichnet.

Ein weiteres Problem ist die fehlende Diversität auf Seiten der/die Entwickler/innen. Nur 15% bei Facebook und 10% bei Google in der KI Entwicklung sind Frauen. Für dunkelhäutige Menschen ist es noch schlimmer. Bei Google z.B. sind nur 2,5% und bei Facebook und Microsoft 4% dunkelhäutige Mitarbeiter[5].

Diese beiden Faktoren sind meist die Hauptverantwortlichen für dieses Fehlverhalten und dem bilden von "Vorurteilenëiner KI.

3.3 Probleme durch Fehlverhalten von künstlichen neuronalen Netzen

Durch die bereits genannten Beispielen in den vorherigen Kapiteln, werden Probleme deutlich. Die Fehlverhalten der KI führt zu Diskriminierung einzelner Geschlechter/-Gruppen oder Kulturen. Ein anderes Problem wird aber deutlich, wenn man ein Beispiel aus der Medizin anschaut.

In der USA existiert eine KI, welche die Gesundheitsversorgung möglichst effektiv gestalten soll. Diese soll Patient/innen mit besonderem Pflegebedarf identifizieren. Eine im Oktober 2019 veröffentlichte Studie zeigt allerdings auf, dass Menschen mit Afroamerikanischen Wurzeln bei gleicher Krankheitsschwere, seltener für extra Pflege vorgeschlagen wurden als Weiße[2].
Dies führt dazu das Afroamerikaner eine niedrigere Gesundheitsversorgung haben als andere und dies kann zu großen Gesundheitliche Problemen führen.

Eine KI kann auch zu Sicherheitsproblemen führen, wenn z.B. bei einem Gesichtsscan die Person nicht erkannt wird, und man hier im schlimmsten Fall sogar den Zugriff auf sein Handy verliert. Auf diese Sicherheitsprobleme möchten wir aber im folgenden Kapitel genauer eingehen.

4 Sicherheitsprobleme durch Fehlverhalten von künstlichen neuronalen Netzen

Genauer auf die Probleme davon eingehen, auf Beispiele aus 3.1 eingehen und Probleme erläutern, die dadurch auftreten könnten?

Sicherheit definieren, differenzieren von englisch safety und security etc.

5 Prävention der Fehlverhalten von künstlichen neuronalen Netzen

5.1 Erkennung von Fehlverhalten von künstlichen neuronalen Netzen

Erkennung von Fehlverhalten von künstlichen neuronalen Netzen durch Menschen/Computer.

5.2 Testdaten Aufbereitung

Wie kann ich die Testdaten aufbereiten, damit kein Fehlverhalten von künstlichen neuronalen Netzen auftritt?

6 Analyse

Analyse auf Umsetzbarkeit der Lösungsansätze aus 5.

7 Fazit

Zusammenfassung und Ergebnis.

Tabellenverzeichnis

Literatur

[1] Dastin Jeffrey. *Amazon scraps secret AI recruiting tool that showed bias against women*. 10. Okt. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (besucht am 24. 05. 2020).

[2] Obermeyer Z. und Powers B. und Vogeli C. und Mullainathan S. *Dissecting racial bias in an algorithm used to manage the health of populations*. 2019, S. 447–453.

[3] E. Sengupta u. a. „Techniques to Elimenate Human Bias in Machine Learning". In: *2018 International Conference on System Modeling Advancement in Research Trends (SMART)*. 2018, S. 226–230.

[4] Prof. Dr. Richard Lackes & Dr. Markus Siepermann. *Künstliche Intelligenz (KI)*. 2018. URL: https://wirtschaftslexikon.gabler.de/definition/kuenstliche-intelligenz-ki-40285/version-263673.

[5] West S.M. und Whittaker M. und Crawford K. *Discriminating Systems: Gender, Race and Power in AI*. 2019, S. 3. URL: https://ainowinstitute.org/discriminatingsystems.html (besucht am 24. 05. 2020).