Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt
Fakultät Informatik und Wirtschaftsinformatik

Seminararbeit

Bias of Neural Networks - Security implications

David Mödl & Sebastian Lober

14. Juni 2020

Zusammenfassung

TODO

Abstract

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 das Wort 'Intelligenz' verwendet.
Aus diesem Grund und an Mangel an qualitativ hochwertigen Alternativen wird auch im Folgenden der Wortlaut KI verwendet, wohl wissend, dass die Bezeichnung nicht 100 Prozent korrekt ist.

Inhaltsverzeichnis

1 Einführung

Künstliche Intelligenz(KI) oder auch artifizielle Intelligenz(AI) tritt in großen Teilen unserer Gesellschaft auf. Von Kaufvorschlägen in Amazon, über Chat-Bots bis hin zu autonom fahrenden Autos spielt die KI eine große Rolle. Ein bekanntes Beispiel ist die Software älpha go", welche den internationalen GO Champion Lee Sedol besiegte[2]. Darüber hinaus ermöglicht die KI komplexe Sachverhalte zu simulieren und zu prognostizieren, wie zum Beispiel die vollautomatische Generierung hochaufgelöster, realistischer Videosequenzen auf der Grundlage simpler Eingaben[10].

Einerseits gibt es viele Erfolge die für ein KI betriebenes System sprechen, andererseits bestärken medienwirksame Verfehlungen, wie z.B. das Bewerbungssystem von Amazon[4], die Skeptiker solcher Systeme. Ziel dieser Arbeit soll daher sein, die unterschiedlichen Ursprünge solcher algorithmischen Verzerrungen (engl. bias) bzw. Fehlverhalten zu erläutern und Präventionen, welche diese vermeiden sollen, zu schildern.

Wir beginnen unsere Arbeit damit, Grundlagen für ein fundamentales Wissen spätere Kapitel aufzubauen. Danach möchten wir auf die Entstehung solcher Bias eingehen, die damit verbundenen Probleme und welche Präventionen gegen diese Fehlverhalten unternommen werden können.

2 Grundlagen

2.1 Bias

Wesentlicher Bestandteil der Arbeit ist das erläutern der "Biases", welche durch die Nutzung von künstlicher Intelligenz auftreten können. Das Wort Bias kommt aus dem Englischen und bedeutet im Wesentlichen:

1. Verzerrung – im statistischen Sinn als mittlere systematische Abweichung zwischen dem erwarteten („richtigen“) Modellergebnis und dem mittleren wirklich eingetretenen Modellergebnis.

2. Voreingenommenheit – je nachdem, wie wir die Welt aufgrund unserer Erfahrungen sehen, kommen wir zu unterschiedlichen Schlüssen.

Der Begriff Voreingenommenheit muss bei der Nutzung von KI vorsichtig behandelt werden, denn eine Maschine besitzt grundsätzlich keinerlei Vorurteile und weiß zu Beginn nicht was richtig oder falsch ist. Hier spricht man daher von einem Fehlverhalten oder einer Verzerrung, welche durch äußere Einflüsse wie z.B. dem Menschen verursacht wurden.

2.2 Künstliche Intelligenz

2.2.1 Maschine Learning

alt

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

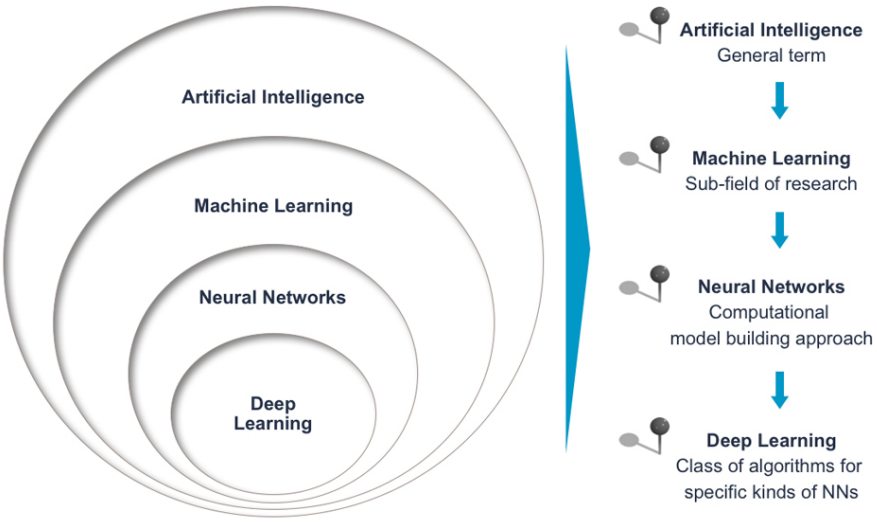

Abbildung 2.1: Verschiedene Abstraktionslevel von Artificial Intelligence in hierarchischer Ordnung

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[9].

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 Datenbestandes 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.

2.2.2 Neuronale Netze

alt:

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

Hochauflösende Bilder bestehen aus mehreren Millionen Pixel mit Position und meist RGB-Werten für die Farbe des Punktes. Nur durch Verbindung der richtigen einzelnen Punkte könnte man Konturen erkennen. Diese müssen dann der richtigen Kategorie zugeordnet werden, um sie bei einem Suchwort wie 'Haus' anzuzeigen. Doch wie kann ein Computer aus dieser riesigen Menge an einzelnen Pixeln Bilder richtig kategorisieren?

Wir Menschen nehmen Bilder mit der Netzhaut des Auges als visuelle Reize auf und senden sie an unser Gehirn. Dieses verarbeitet die Informationen durch einen 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.

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

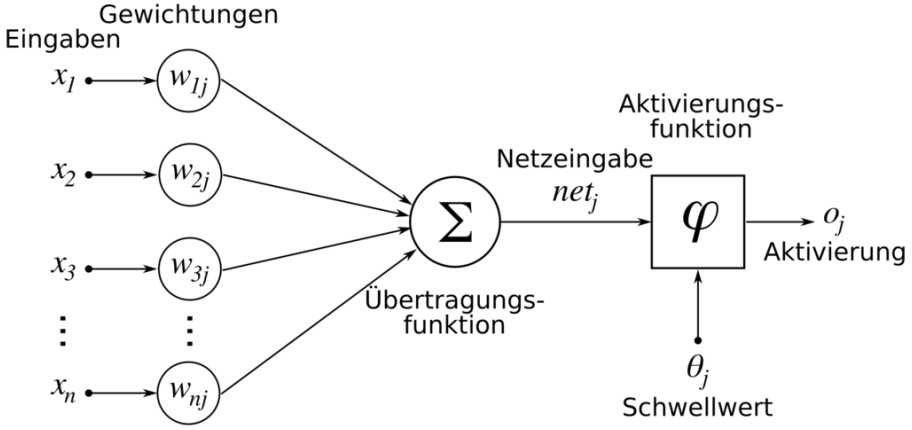

Abbildung 2.2: Künstliches Neuron

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 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 Eingangskanäle ein gewisser Schwellwert summiert anliegen. Wenn dieser Schwellwert erreicht wird, feuert das Neuron. Der Ausgangskanal eines Neuron ist gleichzeitig ein Eingangskanal eines oder mehreren anderen Neuronen.

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 reduziert werden.

Jedoch muss ein künstliches neurales Netzwerk erst trainiert werden. Trainieren bedeutet hier das Einstellen der Gewichtungen von Verbindungen und Schwellwerte der Neuronen für eine spezielle Aufgabe. Diesen Prozess nennt man Deep Learning eine Form von Machine Learning.

2.2.3 Deep Learning

Konzentration auf Deep Learning in dieser Arbeit.

alt

Beim Deep Learning werden die zahlreichen Zwischenschichten, eingeschlossen von Eingabe- und Ausgabe-Schicht, trainiert. Hierbei wird eine umfangreiche und komplexe Struktur der Neuronen-Verbindungen zwischen den Schichten aufgebaut. Wie das Programm endgültig die Aufgabe lösen soll, wird hierbei nicht vorgegeben, sondern wird bei diesem Prozess autonom ermittelt.
Ein künstliches neuronales Netz wird mit dem Zweck aufgebaut, eine bestimmte Aufgabe zu lösen. Extra dafür müssen Testdaten aufbereitet werden. Diese Art der Daten, beispielsweise Bilder, soll das fertig trainierte Netz richtig interpretieren können. Testdaten sind Daten, bei dem das korrekte Ergebnis zu jedem Test bekannt ist.

Am Anfang ist das künstliche neuronale Netz meist mit relativ zufälligen Werten und Verbindungen vorbelegt. Testdaten werden dem zu trainierenden Netzwerk an die Eingangsschicht übergeben. Diese durchlaufen das Netz. Das Ergebnis wird an der Ausgabeschicht überprüft. Die Ausgabeschicht besteht im einfachsten Fall aus zwei Neuronen, beispielsweise 'Gesicht erkannt' oder 'kein Gesicht', an dem gemessen wird, wie viel gewichtete Signale ankommen. Diese summiert, ergeben die Endergebnisse der Berechnungen und das Neuron mit dem höchsten Gewicht, ergibt die Antwort.

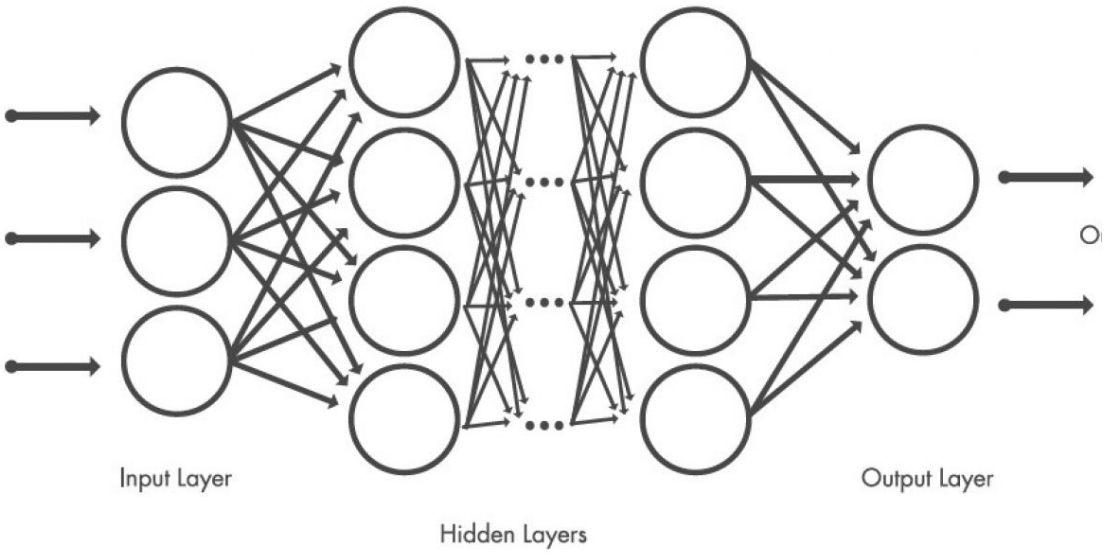

Abbildung 2.3: Hidden Layer

Neuronale Netze sind für Menschen ab einer gewissen Größe nicht mehr nachvollziehbar. Somit kann nur die Eingabe mit der Ausgabe verglichen werden, um auf die Korrektheit der Aufgabenlösung zu prüfen. Ziel ist es durch möglichst viele und umfangreiche Testdaten das Netzwerk so einzustellen, dass diese nicht nur die Testdaten richtig beantwortet, sondern auch unbekannt Daten korrekt interpretiert.

Um eine Aufgabe, wie 'Gesicht in Bild erkennen', zu trainieren, wird nicht nur ein Netz mit Zufallswerten und Verbindungen generiert sondern tausende. Alle werden mit den gleichen Testdaten geprüft und für jedes Netz ein Mittelwert über die Korrektheit der Antworten erstellt. Da alle Netze Initial mit Zufallswerten belegt sind, haben die meisten Netze eine Erfolgsrate von ca. 50 Prozent. Die Netze mit den höchsten Erfolgsraten, mit beispielsweise mehr als 60 Prozent, werden behalten, der Rest wird verworfen. Die erfolgreichsten Netze werden mehrfach kopiert und bei jedem Kopiervorgang individuell leicht verändert und erneut getestet. Die Besten werden wieder genommen und leicht modifiziert kopiert und schlechteren verworfen.

Nach einer gewissen Anzahl an Iterationen entscheidet das Netz nicht mehr willkürlich, sondern scheint intelligent die Aufgabe zu lösen. Dieser Iterationsschritt kann unendlich oft laufen, jedoch empfiehl es sich, je nach Anwendungsfall, ab einer gewissen Erfolgsrate das Training zu beenden oder neue Testdaten zu verwenden. Als Ergebnis des Prozesses erhält man durch Deep Learning ein trainiertes künstliches neuronales Netz, das im Allgemeinen als KI bezeichnet wird.

Die KI kann nun die Testdaten nahe zu perfekt interpretieren. Trotzdem muss nun noch weiter überprüft werden, ob die KI auch unbekannte Daten korrekt interpretiert.

Fehlverhalten einer KI kann leicht übersehene werden, da es fast unmöglich ist, jeden Testfall abzudecken. Beispiele von KIs mit Fehlverhalten und wie sie genau entstehen, werden im nächsten Abschnitt genauer betrachtet.

Lernarten

i. Supervised
ii. Unsupervised
iii. Controlevisional
iv. Reinforcement

2.2.4 Loss-Funktion

Performanz

2.2.5 Informationsverlust zwischen Schichten

1. Erste Schicht verbunden mit letzter Schicht
a. Eingabe hoher Einfluss auf Endergebnis
2. Jede Schicht nur Verbindung zu der Nächsten
a. Hohe Informationsverlust

2.3 Neuronen und Features

2.3.1 Arten von NN

Es gibt drei Arten, wie neuronale Netze Daten verarbeiten können.

3D

Gewichtete Netze

Features

2.4 Architekturen

Arten
1. Full Connected
2. CNN
3. ResNet
4. Natural Network Connection
5. Dropout
6. ...

2.5 Daten

Erst durch eine Kombination aus Algorithmen und Daten wird die Entscheidungsfindung unterstützt. Wie ein menschlicher Entscheider können auch Algorithmen wegen unvollständiger oder fehlerhafter Daten zu fehlerhaften Entscheidungen gelangen. Deswegen sollte bei der Datenaufbereitung bereits darauf geachtet werden die richtige Woge zwischen Quantität und Qualität zu finden.

2.5.1 Quantität

Je breiter, also je mehr Variablen in den Datensätzen existieren umso komplexer wird die Aufgabe. Und diese Komplexität der Probleme erfordert, dass die Menge an Daten entsprechend groß sein muss, damit das zu trainierende System immer besser reagiert.

Ein Beispiel hierfür findet sich in der Autoindustrie. Beim autonomen Fahren müssen Daten von Laser-, Kamera- und Radarsensoren im Auto zuverlässig und schnell verarbeitet und zusammengeführt werden. Dadurch verfügt das Fahrzeug jederzeit über ein präzises Abbild der realen Verkehrsbedingungen, kann sich selbst in diesem Umfeld verorten und darauf basierend in jeder Fahrsituation die richtige Entscheidung treffen [3].

Anhand dieses Beispiels erkennt man die Wichtigkeit der Quantität der Trainingsdaten, da die Anzahl möglicher Situationen im Straßenverkehr prinzipiell unendlich ist. Um gleichartige Strukturen im Verkehrsgeschehen zu erkennen, sind viele Trainingsdaten erforderlich, die ein immer genaueres Bild ergeben.

Da aber nicht immer die Anwendungen komplex sind, sind nicht immer viele Datensätze notwendig. Bei geringerer Komplexität reichen somit weniger Datensätze aus um gute Ergebnisse zu erreichen.

2.5.2 Qualität

Auch die Qualität der Trainingsdaten spielt eine wichtige Rolle. Die Datenqualität zeichnet sich dadurch aus, dass Daten den Zweck in einem bestimmten Zusammenhang erfüllen müssen.

Dass heißt wenn nun die KI auf Bildern z.B. einen Panzer erkennen soll[7], muss die KI mit Bildern von Panzern trainiert werden, welche auch als Panzer gekennzeichnet wurden. Würden nun Bilder von Autos mit in die Testdaten gelangen, welche auch zuvor als Panzer gekennzeichnet wurden, würde die Maschine diese Autos bei unbekannten Bildern auch als Panzer erkennen.

Daher ist es für den Erfolg der Daten die Qualität dieser sehr wichtig. Hierbei sollten somit keine unzweideutigen Stammdaten existieren.

3 Bias Entstehung

Bei der Nutzung von KI System können Verzerrungen(Bias) bzw. Fehlverhalten entstehen, diese können unterschiedlicher Natur sein und an unterschiedlichen Stellen, in der in Abbildung 3.1 gezeigten, vereinfachten Machine Learning Pipeline, auftreten. Dabei möchten wir auf die Daten eingehen, welche bei der Eingabe zu Bias führen können und menschliche Fehler verdeutlichen, welche bei der Verarbeitung und der Ausgabe auftreten können. Zuletzt möchten wir Adversial Attacks ansprechen, welche zu weiteren Verzerrungen führen können.

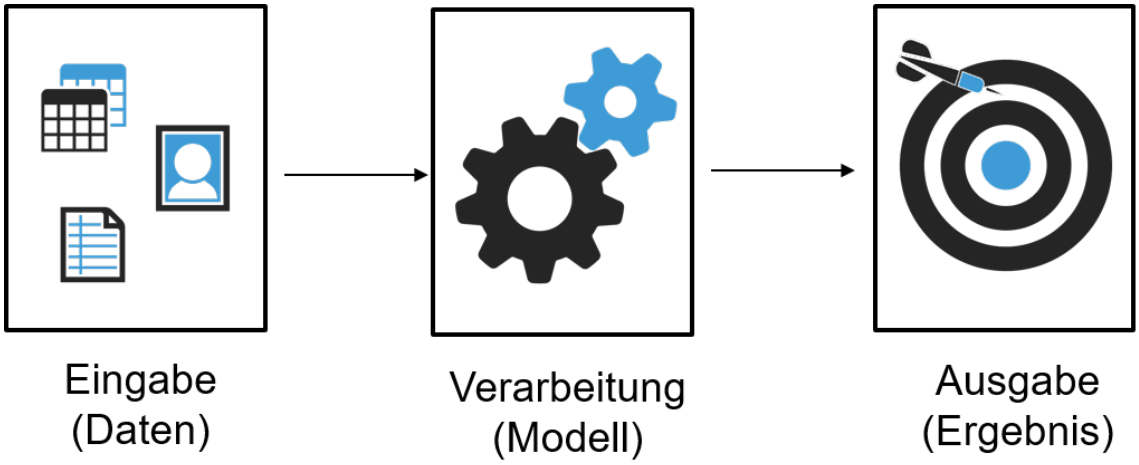

Abbildung 3.1: Machine Learning Pipeline: Eingabe, Verarbeitung, Ausgabe

3.1 Daten

Im ersten Kapitel möchten wir erläutern welche Bias, durch z.B. eine schlechte Datenqualität oder Datenquantität, entstehen können.

3.1.1 Unvollständigkeit der Daten

Zuerst möchten wir auf ein Problem aufmerksam machen, welches zu Verzerrungen führt, anhand des Beispiels aus 2.5.2.

Das Pentagon hatte eine Software angefordert, welche Panzer in der Natur erkennen sollte. KI Forscher haben daraufhin ihr neuronales Netz mit Fotos von getarnten Panzern trainiert, und mit Landschaftsfotos ohne getarnte Panzer. Dadurch sollte gewährleistet werden, dass die Software Panzer auf unbekannten Bildern erkennt.

Bei internen Test funktionierte das System sehr gut, doch bei den realen Test schien die Software nicht zu funktionieren. Das Problem hierbei lag daran, dass diese KI zuvor mit Trainingsdaten gefüttert wurde, welche nur bei schönem Wetter fotografiert wurden (siehe Auch die internen Testdateien erfüllten dieses Kriterium. Die realen Test hingegen wurden bei jedem Wetter ausgetragen.

Abbildung 3.2: Panzer bei bewölkten Wetter vs Landschaft ohne Panzer bei schönem Wetter

Die Software hatte somit trainiert schlechtes und gutes Wetter auseinander zu halten und nicht Panzer zu erkennen. Das Problem hierbei lag an unvollständigen Daten, hier wurden zu wenig unterschiedliche Fälle getestet und dadurch wurde ein Bias erzeugt, welcher die Nutzung der Software unmöglich machte.

3.1.2 Garbage in - Garbage out

Eine Maschine kennt grundsätzlich keinen Unterschied zwischen Schwarz und Weiß, Mann und Frau oder Jung und Alt. Erst durch eine KI lernt eine Maschine Verhalten und Muster kennen. Hierfür werden wie bereits in 2.5 beschrieben Daten benötigt, welche die richtige Qualität benötigen um die Ergebnisse zu bestimmen.

Bleiben fehlerhafte Daten unentdeckt, wird ein System trainiert, welches in Zukunft falsche Ergebnisse liefern wird. Das Beispiel aus 2.5.2 erläutert dieses Problem. Möchte ich ein System trainieren, welches Panzer identifizieren kann, muss ich diesem System beibringen Panzer zu erkennen. Füttere ich dieses nun mit Autos und markiere diese versehentlich als Panzer, identifiziert das System daraufhin diese nicht als Autos sondern als Panzer.

Bei der traditionellen Datenanalyse können solche schlechte Daten nachträglich entfernt werden. Hat allerdings eine Maschine durch maschinelles Lernen etwas gelernt, wird es schwer dies wieder zu verlernen. Denn ab einem Gewissen Grad wird es nahezu unmöglich, herauszufinden, auf welche Datenelemente die Vorhersagen basieren. Ähnlich wie beim menschlichen Gehirn.
Baut unser erlerntes Wissen in Teilen auf falsche Grundannahmen oder Informationsbausteinen auf, verliert der ganze Komplex seinen Wert und wir müssen von neu alles erlernen.

Diese Problem wird in der KI als "Garbage in - Garbage out"(Müll rein, Müll raus) bezeichnet.

3.1.3 Bias in Trainings-/Testdaten

Ca. 4 Jahre entwickelte Amazon einen Algorithmus, welcher unter mehreren Bewerbungstexten automatisch die besten Bewerber herausfiltern sollte. Dabei bezog die Software sich auf voran gegangene Bewerbungen, 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 daraus 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 ohne die Angabe eines Geschlechtes und dieses z.B. nur durch Frauenvereine erkennbar wurde. Die KI blieb diesen Auswahlkriterien treu und bevorzugte vorwiegend Männer.[4]

Die Hoffnung solcher Anwendungen liegt eigentlich darin, Vorurteile zu vermeiden und Prozesse fairer zu gestalten, da eine Maschine wie in 3.1.2 bereits genannt keine Unterschiede kennt. Doch in diesem Beispiel beinhalteten die Trainingsdaten bereits Vorurteile und führten somit zu einem Fehlverhalten des Systems.

An diesem Beispiel wird deutlich wie zentral die Daten für eine KI sind. Meist ist es nicht möglich Daten zu finden, welche nicht bereits menschliche Bias enthalten. Solch verzerrte Trainingsdaten, werden unter Bezug auf ihre Zusammensetzung auch als WEIRD Samples (western, educated, industrialized, rich and democratic societies) bezeichnet[5].

3.1.4 Under-/ Overfitting

Wie in 2.5.1 bereits beschrieben, spielt die Datenquantität eine große Rolle beim maschinellen Lernen. Die beiden Fehler die allerdings dadurch entstehen können, sind nutzloses Wissen aufzubauen oder aus einem vorhandenem Trainingsdatensatz keine Relevanten Lerninformationen ziehen zu können. Diese Phänomene werden als Overfitting (Überanpassung) und Underfitting(Unteranpassung) bezeichnet.

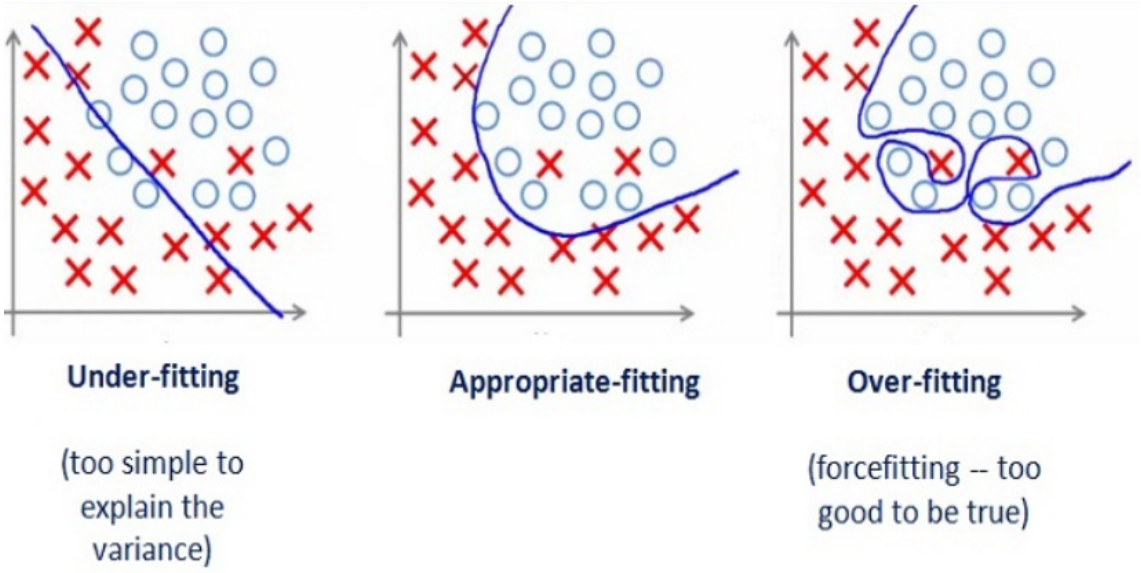

Abbildung 3.3: Over- und Underfitting

Der linke Graph aus 3.3 zeigt eine Linie, welche viele Punkte falsch abdeckt, da die Linie zu einfach ist. Solche Modelle neigen somit zur Unteranpassung, bzw. zu einer hohen Verzerrung der Vorhersagen.

Der Graph auf der rechten Seite hingegen sagt alle Punkte richtig vorher. Unter dieser Annahme könnte man denken es wäre ein sehr guter Graph. Das Problem hierbei ist

allerdings, dass alle Punkte mit vorher gesagt werden, auch diese die Grundrauschen oder Outliner sind. Aus diesem Grund lernt die Maschine nicht vorhandene Muster kennen und liefert aufgrund ihrer Komplexität schlechte und abweichende Vorhersagen aus.[1]

Ein Beispiel für Overfitting ist, wenn durch maschinelles Lernen ein System einen Schrank auf einem Bild erkennen soll. Diesem System werden nun alle Möglichkeiten an Schränke antrainiert. Ein Schrank kann sehr ausgefallen sein und somit lernt das System sehr viele unterschiedliche Schränke kennen und wenn nun ein Schrank z.B. aus Autoteilen hergestellt wurde. Könnte es passieren, dass auf einem Bild ein Auto als Schrank erkannt wird, obwohl in diesem Bild kein Schrank vorhanden ist.

Am besten ist somit der mittlere Graph, welcher Outliner und das Grundrauschen ignoriert und eine gute Balance zwischen Verzerrung und Abweichungen besitzt.

3.1.5 Ähnlichkeit der Daten

Abbildung 3.4: Hund oder Bagel?

Es gibt Bilder, welche selbst für den Menschen schwer zu differenzieren sind (Beispiel Abbildung 3.4). Ist nun ein Mensch nicht aufmerksam und sagt der Maschine das ein Bagel ein Hund ist, erlernt die Maschine falsche Bilder und sagt zukünftig auch Bagel als Hunde vorher. Aber auch durch richtiges Kennzeichnen der Bilder durch den Menschen, können solche Bilder zu Verzerrungen führen.

Ein weiters Beispiel hierfür kommt aus der Autoindustrie. In Testreihen für autonomes Fahren stuften die Probanden (weil sie nicht aufmerksam waren) immer wieder das bestimmte Bild eines Menschen als Bild einer Tonne ein. Das System reagiert folgerichtig und wertet in einer kritischen Verkehrssituation das Überfahren einer (vermeintlichen) Tonne als verhältnismäßige Alternative, die möglichst wenig Schaden anrichtet [6].

3.2 Menschliche Fehler

3.2.1 Falsches Ziel

3.2.2 Falsche Architektur

3.2.3 Falsches Lernen

KI lernt einfachste Unterschiede
i. Nicht Unterscheid zwischen Auto und Boot sondern Untergrund(Wasser/Land)
ii. Sehr Fehleranfällig z.B. Auto fährt durch flaches Wasser (KI -¿ Boot)

3.3 Angriff auf KI

Adversarial Attacks

4 Sicherheitsprobleme durch BIAS

4.1 Gefahren für Maschinen

Google KI -¿ Kühlung von Maschinen

4.2 Gefahren für Menschen

Tesla Autopilot
Etc.

5 Prävention

5.1 Passende Architektur zu Daten

5.2 Nur ein Ziel

Viele Ziele = Komplex -¿ Fehleranfällig -¿ Bias
Beispiel: Baidu Gesichtserkennung (erkennt nur Asiaten)

5.3 Verfahren zum Validieren

Unterschiedliche Personen (Entwickler/Tester)

An echte Daten Testen (Überwachtes Demo Live Betrieb)

Beispiel:
Polizei Berlin Gesichtserkennung bei Überwachungskamera
3 verschiedene KIs

5.4 Test-/Trainingsdaten Aufbereiten

Wie bereits in Abschnitt 3.1 beschrieben, gibt es viele Fehler die gemacht werden können um zu Bias einer KI zu führen. In diesem Kapitel möchten wir darauf eingehen, was dagegen unternommen werden kann.

Die Datenaufbereitung ist ein wesentlicher Teil für die Nutzung einer KI. Wichtig ist das man die Daten generalisiert. Nutzt man wie im Beispiel aus dem Kapitel 3.1.4 zu viele Daten, nur um alle Fälle abzudecken, wie ein Schrank aussehen kann, führt dies zu Verzerrungen bei den Vorhersagen. Wichtig ist daher die Daten auf die Standard

Schränke zu begrenzen, und Outliner zu ignorieren. da Bilder, welche anderen Objekten zu ähnlich aussehen, lernt die KI falsche Muster kennen und es entstehen Fehlverhalten. Generalisiert man nun die Daten, und lässt die Maschine mit offensichtlichen Schränken trainieren, führt dies zu weniger Fehlern, auch wenn eventuell nicht alle Schränke als solche erkannt werden.

Damit Vorurteile, Outliner oder falsche Bilder in den Daten nicht zu Problemen führen sollte diese zuvor validiert werden. Beispiele für eine Validierung könnten folgende sein:

1. Lektorat oder Peer-Review

2. Das Vier-Augen-Prinzip (gegenseitige Kontrolle)

3. Mehrheitsentscheide bei unterschiedlichen Ergebnissen

Mit diesen Maßnahmen können zuvor einige Fehler vermieden werden und unnötig hohe Kosten vermieden werden.

6 Fazit

Thema ist größer als hier beschreibbar
Evtl. Deep Fake http://iphome.hhi.de/samek/pdf/LapNCOMM19.pdf https://ujjwalkarn.me/2016/08/explanation-convnets/

7 Alt: Problemstellung Fehlverhalten von künstlichen neuronalen Netzen

7.1 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 voran gegangene Bewerbungen, 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 daraus 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. Die KI blieb diesen Auswahlkriterien treu und bevorzugte vorwiegend Männer.[4]

Dieses Fehlverhalten der KI führte dazu, dass die Software nicht genutzt werden konnte, bzw. nach Anpassungen nur eine beratende Funktion besaß.

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.

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

Wie in vorangegangen Beispielen bereits erwähnt wurde, wird Künstliche Intelligenz in vielen Bereichen eingesetzt. Die Hoffnung solcher Anwendungen, liegt eigentlich darin,

Vorurteile zu vermeiden und Prozesse fairer zu gestalten, da eine Maschine grundsätzlich keinen Unterschied zwischen Schwarz und Weiß, Mann und Frau oder Jung und Alt erkennt.

Anhand des Beispiels aus 7.1 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[11].

Daher kommt der Ausdruck "Garbage in - Garbage Out", dass heißt benutzt man fehlerhafte Daten oder welche, die Vorurteile beinhalten, erhält man fehlerhafte oder mit Vorurteilen belastete Ergebnisse.

7.3 Probleme durch Fehlverhalten von künstlichen neuronalen Netzen

Durch die bereits genannten Beispielen in den vorherigen Kapiteln, werden Probleme deutlich. Das Fehlverhalten einer KI kann zu Diskriminierung einzelner Geschlechter/-Gruppen oder Kulturen führen. 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 eine Person 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[8].
Dies führt dazu das Afroamerikaner eine niedrigere Gesundheitsversorgung haben als andere und dies kann zu großen Gesundheitliche Problemen führen.

Ein weiteres Beispiel nannte das Heise Magazin 2017[12]. Das Online Magazin erwähnte die Software COMPAS, welche auch in der USA verwendet wird und dort in mehreren Bundesstaaten zum Einsatz kommt. COMPAS steht für „Correctional Offender Management Profiling for Alternative Sanctions“ und gibt vor, das Rückfallrisiko von Straftätern verlässlich berechnen zu können.

COMPAS errechnet für jeden Delinquenten einen individuellen Risk Score, welcher auf das spätere Strafmaß Auswirkungen hat. In die Risikobeurteilung des Algorithmus fließen eigene Vorstrafen, eventuelle Vorstrafen naher Verwandter, Alkohol- und Drogenmissbrauch, soziale Bindungen, usw..

Falls der Risk Score eines Verurteilten nun zwischen 1 und 10 ist, lässt der Richter keine Bewährung mehr zu sondern schickt ihn präventiv hinter Gitter. Das Problem hierbei ist, dass der Algorithmus nicht offen gelegt wird und somit wird dem Algorithmus blind vertraut, ohne überprüfen zu können, wie dieser Score zustande kam.

Wie Heise Magazine weiter berichtete, belegte allerdings 2016 eine Studie der Investigativ-Plattform ProPublica, dass die COMPAS-Algorithmen beispielsweise schwarzen Angeklagten grundsätzlich ein höheres Risiko attestieren, erneut straffällig zu werden, als dies tatsächlich der Fall ist. Bei weißen Angeklagten ist es hingegen genau umgekehrt.

Falls solche Algorithmen blind vertraut werden, kann dies zu gravierenden Folgen eines Verdächtigen führen, welcher zu unrecht ein ganzes Leben hinter Gitter sitzen könnte.

Tabellenverzeichnis

Literatur

[1] Anup Bhande. *What is underfitting and overfitting in machine learning and how to deal with it.* 2018. URL: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76 (besucht am 14. 06. 2020).

[2] S Borowiec. *AlphaGo seals 4–1 victory over go grandmaster Lee Sedol.* 2016. URL: https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol (besucht am 08. 06. 2020).

[3] Irina Hübner. *Wie KI autonomes Fahren sicherer macht.* 2019. URL: https://www.elektroniknet.de/elektronik-automotive/assistenzsysteme/wie-ki-autonomes-fahren-sicherer-macht-171811.html (besucht am 11. 06. 2020).

[4] 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).

[5] Czihlarz Jochen. *Biases in Künstlicher Intelligenz (KI).* 21. Apr. 2020. URL: https://www.anti-bias.eu/allgemein/biases-in-kuenstlicher-intelligenz/ (besucht am 24. 05. 2020).

[6] Jan Knupper. *Trainingsdaten für KI – Loesungen: Nicht nur die Masse macht's.* 2019. URL: https://www.clickworker.de/2019/03/19/trainingsdaten-ki-loesungen/ (besucht am 14. 06. 2020).

[7] Jaromir Konecny. *Von Jaromir KonecnyLesedauer ca. 14 Minuten148 Kommentare Künstliche Intelligenz, künstliche Dummheit und der gesunde Menschenverstand.* 2018. URL: https://scilogs.spektrum.de/gehirn-und-ki/kuenstliche-intelligenz-kuenstliche-dummheit-und-der-gesunde-menschenverstand/ (besucht am 11. 06. 2020).

[8] 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.

[9] 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.

[10] TC Wang u. a. *Video-to-video synthesis.* 2018, S. 1144–1156.

[11] 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).

[12] Peter-Michael Ziegler. *Im Namen des Algorithmus*. 2017. URL: `https://www.heise.de/select/ct/2017/25/1512700333136715` (besucht am 24. 05. 2020).