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[**alphaGo**]. 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[**videoToVideo**].

Einerseits gibt es viele Erfolge die für ein KI betriebenes System sprechen, andererseits bestärken medienwirksame Verfehlungen, wie z.B. das Bewerbungssystem von Amazon[**amazon**], 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

Künstliche Intelligenz (KI) oder englisch artificial intelligence (AI) ist der Oberbegriff für ein Teilgebiet der Informatik. Dieses Gebiet befasst nicht nur mit neuronalen Netzen, sondern generell mit jeglicher Form von maschinellen intelligenten Verhalten und dem maschinellen Lernen, siehe 2.1. Generell wird bei der künstlichen Intelligenz versucht biologische Intelligenz auf einen Computer zu simulieren. Dies basiert meist auf simplen Algorithmen, wodurch die Begrifflichkeit 'Intelligenz' in Bezug auf einer Maschine öfter in Frage gestellt wird.

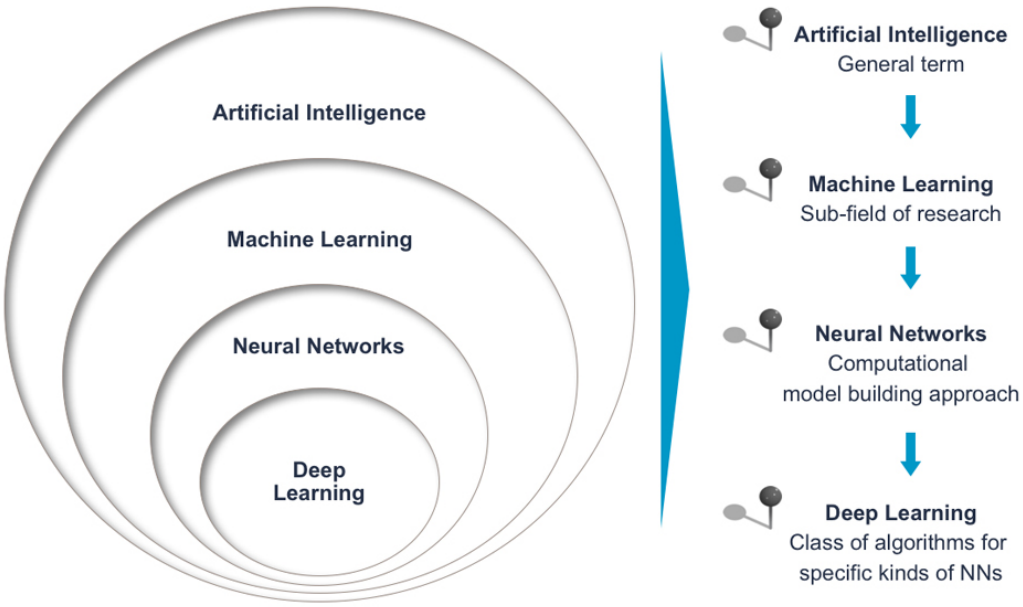

Abbildung 2.1: Verschiedene Abstraktionslevel von Artificial Intelligence in hierarchischer Ordnung

Damit ein Programm den Titel KI tragen darf, muss sie zum einen die Fähigkeit zu lernen besitzen, zum anderen die Fähigkeit Intelligent Lösungen zu finden, auch bei nicht eindeutigen Eingaben.
KIs werden grob in zwei Kategorien aufgeteilt. Einmal die starke KI, die ebenbürtig mit Menschen zusammenarbeiten kann und schwacher KI, die lediglich das Arbeiten von Menschen unterstützen soll.

2.2.1 Maschine Learning

Machine Learning (ML) ist ein Teilgebiet der künstlichen Intelligenz. Es ist der Oberbegriff jeglicher Lernvarianten von KIs. Im Allgemeinen versucht eine KI neue Muster und Gesetzmäßigkeiten in Trainingsdaten zu erkennen, diese zu verallgemeinern und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten zu verwenden[**EliminateHumanBias**].

Diese Arbeit speziell konzentriert auf Deep Learning, welches eine Variante zum Trainieren von neuronalen Netzen darstellt.

Lernansätze

KIs bestehen aus vielen Algorithmen. Damit eine KI lernt, müssen diese Algorithmen angepasst werden. Dies kann überwacht geschehen mittels eines "Lehrers", der den Lernerfolg bei evolutionären Algorithmen-Änderungen überprüft.
Die andere Variante ist das unüberwachte Lernen. Hierbei erzeugt ein Algorithmus ein statisches Model aus den Trainingsdaten und erkennt Zusammenhänge zwischen den Daten und dessen Kategorie, sogenannte Features. Diese System gibt die Wahrscheinlichkeit zurück, zu welcher Kategorie die Eingabe gehört, abhängig von den erkannten Features der Eingabe.

Aus den Konzept dieser zwei Hauptlernvarianten wurden diverse Unterlernvarianten erstellt, wie teilüberwachtes Lernen, bestärktes Lernen oder unüberwachten Lernen ohne Kategorisierung.

2.2.2 Neuronale Netze

Künstliche neuronale Netze (KNN) bestehen aus künstlichen Neuronen, die untereinander verflechtet sind. Diese Konstrukte sind denen der Neuronen-Verbindungen im Nervensystem eines Lebewesens nachempfunden.
KNNs sind nicht dazu da das Nervensystem von Lebewesen nachzubilden, sondern abstrakt die Eigenschaften der Informationsverarbeitung und der Lernfähigkeit zu imitieren.

KNNs sind meist in Schichten mit beliebig vielen künstlichen Neuronen aufgebaut. In der Regel besteht ein KNN aus drei Teilen die Eingangsschicht (grün), verdeckte Schicht (blau) im Englischen Hidden Layer und die Ausgabeschicht (gelb). In der Eingangsschicht fließen die Informationen in das Netz ein und in der Ausgabeschicht das Ergebnis der Berechnungen aus. Jede Schicht besteht aus beliebig vielen Neuronen je nach Komplexität des Zieles, die Hidden Layer sogar aus beliebig viele Schichten.

In der Regel arbeiten KNNs nach dem feedforward-Prinzip, bei dem die Informationen immer nur in eine Richtung fließt. Es gibt jedoch auch rekurrente Netze, bei denen durch rückgerichtete Kanten Rückkopplungen im Netz entstehen.
Die einfachste Netzstruktur ist das einschichtige feedforward-Netz. Dies besteht ohne Rückkopplungen aus nur einer Schicht, der Ausgabeschicht.

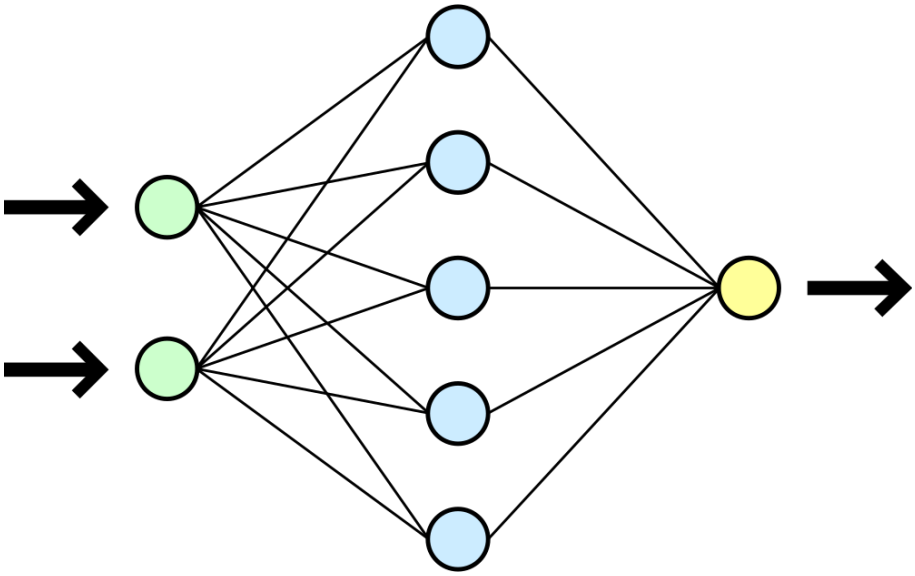

Abbildung 2.2: Vereinfachte Darstellung eines künstlichen neuronalen Netzes

Das künstliche Neuron

Künstliche Neuronen sind die Grundbestandteile eines künstlichen neuronalen Netzes. Ein künstliches Neuron ist die vereinfachte und abstrakte Version einer biologischen Nervenzelle und ist wie folgt aufgebaut.

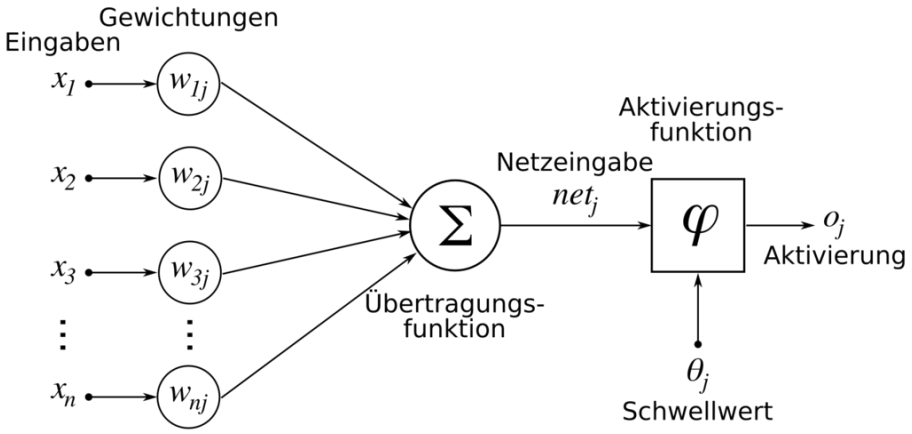

Abbildung 2.3: Ein künstliches Neuron

Ein künstliches Neuron besitzt n Eingangskanäle und einen Ausgangskanal. j repräsentiert hier die eindeutige Nummer des Neuron. Jede Eingabe x_i besitzt ein dazugehörendes Ge-

wicht $w_{1j}..w_{nj}$. Dieser spiegelt die Wichtigkeit der Eingabe wider, diese kann hemmend negativer oder erregend positiver Wert wirken. Die Übertragungsfunktion Σ summiert alle multiplizierten Eingaben mit ihrem Gewicht und geht als Netzeingabe net_j in die Aktivierungsfunktion φ ein.
Ob das Neuron "feuertöder kein Signal sendet, wird hier berechnet. Auf die Netzeingabe net_j wird ein Schwellenwert θ_j addiert. Als mathematisch Vereinfachung wird der Schwellwert θ als w_0 bezeichnet und $x_0 = 1$ eingeführt und somit in der folgenden Formel immer auf die Netzeingabe net_j addiert.

$$a = \sum_{i=0}^{n} x_i w_i$$

Die Variabel a geht in die Aktivierungsfunktion $\sigma(a)$ ein, anhängig von dem Ergebnis wird das Neuron aktiv oder bleibt inaktiv. 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 jedes künstliche Neuron eines KNNs richtig eingestellt werden, damit das KNN dessen Ziel erfüllt. Einstellt wird das Neuron durch Training. Trainieren bedeutet hier das Ermitteln der richtigen Werte für Gewichtungen und Schwellwerte, als auch das Einstellen der richtige Verbindungskombinationen der Neuronen untereinander. Diesen Prozess nennt man Deep Learning, eine Form von Machine Learning.

2.2.3 Deep Learning

Deep Learning ist eine Maschine Learning Variante, die speziell bei künstliche neuronale Netze eingesetzt wird. Beim Deep Learning werden die zahlreichen Zwischenschichten der Hidden Layer trainiert. Dabei wird eine umfangreiche und komplexe Struktur der Neuronen-Verbindungen aufgebaut. Wie das Programm endgültig die Aufgabe lösen soll, wird hierbei nicht vorgegeben, sondern wird bei diesem autonomen Prozess evolutionär ermittelt.

Ein künstliches neuronales Netz wird mit dem Zweck aufgebaut, eine bestimmte Aufgabe zu lösen. Extra dafür müssen Trainingsdaten aufbereitet werden. Diese Art der Daten, beispielsweise Bilder, soll das fertig trainierte Netz richtig interpretieren können. Trainings und Testdaten sind Daten, bei dem das korrekte Ergebnis bekannt ist.

Am Anfang ist das künstliche neuronale Netz meist mit relativ zufälligen Werten und Verbindungen vorbelegt. Trainingsdaten 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.

Neuronale Netze sind für Menschen ab einer gewissen Größe nicht mehr nachvollziehbar. Somit kann nur die Eingabe mit der Ausgabe mit Hilfe von Testdaten verglichen werden, um auf die Korrektheit der Aufgabenlösung zu prüfen. Ziel ist es mit möglichst vollständigen Trainingsdaten das Netzwerk so einzustellen, dass diese nicht nur die Trainingsdaten und 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 Trainings- und Testdaten zu verwenden. Als Ergebnis des Prozesses erhält man durch Deep Learning ein trainiertes künstliches neuronales Netz, das im Allgemeinen als KI bezeichnet wird.

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 [**autonomesFahren**].

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 Datensätze komplex sind und zu viel Trainingsdaten bei weniger Variablen zu Problemen führen können, muss hier eine richtiges Maß gefunden werden. In 3.1.4 möchten wir auf das Problem eingehen, wenn zu viele oder zu wenig Daten genutzt werden.

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[**panzer**], 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.

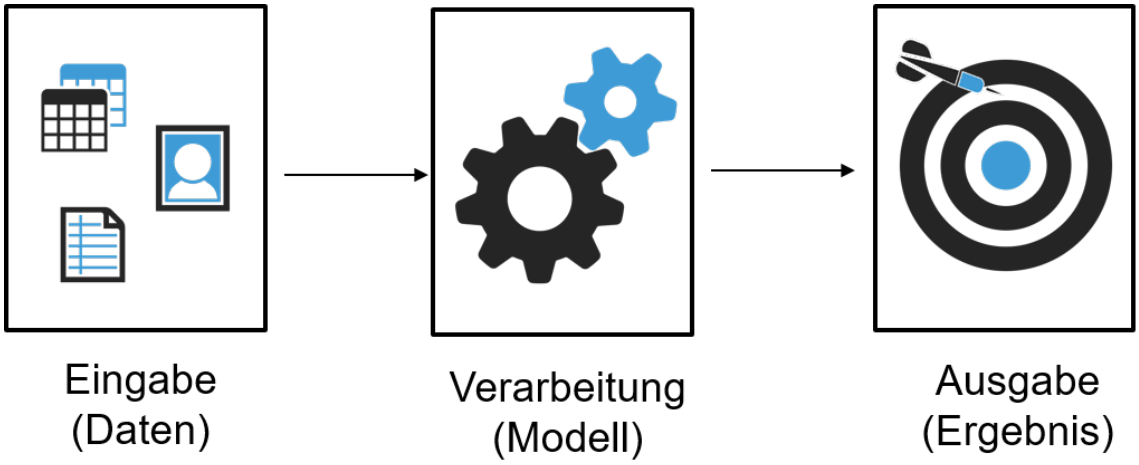

Abbildung 3.1: Machine Learning Pipeline: Eingabe, Verarbeitung, Ausgabe

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.[**amazon**]

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[**BiasInKi**].

3.1.4 Under-/ Overfitting

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

3.1.5 Ähnlichkeit der Daten

Selbst für Menschen auf ersten Blick schwierig zu differenzieren Beispiel Bild Croissant vs Katze. -¿ Daher wichtig gute Qualität der Daten. Solche Verfahren aus 2.5.2 lassen sich für die verschiedensten Aufgaben einsetzen. Im Normalfall kann ein Computer die Qualität von Daten nicht bewerten. Deshalb besteht immer die Gefahr, dass Maschinelles Lernen zu formal logischen, aber praktisch falschen Ergebnissen führt. Ein Beispiel: In Testreihen für autonomes Fahren stufen die Probanden (weil sie nicht aufmerksam sind) 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.
ii. Bild Croissant vs Katze
iii. https://distill.pub/2019/activation-atlas/

3.2 Menschliche Fehler

Den Faktor Mensch darf man bei der Bias Entstehung nicht vergessen. Konzeptionelle Fehler mangels an Wissen oder einem Missgeschick heraus fördern Fehlverhalten. Markante Defizite können ein funktionale KI-Entstehung komplett verhindern, sind in

Abbildung 3.3: Hund oder Bagel?

der Regel noch die besseren Missstände. Denn kleine Mängel, die nicht sofort auffallen, können in Produktion fatale Folgen haben.

3.2.1 Falsche Zielsetzung

Die Zielsetzung eines KNNs ist ein nicht zu unterschätzender Teil bei der KI Entwicklung. Setzt man hier den falschen Grundstein, können sich vermeidbare Fehlverhalten einer KI entwickelt.

Falsche Ziel Definition

Wenn man eine Aufgabe hat und diese Automatisieren möchte, greift man heute zutage gerne zur KI. KI ist modern und in aller Munde. Jedoch sollte einem im Klaren sein, dass künstliche Intelligenz kein Allheilmittel ist.

Zu viele Ziele

Viele Köche verderben den Brei. Diese Weisheit kann auch auf die Ziele von KNNs umgemünzt werden. Denn je mehr Ziele ein KNN hat, desto größer und komplexer muss ein KNN sein, um alle Fälle abdecken zu können. Je komplexer ein KNN ist, desto größer ist die Wahrscheinlichkeit, dass Fehler passieren. Auch ist sich das Netz deutlich unsicherer bei seinen Entscheidungen.

Wenn man eine nahe zu perfekte KI mit einer Aufgabe erweitern möchte, hat das meist zu Folge, dass die KI nach der Erweiterung zwar mehr kann, jedoch seine ehemalige Hauptaufgabe nicht mehr so gut meistert wie zuvor.

3.2.2 Falsche Modearchitektur Wahl

Grundgedanke 'Je mehr Neuronen und Schichten eine Netz hat desto besser' ist falsch.

3.2.3 Falsches Lernverhalten

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

4 Sicherheitsprobleme durch BIAS

4.1 Angriff auf KI

Adversarial Attacks

4.2 Gefahren für Maschinen

Google KI -¿ Kühlung von Maschinen

4.3 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

Vollständigkeit prüfen
-¿ Fehler hier = Bias

Bias aus Daten entfernen

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.[**amazon**]

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[**Discriminating**].

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[**Gesundheitsversorgung**].
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[**heise**]. 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