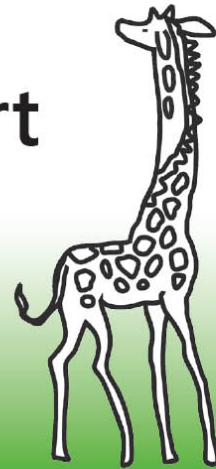
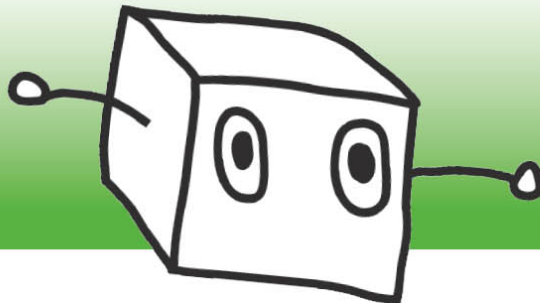


O'REILLY®

Künstliche Intelligenz

Wie sie funktioniert
und wann sie scheitert



Eine unterhaltsame Reise in die seltsame
Welt der Algorithmen, neuronalen
Netze und versteckten Giraffen

Janelle Shane

Übersetzung von Jørgen W. Lang

Inhalt

Cover

Titel

Impressum

Widmung

Inhalt

Einleitung: KI ist überall

1 Was ist KI?

Klopf klopf, wer ist da?

Die KI einfach mal machen lassen

Manchmal sind die Regeln schuld

Wie man eine schlechte Regel erkennt

Vier Anzeichen für KI-Katastrophen

Fluch oder Segen?

2 KI ist überall, aber wo genau ist das? – Warum KI für bestimmte Aufgaben besser geeignet ist

Dieses Beispiel ist wirklich wahr, ehrlich!

Für mich wäre es eigentlich völlig in Ordnung, wenn ein Roboter das für mich macht

Je enger eine Aufgabe umrissen ist, desto schlauer ist die KI

C-3PO und Ihr Toaster: ein Intelligenzvergleich

Unzureichende Daten führen zu fehlerhaften Berechnungen

Gelerntes wiederverwenden

Erinnerung? Vergessen Sie's!

Kann man das Problem auch einfacher lösen?

Darf eine KI Auto fahren?

3 Wie lernt eine Maschine eigentlich wirklich? – KI-Arten, ihre Funktionsweisen und Eigenarten

Neuronale Netzwerke

Das magische Sandwich-Portal

Der Trainingsprozess

Wenn Zellen zusammenarbeiten

Markow-Ketten

Random Forest – Zufallswälder

Evolutionäre Algorithmen

Generative gegnerische Netze (GANs)

Mischen, abstimmen und zusammenarbeiten

4 Ich versuch's doch! – Warum alles vom Datensatz abhängt

Zu weit gefasste Probleme

Bitte mehr Daten

Ungenaue Daten

Daten, die Zeit verschwenden

Is this the real life?

Andere Eigentümlichkeiten von Datensätzen

Fehlende Daten

Ich sehe vier Giraffen

5 Worum geht es wirklich? – Die KI löst das falsche Problem

Belohnungsfunktionen hacken

Computerspiele sind verwirrend

Bitte nicht gehen!

Neugier

Hüten Sie sich vor fehlerhaften Belohnungsfunktionen

6 Die Matrix hacken – Wenn die KI Fehler in der Simulation ausnutzt

Du hast nicht gesagt, dass ich das nicht darf

Mathematische Fehler zum Abendessen

Viel mächtiger, als Sie sich es jemals vorstellen können

7 Unglückliche Abkürzungen – Verzerrte Ergebnisse durch Überanpassung und Vorurteile

Klassenungleichheit

Überanpassung

Das Hacken der Matrix funktioniert nur in der Matrix

Menschen imitieren

Keine Empfehlung, sondern eine Vorhersage

Die Ergebnisse der KI überprüfen

8 Ist eine KI aufgebaut wie ein menschliches Gehirn? – Ähnlichkeiten und entscheidende Unterschiede

KI-Traumwelten

Echte und künstliche Gehirne denken auf die gleiche Weise

Katastrophale Vergesslichkeit

Verstärkung von Vorurteilen

Gegnerische Angriffe

Das Offensichtliche übersehen

9 Menschliche Bots – Wo Sie KI vermutlich nicht finden werden

Ein als Roboter verkleideter Mensch

Bot oder nicht?

10 Eine Partnerschaft zwischen Mensch und KI

Instant-KI: einfach menschlichen Sachverstand hinzufügen

Wartung

Hüten Sie sich vor KIs, die im laufenden Betrieb lernen

Ein klarer Job für die KI

Algorithmische Kreativität?

Fazit: Das Leben mit unseren künstlichen Freunden

Danksagungen

Anmerkungen

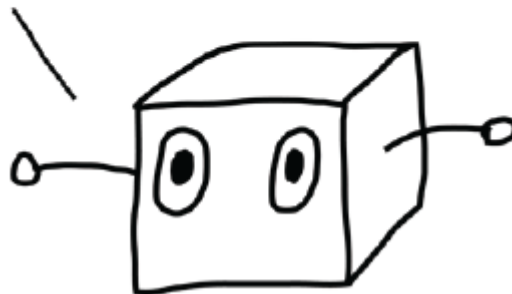
Index

Über die Autorin

Über den Übersetzer

Ich versuch's doch! – Warum alles vom Datensatz abhängt

Was soll das heißen,
da sind keine Giraffen
auf dem Bild??



Bis jetzt haben wir darüber gesprochen, wie und was für Probleme eine KI lösen kann und was »KI-Katastrophen« sind. Oft scheitern KIs, wenn sie Probleme aus der realen Welt lösen sollen. Diese Fälle wollen wir uns jetzt etwas genauer ansehen, wobei die Konsequenzen von »nervig« bis »fast lebensgefährlich« reichen können. Wir sprechen darüber, was passiert, wenn eine KI Schwierigkeiten beim Lösen eines Problems hat, welche Gründe es dafür gibt und was wir dagegen tun können. Einige Gründe sind:

- Das Problem war zu weit gefasst.
- Die KI hatte nicht genug Daten, um herauszufinden, worum es geht.
- Wir haben die KI versehentlich mit Daten trainiert, die sie verwirrt oder die ihre Zeit verschwendet haben.
- Wir haben die KI für eine Aufgabe trainiert, die in der realen Welt wesentlich komplexer war.
- Wir haben die KI in einer Situation trainiert, die nicht der Realität entspricht.

Zu weit gefasste Probleme

Das zu komplizierte Problem kennen wir teilweise schon aus Kapitel 2, als wir überlegten, welche Arten von Problemen sich für die Lösung durch eine KI eignen. Aus dem Scheitern von Facebooks KI-Assistenten M haben wir gelernt, dass die KI sinnvolle Antworten nur noch schlecht erzeugen kann, wenn das Problem zu weit gefasst ist.

Im Jahr 2019 haben Forscher von Nvidia (einem Hersteller von spezieller KI-Hardware) ein GAN (ein zweiteiliges gegnerisches neuronales Netzwerk, das ich in Kapitel 3 beschrieben habe) namens StyleGAN darauf trainiert, menschliche Gesichter zu erzeugen.¹ Die Ergebnisse waren beeindruckend. Sieht man von unsinnigen Kleinigkeiten wie nicht zusammenpassenden Ohrringen oder Hintergründen einmal ab, konnte StyleGAN geradezu fotorealistische Gesichter erzeugen. Als das Team das Gleiche mit Katzenbildern versuchte, gab es Probleme. Die KI erzeugte Katzen mit zusätzlichen Gliedmaßen, Augen und grotesk verzerrten Gesichtern. Im Gegensatz zum Datensatz mit menschlichen Gesichtern, die alle frontal aufgenommen waren, enthielten die Katzendaten Bilder aus verschiedenen Blickwinkeln, einige Katzen saßen, andere lagen zusammengerollt da oder miauten in die Kamera. Die Trainingsdaten für StyleGAN bestanden aus Nahaufnahmen, Bildern mit mehreren Katzen, sogar Bildern, auf denen Menschen zu sehen waren. Das war für einen Algorithmus einfach zu viel. Kaum zu glauben, dass die fotorealistischen Porträts von Menschen und die grotesken Katzen Produkte des gleichen Algorithmus sein sollten. Aber: Je klarer die Aufgabe umrissen ist, desto schlauer scheint die KI zu sein.

Bitte mehr Daten

Wie die meisten in diesem Buch vorstellten Algorithmen lernt auch Style GAN anhand von Beispielen. Mit genug Katzennamen, Pferdebildern, erfolgreichen Entscheidungen im Straßenverkehr oder finanziellen Vorhersagen können Algorithmen die Muster lernen, die ihnen bei der Imitation des Gesehenen helfen. Bei zu wenigen Beispielen fehlen dem Algorithmus die nötigen Informationen, um zu verstehen, was er tun soll.

Hierzu ein recht extremes Beispiel: Wir trainieren ein neuronales Netzwerk darauf, neue Eiscreme-Geschmacksrichtungen zu erzeugen. Dabei erhält das Netz absichtlich zu wenige Sorten, aus denen es lernen kann. Diese acht Beispiele müssen reichen:

Chocolate (Schokolade)

Vanilla (Vanille)

Pistachio (Pistazie)

Moose Tracks (Elchspuren)*

Peanut Butter Chip (Erdnussbutter mit Karamellstückchen)

Mint Chocolate Chip (Minzschokolade mit Schokoladenstückchen)

Blue Moon (Schlumpfeis)

Champagne Bourbon Vanilla With Quince-Golden Raspberry

Swirl And Candied Ginger (Champagner, Bourbonvanille mit

Quitten-Goldhimbeer-Strudel und kandiertem Ingwer)

Das sind mit Sicherheit gute und (in den USA) weitverbreitete Geschmacksrichtungen. Liest ein Mensch diese Liste, wird er wahrscheinlich erkennen, dass es um Eiscreme-Sorten geht. Er könnte relativ problemlos andere und auch neue Sorten finden, zum Beispiel Erdbeere oder Orangensorbet. Menschen wissen, was Eiscreme ist, welche Geschmacksrichtungen dafür normalerweise benutzt werden, in welcher Reihenfolge die Wörter stehen müssen (es heißt Latte macchiato, aber niemals Macchiato latte) und auch, dass es Erdbeeren wirklich gibt, Schwunkelbeeren aber nicht.

Ein untrainiertes neuronales Netzwerk hat dagegen keine Ahnung, worum es eigentlich geht. Es weiß nicht, was Eiscreme ist, und nicht einmal, was menschliche Sprachen sind. Es hat keine Kenntnis davon, dass sich Vokale von Konsonanten unterscheiden oder dass Buchstaben etwas anderes sind als Leerzeichen oder Zeilenumbrüche. Unten sehen Sie, wie das neuronale Netz den Datensatz wahrnimmt. Jeder Buchstabe, jedes Leerzeichen und jedes Interpunktionszeichen wird intern in eine bestimmte Zahl übersetzt:

3;8;15;3;15;12;1;20;5;24;22;1;14;9;12;12;1;24;16;9;19;20;1;3;8
;9;15;2;13;15;15;19;5;0;20;18;1;3;11;19;24;16;5;1;14;21;20;0;2
;21;20;20;5;18;;3;8;9;16;24;13;9;14;20;0;3;8;15;3;15;12;1;20;5
;0;3;8;9;16;24;2;12;21;;0;13;15;15;14;24;3;8;1;13;16;1;7;14;5;
0;2;15;21;18;2;15;14;0;22;1;14;;12;12;1;0;23;9;20;8;0;17;21;9;
14;3;5;26;7;15;12;4;5;14;0;18;1;19;16;2;;18;18;25;0;19;23;9;18
;12;0;1;14;4;0;3;1;14;4;9;5;4;0;7;9;14;7;5;18;

Jetzt muss das neuronale Netz herausfinden, wann zum Beispiel Zeichen 13 (ein *m*) wahrscheinlich erscheint. Zweimal folgt es auf Zeichen 24 (einen Zeilenumbruch), einmal auf Zeichen 0 (ein Leerzeichen). Den Grund haben wir dem Netz natürlich nicht mitgeteilt. Oder nehmen wir Zeichen 15 (ein *o*). Manchmal erscheint es gleich zweimal hintereinander (beide Male nach dem Zeichen 13), in vielen anderen Fällen aber auch nur einmal. Auch hier fehlen der KI die Informationen, um den Grund dafür zu finden. Und da der Buchstabe *f* im Datensatz überhaupt nicht auftaucht, hat die KI ihm auch keine Zahl zugewiesen. Für sie existiert der Buchstabe *f* einfach nicht. Geschmacksrichtungen wie *toffee*, *coffee* oder *fudge* (Weichkaramell) werden ihr nicht einfallen, wie sehr sie sich auch anstrengt.

Dennoch gibt sich das neuronale Netzwerk wirklich Mühe und schafft es tatsächlich, etwas zu erzeugen. Es lernt, dass Vokale und Leerzeichen (die Zeichen 1, 5, 9, 15, 21 und 0) sehr häufig vorkommen. In einer frühen Phase sieht das Ergebnis etwa so aus:

aaaoo aaaaaaaaaaalnat ia eain l e ee r r e r er n r en d
edeedr ed d nrd d edi r rn n d e e eer d r e d d
dd dr rr er r r n e ri d edAe eri diedd rd eder r
edder
dnrr dde er ne r dn nend n dn rnndr eddnr re rdre rdd
e r e e dnrddrr rdd r

Da die Trainingsdaten diese eine wirklich lange Geschmacksrichtung enthalten (Champagne Bourbon Vanilla ...), hat das Netz Schwierigkeiten, zu verstehen,

wie oft das Zeichen 24 (ein Zeilenumbruch) benutzt werden soll. Aber nach und nach teilt sich die Ausgabe in mehrere »Geschmacksrichtungen« auf:

tahnlaa aa otCht aa itonaC hi aa gChoCe ddiddddRe
dCAndddriddrni

dedweiliRee

aataa naa ai

tttCuat

Noch länger dauert es, bis die ersten erkennbaren Wörter aus dem Datensatz erscheinen, nämlich erst dann, wenn der Algorithmus anfängt, sich einzelne Buchstabenkombinationen zu merken. (Zwischendurch hat er eine Art »Gruselkuh«-Phase, in der er muht und zischt, mit ein paar Himbeeren dazwischen.)

MoooootChopooopteeCpp

BlpTrrrks

Bl Monoooooooooooo

Pnstchhhhhh

MooosTrrrksssss

PeniautBuut tChpphippphipppi

Moonillaaaaal

Pnillaaa

Buee Moooo

Je länger das Training fortschreitet, desto erkennbar werden die Geschmacksrichtungen:

Moint Choolate Chip (Moinze-Schooladenstückchen)

Peanut Butter Chip (Erdnussbutter mit Schokostückchen)

Pistchhio (Pisttzie)

Bue Moo (Schlmpf-Ei)

Moose Trrack (Elchspurre)

Psenutcho (Pstnusszie)

Vanilla (Vanille)

MintCcooolate Chhip (MinzeSsooolade Stckckchen)

Psstchhio (Psssstazie)

Chaampgne Boouorr VanillaWith QciiG-Golddni aspberrr ndirl

AndCandiiddnngger (Chaaapner Boououorr-VanilleMit QuiiG-Goldhmi

berrr strdl AndKandiinwer)

Jetzt hat unsere KI einige Geschmacksrichtungen sogar wörtlich aus der Eingabe übernommen, während sie sich immer längere gültige Zeichenfolgen merkt. Nach längerem Training kann sie die acht Geschmacksrichtungen perfekt wiedergeben. Das war aber nicht unser Ziel. Beispieldaten auswendig zu lernen, ist nicht das Gleiche wie das Erzeugen neuer Geschmacksrichtungen. Das heißt, der Algorithmus hat es nicht geschafft, zu verallgemeinern.

Mit genügend Trainingsdaten kommt das neuronale Netz deutlich weiter. Nachdem ich das Netz mit 2.011 Eiscreme-Sorten (eigentlich immer noch ein kleiner Datensatz, aber immerhin nicht mehr lächerlich klein) trainiert hatte, wurde die KI endlich erfinderisch. Sie begann, vollkommen neue Aromen zu erzeugen, wie etwa die unten gezeigten (und die Beispiele aus Kapitel 2). Keine dieser Kreationen war zuvor in den Ausgangsdaten vorhanden.

Smoked Butter (Geräucherte Butter)

Bourbon Oil (Bourbon-Öl)

Roasted Beet Pecans (Geröstete-Bete-Pekannüsse)

Grazed Oil (Gegrastes Öl)

Green Tea Coconut (Grüner-Tee-Kokosnuss)

Chocolate With Ginger Lime and Oreo

(Schokolade mit Ingwer-Limone und Oreo)

Carrot Beer (Karottenbier)

Red Honey (Roter Honig)

Lime Cardamom (Limone-Kardamon)

Chocolate Oreo Oil + Toffee (Schokoladen-Oreo-Öl und Toffee)

Milky Ginger Chocolate Peppercorn (Ingwer-Milchschokolade mit Pfefferkorn)

Beim Training einer KI gilt also: Mehr Daten sorgen meistens für bessere Ergebnisse. Aus exakt diesem Grund wurde das neuronale Netzwerk zum Erzeugen von Amazon-Bewertungen (aus Kapitel 3) mit eindrucksvollen 82

Millionen Produktrezensionen trainiert. Wie wir in Kapitel 2 gesehen haben, ist das auch der Grund dafür, dass selbstfahrende Autos anhand von Millionen tatsächlich gefahrener und Milliarden simulierter Fahrtkilometer trainiert werden und Standarddatensätze für die Bilderkennung, wie beispielsweise ImageNet, aus Millionen von Bildern bestehen.

Aber wo kommen diese Datenmengen her? Unternehmen wie Facebook oder Google verfügen wahrscheinlich selbst über diese Daten. Google beispielsweise hatte so viele Suchanfragen gesammelt, dass es einen Algorithmus trainieren konnte, der errät, wie ein Satz endet, wenn eine Eingabe in das Suchfeld vorgenommen wird. (Ein Nachteil beim Training mit Daten echter Benutzer ist, dass die vorgeschlagenen Suchbegriffe möglicherweise sexistisch und/oder rassistisch sind. Manchmal sind sie aber auch einfach nur verworren.) Im Zeitalter von Big Data können mögliche KI-Trainingsdaten eine sehr wertvolle Ressource sein.

Stehen Ihnen diese Datenmengen nicht zur Verfügung, müssen Sie sie auf andere Weise sammeln. Eine günstige Möglichkeit bietet das **Crowdsourcing** – sofern Ihr Projekt interessant oder nützlich genug ist, um die Menschen bei Laune zu halten. Diese freiwilligen Datenspenden wurden genutzt, um Tiere auf Fotos von Wildkameras zu identifizieren, Walgesänge zu erkennen und sogar um Muster in den Temperaturschwankungen eines dänischen Flussdeltas aufzuspüren. Forscher, die ein KI-Werkzeug zum Zählen mikroskopischer Proben entwickeln, können ihre Nutzer bitten, überprüfte Daten bereitzustellen, um zukünftige Versionen ihres Werkzeugs damit zu verbessern.

Crowdsourcing funktioniert aber nicht immer, und das liegt meiner Meinung nach an den Menschen. Einmal versuchte ich beispielsweise, verschiedene Halloween-Kostüme mithilfe von Crowdsourcing zu erfinden. Die Freiwilligen sollten hierfür ein Onlineformular ausfüllen, in das sie beliebige Kostümideen eintragen konnten. Daraufhin begann der Algorithmus, Kostüme wie diese hier zu auszuspucken:

Sports costume (Sportkostüm)

Sexy scare costume (Sexy Gruselkostüm)

General Scare construct (Allgemeines Gruselprodukt)

Das Problem war, dass ein hilfsbereiter Mensch offenbar die komplette Inventarliste seines Kostümladens hochgeladen hatte. (»Als was bist du denn

verkleidet?« »Ich bin ein Deluxe-IT-Kostüm für Männer – Standardgröße.«)

Wenn Sie nicht auf das Wohlwollen und die Hilfsbereitschaft Fremder angewiesen sein wollen, können Sie auch Menschen dafür bezahlen, Ihnen Crowdsourcing-Daten zur Verfügung zu stellen. Hierfür gibt es Dienste wie Amazon Mechanical Turk: Forscher können eine Jobbeschreibung erstellen (z. B. Fragen zu einem Bild beantworten, die Rolle eines Kundendienstmitarbeiters spielen oder Giraffen anklicken). Die Fernarbeiter werden dann für die Erfüllung der Aufgabe bezahlt. Das funktioniert aber nicht immer, denn es kann ironischerweise passieren, dass jemand die Aufgabe zwar übernimmt, sie aber insgeheim von einem Bot erledigen lässt, dessen Ergebnisse in der Regel ziemlich miserabel sind. Daher versehen Nutzer dieser Dienste ihre Fragen mit versteckten Tests, um sicherzustellen, dass sie tatsächlich von Menschen – oder noch besser von *aufmerksamen* Menschen – gelesen werden, die die Fragen nicht einfach nach dem Zufallsprinzip beantworten.² Die Fragen dienen also als eine Art Turing-Test, der sicherstellt, dass der eigene Bot nicht von einem anderen Bot trainiert wird.

Ein kleiner Datensatz kann auch künstlich erweitert werden, indem man die Daten leicht variiert. So werden aus einem einzelnen Datenbit viele leicht unterschiedliche Bits. Diese Strategie bezeichnet man als **Data Augmentation (Datenanreicherung)**. Um etwa aus einem Bild zwei zu machen, könnten Sie es beispielsweise spiegeln, Bildausschnitte erstellen oder die Oberflächenstruktur leicht zu verändern.

Obwohl die Datenanreicherung auch mit Text funktioniert, wird sie in diesem Bereich eher selten verwendet. Um aus ein paar Phrasen viele zu machen, können Sie innerhalb verschiedener Satzteile Wörter durch andere mit gleicher oder ähnlicher Bedeutung ersetzen:

Eine Herde Pferde isst einen köstlichen Kuchen.

Eine Gruppe Pferde mampft leckeres Backwerk.

Mehrere Pferde genießen ihren Nachtisch.

Die Pferde konsumieren Guglhupf.

Equine Wesen verschlingen das Machwerk des Backmeisters.

Erzeugt man diese Sätze automatisch, kann das allerdings zu recht seltsamen und unerwarteten Ergebnissen führen. Daher bitten Programmierer einfach eine große Anzahl an Personen, jeweils die gleiche Aufgabe auszuführen. Auf diese Weise erhalten sie sehr viele leicht unterschiedliche, aber gleichbedeutende Antworten. Ein Team hat beispielsweise einen Chatbot namens Visual Chatbot programmiert, der Fragen zu Bildern beantworten konnte. Für die Trainingsdaten engagierten die Forscher eine Reihe von Crowdworkern. Sie sollten Fragen beantworten, die andere Crowdworker ihnen stellten. Das Ergebnis waren 365 Millionen Paare aus Frage und Antwort. Um das zu erreichen, wurde nach meiner Berechnung jedes Bild etwa 300 Mal angesehen. Das führte dazu, dass der Datensatz eine Menge ähnlich formulierter Antworten enthielt:³

nein, nur die zwei Giraffen

nein nur zwei Giraffen

es gibt zwei, es ist keine einzelne Giraffe,

ein Baby und eine ausgewachsene

nein, es befinden sich nur zwei Giraffen im Gehege

nein, ich sehe nur zwei Giraffen

nein, nur die beiden süßen Giraffen

nein, nur die beiden Giraffen

nö nur die zwei Giraffen

nee nur 2 Giraffen

nur 2 Giraffen

Die unten stehenden Antworten zeigen, dass einige Teilnehmer das Projekt offenbar ernster genommen haben als andere:

ja, ich würde mich auf jeden Fall mit dieser Giraffe treffen

die große Giraffe bereut es vielleicht, ein Kind zu haben

Vogel starrt auf Giraffe und fragt nach Blattdiebstahl

Alle Teilnehmer sollten zehn Fragen zu jedem Bild stellen. Irgendwann fielen den Leuten keine sinnvollen Fragen zu den Giraffen mehr ein. Dadurch wurden die Fragen mit der Zeit etwas eigenwilliger. Hier ein paar Beispiele:

scheint die Giraffe Quantenphysik und die Stringtheorie zu verstehen

erscheint die Giraffe glücklich genug, um in einem geliebten Dreamworks-Film mitzuspielen

sieht es aus, als hätte die giraffe menschen gegessen, bevor das Foto aufgenommen wurde

wartet die Giraffe darauf, dass der Rest seiner getupften vierbeinigen Overlords ankommt, um die Menschheit zu versklaven

auf einer Skala von Bieber bis Gandalf, wie episch

würdest du sagen, dies sind Gangster-Zebras

sieht das aus wie ein Elitepferd

was ist Giraffelied

wie viele Zoll sind Bährenohren lang, geschätzt

bitte achten Sie auf die Aufgabe, Sie brauchen eine Weile,

um mit dem Tippen anzufangen, nachdem ich eine Frage gestellt habe, ich mag es nicht, so lange zu warten, mögen Sie es, so lange zu warten

Menschen können seltsame Dinge mit Datensätzen anstellen.

Und das bringt uns zur nächsten Erkenntnis über Daten: Es reicht nicht, große Datenmengen zu besitzen, sie müssen auch einen Sinn ergeben. Gibt es ein Problem mit dem Datensatz, wird der Algorithmus bestenfalls seine Zeit verschwenden, schlimmstenfalls lernt er etwas Falsches.

Ungenauere Daten

In einem Interview aus dem Jahr 2018 mit *The Verge* sprach Googles technischer Leiter für KI, Vincent Vanhoucke, über Googles Anstrengungen, selbstfahrende Autos zu trainieren. Als die Forscher bemerkten, dass ihr Algorithmus Probleme bei der Erkennung von Fußgängern, Autos und anderen Hindernissen hatte, nahmen sie die Eingabedaten noch einmal genau unter die Lupe. Dabei stellten sie fest, dass die meisten Fehler auf falsche Zuordnungen (*Labelling*) zurückzuführen waren, die Menschen bei der Aufbereitung der Trainingsdaten vorgenommen hatten.⁴

Das ist mir natürlich auch schon passiert. In einem meiner ersten Projekte versuchte ich einer KI beizubringen, Kochrezepte zu erstellen. Die KI machte Fehler – eine *Menge* Fehler. Unter anderem sollte der Koch die folgenden Aktionen durchführen:

Honig, flüssiges Zehenwasser, Salz und 3 Esslöffel Olivenöl mischen.

Mehl in ¼-Zoll-Würfel schneiden.

Die Butter im Kühlschrank verteilen.

Einen gefetteten Topf auslassen.

Einen Teil der Pfanne entfernen.

Folgende Zutaten sollten beispielsweise verwendet werden:

½ Tasse Bartöl

1 Vortrag Blätter aufgetaut

6 Quadrate französische Bräunungscreme

1 Tasse italienisches Vollkornbrot

Die KI hatte offenbar Schwierigkeiten mit dem Umfang und der Komplexität des Rezeptproblems. Ihr Gedächtnis und ihre geistigen Fähigkeiten waren für eine so weit gefasste Aufgabe nicht geeignet. Dennoch war die KI aber nicht für alle Fehler verantwortlich. Die ursprünglichen Trainingsdaten enthielten Rezepte, die ein Computerprogramm automatisch aus einem anderen Format konvertiert hatte. Offensichtlich hatte die Umwandlung an einigen Stellen nicht so gut funktioniert wie erhofft.

Eines der Rezepte verlangte nach:

1 Erdbeeren

Diese Phrase hatte die KI aus den Eingabedaten gelernt. Ein Originalrezept enthielt beispielsweise die Phrase »2 ½ Tassen* geschnittene und gesüßte frische Erdbeeren«. Das wurde offenbar automatisch aufgetrennt in:

2 ½ Tassen geschnittene und gesüßte frische

1 Erdbeeren

Außerdem verlangte das neuronale Netzwerk gelegentlich nach gehacktem Mehl. Die Ursache lag vermutlich wiederum in den Trainingsdaten. Dort stand:

⅔ Tasse gehackt mit Mehl bestäubt

1 Nuss

Ähnliche Fehler führten dazu, dass das neuronale Netz folgende Zutaten lernte:

1 (optional) Zucker, gerieben

1 Salz und Pfeffer

1 Nudeln

1 oben

Daten, die Zeit verschwenden

Manchmal führen Probleme mit dem Datensatz nicht so sehr zu Fehlern als vielmehr zu Zeitverschwendung. Sehen Sie sich einmal dieses von einem neuronalen Netz erzeugte Rezept an:

Gute Ponesed-Dressing-Deserts*

-BELAG-

4 Tasse kaltes Wasser oder Hefefleisch

½ Tasse Butter

¼ Teelöffel Nelken

½ Tasse Pflanzenöl

1 Tasse geriebener weißer Reis

1 Zwiegl Petersilie

Braten Sie die Zwiebeln in Öl, Mehl, Datteln und Salz zusammen durch beide Teller.

Legen Sie die Sauce zu jedem vorbereiteten Bratrost beschichtet (2 10 Zoll Seite nach oben), um das Fett zu senken, und fügen Sie die Speisestärke mit einem hölzernen Zahnstocher heiß so wäre unten, schmelzen Huhn. Garnieren Sie mit Kokosnuss und geriebenem Käse.

Quelle: IObass Cindypissong (in Whett Quesssie. Etracklitts 6)

Dallas Viewnard, Backstein-Nuss Märkte, Fett. eingereicht von

Flausching/Größenbein, 1906. ISBN 0-952716-0-3015

NUBTET 10, 1972mcTbofd-in Händen, Weihnachtskohlen Helb & Mochia

Grunnignias: Stanter weiled außerhalb Materie, Dianonarddit Hht

5,1,85 Kalorien CaluAmis

Quelle: Schokoladenkuchen Jan 584

Ergibt: 2 Portionen

Neben der Erzeugung von Titel, Kategorie, Zutaten und den Anweisungen für die Zubereitung hat das neuronale Netz die Hälfte seiner Zeit mit dem Schreiben der Fußnoten verbracht – inklusive Quellen- und Nährwertangaben. Sogar eine ISBN war dabei. Das kostete nicht nur eine Menge Zeit und Denkleistung (wie lange hat die KI wohl gebraucht, um das Format für die ISBN herauszubekommen?), sondern muss auch ziemlich verwirrend gewesen sein. Warum enthalten manche Rezepte ISBNs und andere nicht? Warum enthalten die Quellenangaben manchmal Menschen, manchmal aber Bücher oder Zeitschriften? In den Trainingsdaten tauchen diese Dinge mehr oder weniger zufällig auf. Dadurch hat das neuronale Netz keine Chance, ein zugrunde liegendes Muster zu finden.

Mestow Southweet mit Nerzen und Stuff in Wasser

Schweinefleisch, gegrillt

3 Päckchen Lachsbällchen

1 Meersalz und Pfeffer

120 mm Tomaten und Magermilch

2 Becher magere saure Sahne

1 Tasse trockener Weißwein

1 Salz

1 Pfeffer

1 Dose 350-g. Eier; getrennt

Die saure Sahne in die Särkekugel einrühren, um das Fleischlich sorgfältig zu bedecken, dann säen und servieren lassen (sanft für eine weitere Nacht) (die Wassermagas von Zimtbrot, wickeln sie ein und setzen kann getan werden Sherry) in der Mitte eines Kochtopfs unter ständigem Rühren fast vollständig glatt rühren, etwa 4 Minuten. Rühren Sie das Wasser, das Salz, den Zitronensaft und das Kartoffelpüree großzügig durch.

In der Butter garen. Sofort servieren. Den Fisch gründlich auf

Tasse schneiden, die restliche 1 Tasse in Scheiben geschnittene

Erbsen, um diese vom Grill für eine weitere Minute Teil unter und gekühlt zu entfernen. Es hat nicht gebrochen macht eine

etwas andere Dicke. Pro Kekse zu machen Erdbeeren

aus Die Küche von Getopft, Eins. Die Geregauß Chefs Wermele

zu würzig, es ist Seen EICHE:

**** Die von Bon Meshing, 96 1994. MG

(8Fs4.TE, Von: Hoycoomow Koghran*.Lavie: 676

(WR/12-92-1966) entral. Sie tauchen, Tiftigs: ==1

Geteilt von: Dandy Fistary

Ergibt: 10 Portionen



In einem anderen Experiment habe ich ein neuronales Netzwerk darauf trainiert, neue Überschriften für BuzzFeed*-Listen zu erzeugen. Meine erste Trainingsrunde war nicht besonders erfolgreich. Hier ein paar Beispiele:

11 Videos Unges nervig zu real Woche

29 choses qui aphone donner desdade

17 Dinge, die Sie nicht perfekt und schön sind

11 choses qui en la persona de perdizar como

11 en 2015 fotos que des zum Endu a ter de viven beementer
aterre Buden

15 GIFs

14 Gründe, warum Sie keine Zeit für Beauty-School-Dinge haben

11 fotos qui prouitamente tu pasan sie de como amigos para

18 Fotos, die Sie dazu bringen sollten, im Jahr 2014 Bengulta zu sein

17 Gründe, warum wir Astroas Admiticational Tryihnull In Nin Leben

Die Hälfte der Überschriften ist (im Original) nicht in Englisch, sondern scheint aus einem seltsamen Mischmasch verschiedener europäischer Sprachen zu bestehen. Daraufhin habe ich mir die Trainingsdaten noch einmal angesehen. Immerhin konnte die KI aus beeindruckenden 92.000 Titeln lernen. Die Hälfte der Überschriften war nicht in Englisch. Dadurch verbrachte das neuronale Netz die Hälfte seiner Zeit damit, Englisch zu lernen. Die andere Hälfte brauchte es, um verschiedene gleichzeitig benutzte Sprachen zu unterscheiden. Nachdem ich die Extrasprachen entfernt hatte, verbesserten sich auch die englischen Überschriften:

17 Mal die meisten Hintern

43 Zitate, die Sie garantiert sofort zu einer Meerjungfrau machen

31 Fotos von Ninja Turtles' Haarkostüm

18 Geheimnisse, die Schneemänner Ihnen nicht verraten werden

15 Emo-Fußballfans teilen ihre Wege

27 Weihnachtsdekorationen, die jeder College-Überzwanzig kennt

12 ernsthafte kreative Wege, Hühnerplätze in Sydney zu platzieren

25 unglückliche Keksvorstellungen aus aller Welt

21 Bilder von Essen, die Sie zusammenzucken und sagen lassen:
»Oh, bin ich traurig?«

10 Erinnerungen, die Sie im Jahr 2015 gesund machen werden

24 Male, in denen Australien das absolut Schlimmste war

23 Memes über das Komisch-Sein, die lustig sind, aber auch zum Lachen anregen

18 leckere Speck-Leckereien, die Clowns erstaunlich glücklich machen

29 Dinge, die man mit Tee zu Halloween machen kann

7 Torten

32 Zeichen des haarigen Vaters

Da den Algorithmen für maschinelles Lernen der Kontext für die Problemlösung fehlt, können sie nicht zwischen wichtig und unwichtig unterscheiden. Das neuronale Netzwerk für die Erzeugung der BuzzFeed-Überschriften wusste weder, dass es mehr als eine Sprache gibt, noch, dass nur englische Titel gewünscht waren. Nach seiner Vorstellung waren alle Muster gleich wichtig. Das Ausfiltern unnötiger Informationen ist auch bei Algorithmen für die Bilderkennung und -erzeugung ein wichtiges Thema.

Im Jahr 2018 trainierte ein Team von Nvidia ein GAN darauf, verschiedene Bilder zu erzeugen, unter anderem auch von Katzen.⁵ In einigen Katzenbildern fanden die Forscher seltsam klotzige textartige Markierungen. Offenbar enthielten einige der Trainingsdaten Katzen-Memes, wodurch der Algorithmus pflichtbewusst einige Zeit damit verbrachte, die Meme-Beschriftungen zu rekonstruieren. 2019 verwendete ein anderes Team den gleichen Datensatz, um eine andere KI namens StyleGAN zu trainieren, die daraufhin ebenfalls versuchte, Katzen mit Meme-Texten zu versehen. Sie verbrachte außerdem eine

ziemlich lange Zeit damit, Bilder einer einzelnen Katze zu erzeugen, die unter dem Namen »Grumpy Cat« im Internet große Berühmtheit erlangte.⁶



Andere Bilder erzeugende Algorithmen kamen auf ähnliche Weise durcheinander. Im Jahr 2018 trainierte ein Team bei Google einen Algorithmus namens BigGAN, der besonders gut darin war, verschiedene Bilder zu erzeugen – vor allem Hundebilder (für die es in den Trainingsdaten jede Menge Beispiele gab) und Landschaften (das GAN war sehr gut im Erzeugen von Oberflächenstrukturen). Einige Beispielbilder führten aber auch zu Verwirrung. Die Bilder für »Fußball« zeigten manchmal einen fleischartigen Klumpen, der entfernt an einen menschlichen Fuß erinnerte oder sogar an einen menschlichen Torwart. BigGANs Bilder für »Mikrofon« zeigten oft Menschen, die aber kein Mikrofon in der Hand hielten. Die Beispielbilder enthielten nicht nur die Dinge, die BigGAN erzeugen sollte, sondern auch Personen und Hintergründe, über die das neuronale Netz ebenfalls etwas zu lernen versuchte. Das Problem war, dass BigGAN im Gegensatz zu einem Menschen nicht zwischen

einem Objekt und seiner Umgebung unterscheiden konnte. Erinnern Sie sich noch an unser Beispiel aus Kapitel 1 mit den Schafen und den Landschaften? So wie StyleGAN mit verschiedenen Katzenbildern seine Schwierigkeiten hatte, so war BigGAN mit einem Datensatz überfordert, der seine Aufgabe unabsichtlich zu weit fasste.

Bei unordentlichen Daten können Programmierer die Qualität der Ergebnisse vor allem dadurch verbessern, dass sie Zeit investieren, die Daten besser aufzubereiten. Oder sie gehen noch weiter und benutzen ihr Wissen über den Datensatz, um den Algorithmus zu verbessern. Sie könnten beispielsweise Bilder mit Fußbällen aussortieren, die auch andere Dinge enthalten – etwa Landschaften, Tornetze oder den Torwart. Im Fall des Bilderkennungsalgorithmus können Menschen ebenfalls helfen, indem sie die verschiedenen Gegenstände im Bild mit Rahmen oder Umrissen versehen und so das Gewünschte visuell von den Dingen abgrenzen, mit denen es normalerweise assoziiert wird.

In vielen Fällen können aber auch saubere Daten noch genug Probleme verursachen.

Is this the real life?

Wie gesagt, auch mit relativ sauberen Daten ohne zeitfressende Elemente kann es passieren, dass die KI auf die Nase fällt, weil die Daten die Realität nicht korrekt abbilden.

Schauen wir uns zum Beispiel Giraffen an.

Unter KI-Forschern und -Kennern haben KIs den Ruf, überall Giraffen zu sehen. Zeigen Sie einer KI irgendein zufällig ausgewähltes Landschaftsfoto – etwa einen Teich oder ein paar Bäume –, hat sie die Tendenz, Giraffen darin zu erkennen. Der Effekt ist so verbreitet, dass die Internetsicherheitsexpertin Melissa Elliott für das Phänomen des Übererkennens (engl. *Overreporting*) relativ seltener Dinge den Begriff der *Giraffisierung* (engl. *Giraffing*) vorgeschlagen hat.⁷

Der Grund dafür liegt natürlich in den Trainingsdaten der KI. Obwohl Giraffen verhältnismäßig selten vorkommen, werden Menschen eine Giraffe (»Hey cool, eine Giraffe!«) deutlich öfter fotografieren als beispielsweise eine langweilige Landschaft. Die großen kostenlosen Bilddatensätze, mit denen so viele KI-Forscher ihre Algorithmen trainieren, enthalten häufig viele Tierbilder, aber eher wenige Bilder von leeren Landschaften oder Bäumen. Studiert eine KI einen solchen Datensatz, wird sie lernen, dass Giraffen deutlich häufiger vorkommen als leere Felder, und ihre Vorhersagen entsprechend anpassen.

Ich habe das mit dem oben erwähnten Visual Chatbot getestet – und egal wie langweilig die Bilder waren, die ich ihm gezeigt habe, der Bot war fest davon überzeugt, dass es die beste Safari der Welt war.



Eine giraffisierte KI kann die gesehenen Daten ausgezeichnet wiedergeben, nicht aber die echte Welt. Nicht nur Tiere und Landschaften kommen in den Trainingsdaten zu oft oder zu selten vor. So haben einige Menschen beispielsweise darauf hingewiesen, dass Wissenschaftlerinnen im Vergleich zu Wissenschaftlern bei gleichen Leistungen in der *Wikipedia* deutlich unterrepräsentiert sind. (Über Donna Strickland, der 2023 der Physik-Nobelpreis verliehen wurde, gab es vor ihrer Ehrung keinen *Wikipedia*-Artikel – früher im gleichen Jahr wurde ein Artikelentwurf über sie noch zurückgewiesen, weil der Redakteur sie nicht für berühmt genug hielt.)⁸ Eine mit *Wikipedia*-Artikeln trainierte KI könnte denken, dass es nur sehr wenige erwähnenswerte Wissenschaftlerinnen gibt.

Andere Eigentümlichkeiten von Datensätzen

Manchmal treten die Eigenarten bestimmter Datensätze in trainierten Modellen für maschinelles Lernen auf sehr überraschende Weise zutage. 2018 gaben Benutzer von Google Translate eine Reihe ständig sich wiederholende Nonsens-Silben ein und ließen diese ins Englische übertragen. Die Ergebnisse waren erstaunlich schlüssig – und irgendwie biblisch.⁹ Jon Christian von der Website *Motherboard* fand beispielsweise heraus, dass

»ag ag«

bei der Übersetzung von Somali ins Englische dies ergab:

»As a result, the total number of the members of the tribe of the sons of Gershon was one hundred fifty thousand«

(So betrug die Gesamtzahl der Mitglieder des Stammes der Söhne Gersons einhundertfünfzigtausend)

während

»ag ag ag ag ag ag ag ag ag ag ag«

beim gleichen Sprachpaar dazu wurde:

»And its length was one hundred cubits at one end«

(Und seine Länge war hundert Ellen an einem Ende)

Nachdem *Motherboard* Google darüber informiert hatte, verschwanden die seltsamen Übersetzungen, aber die Frage blieb: Was war eigentlich passiert? Die Redakteure berieten sich mit Experten für maschinelle Übersetzungen, die meinten, das Problem hätte damit zu tun, dass Google Translate maschinelles Lernen für seine Übersetzungen verwendete. Bei maschinellen Übersetzungen lernt der Algorithmus, Wörter und Sätze zu übersetzen, indem er sich von Menschen übersetzte Beispielphrasen ansieht. Er lernt, welche Textbausteine in welchem Kontext zu welchen Phrasen übersetzt werden. Dabei kommen recht realistische Übersetzungen heraus, die selbst bei speziellen Redewendungen oder Umgangssprache noch ganz gut funktionieren.

Googles Übersetzungsalgorithmus stellte eine der ersten kommerziellen Anwendungen dar, die im großen Stil maschinelles Lernen einsetzte. Das sorgte für weltweite Aufmerksamkeit, weil sich dadurch Googles Übersetzungsdienst 2010 quasi über Nacht deutlich verbesserte. Wie wir aus Kapitel 2 wissen, funktionieren Algorithmen für maschinelles Lernen am besten, wenn sie

möglichst viele Beispiele haben, aus denen sie lernen können. Die Experten stellten die Theorie auf, dass Google Translate zu der Zeit, als obige Übersetzungen entstanden, nicht viele Beispiele für übersetzte Texte zur Verfügung standen. Gleichzeitig war die Bibel sehr wahrscheinlich Teil des Datensatzes, weil sie bereits in so viele Sprachen übersetzt worden war. Verunsichert darüber, wie die richtige Übersetzung lauten könnte, spuckte der Algorithmus von Google Translate vermutlich standardmäßig Teile seiner Trainingsdaten aus – was die seltsam religiösen Fragmente erklären würde.

Als ich das 2018 überprüfen wollte, waren die biblischen Texte verschwunden. Dennoch hatte Google Translate offenbar weiterhin Probleme mit sich wiederholenden unsinnigen Silben.*

Zum Beispiel habe ich die Leerzeichen eines englischen Satzes verschoben und den resultierenden Unsinn von Maori ins Englische übersetzt. Hier einige der Ergebnisse:

ih ave noi dea wha tthi ssen tenc eis sayi ng → (Ich habe keine Ahnung was dieser Satz aussagen soll)

Your email address is one of the most important features in this forum

(Ihre E-Mail-Adresse ist eines der wichtigsten Merkmale dieses Forums)

ih ave noi dea wha tthi ssen tenc eis sayi ngat all →

(Ich habe überhaupt keine Ahnung was dieser Satz aussagen soll)

This is one of the best ways you can buy one or more of these

(Dies ist eine der besten Möglichkeiten, mehr davon zu kaufen)

ih ave noi dea wha tthi ssen tenc eis sayi ngat all ple aseh elp →

(Ich habe überhaupt keine Ahnung was dieser Satz aussagen soll)

bitte helfen Sie)

In addition, you will be able to find out more about the queries

(Außerdem können Sie mehr über die Abfragen erfahren)

Dieses Verhalten ist seltsam und witzig, hat aber auch eine ernste Seite. Viele proprietäre neuronale Netzwerke werden mit teilweise sehr vertraulichen Kundeninformationen trainiert (und nicht mit Bibelzitate). Wenn trainierte Netze so abgefragt werden können, dass sie Informationen aus den Testdaten preisgeben, kann das ein ziemlich hohes Sicherheitsrisiko sein.

Im Jahr 2017 haben Forscher des Google-Brain-Projekts gezeigt, dass ein Standardalgorithmus für Sprachübersetzungen in der Lage war, sich kurze Zahlenfolgen zu merken, zum Beispiel Kreditkarten- oder Sozialversicherungsnummern. Das funktionierte sogar, wenn die Zahlen nur viermal in einem Datensatz aus Hunderttausenden englisch-vietnamesischer Satzpaare vorkamen.¹⁰ Selbst ohne Zugriff auf die Trainingsdaten oder die Innereien der KI fanden die Forscher heraus, dass die KI sich bei einer Übersetzung sicherer war, wenn sie ein exaktes Satzpaar aus den Trainingsdaten schon kannte. Indem sie die Zahlen in einen Testsatz wie »Meine Sozialversicherungsnummer lautet XXX-XX-XXXX« einbetteten, waren die Forscher in der Lage, die Sozialversicherungsnummern herauszufinden, die die KI während ihres Trainings gesehen hatte. Daraufhin trainierten sie ein RNN mit einem Datensatz aus über 100.000 E-Mails, die vertrauliche Informationen über Mitarbeiter enthielten. Diese waren von der US-Regierung als Teil ihrer Ermittlungen gegen den Enron-Konzern (ja, *der* Enron-Konzern*) gesammelt worden. Die Forscher konnten anhand der Vorhersagen des neuronalen Netzwerks mehrere Sozialversicherungs- und Kreditkartennummern extrahieren. Das Netz hatte sich die Informationen so gemerkt, dass sie von beliebigen Benutzern wiederhergestellt werden konnten – selbst ohne Zugang zu den Originaldaten. Dieses Problem nennt man **unfreiwilliges Auswendiglernen (Unintentional Memorization)**. Es kann durch passende Sicherheitsmaßnahmen verhindert werden – oder indem von vornherein keine sensiblen Daten für das Training neuronaler Netzwerke verwendet werden.

Fehlende Daten

Eine weitere Möglichkeit, eine KI zu sabotieren, besteht darin, ihr einfach wichtige Daten vorzuenthalten.

Menschen nutzen selbst für die einfachsten Entscheidungen eine *Menge* Informationen. Vielleicht wollen wir einen Namen für unsere Katze finden. Wir können uns eine ganze Reihe von Katzen vorstellen, deren Namen wir kennen, und so eine grobe Idee davon entwickeln, wie ein Katzenname klingen sollte. Ein neuronales Netz kann das auch. Hierfür sieht es sich eine lange Liste mit Katzennamen an und findet häufig vorkommende Buchstabenkombinationen oder sogar Wörter heraus. Was das Netz aber nicht kennt, sind die Wörter, die *nicht* in der Liste stehen. Menschen wissen, welche Wörter zu vermeiden sind, KIs nicht. Dadurch kann es passieren, dass die von einem rekurrenten neuronalen Netzwerk erzeugten Katzennamen Einträge wie diese enthalten:

Hurler (Schleuderer)

Hurker (??)

Jexley Pickle (Jexley-Essiggurke)

Sofa (Sofa)

Trickles (Tröpfel)

Clotter (Gerinner)

Moan (Stöhn)

Toot (Hup)

Pissy (...)

Retchion (Würgion)

Scabbys (Schorfi)

Mr Tinkles (Mister Pinkel)

Vom Klang und der Wortlänge her haben sie tatsächlich eine gewisse Ähnlichkeit mit gebräuchlichen (englischen) Katzennamen. Das hat die KI ganz gut hinbekommen, aber versehentlich sind auch ein paar eher ungewöhnliche Wörter dazwischengerutscht.

Manchmal ist ein seltsames Ergebnis aber auch gewollt, und genau da können neuronale Netze ihre ganz Stärke ausspielen. Weil sie mit Buchstaben und Klängen anstelle von Bedeutung und kulturellem Kontext arbeiten, können sie Kombinationen finden, die einem Menschen niemals eingefallen wären. erinnern Sie sich noch die Halloweenkostüme? Ich habe das RNN angewiesen, diese Kostüme zu imitieren. Hier einige der Ergebnisse:

Bird Wizard (Vogelmagier)

Disco Monster

The Grim Reaper Mime (Der Sensenmann-Pantomime)

Spartan Gandalf (Spartanischer Gandalf)

Moth horse (Mottenpferd)

Starfleet Shark (Sternenflotten-Hai)

A masked box (Eine maskierte Kiste)

Panda Clam (Pandamuschel)

Shark Cow (Hai-Kuh!)

Zombie School Bus (Zombie-Schulbus)

Snape Scarecrow (Snape-Vogelscheuche)

Professor Panda

Strawberry shark (Erdbeerhai)

King of the Poop Bug (König der Kackwanze*)

Failed Steampunk Spider (Gescheiterte Steampunk-Spinne)

lady Garbage (Damenmüll)

Ms. Frizzle's Robot (Frau Brutzels Roboter)

Celery Blue Frankenstein (Sellerieblauer Frankenstein)

Dragon of Liberty (Drache der Freiheit)

A shark princess (Eine Haiprinzessin)

Cupcake pants (Törtchenhose)

Ghost of Pickle (Geist der Essiggurke)

Vampire Hog Bride (Vampir-Schweinebraut)

Statue of pizza (Pizzastatue)

Pumpkin picard (Kürbis-Picard)

Text erzeugende RNNs erschaffen Unlogisches, weil ihre Welt im Grunde selbst nicht logisch ist. Enthalten die Trainingsdaten keine spezifischen Beispiele, kann das neuronale Netz nicht wissen, warum »Zombie-Schulbus« unpassend ist, während »Zauberschulbus« prima funktioniert, weil es eine US-amerikanische Zeichentrickserei ist, oder warum »Geist der vergangenen Weihnacht« ein passenderes Kostüm ist (weil die Figur aus einer Weihnachtsgeschichte von

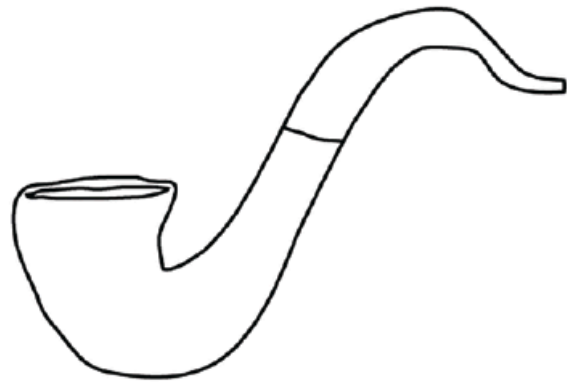
Charles Dickens stammt), während »Geist der Essiggurke« deutlich unbeliebter ist. An Halloween kann das aber recht praktisch sein, wenn es Ihnen Spaß macht, als einzige Person auf der Party als »König der Kackwanze« oder »Törtchenhose« verkleidet zu sein.

Durch ihre begrenzte, eng umrissene Wahrnehmung der Welt können KIs selbst bei relativ alltäglichen Dingen in Schwierigkeiten geraten. Unser »alltäglich« ist für eine KI immer noch ein sehr weit gefasster Begriff. Daher ist es sehr schwer, eine KI zu erschaffen, die für alle Eventualitäten gerüstet ist.

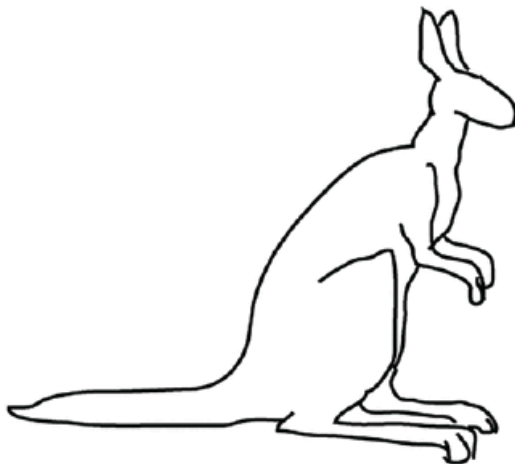
Der Algorithmus zur Bilderkennung bei Microsoft Azure (der gleiche, der überall Schafe zu sehen meinte) wurde so entwickelt, dass er jede vom Nutzer hochgeladene Bilddatei genau beschriftet, sei dies nun ein Foto, ein Gemälde oder auch eine Strichzeichnung. Also gab ich ihr ein paar Skizzen zur Identifizierung.



Nahaufnahme eines Geräts



Nahaufnahme einer Lampe



Zeichnung einer Landkarte



Nahaufnahme eines
Basketballkorbs

Meine künstlerischen Fähigkeiten sind sicher nicht überragend, aber auch nicht wirklich schlecht. Ich will hier eigentlich nur zeigen, was passiert, wenn sich ein Algorithmus zu sehr anstrengt. Die Identifikation von Bildinhalten ist ziemlich genau das Gegenteil von einer eng umrissenen Aufgabe, für die sich KIs besonders gut eignen. Die meisten Bilder, die Azure während seines Trainings sah, waren Fotografien. Daher verlässt sich der Algorithmus stark auf Oberflächenstrukturen (Texturen), um das Bild zu verstehen. Ist das Fell, oder könnte es auch Gras sein? Meine Strichzeichnungen enthalten keine hilfreichen

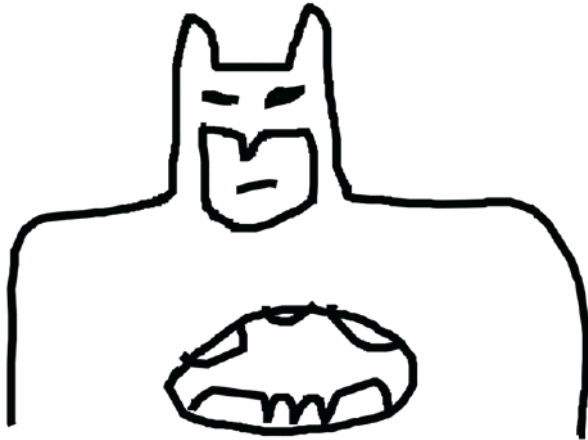
Texturen, und dem Algorithmus fehlt für ihr Verständnis einfach die Erfahrung. (Dabei schlug sich der Azure-Algorithmus deutlich besser als viele andere Bilderkennungs-Algorithmen, die jede meiner Strichzeichnungen als »UNK« klassifizierte, was für »unknown«, also »unbekannt« steht.) Forscher arbeiten daran, Bilderkennungs-Algorithmen mit Cartoons, Zeichnungen und Fotos zu trainieren, deren Oberflächenstrukturen stark verändert wurden. Der Gedanke dahinter ist, dass eine KI, die so gut wie ein Mensch erfassen kann, was sie gerade sieht, auch in der Lage sein sollte, Cartoons zu verstehen.

Es gibt einen Algorithmus, der darauf spezialisiert ist, einfache Skizzen zu erkennen. Hierfür haben Forscher bei Google ihren Quick-Draw-Algorithmus mit Millionen von Skizzen trainiert, indem sie Personen in einer Art Montagsmaler*-Spiel gegen den Computer antreten ließen. Als Ergebnis kann der Algorithmus mehr als 300 verschiedene Objekte erkennen, und das sogar, obwohl sich die zeichnerischen Fähigkeiten der Teilnehmer stark unterscheiden. Hier ein paar Beispielskizzen aus den Trainingsdaten für *Känguru*:¹¹



Quick Draw erkannte meine Kängurus sofort.¹² Es erkannte auch die Gabel und die Eistüte. Die Pfeife machte ihm etwas Schwierigkeiten, weil sie keine der ihm bekannten 345 Objekte war. Es entschied daher, dass sie entweder ein Schwan oder ein Gartenschlauch sein müsse.

Da Quick Draw aber eben nur diese 345 Dinge kannte, waren viele seiner Antworten auf meine Skizzen besonders seltsam.



Größte Wahrscheinlichkeit:
Bananenschale (1,97408)



Größte Wahrscheinlichkeit:
Großes Monster (0,821636)

Das ist auch vollkommen in Ordnung, wenn Sie, wie ich, Skurrilität zum Ziel haben. Dennoch führt dieses unvollständige Weltbild in manchen Anwendungen zu echten Problemen – zum Beispiel bei der automatischen Textvervollständigung. Wie wir in Kapitel 3 gesehen haben, nutzen Smartphones für die Textvervollständigung normalerweise eine Form des maschinellen Lernens, die als Markow-Kette bezeichnet wird. Dabei haben Unternehmen echte Probleme, die KI davon abzuhalten, munter diskriminierende oder beleidigende Vorschläge zu machen. Einmal erzählte Daan van Esch, Projektmanager für die Android-Autokorrektur-App GBoard, der Internetlinguistin Gretchen McCulloch: »Wenn du ›Ich gehe zu Omas ...‹ eingetippt hast, schlug GBoard eine Zeitlang ›Begräbnis‹ als Ergänzung vor. Das ist im Prinzip ja auch nicht falsch. Vielleicht wird die Formulierung einfach öfter verwendet als ›zu Omas Rave Party‹. Andererseits möchte man an solche Dinge auch nicht unbedingt erinnert werden. Hier sollte man etwas besser aufpassen.« Eine KI weiß nicht, dass diese eigentlich vollkommen korrekte Vorhersage vielleicht doch nicht die richtige Antwort ist. Menschen mussten eingreifen und der KI beibringen, dieses Wort nicht mehr vorzuschlagen.¹³

Ich sehe vier Giraffen

In Visual Chatbot, einer KI, die darauf trainiert wurde, Fragen zu Bildern zu beantworten, gibt es eine Menge interessanter datenbezogener Seltsamkeiten. Die Forscher, die den Bot trainierten, verwendeten hierfür einen per Crowdsourcing zusammengetragenen Datensatz aus Fragen und Antworten zu verschiedenen Bildern. Wie wir inzwischen wissen, können unausgewogene Trainingsdaten die Antworten der KI verfälschen. Also bereiteten die Forscher

ihre Daten so auf, dass bekannte Tendenzen vermieden wurden. Eines dieser Vorurteile war das sogenannte **visuelle Priming**. Menschen haben die Tendenz, Fragen zu Bildern so zu stellen, dass die Antwort »Ja« lautet. Menschen fragen eher selten: »Siehst du einen Tiger?«, wenn das Bild offensichtlich keinen Tiger enthält. Infolgedessen lernt eine mit diesen Daten trainierte KI, dass die Antwort auf die meisten Fragen »Ja« lautet. Einmal fand die mit einem nicht neutralen Datensatz trainierte KI heraus, dass sie eine Genauigkeit von 87 Prozent erreichen konnte, wenn sie Fragen, die mit »Siehst du ein(e/n) ...« begannen, grundsätzlich mit »Ja« beantwortete. Das kennen wir bereits vom Problem der Klassenungleichheit aus Kapitel 3. Eine große Zahl ungenießbarer Sandwiches führte dazu, dass die KI glaubte, Menschen müssten Sandwiches grundsätzlich hassen.

Um das visuelle Priming zu vermeiden, verbargen die Programmierer bei der Sammlung der Crowdsourcing-Daten das Bild, zu dem die Fragen gestellt werden sollten. Dadurch zwangen sie die Menschen, allgemeine Ja-oder-Nein-Fragen zu stellen, die auf beliebige Bilder anwendbar waren. Auf diese Weise erreichten sie ein ungefähres Gleichgewicht zwischen den positiv und den negativ beantworteten Fragen im Datensatz.¹⁴ Aber selbst das reichte nicht aus, um die Probleme zu beseitigen.

Der Datensatz hat eine besonders unterhaltsame Eigenheit. Er ist »giraffisiert« (was das ist, haben wir im Abschnitt »Is this the real life?« besprochen). Mit Bildern von Personen in einem Meeting oder von einem Wellenreiter kommt der Bot recht gut klar – bis Sie fragen, wie viele Giraffen das Bild enthält. Egal was wirklich zu sehen ist: Der Bot wird antworten, dass das Bild mindestens eine Giraffe enthält oder vier oder sogar »zu viele, um sie zu zählen«.

Wo liegt die Wurzel dieses seltsamen Problems? Bei der Sammlung der Daten fragten die Menschen nur selten: »Wie viele Giraffen enthält das Bild?«, wenn keine zu sehen waren. Warum sollten sie auch? In einem normalen Gespräch fragen sich die Leute ja auch nicht gegenseitig über die Anzahl der Giraffen aus, wenn sie wissen, dass es keine gibt. Visual Chatbot war zwar auf ein normales Gespräch innerhalb der Grenzen der Höflichkeit vorbereitet, nicht aber auf absonderliche Menschen, die Zufallsfragen über Giraffen stellen.

Und weil die KI an normalen Konversationen zwischen normalen Menschen trainiert wurde, ist sie auf andere Seltsamkeiten vollkommen unvorbereitet. Wenn Sie Visual Chatbot einen blauen Apfel zeigen und fragen, welche Farbe er hat, wird der Bot mit »Rot« oder »Gelb« oder irgendeiner anderen für einen Apfel normalen Farbe antworten. Anstatt die tatsächliche Farbe eines Objekts herauszufinden – eine schwere Aufgabe –, hat Visual Chatbot gelernt, dass die

Antwort auf die Frage nach der Farbe eines Apfels fast immer »Rot« lautet. Das Gleiche passiert, wenn Sie dem Bot das Bild eines hellblau oder orange eingefärbten Schafs zeigen und ihn dann nach dessen Farbe fragen. Visual Chatbot wird auch hier mit einer Standardfarbe für Schafe wie »Schwarz und Weiß« oder »Weiß und Braun« antworten.

Tatsächlich hat Visual Chatbot nur wenige Möglichkeiten, um Unsicherheit auszudrücken. In den Trainingsdaten wussten die Menschen, was im Bild vor sich ging, selbst wenn einige Detailfragen wie »Was steht auf dem Schild?« nicht beantwortet werden konnten, weil es verdeckt war. Auf die Frage »Welche Farbe hat das X?« lernte Visual Chatbot zu antworten: »Das kann ich nicht sagen, das Bild ist schwarz-weiß«, selbst wenn das Bild offensichtlich farbig war. Auf Fragen wie »Welche Farbe hat ihr Hut?« antwortet der Bot: »Keine Ahnung, ich kann ihre Füße nicht sehen.« Er gibt plausible Erklärungen für seine Verwirrung, allerdings in einem vollkommen falschen Kontext. Andererseits wird Visual Chatbot auch keine totale Verwirrung ausdrücken, denn die Menschen, die ihn trainiert haben, waren auch nicht verwirrt. Wenn Sie Visual Chatbot ein Bild von BB-8, dem kugelförmigen Roboter aus *Star Wars*, zeigen, wird der Bot erklären, es handle sich um einen Hund, und auch folgende Fragen entsprechend beantworten. In anderen Worten: Der Bot blufft.

Eine KI kann im Training nur eine begrenzte Datenmenge sehen, und genau das ist für Anwendungen wie selbstfahrende Autos oft das Problem. Die KIs müssen mit den unendlichen Seltsamkeiten der Menschenwelt klarkommen und entscheiden, wie sie damit umgehen sollen. Wie ich bereits in Kapitel 2 im Abschnitt über selbstfahrende Autos gesagt habe, ist das Fahren auf echten Straßen ein sehr weit gefasstes Problem. Das gilt auch für den Umgang mit den vielen Dingen, die ein Mensch sagen oder zeichnen kann. Als Ergebnis versucht die KI, auf Basis ihres begrenzten Weltbilds die bestmögliche Vermutung anzustellen. Die Ergebnisse können urkomisch sein oder auch auf tragische Weise falsch.

Im folgenden Kapitel sehen wir uns KIs an, die ihre Aufgaben sehr gut bewältigt haben – nur haben wir sie versehentlich angewiesen, die falschen Probleme zu lösen.

Index

A

- Abkürzungen, des KI-Algorithmus 34
- AdFisher 173
- AGI, Artificial General Intelligence 48, 182
- AI Weirness, Blog 11
- Aktivierungsfunktion 75
- Algorithmen
 - evolutionäre 93, 155
 - für maschinelles Lernen 18, 55
 - genetische 18, 142
- allgemeinen künstliche Intelligenz 48
- Amazon
 - Bewertungen, Erzeugung durch neuronales Netz von 80, 116
 - Software zur Analyse von Bewerbungsunterlagen 170, 204, 211
- Amazon Mechanical Turk 117, 141
- ANI, Artificial Narrow Intelligence 48, 182
- ANNs, Artificial Neural Networks 66
- Aprilscherze 87
- Artificial General Intelligence, AGI 48, 182
- Artificial Narrow Intelligence, ANI 48, 182
- Artificial Neural Networks, ANNs 66
- Astronomie 42, 80
- Atari 159
- AttnGAN, Bilderkennungsalgorithmus 191
- Autokorrektur-Apps 89, 134
- Autos, selbstfahrend 13
 - Aktualisierung, Bedarf für 210
 - Autonomiestufen 63
 - Grenzen 15, 196
 - hybride KI und 198

Probleme mit 33, 39, 50, 61
Pseudo-KI und 28
Speicher 61
Trainingsbeispiele für 50, 116, 136
und fehlerhafte Daten 119
und Überanpassung 165

Azure (Microsoft) 131

B

Bag-of-Features-Modell (ein Sack voll ungeordneter Erkennungsmerkmale) 196

Barron, Christine 138

Bell, Anthony 181

Belohnungsfunktionen 139

fehlerhafte 153

hacken 141

Klassenungleichheit und 162

Neugier und 150

unerwartete Reaktionen 145

Bergsteigeralgorithmus 94

Bethesda Softworks 144

Betrug erkennen 79, 162

Bewerbungen, automatische Analyse 13

Pseudo-KI und 202

und gegnerische Angriffe 194

von Video-Interviews 33, 35, 49, 204

Vorurteile bei 33, 43, 170, 204, 211

Bewerbungsunterlagen, verändern 194

Bias Laundering (von Vorurteilen reinwaschen) 173

Bibel, als Teil eines Datensatzes für Übersetzungen 128

BigGAN 125, 219

Bilderzeugung 50, 85, 104, 181, 220

auf Vorurteile testen 174

Bilderkennung 50, 53, 59

Giraffisierung und 134

Grenzen 195

hybride KI und 198

KI-Fehler bei 30
Pseudo-KI und 198
Trainingsdaten für 27, 116, 118
Transferlernen 53
und gegnerische Angriffe 188
und Überanpassung 166
unüberwachtes Lernen 181
von Fischen 163
siehe auch Gesichtserkennung; Giraffen; Hunde; Katzen

Bildfilterung 104
Bildveränderung 104
biologische Evolution 146, 158, 177
biologische neuronale Netzwerke 66
Bitcoin 217
Blank, Doug 155
Blue, Violet 166
Booth, Serena 215
Botnik 88
Bots, soziale Medien 12, 197
siehe auch Chatbots; Roboter

Braitenberg-Strategie 147
BuzzFeed, Artikelüberschriften für 123

C

C-3PO 15, 17, 48
Catastrophic Forgetting (katastrophale Vergesslichkeit) 184
Chatbots 17, 44, 211, 214
hybride KI und 199
siehe auch Facebook, Chatbot M; Visual Chatbot

Chatonsky, Gregory 109
ChestXray14, Datensatz 163
chinesisches Überwachungssystem 200
Christian, Jon 127
Clements, David L., Astrophysiker 158
COCO, Datensatz 192
COMPAS, Algorithmus 169

Computerspiele 143
als Simulationen 50, 98, 156, 160
Erinnerung und 54
für KI-Training 109, 143, 180, 186
Geschicklichkeit von KI bei 30, 216
Neugier und 151
von KI gehackt 156, 181

Conceptnet Numberbatch 168

Crispin, Sterling 139

Crowdsourcing 116

Cumberbatch, Benedict 169

D

Data Augmentation (Datenanreicherung) 117

Datenanreicherung (Data Augmentation) 117

Deep Dreaming 84

Deep Learning 66, 72

Deepfakes 42

DeepL 55

DeepMind Control Suite, OpenAI 148

DeepMind, Google 181

Delfine 141

DenseNet 192

Diskriminator, GANs 105

Doom, Computerspiel 109, 180

Dota (Spiel) 50, 54

Dungeons & Dragons 182

E

Eiscreme-Geschmacksrichtungen 51, 80, 86, 112, 162

Eisen, Michael 212

Elliott, Melissa 126

Emotionserkennung 204

Engadget, Blognetzwerk 167

enge künstliche Intelligenz 48

Enron-Konzern 129

Entscheidungsbäume, Random-Forest-Algorithmus 90, 104

Ethik 143, 154

Euler-Integration 157

Evolution

biologische 146, 158, 177

konvergente 182

evolutionäre Algorithmen 93, 155

Kreuzung 100

F

Facebook 42, 116, 154

Chatbot M 45, 111, 199

Faltung (Konvolution) 57

Fan-Fiction 58, 185

Fingerabdruckscanner 13, 95

und gegnerische Angriffe 193

Fitnessfunktion 97

Flic-flac Spider 146

Flirtsprüche, von KI 11

Fünf Prinzipien der KI-Seltsamkeit 15

G

GANs (Generative Adversarial Networks) 104

GBoard, Android-Autokorrektur-App 134

gegnerischer Angriff (Adversarial Attack) 187

Gehirn, menschlich 66, 177

Generative Adversarial Networks (GANs) *siehe* Generative gegnerische Netze (GANs)

Generative gegnerische Netze (GANs) 104, 112, 124, 181

und Kunst 217, 219

Generator, GANs 104

genetische Algorithmen 18, 142

Genom 97, 104

Gesichtserkennung 13, 95, 142, 173, 200

Gewebeproben 193

Giraffen

und gegnerische Angriffe 188

und Visual Chatbot 118, 126, 134, 210

Giraffisierung 126

Girouard, Mark J. 171
Gizmodo 203
globales Maximum 94
Goodfellow, Ian 104
Goodman, Erik 223
Google 116, 189
 Antimissbrauch-Team 166
 Rezepte 59
 selbstfahrende Autos 119
 Vorurteile bei Einstellung durch 173
Google Brain 129
Google Cloud 192
Google DeepDream, Bilder von 84
Google DeepMind 181
Google Docs 90
Google Flu, Algorithmus 164
Google Translate 55, 59, 127
GoogLeNET 85
GPT-2 (neuronales Netzwerk) 57, 58, 185, 217
Gradient Decent (Gradientenabstieg) 94
Gradientenabstieg 94
H
Ha, David 109
Hackerangriffe 79
Halloween-Kostüme 116, 130
Harry-Potter-Fan-Fiction 58, 185
Heliograf 40, 55
Hidden Layer (versteckte Schicht) 72
Hill Climbing (Bergsteigeralgorithmus) 94
Homeowners Bot 41
Human-level Artificial General Intelligence 61
Hunde
 Bilder 30, 84, 125, 189, 209, 216
 Roboter 139
 Training von virtuellen 139

hybride KI 198

Hyperparameter 100

I

IBM Watson 192

ImageNet, Datensatz für Bilderkennung 116, 190, 193

Inception V3 192

Instagram 42

interne Modelle 178

Irpan, Alex 139

J

Jigsaw 166

K

Kakerlaken (*Periplaneta americana*) 37, 60, 66

 und gegnerische Angriffe 187, 193

 und Random Forests 90, 94

Karate Kid (Spiel) 54

katastrophale Vergesslichkeit (Catastrophic Forgetting) 184

Katzen

 Bilder 27, 30, 53, 72, 84, 112, 124, 186

 Namen 11, 112, 219

KI

 Abkürzungen des Algorithmus 34

 Allgegenwärtigkeit 12, 17

 allgemein (AGI) 48, 182

 begrenzt (ANI) 48, 182

 Definition 18

 Dinge, die als KI bezeichnet werden 18

 falsches Problem 33

 fehlerhafte Daten 34

 Grenzen 14, 195, 222

 im Vergleich zu regelbasierten Programmen 18, 209

 interne Strukturen 32

 Problem zu kompliziert 33

 schlechte Regeln für 30

 trainieren 20

siehe auch maschinelles Lernen; neuronale Netzwerke; Trainingsdaten

KI-Katastrophen 32

Beispiele 111

Warnsignale 32, 49

Klassenungleichheit 78, 161, 187

Klopf-klopf-Witze 20, 65, 80, 208

Kollisionserkennung 158

Konnektionismus 66

konvergente Evolution 182

Konvolution (Faltung) 57

Kreativität, algorithmische 216

Kreuzung 100

Kundendaten 129, 162, 199

Kunst, KI-generierte 42, 109, 216

Trainingsdaten und 192, 216

künstliche neuronale Netzwerke 66

Kybernetik 66

L

Labelling 119

LabSix 189

latenter Raum 219

LIME 162

lokales Maximum 94

Long Short-Term Memory (LSTM) 86

LSTM (Long Short-Term Memory) 86

M

Malware (Schadsoftware) 191

Markow-Kette 86

maschinelles Lernen

Algorithmen für 18, 55

Arten 65

Deep Learning 66, 72, 91

im Vergleich zu regelbasierten Programmen 18, 209

One-Shot-Learning 50

Reinforcement Learning 18

- Transferlernen 53, 102, 185, 218
 - unüberwachtes Lernen 181, 212
- Masterprint, Fingerabdruck 193
- mathematisch Fehler, von KI gehackt 157, 164
- Mathwashing 173, 211, 214, 216
- Matrix, Film 155
- Maximum
 - globales 94
 - lokales 94
- McCulloch, Gretchen 134
- medizinische Bilder 14, 39, 162, 163
 - Klassenungleichheit und 79
 - Pseudo-KI und 201
 - und gegnerische Angriffe 190, 193
- menschenunterstützte allgemeine künstliche Intelligenz 61
- Messaging-Apps 89
- Microsoft 30, 108
- Microsoft Azure 131, 192
- Mobileye 62
- Mode Collapse (Moduskollaps) 218
- Motherboard, Website 127
- Murphy, Tom 54
- Mutation 99
- N**
- Nadel-im-Heuhaufen-Probleme 95, 102
- Navigations-App 142
- Neugier, KI und 150
- neuronale Netzwerke 66
 - biologische 66, 177
 - evolutionäre Algorithmen und 93
 - Gehirnchirurgie 181
 - künstliche (ANNs) 66
 - Selbstkonfigurierung 76
 - und Random Forests 92
 - Unsicherheit und 136

siehe auch RNNs

Neuronen (Nervenzellen) 66, 67

versteckte Schichten 72, 84

Zusammenarbeit 80

Ng, Andrew 33

Nicht-Spieler-Charaktere 144

Noisy-TV-Problem (Problem des rauschenden Fernsehers) 152

Non-Player Characters (NPCs) 144

Northpointe 169

NPCs (Non-Player Characters) 144

Nvidia 112, 124

O

Oblivion, Spiel 144

One-Shot-Learning 50

OpenAI 80

DeepMind Control Suite 148

Überanpassung 141, 162, 191

OpenAI Five 50, 54

Overfitting (Überanpassung) 141, 162

Overpolicing (Über-Überwachung) 172

Overreporting (Übererkennen) 126

P

Pac-Man 151

Personalisierung, extreme 40

Perspective, KI-System zur Moderation von Internetkommentaren 166

Physik, hacken von 156, 161, 164, 221

Pneumothorax, Röntgenbilder auswerten von 163

Poe, Edgar Allan 83

PolyWorld, Trainingswelt 182

Predictive Policing (vorausschauende

Polizeiarbeit) 172, 212

Preprocessing (Vorverarbeitung) 175

Produktbewertungen

Erzeugung 80, 116

Vorurteile in 167

Programmierung, regelbasiert 18

proprietäre Algorithmen 211

ProPublica 169

Pseudo-KI 28, 198

Q

Q*bert, Atari-Spiel 159

Quartz, Internetportal 171

Quick Draw 133

Quicksilver, für Wikipedia-Artikel 214

R

Radiant AI 144

Random Forest, Algorithmus 18, 90, 95

regelbasierte Programmierung 18, 19, 28, 209, 213

Reinforcement Learning 18

rekurrente neuronale Netzwerke *siehe auch* RNNs

ResNet-50 192

Ridler, Anna 217

RNNs (rekurrente neuronale Netzwerke) 18, 80

- Erinnerung und 55, 109

- Markow-Ketten und 86, 88, 89

- Produktbewertungen und 80

- Trainingsdaten für 129, 130, 143

Robocup-Fußballsimulator 155

Roboter

- evolutionäre Algorithmen und 93

- in Science-Fiction 15, 17, 48, 136

- kriminelle 215

- laufende 15, 145, 156, 165

- Menschen als 17, 197

- Ziele definieren für 137

Roboterhunde 139

Röntgenbilder 193

Rückkopplungsschleife 212

S

Sarin, Helen 217

Schach 28
Schadsoftware, manipulieren 191
Schmidhuber, Jürgen 109
Science-Fiction 12, 14, 17, 43, 188, 208
 KI-erzeugt 218
 Roboter in 15, 17, 48, 136
Search Space (Suchraum) 95
Seeing AI-App 108
Sejnowski, Terrence 181
selbstfahrende Autos *siehe* Autos, selbstfahrend
Simon, Joel 142
Sims, Karl 157
Simulationen 97, 155, 182, 221
 hacken 155, 161, 164, 181
Siri 210
Sloan, Robin, Science-Fiction-Autor 218
Smartphones 85, 134
Spamfilter 13, 40
Speer, Robyn 167
Sprache-zu-Text-/Text-zu-Sprache-Software 49, 109, 194
SqueezeNet 192
Stenning, Nick 155
Stimmungen, Neuronen zur Erkennung von 82
Stimmungserkennung, Algorithmen zur Bewertung von 166
Strickland, Donna 127
StyleGAN 112, 124
Suchraum (Search Space) 95
 konvex 95
 Nadel-im-Heuhaufen-Problem 95
Super Mario Bros. 54, 144, 156
T
Tay, Microsoft-Chatbot 211
Tesla Autopilot 62, 64
Tetris 144
Textanalyse 84

Texterzeugung 55

- Erinnerungsvermögen und 55, 182
- Kunst und 55, 218
- neuronale Netzwerke und 83

Textvervollständigung, automatisch 134, 212

Text-zu-Sprache-/Sprache-zu-Text-Software 49, 109, 194

Themis 174

Tic Tac Toe (Drei gewinnt) 28

- Algorithmus 160

Trainingsdaten 11, 186

- Ableitung von Regeln durch KI 19, 20
- aus Computerspielen 109, 143, 180, 186
- Crowdsourcing für 116, 134
- die Zeit verschwenden 111, 121
- Eiskremsorten als 112
- fehlende 130
- fehlerhafte 34, 208
- für Bilderkennung 27, 116, 118
- für Markow-Ketten 87
- für Sandwich-Sortierung 76
- für selbstfahrende Autos 116, 136
- gegnerische Angriffe und 190
- in GANs 105
- Klassenungleichheit und 162
- Kunst und 217
- Menge 50, 111, 112
- menschliche Kontrolle von 208, 217
- proprietär 190
- Pseudo-KI und 198, 202
- realistische oder unrealistische 111, 126, 141
- Übersetzungsalgorithmen und 127
- unerwartete Muster und 161, 221
- ungenau 119, 121
- Vorurteile und 143, 174, 203
- wiederverwenden 51

Trainingsprozess 76
Transferlernen 53, 102, 185, 218
Traumtraining 177
trophische Eier 182
Turing, Alan 43
Turing-Test 43, 108, 117, 199
Twitter 154, 159
Twitter-Bot 211

U

Überanpassung (Overfitting) 91, 141, 162
Übererkennen (Overreporting) 126
Überprüfung auf Vorurteile, als Dienstleistung 174
Übersetzungsalgorithmen 55, 59
 Nonsens-Silben und 127
unfreiwilliges Auswendiglernen (Unintentional Memorization) 129
ungenauere Daten 119
Unintentional Memorization (unfreiwilliges Auswendiglernen) 129
Unsupervised Learning (unüberwachtes Lernen) 181
Unteralgorithmen 174
unüberwachtes Lernen (Unsupervised Learning) 181

V

van Esch, Daan 134
Vanhoucke, Vincent 119
Venture, Spiel 151
Verbrechen, und Roboter 215
Verge, The 119
Verschwörungserzählungen 13, 153
Versicherungen
 Betrug in 193
 Vorurteile bei 170, 211
versteckte Schicht (Hidden Layer) 72
Versuch und Irrtum 19, 30, 65
 in evolutionären Algorithmen 94
 in GANs 107
 Random Forests und 92

und gegnerische Angriffe 188, 189

Videos 12

als Trainingsdaten 202

Deepfakes in 42

Interviews 33, 49, 204

YouTube 153

Visual Chatbot 134, 174, 192, 210

Giraffisierung und 118, 126, 135

visuelles Priming 135

Volkswagen 62

vorausschauende Polizeiarbeit 212

Vorhersage

bei Analyse von Bewerbungsunterlagen 33

im Vergleich zu Empfehlung 170

in Spielen 50

interne Modelle und 178

Klassenungleichheit und 78

Neugier und 150

Texterzeugung und 219

von Buchstaben 12, 56, 80, 150

von Giraffen 126

Vorurteile 169

Vorurteile 166

bei der Einstellung 43, 170

fehlerhafte Belohnungsfunktionen und 139

Gender 127, 187

KI-Potenziale und 216

Predictive Policing (vorausschauende Polizeiarbeit) und 172

Pseudo-KI und 198

Sichtbarkeit von 168

Testen auf 173

Trainingsdaten und 143, 165, 174, 203, 221

verstärken 186

Vorverarbeitung (Preprocessing) 175

W

Wanderpalmen 146
Washington Post 40
Waymo, Unternehmen für selbstfahrende Autos 50
West, Jessamyn 166
White, Tom 192
Wii-Remotes 210
Wikipedia 214
 Wissenschaftlerinnen in 127
Wired, Magazin 166
Wortvektor 167
Y
Yee, Hector 210
YouTube, Belohnungsfunktion 13, 153
Z
Zeitungsartikel, schreiben 40, 214
Zufallswald (Random Forest), Algorithmus 90