Lernen wie Gehirne

Stellen Sie sich vor, Sie erforschen Insekten und beobachten auf Ihrer Forschungsreise das Zusammenleben von zwei Käferarten. Die einen Käfer haben einen einfarbigen, braunen Körper. Die anderen sehen sehr ähnlich aus, lassen sich aber gut an den glänzenden Punkten auf ihrem sonst braunen Rücken erkennen. Wie Ameisen laufen diese Käfer geschäftig auf Duftpfaden zwischen verschiedenen Futterquellen hin und her. Treffen zwei Exemplare aufeinander, drängeln sie sich einfach aneinander vorbei. Doch dann sehen Sie plötzlich ein ungewöhnliches Schauspiel: Einer der Käfer macht so etwas wie einen Knicks und bleibt so lange regungslos stehen, bis der andere außer Sichtweite ist. Nachdem Sie die Käfer geduldig über eine längere Zeit beobachtet haben, erkennen Sie die Regel, die diesem ungewöhnlichen Verhalten zugrunde liegt. Wenn einer der einfarbigen Käfer auf einen größeren Käfer mit glänzenden Punkten trifft, dann macht er einen Knicks. Sonst nicht.

Als Mensch kann man in dieses Verhalten viel hineininterpretieren. Vielleicht ist der Knicks eine Ehrerbietung gegenüber den größeren Käfern. Oder die einfarbigen Käfer haben schlicht Angst und der Knicks ist eine Unterwerfungsgeste. Mit solchen Erklärungen muss man aber vorsichtig sein. Wir Menschen neigen dazu, zu viel zu psychologisieren. Versuchsteilnehmer, denen man zum Beispiel einen Film zeigt, in dem sich ein Dreieck und ein Quadrat kreisförmig umeinander drehen oder in dem das Dreieck und das Quadrat sich hintereinander herbewegen, sprechen ganz natürlich über die geometrischen Formen, als ob sie Personen wären: Sie haben zusammen getanzt und sich gefreut oder der eine hatte Angst und ist vor dem anderen weggelaufen. Obwohl die Versuchsteilnehmer nur sich bewegende geometrische Formen sehen, schreiben sie den Formen spontan ein psychisches Innenleben zu, um ihr Verhalten zu erklären. Genauso wie bei tanzenden Dreiecken und bei sprechenden Computerprogrammen ist die Versuchung groß, auch

das Verhalten unserer Käfer durch psychische Zuschreibungen zu erklären.[1]

Tatsächlich gibt es aber eine viel einfachere, mechanistische Erklärung für das sonderbare Knicksverhalten der Tiere. In Australien lebt ein Käfer, der den Käfern in unserem Gedankenexperiment ähnlich ist. Die Deckflügel von weiblichen ›Julodimorpha Bakewelli‹ sind braun und auffällig glänzend gepunktet. Immer wenn ein Männchen diese Punkte sieht, versucht es das Weibchen zu begatten. Das Insekt hat einen Detektor für glänzende Punkte, um eine bestimmte Verhaltensweise auszulösen. Dummerweise ist die Farbe des Weibchens dem Braun einer Bierflasche sehr ähnlich und die Knubbel am Boden der Bierflasche glänzen noch verlockender als die Punkte des Weibchens. Achtlos weggeworfene Bierflaschen in der australischen Wüste werden so zur Liebesfalle für männliche Käfer, die wieder und wieder Bierflaschen besteigen, bis sie entweder verhungern oder von räuberischen Wüstenameisen aufgefressen werden.[2]

Die Käfer, die Sie beobachtet haben, besitzen also so einen Detektor für glänzende Punkte. Außerdem haben sie einen weiteren Detektor, der feststellt, ob der Käfer, der ihnen entgegenkommt, größer ist als sie selber (zum Beispiel, weil sie hochschauen müssen). Im Verlauf Ihrer Käferstudie vermuten Sie, das Knicksverhalten des Insekts wird gerade immer dann ausgelöst, wenn der Punkt-Detektor und der Größer-Detektor anschlagen. Aber wie würde das im Detail funktionieren?

1 Die klassische Studie dazu stammt von Heider & Simmel (1944). Im Internet finden sich viele Videos dieser Studie und es lohnt sich, diese anzuschauen, um einen Eindruck davon zu bekommen, wie leicht man einfachen geometrischen Formen ein komplexes, psychisches Innenleben zuschreibt. Braitenberg (1984) beschreibt verschiedene Gedankenexperimente, die zeigen, dass sich Verhalten oft auch viel einfacher erklären lässt. Das Käferbeispiel ist von seinen Gedankenexperimenten inspiriert.

2 Siehe Gwynne & Rentz (1983) für den Käfer und die Bierflaschen und Lettvin, Maturana, McCulloch & Pitts (1959) für ähnliche Auslöser beim Frosch. Ich habe zuerst bei Hoffman (2009) von Julodimorpha Bakewelli gelesen.

Wie Nervenzellen rechnen

Käfer haben – so wie wir Menschen – ein Nervensystem, das ihr Verhalten steuert. Nervensysteme bestehen aus spezialisierten Nervenzellen, auch Neurone genannt, die untereinander elektrische Signale austauschen. Diese Signale sind digitale Signale: Jedes Neuron kann nur an oder aus sein. Eins oder Null. Wenn ein Neuron an ist, dann sendet es einen elektrischen Impuls, ein sogenanntes Aktionspotenzial, an andere Neurone, mit denen es verschaltet ist.[3] Man sagt: Das Neuron feuert.

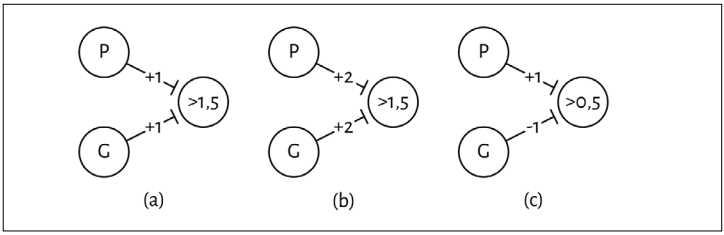

Abb. 4: Drei neuronale Netze

In Abbildung 4a ist eine einfache Verschaltung von drei Neuronen zu sehen. Eine solche Verschaltung von mehreren Neuronen nennt man ein ›neuronales Netz‹. Das Neuron P ist der Punkt-Detektor. Sieht das Insekt glänzende Punkte, schaltet sich der Punkt-Detektor an und das Neuron sendet ein elektrisches Signal an das Ausgabeneuron, mit dem es verschaltet ist. Neuron G, der Größer-Detektor, schaltet sich wiederum nur an, wenn der entgegenkommende Käfer größer ist. Dann sendet auch Neuron G ein elektrisches Signal an das Ausgabeneuron. Doch das Ausgabeneuron schaltet sich nur an, wenn die Summe aller Signale, die bei ihm ankommen, größer als der Schwellenwert 1,5 ist. Falls das passiert, wird das Knicksverhalten des Insekts ausgelöst. Solche einfachen Neurone nennt man auch McCulloch-Pitts-Zellen – nach den zwei theoretischen Hirnforschern, die sie zuerst untersucht haben.[4]

3 Wobei jedes Aktionspotenzial aussieht wie jedes andere, ganz wie bei digitalen Signalen in technischen Systemen. Obwohl die allermeisten Neurone diesem Alles-Oder-Nichts-Gesetz folgen, gibt es bei Insekten recht häufig auch graduierte Potenziale bei denen die Stärke des Signals variiert. Das deutet auf eine analoge Signalverarbeitung hin.

4 Bei Piccinini (2004) findet sich eine hervorragende Darstellung der Originalarbeit von McCulloch & Pitts (1943) im historischen und philosophischen Kontext.

Schickt nun nur Neuron P ein Signal an das Ausgabeneuron (P=1 und G=0), kommt beim Ausgabeneuron nur ein Signal der Stärke 1 an. Und da das kleiner als der Schwellenwert 1,5 ist, schaltet sich das Ausgabeneuron nicht an, und der Knicks wird nicht ausgelöst. Wenn aber P und G beide feuern (P=1 und G=1), dann ist die Eingabe 1+1=2 größer als 1,5 und das Ausgabeneuron schaltet sich an. Dieses neuronale Netz in Abbildung 4a ist eine UND-Verschaltung: Das Ausgabeneuron schaltet sich nur an, wenn Neuron P und Neuron G angeschaltet sind. Der Knicks passiert nur, falls der entgegenkommende Käfer Punkte hat und größer ist.

Neurone können auch so verschaltet sein, dass sie sich wie ein logisches ODER verhalten. Die effektive Verschaltung in einem neuronalen Netz ändert sich mit den ›Verbindungsstärken‹. Ist die Verbindungsstärke zwischen Neuron P und dem Ausgabeneuron doppelt so groß (+2, wie in dem Netz in Abbildung 4b), ist das Signal, das P schickt, sobald es angeschaltet ist, doppelt so stark. Bei einem Schwellenwert von 1,5 reicht dann die Aktivität von P alleine aus, um das Ausgabeneuron anzuschalten (denn bei P=1 und G=0 ist mit einer Verbindungsstärke von 2 die Summe 2+0=2 größer als 1,5). Da die Verbindungsstärke von G zum Ausgabeneuron in Abbildung 4b auch verdoppelt ist, feuert das Ausgabeneuron ebenso, falls nur G an ist. Senden beide Eingabeneurone Signale, schaltet sich das Ausgabeneuron sowieso an (denn 2+2=4>1,5). Lediglich wenn gar kein Signal ankommt, bleibt es ausgeschaltet. Das Insekt macht also einen Knicks, wenn der entgegenkommende Käfer Punkte hat oder größer ist (das logische ODER schließt den Fall, dass der Käfer Punkte hat und größer ist, mit ein).

Andere Verbindungsstärken und Schwellenwerte führen dazu, dass das neuronale Netz sich nach anderen logischen Regeln verhält. In dem Netz in Abbildung 4c ist die Verbindungsstärke von P zum Ausgabeneuron +1 und die von G zum Ausgabeneuron -1. Der Schwellenwert am Ausgabeneuron beträgt in diesem Beispiel 0,5. Hier schaltet sich das Ausgabeneuron nur ein, wenn allein P ein Signal sendet (P=1 und G=0 ergibt 1 und ist größer 0,5). Feuert G, wird die Eingabe so stark gehemmt, dass P keinen Effekt hat (P=0 und G=1 führt zu 0-1=-1<0,5 und P=1 und G=1 zu 1-1=0<0,5). Das Insekt macht nur dann einen Knicks, falls der entgegenkommende Käfer Punkte hat, nicht aber, wenn er größer ist. Die logische Regel ist also P UND NICHT G.

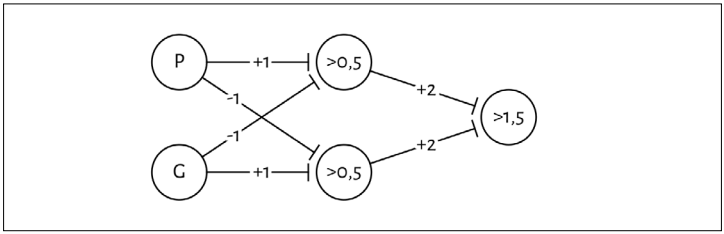

Abb. 5: Ein mehrschichtiges neuronales Netz

Bauen wir neuronale Netze künstlich nach, können wir die Verbindungen und Schwellenwerte so einstellen, dass die Netze sich nach beliebigen logischen Regeln verhalten. Mit größeren Netzen lassen sich kompliziertere logische Regeln ausführen, die wiederum komplizierteres Verhalten erzeugen. Je komplizierter die Regel, desto größer muss jedoch das Netz sein. Ein Netz, in dem Neuronen in mehreren Schichten hintereinander geschaltet sind, nennt man ein ›mehrschichtiges‹ oder ›tiefes neuronales Netz‹.[5] (Hausaufgabe: Was berechnet das mehrschichtige Netz in Abbildung 5?) In Fällen, in denen ein Ausgabeneuron Signale von einem Neuron empfängt, aber auch selber (eventuell indirekt) Signale an dieses Neuron sendet, spricht man von einem ›rekurrenten‹ Netzwerk. Mit rekurrenten neuronalen Netzen lassen sich noch kompliziertere Regeln implementieren.

Es ist kein historischer Zufall, dass ähnliche logische Verschaltungen in Computern verbaut sind. Auch Computer rechnen digital mit Einsen und Nullen: Der Strom ist entweder an oder aus. Statt aus Neuronen bestehen die Schaltkreise zwar aus Transistoren, das Grundprinzip ist aber dasselbe wie bei McCulloch-Pitts-Zellen. Als Warren McCulloch und Walter Pitts in den frühen 1940er Jahren an ihrer neuen Theorie des Nervensystems gearbeitet haben, waren sie natürlich bestens mit Turings Arbeiten zur Turingmaschine vertraut. Sie erkannten sofort, dass man mit einem neuronalen Netz die Verhaltensregeln implementieren kann, mit denen sich eine Papier-und-Bleistift-Maschine steuern lässt. Ein künstliches neuronales Netz, das mit Sensoren und Motoren auf Papier liest und schreibt, ist eine Turingmaschine – und diese Maschine ist genauso mächtig wie ein moderner Computer mit genügend Speicher. Aber was genau mehrschichtige, rekurrente neuronale Netze ohne zusätzlichen Speicher berechnen können, war zunächst noch unklar und stimulierte wichtige Grundlagenforschung zur

5 Auf Englisch ›multilayer‹ oder ›deep neural network‹.

Automatentheorie (zur Erinnerung: der Begriff ›KI‹ war von Anfang an umstritten und der Informationstheoretiker Claude Shannon bevorzugte den langweiligeren Begriff ›Automatenstudien‹). Die Computertechnik beeinflussten McCulloch und Pitts aber auch deshalb, weil ihre Ideen in den bahnbrechenden Bericht einflossen, in dem John von Neumann 1945 die grundlegende Architektur heutiger Computer entwarf und dabei eine Parallele zwischen logischen Schaltkreisen in Rechenmaschinen und neuronalen Netzwerken im Gehirn zog.[6]

Als wir mit unserer Papier-und-Bleistift-Maschine einfache Denkprozesse – zum Beispiel das Addieren von Zahlen – mit einem mechanischen Apparat nachgebildet haben, dachten wir nicht darüber nach, wie diese Denkprozesse im Gehirn tatsächlich ablaufen. Sicher nicht so wie bei unserer mechanischen Papier-und-Bleistift-Maschine. Ihre inneren »Organe« sind nicht genauso aufgebaut wie ein Gehirn. Auch moderne Mikroprozessoren sind in ihrer Struktur Gehirnen nicht sonderlich ähnlich. Obwohl diese verschiedenen Mechanismen völlig unterschiedlich aussehen, so sind sie doch alle Computer, die irgendwie rechnen.

Sicher, McCulloch-Pitts-Zellen sind eine starke Vereinfachung der Funktionsweise echter, biologischer Neurone. Wir wissen heute viel mehr über Nervenzellen als Warren McCulloch und Walter Pitts in den 1940er Jahren. Nervenzellen sind wesentlich komplexer, als die beiden annahmen. Die Idee, dass Neurone so etwas wie logische Schaltelemente sind, übte dennoch auf die theoretische Hirnforschung historisch einen enormen Einfluss aus. Noch entscheidender war allerdings der Einfluss dieser Idee auf die KI-Forschung. Da Gehirne ähnlich funktionieren wie elektronische Computer, kann man vielleicht nicht nur das menschliche Rechnen, sondern auch all die anderen fantastischen Fähigkeiten von Gehirnen in Computern nachbilden.

6 Siehe Piccinini (2004). Kleene (1951) entwickelt ausgehend von neuronalen Netzen den wichtigen Begriff des ›endlichen Automaten‹. Siehe Neumann (1993) für den Bericht. Ein aktueller Trend in der KI-Forschung nutzt die alte Einsicht aus, dass künstliche neuronale Netze in Kombination mit einem externen Speicher Turingmaschinen sind, um Computerprogramme zu lernen (Graves et al., 2016).

Wie Menschen und Computer Bilder erkennen

Wie erkennt ein Gehirn zum Beispiel den Buchstaben ›A‹ auf den Seiten dieses Buches? Die Linse des Auges stellt das Bild auf der Netzhaut scharf. Dort sitzen Fotorezeptoren, die das Licht in elektrische Signale umwandeln, so wie das auch in einer digitalen Kamera passiert.

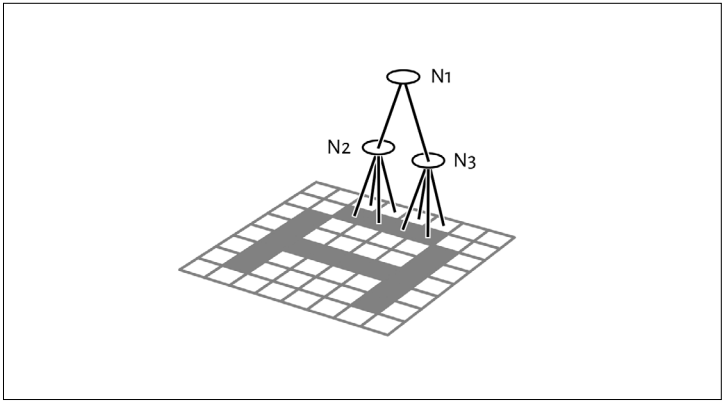

Abb. 6: Musterkennung mit neuronalen Netzen

In Abbildung 6 stellen die Quadrate die Bildpunkte auf der Netzhaut dar. Vom Auge werden die elektrischen Signale an den visuellen Kortex weitergeleitet, der sich im Hinterkopf befindet. Die Neurone dort detektieren Linien und Kanten verschiedener Orientierungen. Es gibt beispielsweise Neurone, die immer aktiv sind, sobald sich an einer bestimmten Stelle in einem Bild eine horizontale Linie befindet. Zwei solche Beispielneurone (N2 und N3) sind in der Abbildung über dem ›A‹ zu sehen. Genauso gibt es im visuellen Kortex Neurone für vertikale Linien oder Linien jedes anderen Winkels. In der nächsten Schicht des neuronalen Netzes können dann kleine Linienstücke zu längeren Linien kombiniert werden. Diese Neuronen werden in der Abbildung von dem obersten Neuron (N1) symbolisiert. Der Querbalken ist vielleicht manchmal etwas schief, sodass der Detektor für horizontale Linien nicht anschlägt. Mit einer ODER-Verschaltung von Liniendetektoren leicht unterschiedlicher Winkel kann der Querbalken aber trotzdem erkannt werden. In einer weiteren Schicht können dann die Detekto-

ren für den Aufstrich, den Abstrich und die zwei Querbalken mit einer UND-Verschaltung zu einem A-Detektor zusammengebaut werden.[7]

So erkennt das Gehirn mit einfachen Liniendetektoren und durch mehrschichtige, logische Verschaltungen komplexe Muster. Das funktioniert nicht nur für Buchstaben, sondern auch für ganze Worte, Objekte oder Personen. Die Neurone, die ganz am Ende eines solchen tiefen neuronalen Netzes sitzen, nennen Hirnforscher auch augenzwinkernd Großmutterneurone – denn falls diese Theorie über die Funktionsweise des Gehirns stimmen sollte, müsste es für jede Person, die man kennt, ein solches Neuron geben, und eben auch für die eigene Großmutter. Dieses Großmutterneuron wird gerade genau dann aktiv, wenn Rotkäppchen die Augen, die Ohren und den Mund seiner Großmutter zusammen sieht.[8]

Lange war die Existenz von Großmutterneuronen nur eine theoretische Hypothese. Hinweise darauf, dass die Hypothese tatsächlich stimmen könnte, verdanken wir unter anderem Epilepsie-Patienten. Um den Herd der Anfälle zu finden, werden Patienten in besonders schweren Fällen manchmal Elektroden implantiert, mit denen man die elektrische Aktivität einzelner Gehirnareale über einen längeren Zeitraum messen kann. Während dieser Beobachtungszeit ließen sich einige Patienten eine große Zahl an Bildern von Prominenten zeigen und bei einem Patienten wurde ein Jennifer-Aniston-Neuron gefunden.[9] Das ist eine Nervenzelle im Gehirn, die aktiv wird, sobald der Patient ein Bild von Jennifer Aniston sieht. Die Forscherinnen und Forscher zeigten dem Patienten auch Bilder von anderen Prominenten, Häusern und Tieren, aber diese Zelle reagierte nur auf Bilder von Jennifer Aniston. Es scheint also tatsächlich Großmutterneurone im menschlichen Gehirn zu geben. Komischerweise reagierte die Nervenzelle aber nicht, wenn neben Jennifer Aniston auch Brad Pitt auf dem Bild zu sehen war. Vielleicht weil die zwei sich bald trennen sollten? Ob der Patient auch eine Brangelina-Zelle und eine Vaughniston-Zelle hatte, wissen wir leider nicht.

7 Der Aufbau und die Funktionsweise des visuellen Systems sind genauer bei Hubel & Wiesel (1979) beschrieben.

8 Gross (2002) beschreibt die Geschichte der Großmutterneurone.

9 Quian Quiroga, Reddy, Kreiman, Koch & Fried (2005) haben außerdem ein Halle-Berry-Neuron und ein Bill-Clinton-Neuron gefunden.

Die eigene Großmutter zu erkennen ist so mühelos. Wie konnte sich Rotkäppchen nur täuschen lassen? Schauen Sie sich ein Bild im Familienalbum an, wissen Sie sofort, ob Ihre Großmutter darauf zu sehen ist. Aber versuchen Sie mal zu erklären, wie genau Sie das machen. Woher wissen Sie, dass das Ihre Großmutter ist und nicht der böse Wolf? Okay, die Augen, die Ohren und der Mund sind nicht so groß. Aber ist das alles? Was ist mit der Farbe der Augen? Würden Sie merken, wenn die Ohren eine andere Form hätten? Da Sie nicht genau erklären können, wie Sie Ihre Großmutter erkennen, handelt es sich dabei um implizites Wissen – genauso wie beim Fahrradfahren oder bei der Intuition von Schachexperten.

Dementsprechend ist es in den Anfangstagen der KI nicht gelungen, Computersysteme zu bauen, die Gesichter erkennen können. Um ein künstliches neuronales Netz zu programmieren, das Gesichter erkennt, müssen wir genau wissen, welche Detektoren wir dafür brauchen und wie man diese in mehreren Schichten verschaltet. Anders als beispielsweise bei der schriftlichen Addition kennen wir den Algorithmus nicht, mit dem wir Gesichter erkennen. Wenn wir nicht wissen, wie das geht, können wir das einem Computer auch nicht beibringen.

Wie schafft es unser Gehirn, die ganzen Nervenzellen genau so zu verschalten, dass wir unsere Lieben erkennen? Babys kommen nicht schon mit fest verdrahteten Neuronen zur Welt, um ihre Großmütter zu erkennen. Irgendwie muss ihr Gehirn diese Fähigkeit erst erlernen. Die Regeln, nach denen sich ein neuronales Netz verhält, können sich mit den Verbindungsstärken zwischen den Neuronen ändern. Zur Erinnerung: Das Netz in Abbildung 4a ist eine UND-Verschaltung und das Netz in 4b eine ODER-Verschaltung. Der einzige Unterschied liegt in den Verbindungsstärken der Punkt- und Größer-Detektoren beim Ausgabeneuron. Wenn es also im Gehirn einen Mechanismus gibt, mit dem die Verbindungsstärken in so einem Netzwerk automatisch verändert werden, dann passt das Gehirn auch sein Verhalten an. Das neuronale Netz kann auf diese Art lernen.

Neuronale Netze lernen durch Korrektur

Die einfachste Form des Lernens ist das sogenannte ›überwachte Lernen‹. Beim überwachten Lernen gibt es einen Lehrer, der den Lernprozess begleitet und das Verhalten korrigiert. Das Feedback des Lehrers

muss allerdings nicht besonders elaboriert sein. Es reicht, dass der Lernende durch den Lehrer mitbekommt, ob sein Verhalten richtig oder falsch war.

Wie genau könnte ein solches überwachtes Lernen in neuronalen Netzen ablaufen? Zurück zu den Käfern mit ihrem außergewöhnlichen Knicksverhalten, die Sie auf Ihrer Forschungsreise entdeckt haben: Ein paar Kilometer weiter beobachten Sie wieder die gleichen Käferarten, doch überraschenderweise sieht das Knicksverhalten ganz anders aus. Diesmal knicksen die einfarbigen Käfer, sobald sie Käfer mit Punkten sehen, die aber nicht größer als sie selber sind. Das neuronale Netz dieser Käfer muss also dem Netz in Abbildung 4c entsprechen (statt dem Netz in 4a wie bei der vorherigen Kolonie). Das Knicksverhalten kann also nicht angeboren sein. Nachdem Sie die Insekten wieder über einen längeren Zeitraum beobachtet haben, fällt Ihnen auf, dass die gerade aus ihrer Puppenhaut geschlüpften, frisch erwachsenen Käfer sich tatsächlich noch nicht an die Knicksregel halten. Sie lernen sie erst nach ein paar Zusammentreffen mit anderen Käfern. Ihnen fällt außerdem auf, dass Käfer, die sich nicht an die richtige Knicksetikette halten, sofort angegriffen werden. Und Sie schließen daraus, dass junge Käfer offenbar durch diese recht direkte Rückmeldung erfahren, dass sie einen Fehler gemacht haben und so lernen, ihr Verhalten entsprechend anzupassen.

In unserem Gedankenexperiment besteht das Nervensystem der Käfer dank Evolution aus den zwei Eingabeneuronen, die Punkte und Größe detektieren, und dem Ausgabeneuron, das den Knicks auslösen kann. Die Verbindungsstärken zwischen den Detektor-Neuronen und dem Knicks-Neuron sind aber nicht durch die Evolution fest voreingestellt, sondern passen sich erst durch Erfahrung an die Umwelt an. Das Gleiche gilt für den Schwellenwert. Zu dem Zeitpunkt, an dem die Käfer aus ihrer Puppenhaut schlüpfen, sind die Verbindungsstärken in ihrem neuronalen Netz alle 0 und der Schwellenwert bei 1,5, so wie in Abbildung 7a. Dann trifft der Jungkäfer auf einen Käfer mit Punkten, der nicht größer als er selber ist (P=1 und G=0). Neuron G bleibt inaktiv und Neuron P sendet ein Signal an das Ausgabeneuron, aber da die Verbindungsstärke 0 ist, kommt beim Ausgabeneuron nichts an und der Schwellenwert von 1,5, bei dem ein Knicks ausgelöst wird, wird nicht überschritten. Der Käfer macht also fälschlicherweise keinen Knicks und die anderen Käfer attackieren ihn deshalb. Doch aus diesem Fehler kann er lernen.

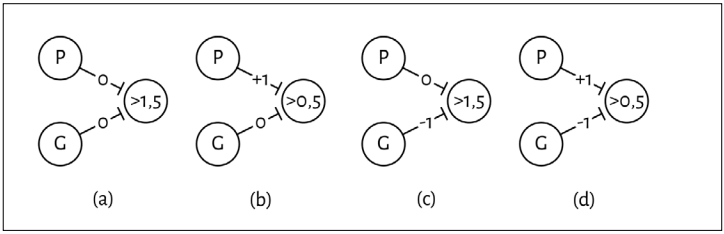

Abb. 7: Veränderung eines Netzes durch Korrektur

Was ist also im neuronalen Netz schiefgegangen? Und wie kann ein Lernalgorithmus das Netzwerk so anpassen, dass zukünftig keine Fehler mehr gemacht werden?[10] Die aggressive Reaktion der anderen Käfer bedeutet, dass es falsch war, dass das Ausgabeneuron inaktiv blieb. Es hätte feuern und einen Knicks auslösen müssen. Entweder war die Verbindungsstärke zwischen Neuron P und dem Ausgabeneuron also zu schwach oder der Schwellenwert zu hoch. Versuchsweise erhöht der Lernalgorithmus deshalb die Verbindungsstärke von Neuron P von 0 auf 1 und senkt den Schwellenwert von 1,5 auf 0,5. Das neue Netz sieht jetzt aus, wie das Netz in Abbildung 7b. Das nächste Mal, wenn der Käfer auf einen gepunkteten Kollegen trifft, der nicht größer ist (P=1 und G=0), sendet Neuron P ein Signal der Stärke 1. Das Signal übersteigt den Schwellenwert von 0,5 und der Käfer macht diesmal richtigerweise einen Knicks.

Unglücklicherweise trifft der Käfer danach auf einen größeren Käfer mit Punkten (P=1 und G=1) und Neuron P löst einen Knicks aus, egal was Neuron G macht, weil die Verbindungsstärke für Neuron G immer noch 0 ist. Wieder wird der Käfer angegriffen. Diesmal, weil er einen Knicks macht, den er nicht hätte machen sollen. Entweder war also die Verbindungsstärke zu groß oder der Schwellenwert zu klein. Der Lernalgorithmus macht daher im nächsten Schritt alle Verbindungsstärken der Detektoren, die aktiv waren, um 1 kleiner und erhöht auch den Schwellenwert um 1. Das Netz sieht jetzt so aus wie in Abbildung 7c.

Durch diese Änderung macht der Käfer aber wieder einen Fehler, sobald er auf einen gepunkteten Kollegen trifft, der nicht größer als er selber ist (P=1 und G=0). Er macht keinen Knicks und wird deshalb angegriffen. Der Lernalgorithmus erhöht daraufhin die Verbindungsstärke von Neuron P wieder von 0 auf 1 und senkt den Schwellenwert

10 Im Folgenden wird der ›Perzeptronlernalgorithmus‹ beschrieben, der auf Rosenblatt (1958) zurückgeht.

wieder von 1,5 auf 0,5 (die Verbindungsstärke von G wird nicht angefasst, weil G=0 war und daher keinen Einfluss hatte). Mit diesen Änderungen sieht das neuronale Netz so aus wie das Netz in Abbildung 7d. Dieses Netz hat die Regel gelernt, dass nur ein Knicks gemacht wird, falls der andere Käfer Punkte hat, nicht aber, wenn er größer ist. Der Käfer macht jetzt keine Fehler mehr.

Stellen Sie sich vor, Sie besuchen eine fremde Kultur, in der jeder vor kleineren Menschen einen Knicks macht. Man muss Ihnen diese Regel nur einmal erklären und Sie wissen sofort, wie Sie sich verhalten müssen. Was aber, wenn Ihnen niemand die Regel verrät und Sie stattdessen böse Blicke ernten, wenn Sie sich falsch verhalten? Sie würden sich wahrscheinlich so ähnlich verhalten wie die Käfer in unserem Gedankenexperiment. Sie würden verschiedene Regeln ausprobieren und so lange bei einer Regel bleiben, bis Sie einen Fehler machen. Dann würden Sie versuchen, die Regel anzupassen, um Ihren Fehler zu korrigieren. Nichts anderes machen Lernalgorithmen: Sie suchen Regeln, die möglichst wenige Fehler machen. Auch Lernalgorithmen sind Suchalgorithmen![11]

Traditionellen Computerprogrammen werden die Regeln, nach denen sie sich verhalten sollen, fest einprogrammiert. Lernende Computerprogramme können ihr Verhalten – oft mit Unterstützung eines Lehrers – selbständig anpassen. Man spricht dann von ›maschinellem Lernen‹. Maschinelles Lernen ist gerade dann von Vorteil, wenn man zwar das richtige Verhalten kennt, aber nicht weiß, wie es eigentlich zustande kommt. Zum Beispiel, wenn es darum geht, Ihre Großmutter auf einem Bild zu erkennen. Da es sich dabei um implizites Wissen handelt, können Sie keinem Computer erklären, wie er das machen soll. Ihr Gehirn macht das zwar nach bestimmten Regeln, Sie kennen diese Regeln aber nicht. Sie können allerdings einem Lernalgorithmus viele Beispielfotos zeigen, auf denen Ihre Großmutter mal zu sehen ist und mal nicht. Der Lernalgorithmus findet durch dieses Training von alleine Regeln, mit denen er Ihre Großmutter erkennen kann. Diese Regeln werden äußerst kompliziert sein und viele Ausnahmen haben.

In Fällen, in denen Regeln nicht perfekt sind, spricht man von ›statistischen Regeln‹, die nicht immer, aber oft gelten. Der Lernalgo-

11 Ein ganz ähnliches Beispiel wird auch von Bruner, Goodnow & Austin (1956) diskutiert, die die Algorithmen untersucht haben, nach denen Menschen in solchen Situationen Regeln lernen.

rithmus findet in solchen Fällen keine fehlerlose Regel, sondern eine, die möglichst wenige Fehler macht. Das ändert aber nicht viel für die Lernalgorithmen, deren Ziel es immer noch ist, die Anzahl der Fehler zu minimieren. Statistische Regeln, die auf diese Weise gelernt wurden, erkennen Muster in der Eingabe, die helfen, die richtige Antwort vorherzusagen. Alle modernen Programme zur Mustererkennung – sei es zur Gesichtserkennung, Buchstabenerkennung, Objekterkennung oder Spracherkennung – funktionieren genau so. Zwar gibt es viele verschiedene Arten von maschinellen Lernalgorithmen in der KI, aber meist basieren diese zurzeit auf tiefen neuronalen Netzen.

Diese künstlichen neuronalen Netze haben nicht selten Tausende von Neuronen in vielen Schichten mit Millionen von Verbindungen. Und mit leistungsfähigeren Computern werden es immer mehr. Dadurch ist es äußerst schwer zu verstehen, warum ein Netz sich so verhält, wie es sich verhält. Zwar waren es die Entwickler, die den Lernalgorithmus programmiert und dem Netzwerk im Training viele Beispiele gezeigt haben, wie es sich richtig verhalten soll, aber auch die Entwickler können sich unmöglich die unzähligen Neuronen und alle Verbindungsstärken ansehen, um das Netz wirklich zu verstehen (wie wir das in unserem Gedankenexperiment gemacht haben, in dem es nur drei Neurone gab). Obwohl das Netz bestimmten Regeln folgt, die präzise im Computer spezifiziert sind, können wir diese Regeln aufgrund ihrer Komplexität nicht nachvollziehen. Wir können also nicht darauf hoffen, dass so ein künstliches neuronales Netz uns helfen kann zu erklären, wie wir unsere Großmutter erkennen. Dieses Wissen bleibt auch in künstlichen neuronalen Netzen implizit!

Obwohl künstliche neuronale Netze ursprünglich mal von Gehirnen inspiriert waren und sie inzwischen einige menschliche Fähigkeiten imitieren können, darf man sie nicht vorschnell mit menschlichen Gehirnen gleichsetzen. Nur weil ein künstliches neuronales Netz auf ein paar Fotos meine Großmutter erkennen kann, heißt das nicht, dass das Netz das genauso macht wie ich. Vielleicht erkennt das Netz nur die Nase. Solange man dem Netz keine Bilder zeigt, auf denen die Nase verdeckt ist oder jemand anderes eine ähnliche Nase hat, bemerkt man den Unterschied nicht. Tatsächlich verhalten sich neuronale Netze auf bestimmten Bildern, auf die sie nicht trainiert worden sind, oft ganz anders als Menschen. Eine meterhohe Zahnbürste wird zum Beispiel von Menschen trotz ihrer Größe leicht übersehen. Menschen rechnen einfach nicht mit Riesenzahnbürsten. Neuronale Netze sind aber extra

so konstruiert, dass sie Objekte erkennen, egal wie groß sie sind. Umgekehrt lassen sich neuronale Netze von verzerrten und verschmutzten Bildern mehr und anders verwirren als Menschen. Die Lektion, die wir von ELIZA gelernt haben, gilt auch hier: Man sollte Computern nicht vorschnell menschliche Intelligenz zuschreiben. Auch wenn ihre Fähigkeiten beeindruckend sind, ist ihre Intelligenz eine andere als unsere.[12]

Nichtsdestotrotz sind die Netze darauf trainiert, auf den meisten Bildern möglichst die Objekte zu erkennen, die wir Menschen erwarten. Damit ein neuronales Netz lernen kann, die Großmutter zu erkennen, muss erst vorher ein Mensch Beispielbilder markieren, auf denen die Großmutter zu sehen ist. Und wenn das Netz einen Fehler macht, wird die Antwort korrigiert und der Lernalgorithmus passt das Netz entsprechend an.

Für die Spracherkennung von Google, Amazon und Apple hör(t)en deshalb Menschen die Mitschnitte von Unterhaltungen mit Siri, Alexa und Co ab, um verbliebene Fehler der Spracherkennung zu korrigieren. Als die Presse über die Mitschnitte berichtete, war es für viele Nutzer ein Schock zu erfahren, dass ihre Gespräche mit KI-Systemen nicht vertraulich sind.[13] Aber wenn man weiß, wie die Technik funktioniert, ist das keine Überraschung. Hinter den meisten Anwendungen neuronaler Netze stecken viele, viele Stunden Arbeit von Menschen, die mit ihrem impliziten Wissen die Fehler markieren, aus denen die Maschinen lernen. Mittlerweile gibt es eine ganze Industrie, die diese kleinteilige und mühselige Arbeit auf viele Menschen in der ganzen Welt – aus Kostengründen oft in Entwicklungsländern – verteilt.[14]

Neuronale Netze lernen auch ohne Lehrer

Kleine Kinder lernen sicher nicht, wie eine Katze oder ein Hund aussieht, indem ihnen die Eltern nacheinander tausende von Katzen- und Hundebilder zeigen und jedes Mal fragen, ob ein Hund oder eine Katze zu sehen ist, um entweder zustimmend zu nicken oder das Kind zu

12 Siehe Eckstein, Koehler, Welbourne & Akbas (2017) für die Riesenzahnbürste und Geirhos et al. (2018) für verschiedene Arten von veränderten Bildern.

13 Erst nachdem die Presse darüber berichtet hatte, wurde diese Verfahrensweise transparent gemacht oder abgestellt (Hurtz, 2019).

14 Siehe Dzieza (2023).

korrigieren. Tiere im Bilderbuch benennen ist eine tolle Beschäftigung, aber wahrscheinlich können schon Babys, die noch nicht sprechen, den Unterschied zwischen Hunden und Katzen sehen – unabhängig davon, ob jemand sie darauf hinweist. Hunde und Katzen sehen nicht nur unterschiedlich aus, sondern verhalten sich auch anders. Das macht es möglich, dass die Unterscheidung zwischen Hunden und Katzen von alleine entdeckt werden kann, ohne dass es dazu einen Lehrer braucht. Den Lehrer braucht es dann nur noch, um die Wörter zu lernen, nachdem das Kind schon erkannt hat, dass es sich um zwei unterschiedliche Tierarten handelt.[15]

Ein Kind, das Hunde und Katzen beobachtet, lernt vielleicht, dass bestimmte Merkmale oft zusammen auftreten. Hunde bellen und wedeln mit dem Schwanz, Katzen schnurren, sobald man sie streichelt. Viele Hunde haben Schlappohren, aber Katzen nicht, und so weiter. Das Kind lernt diese statistischen Regelmäßigkeiten, indem es die Merkmale miteinander assoziiert. Wenn es nun einen Hund mit dem Schwanz wedeln sieht, erwartet das Kind, dass er auch bellt.[16]

Sie erkennen das ›A‹ in Abbildung 8, obwohl das Bild verschmutzt ist, und nur Teile davon zu sehen sind. Sie wissen genau, was sich hinter dem Tintenklecks verbirgt, weil in Ihrem Gedächtnis alle Merkmale des Buchstabens miteinander assoziiert sind und Ihr Gedächtnis die fehlenden Teile deshalb ergänzen kann. So wie das Kind das Bellen erwartet, wenn es ein Schwanzwedeln sieht.

Lernen ohne Lehrer wird im maschinellen Lernen ›unüberwachtes Lernen‹ genannt und ein Beispiel dafür sind sogenannte ›autoassoziative neuronale Netzwerke‹. Das sind künstliche neuronale Netze, die die Eingabe mit sich selber assoziieren (daher der Name). Damit ist gemeint, dass das Netz lernt, welche Merkmale der Eingaben mit welchen anderen Merkmalen der gleichen Eingaben oft zusammen auftreten.

---

15 In Wirklichkeit ist das etwas komplizierter als hier dargestellt. Quinn, Eimas & Rosenkrantz (1993) haben untersucht, wie 3-4 Monate alte Kinder Hunde- und Katzenbilder kategorisieren. Beide sind leicht von Vögeln zu unterscheiden. Katzen sind untereinander sehr ähnlich, sodass sie auch leicht zusammengruppiert werden können. Hunde können allerdings äußerst verschieden aussehen, daher ist es für die Kinder manchmal schwer, Katzen nicht als eine Art von Hund anzusehen.

16 Diese Art von Assoziation, insbesondere zwischen den verschiedenen Sinnen, ist eine alte Idee von empiristischen Philosophen. George Berkeley stellte sich das schon 1709 in *An Essay Towards a New Theory of Vision* so ähnlich vor. In seinem Beispiel war es aber kein Hund, sondern eine Kutsche.

Wie immer in künstlichen neuronalen Netzen bedeutet Lernen, dass die Verbindungsstärke zwischen Neuronen angepasst werden. Die Grundidee ist, dass jedes Neuron in dem Netzwerk ein Merkmal der Eingabe repräsentiert (zum Beispiel das Bellen des Hundes oder den Querbalken des Buchstabens ›A‹). Diese Neuronen sind untereinander verbunden. Wenn zwei Neuronen oft zusammen aktiv sind, stärkt das die Verbindung zwischen den beiden Neuronen. Wird zu einem späteren Zeitpunkt nur eines der zwei Neuronen durch eine Eingabe aktiviert, sorgt die starke Verbindung zwischen den zwei Neuronen dafür, dass auch das andere Neuron aktiviert wird.[17] Die Aktivierung der Neurone für Schwanzwedeln alleine führt dann zur Aktivierung der Neurone für das Bellen – und umgekehrt. Die Aktivierung der Neurone für einzelne Teile eines ›A‹ führt zur Aktivierung der Neurone, die durch den Tintenfleck verdeckt sind.

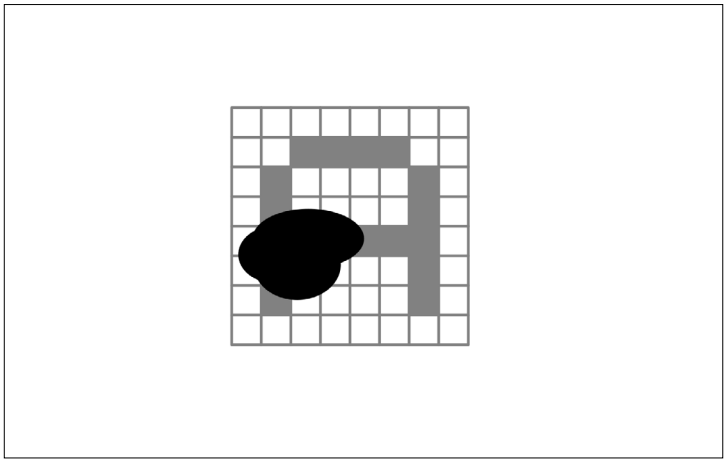

Abb. 8: Ein Bild mit Tintenklecks

17 Diese Art von Lernen nennt man auch ›Hebb'sches Lernen‹, benannt nach dem Psychologen Donald Hebb, der dieses Prinzip wesentlich klarer und neurowissenschaftlich plausibler beschrieben hat als viele Philosophen vor ihm (Hebb, 1949). Hebb'sches Lernen ist allerdings nicht die einzige Möglichkeit, wie man Lernen in autoassoziativen Netzen umsetzen kann.

Neuronale Netze lieben Katzenvideos

Die Kombination von unüberwachtem Lernen und überwachtem Lernen ist sehr mächtig. Überwachtes Lernen kann extrem aufwendig und kostspielig sein. Damit ein künstliches neuronales Netz den Unterschied zwischen Hunden und Katzen lernen kann, müssen Menschen erst eine unglaublich große Menge an Hunde- und Katzenbildern beschriften. Für viele Anwendungen ist das die Mühe nicht wert. Indem das neuronale Netz unüberwacht viele Bilder aus dem Internet durchsieht, kann es allerdings mit deutlich weniger menschlicher Unterstützung trainiert werden. Es hilft, wenn das Netz schon viele Hunde- und Katzenbilder gesehen hat – sogar, wenn es gar nicht weiß, dass es Hunde und Katzen waren. Es hilft aber auch, wenn es viele andere Bilder gesehen hat, weil es so die statistischen Regelmäßigkeiten in Bildern allgemein lernt. Vielleicht wurde ihm auch schon beigebracht andere Tierarten zu unterscheiden, was das Lernen von Hunden und Katzen zusätzlich vereinfacht. Solch ein vortrainiertes Netz, das allgemein etwas über Bilder gelernt hat, lässt sich dann leichter und billiger an verschiedene Aufgaben anpassen, als wenn man für jede Aufgabe ein neues neuronales Netz trainiert. Wir Menschen fangen ja auch nicht bei jeder neuen Aufgabe bei null an, sondern entwickeln unsere vorhandenen Fähigkeiten weiter.

Der beeindruckende Fortschritt in KI-Anwendungen, den man ungefähr seit 2010 beobachten kann, ist größtenteils dadurch angetrieben, dass künstliche neuronale Netze in vielen Fällen überraschend gut lernen, statistische Muster zu erkennen. Auch Menschen sind oft gut darin, Muster zu erkennen, aber unser Wissen darüber ist implizit. Wir können nicht sagen, nach welchen Regeln wir unsere Großmutter erkennen oder wie genau wir Hunde von Katzen unterscheiden. Das macht es für Entwicklerinnen und Entwickler schwer, traditionelle Computerprogramme für solche Mustererkennungsaufgaben zu schreiben, denn sie können dem Computer nicht Schritt für Schritt Anweisungen geben, wie die Aufgaben zu lösen sind. Besser ist es, ein Computerprogramm zu schreiben, das selbständig aus Beispielen lernt. KI-Forscherinnen und -Forscher basteln daher seit den Anfangstagen der Computer an künstlichen neuronalen Netzen, die lernen können, Muster zu erkennen. Warum sehen wir dann erst über 50 Jahre später funktionierende Gesichts- und Spracherkennung? Das lag daran, dass tiefe neuronale Netze schlicht eine extrem große Menge an Beispielen

brauchen, um ihr implizites Wissen zu lernen – mehrere Millionen Beispiele sind keine Seltenheit.[18] Deshalb mussten erst zwei Entwicklungen zusammenkommen: Zum einen mussten Computer schnell genug werden, um solche großen Datenmengen überhaupt verarbeiten zu können. Zum anderen mussten entsprechend große Datenmengen erst einmal zur Verfügung stehen. Diese Datenmengen finden sich im Internet, wo Nutzer jeden Tag Unmengen an Texten, Bildern und Videos teilen. Wer hätte gedacht, dass all diese Katzenvideos zu etwas gut sein würden.[19]

18 Das Training von Netzen zur Objekterkennung geschah zum Beispiel lange gerne mit der Datenbank ImageNet, die etwa 14 Millionen Bilder enthält (Deng et al., 2009).

19 Le et al. (2012) haben zehn Millionen Einzelbilder aus YouTube-Videos extrahiert und ein künstliches neuronales Netz unüberwacht trainiert. Nach dem Training hatte das Netz Neuronen, die selektiv menschliche Gesichter erkennen konnten, weil Gesichter häufig in den Videos vorkamen. Aber Katzen kamen auch häufig vor, weshalb das Netz auch ohne Lehrer gelernt hat, wie Katzen aussehen.