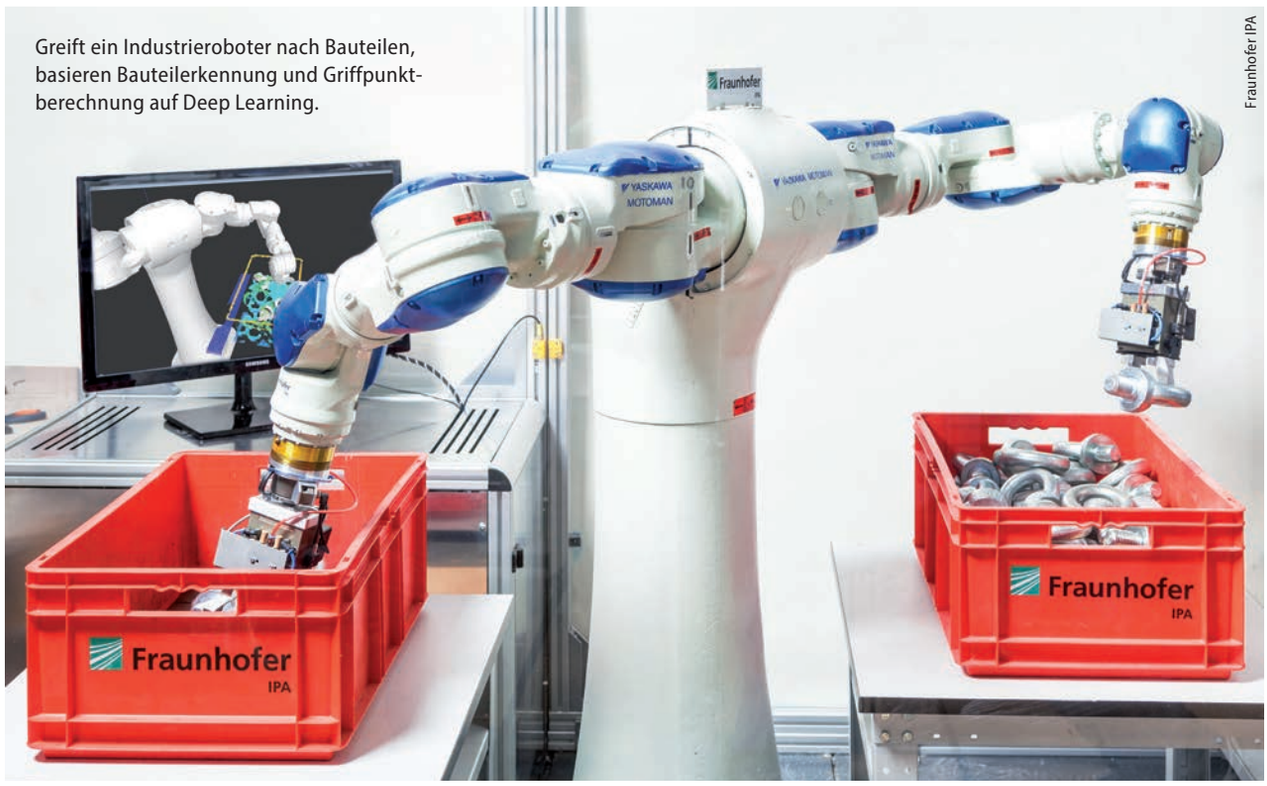


Greift ein Industrieroboter nach Bauteilen, basieren Bauteilerkennung und Griffpunktberechnung auf Deep Learning.

Fraunhofer IPA

KÜNSTLICHE INTELLIGENZ

Lernen aus der Black Box

Cognitive Deep Learning soll neuronale Netze und Wissensverarbeitung kombinieren.

Marco F. Huber

Maschinen sind dem Menschen bei vielen Aufgaben überlegen. Ein Computer führt numerische Berechnungen um Größenordnungen schneller und präziser durch, und auch beim fehlerlosen Speichern und Abrufen großer Datenmengen ist die Maschine ungeschlagen. Was uns auszeichnet, sind die kognitiven Fähigkeiten: Zwar übertreffen künstliche Sensoren unsere Sinneswahrnehmungen, doch in der Verarbeitung dieser Information, also im Lernen, Erinnern und Anwenden, zeigt unser Gehirn seine wahre Leistungsfähigkeit. Durch das sogenannte Deep Learning gelang es in den letzten Jahren, den Rückstand von Maschinen bei Mustererkennung, Sprachverarbeitung oder Problemlösefähigkeit deutlich zu verkürzen und in Teilen sogar in eine Überlegenheit zu wandeln.

Von allen Sinnesorganen des Menschen ist der Sehsinn für das Wahrnehmen der Umwelt der wichtigste: Er liefert rund 80 Prozent aller Informationen. Sehen bedeutet, dass elektromagnetische Lichtwellen die Hornhaut und das gesamte Auge durchdringen. Auf der Netzhaut angelangt, regen sie Nervenzellen an, welche die Lichtreize über den Sehnerv an das Gehirn weiterleiten. Erst dort entsteht das Bild. Im visuellen Cortex, dem Bereich des Gehirns, der für das Sehen zuständig ist, reagieren Nervenzellen auf unterschiedliche Reize wie bestimmte Farbkombinationen oder Hell-Dunkel-Kontraste. Solche Eindrücke ermöglichen durch den Vergleich mit gespeicherten Bildern das Wiedererkennen eines bekannten Gesichts.

Den wichtigen Sehsinn bilden auch viele technische Anwendungen nach. Anstelle des Auges tritt dabei die Kamera, und ein Computer übernimmt die Aufgabe des Gehirns. Die klassische Bildverarbeitung stützt sich auf „Merkmale“, die den gezeigten Inhalt des aufgenommenen Bilds möglichst gut charakterisieren. Dazu gehören Kanten, Ecken oder andere geometrische Grundformen, die sich durch Filter extrahieren lassen. Mustererkennungsverfahren aus der Statistik oder der Künstlichen Intelligenz überführen diese Merkmale in Aussagen zum Bildinhalt, etwa in die Aussage „Frau“, wenn das Bild eine weibliche Person zeigt.

Die Leistungsfähigkeit der Algorithmen hängt dabei stark von der Wahl und Zusammenstellung der richtigen Merkmale und Verfahren ab, wozu viel Erfahrung und auch „Ausprobieren" beitragen. Deep Learning erleichtert diese teils beschwerliche Arbeit und verarbeitet nun die Rohdaten direkt, indem es die Merkmale – etwa die richtigen Filter – automatisch bestimmt.

Vom Perzeptron zum Deep Learning

Deep Learning nutzt künstliche neuronale Netze, um die Daten so zu transformieren, dass die gegebene Aufgabe möglichst genau gelöst wird – etwa das Erkennen von Bildinhalten. Die Arbeitsweise neuronaler Netze orientiert sich an den Vorgängen im menschlichen Gehirn, welches schätzungsweise aus 86 Milliarden Nervenzellen besteht. Jede Nervenzelle, auch Neuron genannt, ist durch seine Dendriten (**Abb. 1a**) mit durchschnittlich zehntausend anderen Neuronen verbunden. Die Kommunikation zwischen Neuronen geschieht mittels elektrischer Signale; Synapsen wandeln diese in chemische Botenstoffe bzw. Neurotransmitter um. Die Leitfähigkeit einer Synapse hängt dabei von der Verfügbarkeit der Neurotransmitter ab. Somit repräsentiert eine Synapse die Stärke der Verbindung zwischen zwei Neuronen, und das Lernen im Gehirn erfolgt im Wesentlichen durch das Verstärken oder Abschwächen dieser Verbindungen. Alle Signale, die an einem Neuron gleichzeitig ankommen, addieren sich, sodass das Neuron „feuert", sobald die Summe einen Grenzwert überschreitet: Ein elektrischer Impuls schießt dann am Axon entlang.

Der US-amerikanische Psychologe Frank Rosenblatt hat 1957 mit dem Perzeptron ein vereinfachtes mathematisches Modell eines Neurons entwickelt (**Abb. 1b**). Aus einem Satz von Eingaben x_i, jeweils multipliziert mit einem kontinuierlichen Gewicht w_i, das der Stärke der Synapse entspricht, ergibt sich eine gewichtete Summe. Überschreitet diese den Schwellwert oder Bias b, folgt der Funktionswert Eins und das Neuron feuert:

$$y = \Sigma_i w_i x_i - b \quad \text{und} \quad f(y) = \begin{cases} 0, y < 0 \\ 1, y \geq 0 \end{cases}$$

Die Aktivierungsfunktion $f(y)$ entscheidet über den Schwellwert. Aufgrund ihrer Unstetigkeit kommen heute andere Aktivierungsfunktionen zum Einsatz, etwa die Sigmoidfunktion $[1 + \exp(-y)]^{-1}$.

Ein solches Perzeptron ist in der Lage, eine mehrdimensionale lineare Funktion zu repräsentieren. Damit lassen sich beispielsweise lineare Klassifikationsprobleme lösen: Die lineare Funktion trennt die gegebenen Daten derart in zwei Mengen auf, dass eine davon zum Feuern des Perzeptrons führt, während bei der anderen das Perzeptron ruht. „Lernen" bedeutet hier, die Gewichte des Perzeptrons so anzupassen, dass die lineare Funktion die gegebenen Daten möglichst fehlerfrei trennt. Rosenblatts spezieller Lernalgorithmus passt die Gewichte genau dann an, wenn der Ausgabewert des Neurons vom Sollwert abweicht.

Reale Problemstellungen lassen sich in der Regel nicht durch lineare Funktionen repräsentieren. Vielmehr herrschen nichtlineare Probleme vor, für die das Perzeptron ungeeignet ist. Analog zum natürlichen Vorbild kann aber der Ausgang eines Perzeptrons als Eingang eines anderen dienen. Dann entstehen mehrschichtige künstliche neuronale Netze mit einer Vielzahl künstlicher Neuronen in jeder Schicht. Streng genommen spiegelt die erste Schicht die Eingabedaten wider und besitzt als Eingabeschicht keine Neuronen bzw. Perzeptronen. Man kann aber auch in der Eingabeschicht die Datenanzahl mit der Zahl von Neuronen gleichsetzen. Die letzte Schicht ist die Ausgabeschicht, in der die Menge an Neuronen von der gewünschten Anzahl möglicher Ergebnisse abhängt. Dazwischen können beliebig viele sogenannte verdeckte Schichten vorhanden sein. Schon in einem dreischichtigen Netz mit einer verdeckten Schicht lässt sich jede stetige Funktion beliebig genau approximieren, sofern ausreichend viele Neuronen zur Verfügung stehen [1]. Allerdings funktioniert in mehrschichtigen Netzen der Perzeptron-Lernalgorithmus nicht mehr. Das führte vom Ende der 1960er- bis in die 1980er-Jahre zum „AI Winter", während dem die Forschung an künstlichen neuronalen Netzen fast vollständig zum Erliegen kam.

Bei mehrschichtigen Netzen ist nur die Ausgabe der Neuronen der letzten Schicht sichtbar. Abweichungen, die Neuronen innerer, verdeckter Schichten auslösen, ließen sich nicht berechnen und ihre Gewichte nicht anpassen. Dafür bietet der 1986 vorgestellte Backpropagation-Algorithmus eine Lösung [2]. Im Kern handelt es sich hierbei um das Lösungsverfahren eines Optimierungsproblems, bei

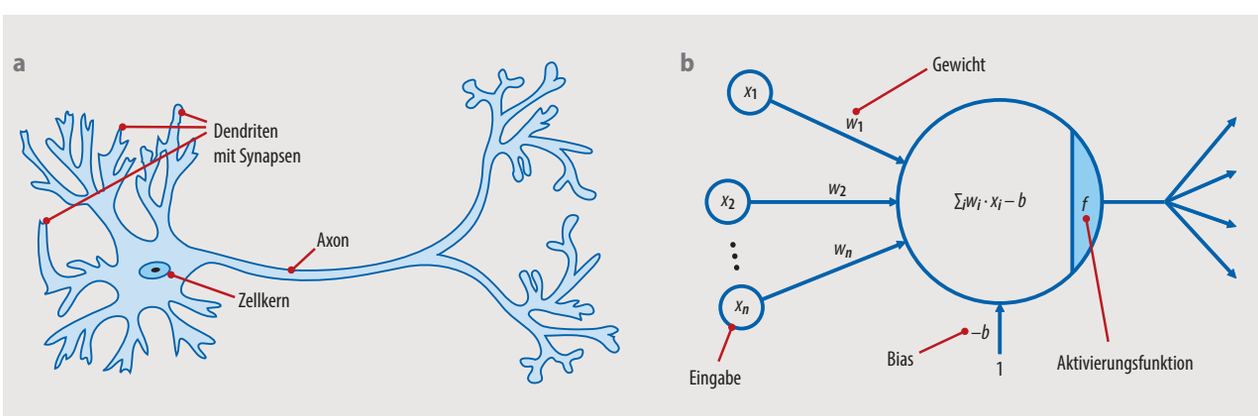


Abb. 1 Eine Nervenzelle des menschlichen Gehirns (a) lässt sich – zumindest prinzipiell – mit einem Perzeptron (b) nachbilden.

dem die Abweichung zwischen der Netzausgabe und dem wahren Wert zu minimieren ist (**Abb. 2**). Da nur Bekanntes verbesserbar ist, dient eine sogenannte Fehlerfunktion dem Bewerten der Abweichung. Mittels Backpropagation lässt sich die berechnete Abweichung Schicht für Schicht an alle Neuronen übermitteln. Zunächst wird die Abweichung der Ausgabeneuronen auf deren eigene Gewichte verteilt. Jedes Neuron der vorangehenden verdeckten Schicht kann damit anteilig eine individuelle Abweichung errechnen, die wiederum anteilig an dessen Vorgänger verteilt wird – so lange, bis jedes Gewicht im Netz, also auch in der ersten Schicht, angepasst wurde.

Mathematisch betrachtet ist Backpropagation ein Gradientenabstiegsverfahren, das die Fehlerfunktion bezüglich der Gewichte des Netzes minimiert. Bei der erforderlichen Ableitung der Fehlerfunktion nach den Gewichten zeigt sich aufgrund der Kettenregel, dass die Ableitung eines Neurons einer bestimmten Schicht von den Ableitungen nachfolgender Schichten abhängt. Sofern ein Neuron also nicht in der Ausgabeschicht liegt, kann die Gewichtsanpassung nur indirekt erfolgen, eben durch Zurückpropagieren der anteiligen Abweichung.

Beim Lernen mittels Backpropagation werden zunächst alle Daten eines Datensatzes nacheinander dem Netz zugeführt und vorwärts durch das Netz propagiert. Die Fehlerfunktion liefert einen Vergleich der Ausgabe des Netzes mit den tatsächlichen Werten. Abweichungen bzw. Fehler werden zurück ins Netz propagiert und dienen der Anpassung der Gewichte aller Neuronen. Dies wird so lange wiederholt, bis der Fehler einen gewünschten Wert unterschreitet oder sich der Fehler nicht weiter reduziert.

„Deep“ bezieht sich also nicht auf ein besonders tiefes Problem- oder Lösungsverständnis. Vielmehr ist damit gemeint, dass viele aufeinander aufbauende Schichten nötig sind, um eine zunehmend wirkmächtige Datentransformation zu erhalten. Nicht wenige Fachleute bezeichnen den Begriff „Deep Learning“ als gutes Marketing für eine eigentlich Jahrzehnte alte, aber wirkungsvolle Idee. Nichtsdestotrotz liegt in der Tiefe, also dem Verwenden vieler Schichten, der Wirkmechanismus von Deep Learning verborgen. Das Anordnen der Neuronen erfolgt in den grundlegenden Netzarchitekturen Feedforward, Convolutional und Recurrent (vgl. dazu den Artikel von Martin Erdmann).

Aus Fehlern über das Lernen lernen

Moderne Deep-Learning-Netze bestehen teilweise aus mehreren hundert Schichten, obwohl in der Theorie eine verdeckte Schicht ausreicht, um beliebige stetige Funktionen zu repräsentieren. Den Nutzen vieler Schichten postulierte Yoshua Bengio, eine der zentralen Persönlichkeiten für das Deep Learning, 2007 so [3]: „We claim that most functions that can be represented compactly by deep architectures cannot be represented by a compact shallow architecture.“

Kürzlich ließ sich nachweisen, dass tiefe künstliche neuronale Netze mit exponentiell weniger Neuronen bestimmte Funktionen besser approximieren als flache Netze

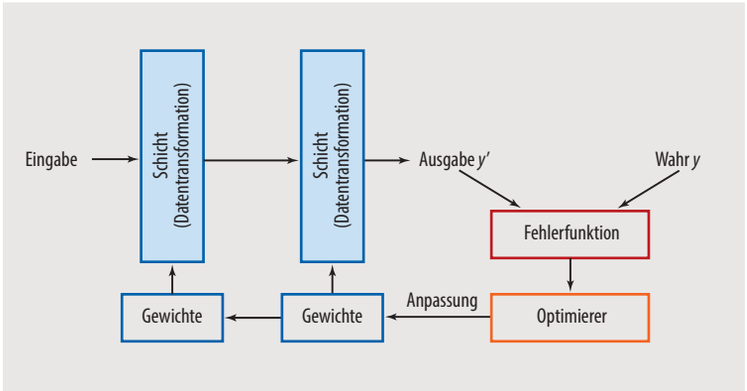


Abb. 2 Ein künstliches neuronales Netz lernt durch das Optimieren der Gewichte in seinen einzelnen Schichten.

mit sehr vielen Neuronen pro Schicht [4]. Die Verwendung vieler Schichten erlaubt dabei die Konstruktion immer komplexerer Datentransformationen bzw. Merkmale, wobei die einzelnen Transformationen aufeinander aufbauen (**Abb. 3**). Die einfachen geometrischen Strukturen aus der ersten Schicht – beispielsweise Kanten extrahiert aus einem Bild – werden in der zweiten Schicht zu Gesichtsteilen und in der dritten Schicht zu kompletten Gesichtern zusammengefügt: Je tiefer das Netz, umso komplexer die extrahierte Datenrepräsentation. Wie viele Schichten und Neuronen erforderlich sind, lässt sich nicht pauschal sagen, sondern hängt von der konkreten Aufgabe ab. Bei zu wenig Neuronen repräsentiert das Netz die Daten nur unvollständig; bei zu vielen Schichten und Neuronen passt sich das Netz zu sehr den Daten an. Dieses Phänomen des „Overfitting“ kommt einem Auswendiglernen der Daten gleich: Das Netz generalisiert schlecht, liefert also bei der Auswertung unbekannter Daten keine guten Ergebnisse.

Einer der wesentlichen Kritikpunkte bei Deep Learning ist der „Black Box“-Charakter. Die Zusammenhänge und Datenrepräsentationen, die ein künstliches neuronales Netz lernt, sind so komplex und abstrakt, dass Menschen – selbst Experten – sie nicht mehr nachvollziehen können. Dieser Umstand verstärkt sich mit zunehmender Tiefe der Netze, ist aber auch bei vielen anderen Verfahren des maschinellen Lernens gegeben. Quantitativ lässt sich die Leistungsfähigkeit anhand verschiedener Metriken bewerten, beispielsweise wie genau einzelne Klassen vorhergesagt werden oder wie hoch die Fehlalarmrate ist. Das Zusammenspiel einzelner Neuronen und die Auswirkungen der gelernten Zusammenhänge bleiben aber immer verborgen. Dieser Umstand hat bereits zu einigen sehr unerwarteten, teils peinlichen Ergebnissen geführt.

So hat sich ein tiefes künstliches neuronales Netz, das Fotos von Pferden perfekt klassifizieren konnte, gar nicht auf spezifische Pferdemerkmale gestützt. Im Nachhinein zeigte sich vielmehr, dass alle Bilder von einem Fotografen stammten und das Netz eine Copyright-Angabe am Rand der Bilder nutzte, um eine Entscheidung zu treffen. Andere Negativbeispiele nährten die Befürchtung, die Physiognomik könnte wiedererstarken, die menschliche Eigenschaften bestimmten Gesichtsmerkmalen zuordnet. So behaupteten Wissenschaftler, von ihnen trainierte Netze könnten

die sexuelle Orientierung oder kriminelle Neigung eines Menschen anhand eines Porträtfotos mit hoher Wahrscheinlichkeit erkennen [5]. Erst eine tiefgehende, teils manuelle Auswertung der Fotos zeigte, dass sich die Netze nicht an Gesichtsmerkmalen, sondern eher an Äußerlichkeiten wie noch sichtbarer Oberbekleidung oder dem Make-up orientierten. Wie fragil die Leistungsfähigkeit tiefer Netze teilweise sein kann, zeigen speziell entworfene Daten, welche die Netze zu Fehlentscheidungen zwingen, beispielsweise Fotos, denen spezielle Rauschmuster überlagert sind. Für das menschliche Auge sind verrauschtes Ergebnis und Originalfoto nicht unterscheidbar, doch das künstliche neuronale Netz liefert ein völlig anderes Klassifikationsergebnis. Im Extremfall reicht die Manipulation sehr weniger Pixel eines Bildes, um das Netz aus dem Tritt zu bringen. Dieser Zusammenhang ist als „Ein-Pixel-Attacke" bekannt.

Diese Fälle machen deutlich, dass sich die Datenverarbeitung der künstlichen neuronalen Netze deutlich von der unseres Gehirns unterscheidet. Solange die Genauigkeit stimmt, scheint das Verstehen der internen Mechanismen eines Netzes auch gar nicht wesentlich zu sein, beispielsweise bei einer Produktempfehlung im Online-Shop. Trifft sie den Geschmack des Nutzers nicht, kann sie mit einem Achselzucken ignoriert werden. Wesentlich kritischer ist das Verständnis beim Einsatz in der Medizin, der Kreditvergabe oder beim autonomen Fahren, wo Fehlentscheidungen sich einschneidend auf das Leben des Menschen auswirken können. Nicht umsonst verlangt Artikel 13 der am 25. Mai 2018 in Kraft getretenen Datenschutzgrundverordnung „aussagekräftige Informationen über die involvierte Logik" bei automatisierten Entscheidungsfindungen, welche gemäß Artikel 12 „in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln" sind [6].

Erklärbare Entscheidungen?

Eine für den Menschen transparente Präsentation und Verarbeitung von Wissen erfordert es, die komplexe Thematik der Erklärbarkeit zu erforschen. Hierbei ist zwischen der Erklärbarkeit des Modells und derjenigen der Daten zu unterscheiden. Ansätze zur Modellerklärbarkeit, auch globale Erklärbarkeit genannt, versuchen, die Wirkmechanismen des Netzes nachvollziehbar zu machen. Beispiele hierfür sind die Extraktion stetig abstrakter werdender gelernter Merkmale eines Netzes (**Abb. 3**) und die Übersetzung eines Netzes in ein „White Box"-Modell [7], also beispielsweise in eine Menge von Regeln oder in einen Entscheidungsbaum. Während es hier um ein ganzheitliches Nachvollziehen des gelernten Modells geht, widmet sich die Datenerklärbarkeit – auch lokale Erklärbarkeit – der Bereitstellung menschenverständlicher Zusatzinformationen für einzelne Eingabedaten. Sie erklärt also, wie das Netz für ein vorliegendes Fallbeispiel zu seiner Aussage gelangt, ohne dabei auf die Details der Wissensverarbeitung des Netzes einzugehen. Ein möglicher Ansatz ist es, den Beitrag einzelner Elemente der Eingabedaten zur Ausgabe des Netzes zu ermitteln [8]. Für Bilder bedeutet dies, Bereiche hervorzuheben, die für die Entscheidungsfindung wesentlich waren; bei Texten wären relevante Wörter oder Textsegmente zu markieren.

Erst eine nachvollziehbare Wissensverarbeitung schafft Vertrauen in die Entscheidungen künstlicher neuronaler Netze. Die vorgestellten Ansätze helfen aber auch, die Leis-

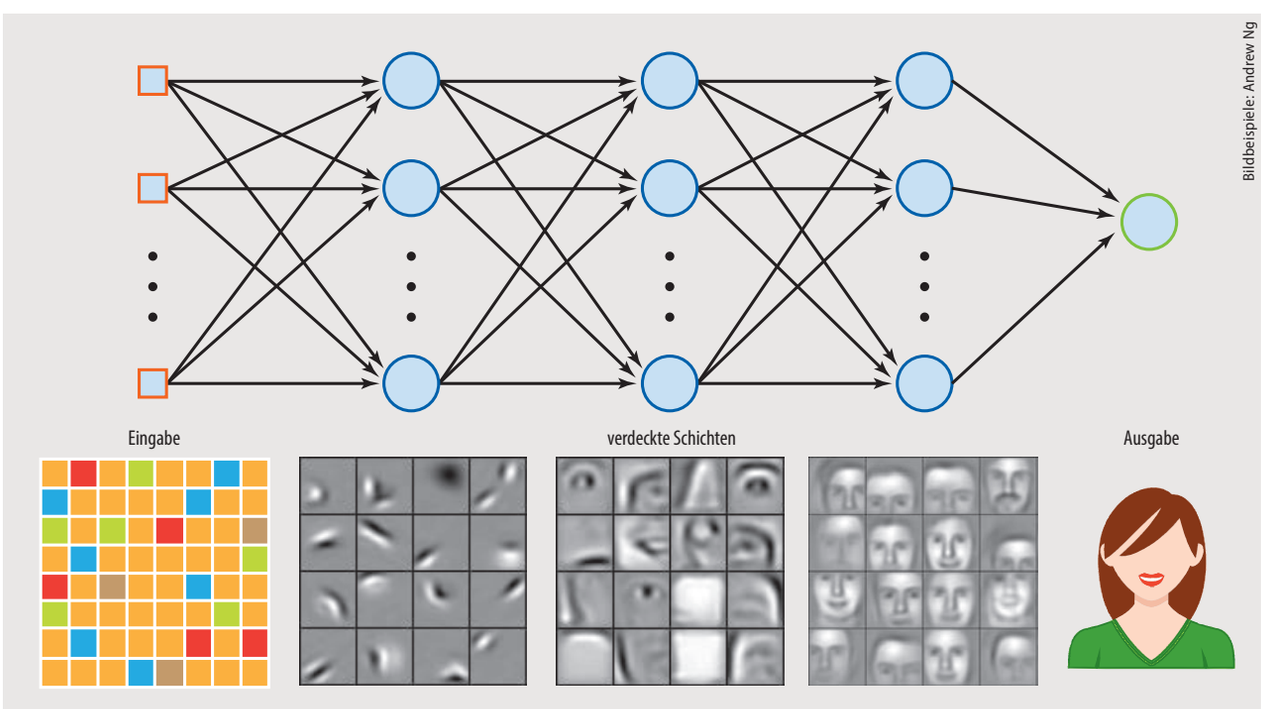


Abb. 3 Die Transformation der Eingabedaten zu einer eindeutigen Ausgabe lässt sich für ein tiefes künstliches neuronales Netz gut am Beispiel der Gesichtserkennung innerhalb von Bildern illustrieren.

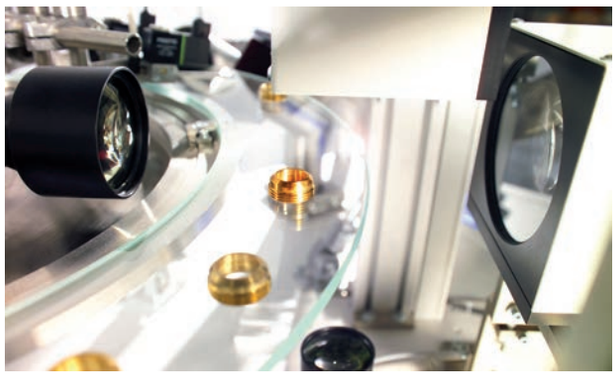

Hörmle GmbH

Abb. 4 Kameras liefern die Daten, um mit Künstlicher Intelligenz die Qualität einer Produktion zu überwachen.

tungsgrenzen der Netze auszuloten, d. h. festzustellen, wie lange das Netz korrekte Ergebnisse liefert. Das ist insbesondere hilfreich, um es zu verbessern und weiterzuentwickeln.

Von Diagnostik bis zu autonomem Fahren

Die Möglichkeit, Deep Learning mit überschaubarem Aufwand zu nutzen, hat zu beachtlichen Leistungen in vielfältigen Anwendungen geführt. In der Medizin spielen zur Diagnostik bildgebende Verfahren wie das Röntgen oder die Sonographie eine wichtige Rolle. Convolutional-Netze erlauben es, mit hoher Genauigkeit beispielsweise Hautkrebs [9] oder durch Diabetes verursachte Augenkrankheiten [10] zu erkennen. Auch beim autonomen Fahren ist das Zusammenspiel von Bildverarbeitung und Deep Learning wesentlich: Bilddaten von Kameras, Radar oder Lasern dienen den tiefen Netzen, um Fahrspuren, Fußgänger, Verkehrszeichen oder Hindernisse zu erkennen und das Fahrzeug autonom zu steuern.

Neben Bildern lässt sich auch Sprache mit Deep Learning erkennen: Systeme wie Alexa, Siri oder Google Now sind dadurch robust und sprecherunabhängig. Einer breiten Öffentlichkeit bekannt ist der Fall des Computerprogramms „AlphaGo“ der Google-Tochter DeepMind, das von 2015 bis 2017 nacheinander die weltbesten Spieler im strategischen Brettspiel Go klar besiegen konnte. Aufgrund der deutlich höheren Komplexität im Vergleich zu Schach hatten Experten dies erst für die 2020er-Jahre erwartet.

Auch in der industriellen Produktion halten kognitive Lernverfahren immer mehr Einzug in vielen Einsatzgebieten. Bei der Qualitätskontrolle produzierter Güter lassen sich Mängel mittels Deep Learning anhand von Bilddaten automatisiert und zuverlässig erkennen (**Abb. 4**). Die sogenannte prädiktive Instandhaltung, bei welcher der Ausfall von Produktionsmaschinen noch vor dessen Eintreten erkannt und somit vermieden wird, sichert eine konstante Produktionsleistung. Für die erforderliche statistische Analyse der Daten kommen verschiedene maschinelle Lernverfahren zum Einsatz, insbesondere Deep Learning [11]. Ein wesentliches Element der industriellen Fertigung sind heute Roboter, für deren Programmierung und Steuerung vermehrt auf Deep Learning zurückgegriffen wird (**Abb.** auf S. 37). Auch für das Optimieren von Produktionsparametern, wie Vorschubgeschwindigkeiten oder Anpressdrücken, lässt sich Deep Learning nutzen.

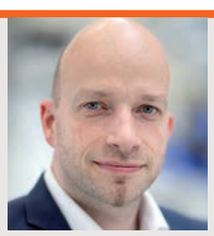
Fazit

Deep Learning zeichnet sich als zurzeit dominantes Teilgebiet der Künstlichen Intelligenz insbesondere bei kognitiven Aufgabestellungen durch eine hohe, teils übermenschliche Leistungsfähigkeit aus. Allerdings lassen sich lediglich Muster erkennen, ohne ein tiefes Verständnis oder Bewusstsein über das Gelernte und Zusammenhänge. Dies erschwert es, die Leistung eines trainierten Netzes auch auf andere Aufgabenstellungen zu übertragen – eine Transferleistung, welche der Mensch sehr oft mit Bravour meistert.

Literatur

[1] *G. Cybenko*, Mathematics of Control, Signals and Systems **2**, 303 (1989)
[2] *D. E. Rumelhart, G. E. Hinton* und *W. J. Ronald*, Nature **323**, 533 (1986)
[3] *Y. Bengio* und Y. *LeCun*, in: Large-Scale Kernel Machines, hrsg. v. *L. Bottou* et al., MIT Press (2007)
[4] *H. Mhaskar, Q. Liao* und *T. Poggio*, in: Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, USA (2017)
[5] *Xiaolin Wu* und *Xi Zhang*, arXiv: 1611.04135v1 (2016); *Xiaolin Wu* und *Xi Zhang*, arXiv: 1611.04135v3 (2017)
[6] Regelungen der DSGVO unter https://bit.ly/391ER0N
[7] *N. Schaaf, M. F. Huber* und *J. Maucher*, in: IEEE International Conference on Machine Learning and Applications (ICMLA), Florida, USA (2019)
[8] *S. Bach* et al., PLoS ONE **10**, Nr. 7, 10. Juli 2015
[9] *A. Esteva* et al., Nature **542**, 115 (2017)
[10] *V. Gulshan* et al., JAMA **316**, 2402 (2016)
[11] *K. Wang* und *Y. Wang*, How AI Affects the Future Predictive Maintenance: A Primer of Deep Learning, in: Advanced Manufacturing and Automation VII, *K. Wang* et al., Springer (2018)

Der Autor



Marco F. Huber forscht zu den Themen maschinelles Lernen, Sensordatenanalyse und Robotik im produktionstechnischen Umfeld. Nach Informatik-Studium und erfolgreicher Promotion an der U Karlsruhe (TH), war er von 2009 bis 2011 Gruppenleiter am Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB) in Karlsruhe. Nach dem Wechsel in die Industrie verantwortete er bis 2018 in unterschiedlichen Leitungsfunktionen die Themenschwerpunkte Künstliche Intelligenz, maschinelles Lernen und Big-Data-Analyse im Bereich Industrie 4.0. Seit Oktober 2018 hat er die Professur für kognitive Produktionssysteme an der U Stuttgart inne und leitet zugleich das Zentrum für Cyber Cognitive Intelligence (CCI) sowie die Abteilung Bild- und Signalverarbeitung am Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA.

Univ.-Prof. Dr.-Ing. Marco F. Huber, Universität Stuttgart, Institut für Industrielle Fertigung und Fabrikbetrieb, Allmandring 35, 70569 Stuttgart und Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, Nobelstr. 12, 70569 Stuttgart