VORWORT ZUR DEUTSCHEN AUSGABE

VON JÜRGEN SCHMIDHUBER

Nach mehr als einem Jahrhundert Forschung zur Künstlichen Intelligenz (KI) hat das Gebiet jüngst enorm an Bedeutung und Beliebtheit gewonnen. Insbesondere die Mustererkennung und das maschinelle Lernen wurden durch das sogenannte *Tiefe Lernen* oder *Deep Learning* (DL) revolutioniert. DL ist dabei nur ein neuer Spitzname für tiefe künstliche Neuronale Netze (NN), die aus Erfahrung lernen. Heute wird DL in der Industrie und im Alltag massiv genutzt, z. B. für die Bild- und Spracherkennung auf Ihrem Smartphone oder für die automatische Übersetzung von einer Sprache in eine andere.

Das vorliegende Buch ist eine Sammlung von Interviews mit Forschern, die an Aspekten der KI gearbeitet haben. Allerdings haben alle Befragten ihren Sitz in der Anglosphäre, obwohl DL und die moderne KI an Orten erfunden wurden, wo Englisch keine offizielle Sprache ist. Deshalb werde ich dieses Vorwort nutzen, um einige der im Rest des Buches nicht berücksichtigten Pioniere des Gebiets hervorzuheben.

Lassen Sie uns zunächst einen Schritt zurücktreten und einen Blick auf die KI-Geschichte im weiteren Kontext der Geschichte des automatischen Rechnens werfen.

Europa ist die Wiege der mechanischen Rechner und der KI. Der Antikythera-Mechanismus Altgriechenlands (aus dem ersten Jahrhundert v. Chr.) wurde an Raffinesse erst eineinhalb Jahrtausende später übertroffen durch Henleins miniaturisierte Taschenuhren (Nürnberg, 1505). 1623 konstruierte Schickard in Tübingen den ersten Rechner für Grundrechenarten. Um 1700 entwickelte Leibniz die heute allgemein übliche binäre Computerarithmetik. Mustererkennung durch lineare Regression oder *Flaches Lernen* begann um 1800 durch Gauß & Legendre. Die ersten programmgesteuerten Maschinen (es waren Webstühle) entstanden ebenfalls zu dieser Zeit in Frankreich durch Jacquard und andere. Bald darauf kamen erste

(allerdings unverwirklichte) Ideen für Allzweckrechner auf (Babbage in Großbritannien um 1830). Im vergangenen Jahrhundert beschleunigten sich die Fortschritte dramatisch. Hier ein kleiner Überblick über Glanzlichter der KI seit 1900:

Meines Wissens war der erste Pionier der *praktischen* KI der Spanier Leonardo Torres y Quevedo, der schon 1914 eine funktionierende Schach-Endspielmaschine schuf. Damals galt Schachspielen noch als intelligente Aktivität.

Der Begründer der KI-Theorie und der theoretischen Informatik im Allgemeinen war der Österreicher Kurt Gödel, der 1931 die erste universelle formale Sprache veröffentlichte (basierend auf den ganzen Zahlen), und damit nicht nur beliebige Rechenprozeduren wie z. B. Theorembeweiser beschreiben konnte, sondern auch selbstbezügliche formale Aussagen aufstellte, die von sich behaupten, dass kein solcher Theorembeweiser sie beweisen kann. So enthüllte er grundlegende Grenzen der Mathematik, der Berechenbarkeit und der KI (1931). Ein Großteil der späteren KI in den 1960er-/70er-Jahren drehte sich in der Tat um Deduktion und das Beweisen von Theoremen im Gödel-Stil durch Expertensysteme. Mehr unter: *http://people.idsia.ch/~juergen/goedel.html*

1935 veröffentlichte Alonzo Church in den USA eine alternative universelle Sprache namens Lambda-Kalkül (Basis von LISP) und erweiterte damit Gödels Ergebnisse auf Hilberts berühmtes Entscheidungsproblem, welches im Allgemeinen unlösbar ist. Im folgenden Jahr formulierte Alan Turing in Großbritannien dieses Resultat mithilfe eines weiteren universellen Konstruktes, der Turing-Maschine (1936), um. Später schlug er auch einen subjektiven KI-Test vor. Mehr unter: *http://people.idsia.ch/~juergen/turing.html*

Von 1935-1941 baute Konrad Zuse in Berlin den ersten praktischen, funktionierenden, programmgesteuerten Computer. In den 1940er-Jahren entwickelte er auch die erste höhere Programmiersprache und schrieb das erste allgemeine Schachprogramm (Schachspielen galt damals immer noch als intelligente Aktivität). Mehr unter: *http://people.idsia.ch/~juergen/zuse.html*

Der Begriff *KI* wurde 1956 durch John McCarthy auf der Dartmouth Konferenz geprägt. Das Thema an sich wurde jedoch bereits fünf Jahre zuvor auf der berühmten Pariser Konferenz zu Rechenmaschinen und menschlichem Denken behandelt. Dort spielte u. a. der Kybernetik-Pionier Norbert Wiener gegen die bereits erwähnte Schachmaschine von Torres. Die entsprechende 570 Seiten umfassende Veröffentlichung heißt *Les Machines à Calculer et la Pensee Humaine*: Paris, 8.-13. Januar 1951, *Colloques internationaux du Centre National de la Recherche Scientifique*; Nr. 37, Paris 1953. Herbert Bruderer bezeichnet die Pariser Konferenz zu Recht als die erste zur KI.

Der erste berühmte Fall von Mustererkennung durch *Flaches Lernen* trug sich meines Wissens schon vor zwei Jahrhunderten zu: Die Wiederentdeckung des Plane-

toiden Ceres um 1800 durch Carl Friedrich Gauß, der aus Datenpunkten früherer Beobachtungen mit verschiedenen Tricks die Parameter eines Prädiktors adjustierte, welcher die Trainingsdaten verallgemeinerte, um den neuen Standort von Ceres richtig vorherzusagen. Damals kamen Gauß und der französische Mathematiker Adrien-Marie Legendre auf die noch immer weit verbreitete Methode der kleinsten Quadrate und der Regression zur Mustererkennung. Frank Rosenblatts einschichtiges lernendes neuronales Netz (NN) der 1950er-Jahre (das Perzeptron) war im Wesentlichen eine Variante der alten linearen Regressoren.

Ein wichtiges Teilgebiet der modernen KI, das heute oft als *Deep Learning* (DL) bezeichnet wird, geht über diese frühen Arbeiten hinaus. Betrachten wir das menschliche Gehirn. Es verfügt über etwa 100 Milliarden Neuronen, jedes mit durchschnittlich 10.000 anderen Neuronen verbunden. Einige sind Eingabeneuronen, die den Rest mit Daten füttern (Gehör, Sicht, Tastsinn, Schmerz, Hunger). Andere sind Ausgabeneuronen, die Muskeln bewegen. Die meisten Neuronen befinden sich dazwischen, wo das Denken stattfindet. Alle lernen, indem sie die Verbindungsstärken ändern, die bestimmen, wie stark Neuronen einander beeinflussen. Dasselbe gilt für unsere tiefen künstlichen NN, die besser lernen als frühere Methoden, Sprache, Handschrift oder Videos zu erkennen, Schmerzen zu minimieren, Lust zu maximieren, Auto zu fahren usw.

Auch DL wurde in Europa geboren, und zwar im Jahre 1965 in der Ukraine, damals Teil der Sowjetunion, welche etliche Bereiche der Wissenschaft und Technologie anführte. Die UdSSR hatte soeben das Weltraumzeitalter eröffnet und die größte Bombe aller Zeiten zur Explosion gebracht. Meines Erachtens noch bedeutsamer war, dass dort viele der besten Mathematiker arbeiteten, u.a. Alexey Ivakhnenko und Valentin Lapa. 1965 veröffentlichten sie den ersten funktionierenden Lernalgorithmus für Netze beliebiger Tiefe, d.h. mit einer beliebigen Anzahl von Schichten. Wenn es einen »Vater des tiefen Lernens« in vorwärtsgerichteten Netzwerken gibt, dann ist es Ivakhnenko. Schon 1970 hatten manche seiner lernenden Netze acht Schichten, was auch nach 2000 noch als tief galt. Wie die heutigen tiefen NNs lernten sie, hierarchische, verteilte, interne Darstellungen eingehender Daten zu finden.

Marvin Minskys & Seymour Paperts berühmtes Buch *Perceptrons* (1969) über die Grenzen flacher NN behandelte also ein »Problem«, das bereits seit vier Jahren gelöst war :-) Das hätten sie eigentlich wissen müssen. Einige behaupten, dass dieses Buch die NN-Forschung zerstört habe, aber das ist natürlich nicht wahr, zumindest nicht außerhalb der Vereinigten Staaten. Vor allem in Osteuropa bauten in den folgenden Jahrzehnten viele Forscher auf Ivakhnenkos Arbeiten auf. Noch in den 2000ern benutzte man seine viel zitierte Methode zum Trainieren tiefer Netze.

Heutzutage verlassen sich die meisten kommerziellen tiefen NN allerdings auf ein gradientenbasiertes Verfahren, das als *Backpropagation* oder *Rückwärtsmodus der Automatischen Differenzierung* bekannt wurde. Seine heutige elegante und effiziente Form wurde erstmals 1970 in einem Anrainerstaat der UdSSR publiziert, nämlich in Finnland, und zwar durch Seppo Linnainmaa in seiner Diplomarbeit (eine Verfeinerung von Kelleys 1960er Arbeit zur Steuerungstheorie). Linnainmaas Methode wurde in den Vereinigten Staaten erstmals 1982 von Paul Werbos auf NN angewendet. Sie dient heute dazu, bestimmte NN-Verbindungen schrittweise zu schwächen und andere zu stärken, sodass sich das NN mehr und mehr wie sein Lehrer verhält.

Eine besonders nützliche NN-Architektur namens Konvolutions-NN oder CNN (*Convolutional Neural Network*) wurde in den 1970er-Jahren von Kunihiko Fukushima in Japan entwickelt, wo NN mit Konvolution 1987 auch vom deutschen Forscher Alex Waibel mit Backpropagation kombiniert wurden. Der Franzose Yann LeCun trug in Amerika viel dazu bei, CNN zum Standard in der Bilderkennung zu machen.

Die DL-bezogenen Interviews des vorliegenden Buches mit LeCun, Bengio & Hinton erwähnen allerdings weder Linnainmaa, den Backpropagation-Erfinder, noch Werbos, den ersten, der das Verfahren auf NN anwendete, noch Fukushima, der 1979 die grundlegende CNN-Architektur veröffentlichte, die sie häufig nutzen. Sie zitieren auch nicht Ivakhnenko & Lapa, die Begründer des DL. Ich begrüße daher die Gelegenheit, in diesem Vorwort diese Urheber der grundlegenden DL-Konzepte zu würdigen.

Die leistungsfähigsten NNs von heute sind in der Regel sehr tief und vielschichtig, mit zahlreichen aufeinander folgenden Rechenstufen (je mehr derartige Stufen, desto tiefer das Lernen). In den 1980er-Jahren funktionierte das gradientenbasierte Training jedoch nicht für tiefe NN, sondern nur für flache.

Dieses DL-Problem offenbarte sich vor allem bei *rekurrenten* bzw. *rückgekoppelten NN* (RNN). Wie das menschliche Gehirn, aber im Gegensatz zu den begrenzteren vorwärtsgerichteten NN (VNN oder *Feedforward-Netze*), haben RNN zyklische Verbindungen. Damit sind RNN universell einsetzbare, parallel-sequentielle Rechner, die Eingabesequenzen beliebiger Länge verarbeiten können (man denke an Sprache oder Videos). Die Verbindungsstärken eines RNNs können grundsätzlich jedes Programm implementieren, das auf einem Laptop läuft. Wenn wir eine Allzweck-KI bauen wollen, dann muss das zugrunde liegende Rechensubstrat so etwas wie ein RNN sein – VNN sind grundsätzlich unzureichend. RNN verhalten sich zu VNN wie Allzweckrechner zu Taschenrechnern.

Insbesondere können RNN im Gegensatz zu VNN grundsätzlich mit Problemen beliebiger Tiefe umgehen. Frühe RNN der 1980er-Jahre funktionierten bei tiefen

Problemen in der Praxis allerdings nicht. Dieser Nachteil wurde in meinem Labor an der TU München im Jahre 1991 zunächst beseitigt durch unüberwachtes Vortrainieren. Damals schufen wir an der TUM auch die ersten unbeaufsichtigten gegnerischen generativen Netzwerke (vgl. *Generative Adversarial Network*), die sich in einem Minimax-Spiel duellierten, um so »künstliche Neugier« und ähnliche, inzwischen weit verbreitete KI-Konzepte zu implementieren. Der eigentliche Grund dafür, dass DL zunächst nicht funktionierte, wurde 1991 ebenfalls an der TUM von meinem ersten Studenten Sepp Hochreiter in seiner Diplomarbeit analysiert: Entweder schrumpfen Fehlersignale bei der Rückpropagierung in typischen tiefen NN extrem schnell, oder sie explodieren. In beiden Fällen scheitert das Lernen. Diese Analyse führte zu den Grundprinzipien des tiefen neuronalen Netzes, das wir später *Long Short-Term Memory* (LSTM) tauften. LSTM wurde weiter verbessert durch meine Studenten am Schweizer KI-Labor IDSIA. Von dort kamen auch die ersten preisgekrönten tiefen GPU-basierten CNN (2011), die erste übermenschlich gute visuelle Mustererkennung (2011) und die ersten funktionstüchtigen sehr tiefen VNN mit mehr als hundert Schichten (2015). Spätestens 2017 beruhten nicht nur die Übersetzungsdienste von Google und Facebook auf LSTM, sondern auch die Spracherkennung von Microsoft, Google, IBM, Samsung usw., sowie die Antworten von Amazon Alexa und viele andere DL-Anwendungen. Mehr unter: *http://people.idsia.ch/~juergen/impact-on-most-valuable-companies.html*

In München wurden in den 1980ern auch die ersten wirklich selbstfahrenden Autos erfunden und entwickelt, und zwar durch das Team von Ernst Dickmanns, dem herausragenden Pionier auf diesem Gebiet. Seine Arbeit wird von zwei der Befragten des vorliegenden Buches erwähnt, nämlich Russell und Brooks. Dickmanns' Team (und das von Uwe Franke) hatte bereits vor über 20 Jahren erste wirklich selbstfahrende Autos im Verkehr. Ein modifizierter Wagen der damaligen Mercedes Benz S-Klasse fuhr 1994 schon dreimal schneller (bis zu 180 km/h auf der Autobahn) als die heutigen Google-Autos, und das ohne GPS, nur mit Kameras (!), also eher wie beim Menschen, trotz der damals 100.000 mal langsameren Rechner. Er fuhr oft hundert Kilometer am Stück ohne Eingreifen des Sicherheitsfahrers, der aus legalen Gründen an Bord sein musste. Selbst heute noch gehören laut FAZ mehr als 50% der Patente für autonomes Fahren deutschen Firmen. Hier eine Übersichtsseite: *http://people.idsia.ch/~juergen/robotcars.html*

Nach dieser Würdigung grundlegender KI- und DL-Durchbrüche aus Ländern, in denen Englisch keine offizielle Rolle spielt, kommen wir zurück zum Hauptthema des Buches: den Meinungen von KI-Wissenschaftlern aus der Anglosphäre! Ich hoffe, dass Sie deren Aussagen aufschlussreich finden und, wo angebracht, mit einem Körnchen Salz genießen werden.

Professor Jürgen Schmidhuber *wird von den Medien oft als »Vater der modernen Künstlichen Intelligenz (KI)« bezeichnet. Der in München geborene Forscher ist Mitgründer und Chefwissenschaftler der Firma NNAISENSE, die die erste praktische Allzweck-KI erschaffen will, und wissenschaftlicher Direktor des Schweizer KI-Forschungsinstituts IDSIA (USI & SUPSI). Die preisgekrönten tiefen neuronalen Netzwerke seiner Forschungsgruppen an der TU München und am IDSIA revolutionierten das Maschinelle Lernen, stecken nun in Milliarden von Computern und werden von den wertvollsten Firmen der Welt jeden Tag vielmilliardenfach genutzt, u.a. für automatische Übersetzung, Spracherkennung, lernende Roboter, Bildbeschreibung, Videospiele, KI-Assistenten, Finanzvorhersage, Gesundheitswesen, usw. Sie waren 2011 die weltweit ersten, die übermenschliche visuelle Mustererkennungsresultate erzielten. Dies weckte bei der Industrie enormes Interesse. Der heutzutage omnipräsente LSTM-Algorithmus seiner Forschungsgruppen an der TU München und am IDSIA nutzte schon 2016 einen signifikanten Teil der Rechenkraft unseres Planeten. So macht LSTM seit 2015 Googles Spracherkennung auf nun über 2 Milliarden Android Smartphones. 2016 wurde auch Google Translate auf LSTM umgestellt, was zu einer dramatischen Verbesserung führte; schon 2016 wurden fast 30% der enormen Inferenzkraft in Googles Datenzentren für LSTM verwendet. Ab 2017 machte auch Facebook Übersetzungen mit LSTM, und zwar 30 Milliarden pro Woche, also über 50.000 pro Sekunde. Man vergleiche: Das erfolgreichste YouTube-Video brauchte zwei Jahre, um auf nur 6 Milliarden Klicks zu kommen. Seit 2016 erzeugt LSTM auch die Frauenstimme von Amazons Alexa, und steckt in Apples Siri & QuickType auf fast einer Milliarde iPhones.* Business Week *nannte LSTM »die wohl kommerziellste Leistung der KI«. Schmidhuber erfand auch künstliche Neugier und meta-lernende Maschinen, die das Lernen selbst lernen. Er erhielt zahlreiche internationale Preise, ist ein höchst gefragter Redner und berät verschiedene Regierungen zur KI.*