1. Intelligente Systeme – Fluch oder Segen? – Grundlegende Konstruktionsprinzipien der Künstlichen Intelligenz aus der Perspektive der Ingenieurwissenschaften

Gerhard Friedrich

1.1. Einleitung

Getrieben von den beachtlichen Erfolgen des maschinellen Lernens (ML) im Bereich der Umwandlung von gesprochenen Wörtern in Symbole[1], der Erkennung und Klassifikation von Objekten in digitalisierten Bildern[2] und der Erlernung von Strategien in Spielen[3] hat sich rund um die Künstliche Intelligenz (KI) ein bemerkenswertes Medienecho entwickelt. Zahlreiche Staaten fördern diese Technologie mit beachtlichen Investitionsprogrammen. China und die USA kämpfen um die technologische Vorherrschaft. Die EU plant, in den nächsten Jahren ein umfangreiches Förderprogramm zu implementieren. Namhafte High-Tech-Unternehmen bilden rund um KI ihre Marketingstrategie. Anerkannte Forscher sehen in KI-Systemen eine Bedrohung, weil sie der Ansicht sind, dass sich eine superintelligente Spezies entwickeln könnte, die mit der Menschheit in Konkurrenz steht.

Für den/die fachfremde/n Beobachter*in kann daher der Eindruck entstehen, dass KI im Zusammenspiel mit Robotern in kurzer Zeit einen großen Teil der kognitiven und mechanischen Fähigkeiten des Menschen automatisiert replizieren wird. Die Anwendung dieser Technologie führt zu zahlreichen

1 *Hinton et al*, Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, IEEE Signal Processing Magazine, Vol. 29, No. 6 (2012).

2 *Krizhevsky/Sutskever/Hinton*, ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25, 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, NV, USA (2012).

3 *Silver et al*, Mastering the game of Go without human knowledge. Nature, Vol. 550, No. 7676 (2017).

Debatten über deren gesellschaftliche Implikationen. Insbesondere führt der Einsatz solcher Systeme in sicherheitskritischen Anwendungen, wie zB im Bereich des autonomen Fahrens, zu Diskussionen in der Rechtswissenschaft. In diesen Diskussionen wird unter anderen argumentiert, dass sich KI von bisherigen Softwaresystemen grundlegend unterscheidet, und daher eine neue Sicht notwendig sei. Wird mit KI ein Paradigmenwechsel in der Informatik eingeläutet, oder ist KI eine evolutionäre Weiterentwicklung der bisherigen Ansätze?

Für die Einordung von KI und die Einschätzung ihres Potentials sowie ihrer aktuellen Grenzen und Herausforderungen ist es notwendig, die grundlegenden Konstruktionsprinzipien zu verstehen. Auf Basis dieses Wissens können die Vor- und Nachteile sowie die Voraussetzungen für die erfolgreiche Anwendung von KI-Methoden schlüssig abgeleitet und zu den bisherigen Ansätzen in Beziehung gesetzt werden.

Wenn zurzeit in der Öffentlichkeit über KI im Allgemeinen oder ML im Speziellen diskutiert wird, dann wird in den meisten Fällen zu einer spezifischen Methode des ML Bezug genommen. Ziel des Beitrages ist eine möglichst einfache Darstellung der Grundlagen dieses Ansatzes sowie dessen Erfolge, die letztlich diese breite öffentliche Wahrnehmung ausgelöst haben. Daraus lassen sich die Stärken und aktuellen Grenzen erkennen. Teildisziplinen der KI sind in der Regel durch praktische Aufgabenstellungen und Defizite von bestehenden Methoden entstanden. Eine Genesis der Teildisziplinen der KI zeigt uns die noch immer zu lösenden Herausforderungen von ML und führt uns zu komplementären Methoden der KI, die auf einer expliziten Formulierung von Wissen, Funktionen und Modellen beruhen. Darauf aufbauend können wir die zukünftigen Forschungsstränge der KI ableiten.

1.2. Maschinelles Lernen – Supervised and Reinforcement Learning

Der in den letzten Jahren bei weitem sichtbarste Teil des ML und der KI sind Artificial Neural Networks (ANNs)[4]. Diese werden sehr erfolgreich für ML auf Basis des Supervised bzw des Reinforcement Learning (RL) eingesetzt. Supervised Learning (SL) ist ein Teilgebiet von ML, das durch vorgegebene Input-Output-Paare eine Funktion generiert (s Abbildung 1).

4 *Russell/Norvig*, Artificial Intelligence – A Modern Approach (2010).

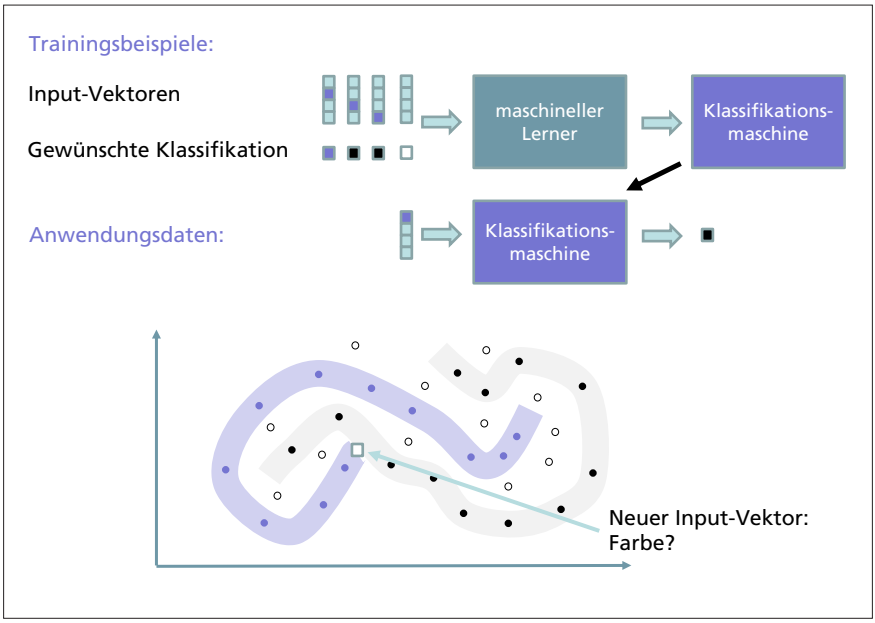

Abbildung 1: Supervised Learning

Ein klassisches Beispiel für SL ist das Erkennen von Objekten (zB Katzen) in Bildern. Die Input-Vektoren sind Bilder. Der zugehörige Output zu einem Bild (dh das gewünschte Klassifikationsergebnis für ein Bild) beschreibt, ob in diesem Bild ein bestimmtes Objekt (zB eine Katze) zu erkennen ist. Diese Input-Output-Paare, die als Trainingsbeispiele bezeichnet werden, dienen dem maschinellen Lernverfahren, um eine möglichst perfekte Klassifikationsmaschine zu erzeugen. Eine Klassifikationsmaschine ist perfekt, wenn alle Inputs, dh alle Trainingsbeispiele und auch alle möglichen anderen Inputs, korrekt klassifiziert werden.

In unserem Beispiel bedeutet dies, dass der maschinelle Lerner eine Klassifikationsmaschine erzeugt, die in allen noch nie gesehenen Bildern korrekt entscheidet, ob im jeweils betrachteten Bild eine Katze zu sehen ist. In Abbildung 1 werden abstrahierend als Klassifikationsergebnis Farben verwendet, und der untere Teil von Abbildung 1 verwendet zur Vereinfachung nur zweidimensionale Input-Vektoren. Die Aufgabe des maschinellen Lerners ist, aus den farbigen Datenpunkten eine Funktion zu erzeugen, sodass allen zukünftig beobachteten Datenpunkten die korrekten Farben (violett, weiß oder schwarz) zugeordnet werden können. In Abbildung 1 sind das zwei kurvige Bereiche und die weiße Fläche, aber es könnten auch andere Aufteilungen der zweidimensionalen Fläche gelernt werden. Der maschinelle Lerner für

SL generiert diese Aufteilung auf Basis der vorgegebenen Beispiele und der eingebauten Heuristik.

Der Zweck von SL ist es, die möglichst richtige Klassifikation aller Inputs zu erraten. In Reinforcement Learning ist das Klassifikationsproblem aufgabenspezifisch abgewandelt. Die Aufgabe von RL ist die Auswahl von möglichst guten Aktionen eines Agenten. Einsatzgebiete sind Steuerungen von Systemen, wie zB Roboter, oder auch Spiele, wie zB Go. RL zielt nicht auf die richtige Klassifikation eines Inputs, sondern auf die Auswahl der „richtigen" (dh einer möglichst guten) Aktion, abhängig vom Zustand des Agenten und seiner Umgebung. Während bei SL für jeden Input eine korrekte Klassifikation zur Verfügung gestellt wird, verwendet RL Belohnungen, die ein Agent zu bestimmten Zeitpunkten erhält. Das heißt, die Auswahl von Aktionen wird durch ein Belohnungssystem gelernt.

1.3. Artificial Neural Networks (ANNs)

Für das Erlernen von Funktionen bzw für das Erzeugen von Klassifikationsmaschinen gibt es in der Literatur eine große Anzahl an Methoden. Die zurzeit prominenteste Methode basiert auf den ANNs, die eine grobe Vereinfachung von Nervenzellen darstellen. Abbildung 2 zeigt das üblicherweise verwendete Modell. Die numerischen Werte der Inputs I_1 bis I_n eines Neurons j werden mit den jeweiligen numerischen Gewichten W_{1j} bis W_{nj} multipliziert und aufsummiert. Solche numerischen Input-Werte könnten zB die Graustufen der Bildpunkte eines digitalisierten Fotos sein. Zusätzlich wird ein sogenannter Bias W_{Bj} zu dieser Summe addiert. Das Ergebnis dient als Eingabe in eine Aktivierungsfunktion, die meistens aus technischen Gründen als eine differenzierbare Sprungfunktion realisiert ist. Diese Funktion liefert in Abhängigkeit der gewichteten Summe des Inputs einen numerischen Output. Das heißt, überschreitet der gewichtete Input einen Schwellenwert, dann wird das Neuron aktiviert – es feuert und liefert einen signifikant größeren numerischen Wert als im inaktiven Zustand.

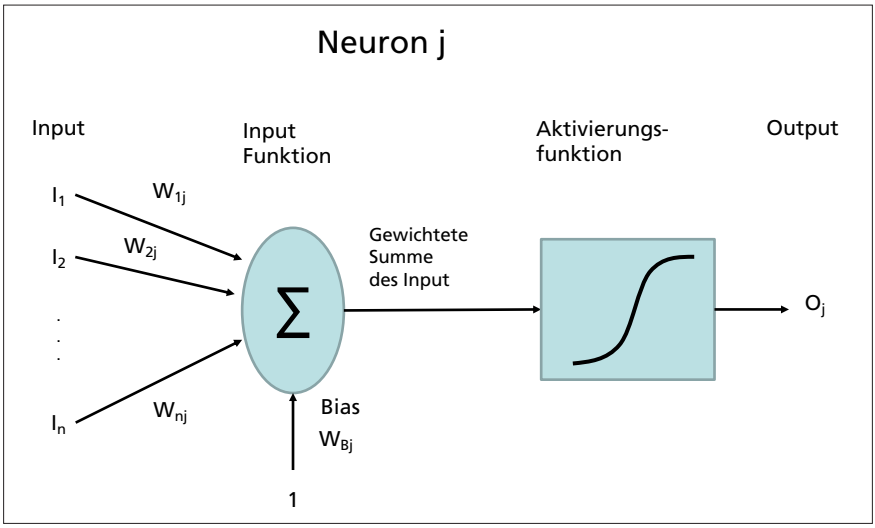

Abbildung 2: Modell eines Neurons

Diese künstlichen Neuronen werden verwendet, um daraus ANNs zu konstruieren. Abbildung 3 zeigt ein dreischichtiges ANN.

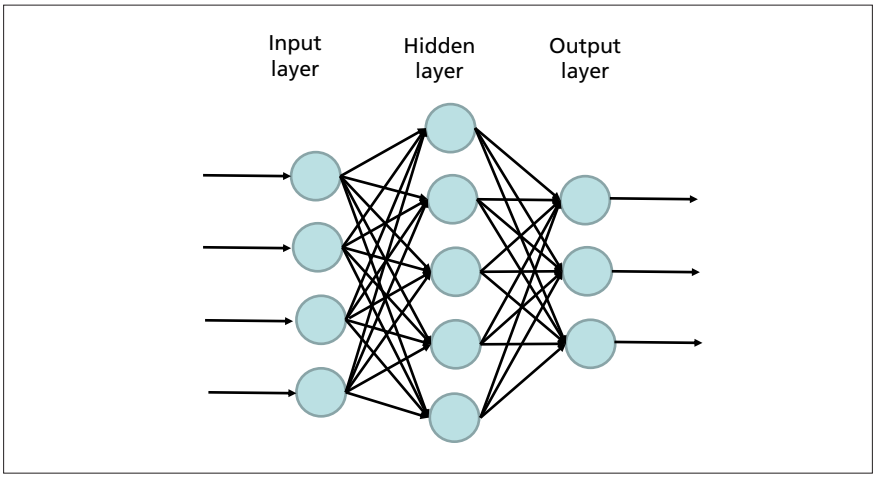

Abbildung 3: Artificial Neural Network

Sogenanntes Deep Learning verwendet ANNs mit vielen Schichten. Abbildung 4 zeigt ein aus zahlreichen Schichten aufgebautes Convolutional Deep Neural Network, wie es für die Objekterkennung verwendet wird.

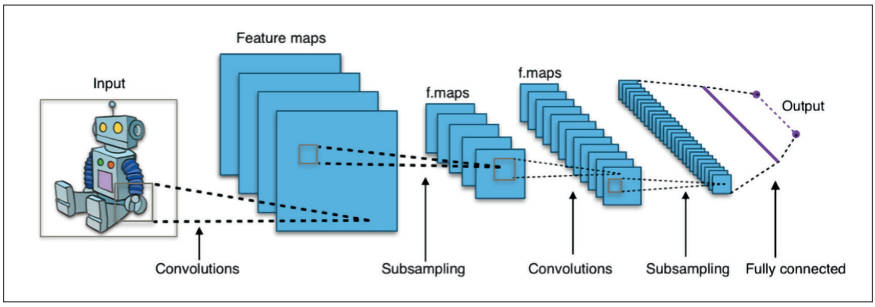

Abbildung 4: Convolutional Deep Neural Network, © Aphex34 – Eigenes Werk, CC BY-SA 4.0

In allen modernen ANN-Ansätzen erfolgt das Lernen durch das Einstellen der „richtigen" numerischen Gewichte in der Trainings- bzw der Lernphase. Daraus folgt, dass in ANNs das Wissen in Form der Netzwerkarchitektur und der numerischen Gewichte kodiert ist. Symbolisches Wissen, wie es der Mensch zB in der Weitergabe des Wissens oder in der Begründung von Schlussfolgerungen verwendet, wird in ANNs nicht dargestellt.

1.4. Geschichtliche Entwicklung von ANNs und ausgewählte Erfolge

Die geschichtliche Entwicklung der ANNs ist sehr bewegt. *Warren McCulloch* und *Walter Pitts*[5] werden als Väter der ANNs angesehen, weil sie 1943 Berechnungsmodelle für neuronale Netzwerke formulierten. In den späten 1950er-Jahren wurde das Perceptron[6] entwickelt, das insbesondere zur Erkennung von Ziffern verwendet wurde.1969 zeigten *Marvin Minsky* und *Seymour Papert*[7] signifikante Einschränkungen des Perceptrons, worauf die Forschung auf dem Gebiet der ANNs eine wesentliche Stagnation verzeichnen musste. Neue, komplizierte Architekturen und Algorithmen zur Anpassung der Gewichte verhalfen den ANNs in den 1990er-Jahren zu einer Renaissance. Die ANNs wurden wieder vermehrt eingesetzt. ANNs stehen aber immer in scharfer Konkurrenz zu alternativen Lernmethoden, wie zB Support Vector Machines oder Hidden Markov Models.

5 *McCulloch/Pitts*, A Logical Calculus of Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics, Vol. 5, No. 4 (1943).

6 *Rosenblatt*, The Perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review, Vol. 65, No. 6 (1958).

7 *Minsky/Papert*, Perceptrons: An introduction to computational geometry (1969).

Der Durchbruch von ANNs gelang ab 2009 mit Siegen bei Wettbewerben im Bereich Mustererkennung und ML, wie zB die Erkennung von handschriftlichen Texten[8]. 2012 wurde mit Hilfe von Deep Learning die „Large Scale Visual Recognition Challenge“[9] gewonnen. 2014 konnte gezeigt werden, dass ANNs nach kurzer Trainingszeit viele Atari-2600-Spiele besser beherrschen als Menschen[10]. 2015 gewann AlphaGo mit einer Kombination von ANNs und einem klassischen Suchverfahren gegen den europäischen Go-Champion[11]. 2016 gelang eine signifikante Verbesserung der automatischen Übersetzung von natürlicher Sprache[12].

Basis dieser Erfolge sind neue ANN-Architekturen, gepaart mit neuen Lernalgorithmen, großen Mengen an Daten für die Trainingsphase sowie signifikanten Steigerungen der zur Verfügung stehenden Rechenleistung.

1.5. Derzeitige Einschränkungen

ANNs haben herausragende Verbesserungen von KI-Systemen erreicht. Allerdings werden ANNs für spezielle Problemstellungen trainiert, um Näherungsfunktionen von kognitiven Fähigkeiten zu erzeugen[13]. ANNs verwenden kein Modell der Welt, wie es Menschen zur Verfügung steht, und können daher gezielt in die Irre geführt werden[14]. Zum Beispiel kann einem Bild eine Störung überlagert werden, die für den Menschen keine Auswirkung hat, aber zu einer Fehlklassifikation des ANN führt[15].

8 *Graves/Schmidhuber*, Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada (2008).

9 *Russakovsky et al*, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, Vol. 115, No. 33 (2015).

10 *Mnih et al.,* Human-level control through deep reinforcement learning, Nature, Vol. 518, No. 7540 (2015).

11 *Silver et al*, Mastering the game of Go with deep neural networks and tree search, Nature, Vol. 529, No. 7587 (2016).

12 *Bojar et al*, Findings of the 2016 Conference on Machine Translation. Proceedings of the First Conference on Machine Translation (2016).

13 *Darwiche*, Human-level intelligence or animal-like abilities? Communications ACM, Vol. 61, No. 10 (2018).

14 *Heaven*, Why deep-learning AIs are so easy to fool, Nature, Vol. 574 (2019).

15 *Goodfellow et al,* Explaining and Harnessing Adversarial Examples, 3rd International Conference on Learning Representations, San Diego, CA, USA, May 7–9, 2015.

Kevin Eykholt et al[16] zeigen verschiedenste Attacken auf ANNs, zB, dass durch einfache Manipulation eines Stoppschildes dieses fehlerhaft als ein Zeichen für eine Geschwindigkeitsbegrenzung klassifiziert wird.

Fehlendes Textverständnis von ANNs wird durch die „Winograd Schema Challenge“[17] deutlich. Basis dieser Aufgaben sind mehrdeutige Pronomina, deren Bezug nur durch sogenanntes „Common Sense Reasoning“ bestimmt werden kann:

„The city councilmen refused the demonstrators a permit because they [feared/advocated] violence“.

Je nachdem, ob „feared“ oder „advocated“ verwendet wird, ändert sich der Bezug von „they“. Der richtige Bezug ist für einen Menschen leicht zu erkennen. Gegenwärtige KI-Systeme können aber keine menschenähnliche Leistungsfähigkeit entwickeln.

Menschen können Wissen in Form von komplexen Aussagen und Beschreibungen anwenden und lehren. Beispiele solcher Beschreibungen sind physikalische oder mathematische Gesetze sowie Verhaltensanweisungen. Diese Beschreibungen können von Computersystemen verwendet werden, um nachweislich korrekte und vollständige Problemlösungssysteme zu realisieren[18].

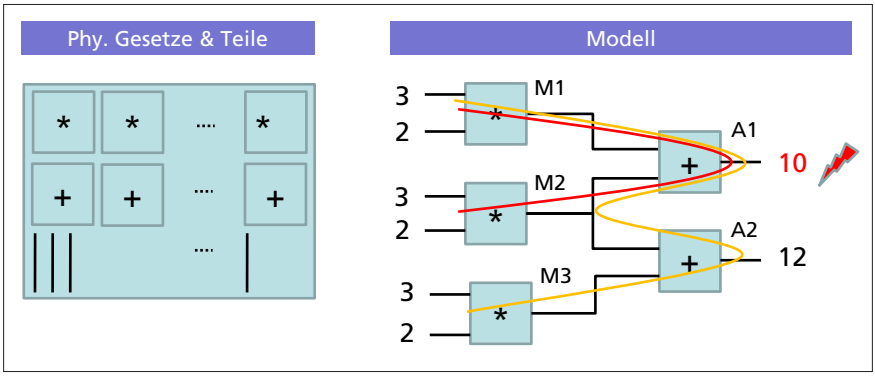

Abbildung 5: Modellbibliothek und Modell eines fehlerhaften Systems

16 *Eykholt/Evtimov/Fernandes/Li/Rahmati/Xiao/Prakash/Kohno/Song*, Robust Physical-World Attacks on Deep Learning Visual Classification, 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18–22, 2018.

17 *Levesque*, The Winograd Schema Challenge. Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11–06, Stanford, CA, USA, March 21–23, 2011.

18 *Reiter*, A Theory of Diagnosis from First Principles. Artificial Intelligence, Vol. 32, No. 1 (1987).