Parallelen zwischen dem menschlichen und maschinellen Lernen

Simon Weitz

Seminar KI: gestern, heute, morgen
Angewandte Informatik, Universität Bamberg

Zusammenfassung. Viele Ansätze des maschinellen Lernens zeigen Parallelen im direkten Vergleich mit den Mechanismen des menschlichen Lernens. Diese Tatsache erfährt allerdings nur wenig Beachtung in der wissenschaftlichen Gemeinschaft. Eine Diskussion der gemeinsamen Merkmale ist notwendig, um die Aufmerksamkeit von Forschern aus dem Bereich Psychologie und maschinellen Lernens auf die Überschneidungen aufmerksam zu machen und deren Austausch anzuregen. Das motiviert eine Gegenüberstellung bekannter Lernalgorithmen mit den Theorien über das Lernen beim Menschen. Es werden zunächst Erkenntnisse über das menschliche Konzeptlernen erläutert. Von diesem Standpunkt aus werden spezifische Gemeinsamkeiten zu Ansätzen des maschinellen Lernens offengelegt. Diese Arbeit soll als eine kompakte Diskussionsgrundlage und Übersichtsartikel dienen. Im Verlauf können zahlreiche Übereinstimmungen identifiziert werden, welche die starke Relation der beiden Bereiche verdeutlicht. Nutzt man das Wissen über die menschliche Kognition im maschinellen Lernen, kann dies die Effizienz lernender Systeme fördern.

Schlüsselwörter: Applied Computing, Machine Learning, Human Concept Learning

1 Einleitung

Im Vergleich zum maschinellen Lernen sind Menschen besonders herausragend im Lernen von Konzepten. Dinge und Situationen erfasst der Mensch in enorm kurzer Zeit. Der Mensch ist in der Lage extrem viel Information von nur einem oder wenigen Beispielen zu extrahieren. Zudem benutzen wir gelernte Konzepte in viel reichhaltiger Art und Weise als konventionelle Algorithmen. Zum Beispiel stellen wir uns mithilfe von Konzepten Dinge vor oder nutzen sie, um mit anderen Personen zu kommunizieren. Sieht sich ein Mensch mit einem unbekannten Gegenstand konfrontiert, so wird dieser komplett unbewusst in seine wichtigsten Bestandteile und Relationen zerlegt. Dabei generalisieren wir über die Unterschiede und Gemeinsamkeiten zu bekannten Kategorien.
Am Beispiel des Konzeptlernens wird die Leistungsfähigkeit menschlicher Kognition deutlich. Prozesse des menschlichen Lernens in Algorithmen für Systeme zu integrieren ist somit ein vielversprechender Ansatz, um effizientes maschinelles

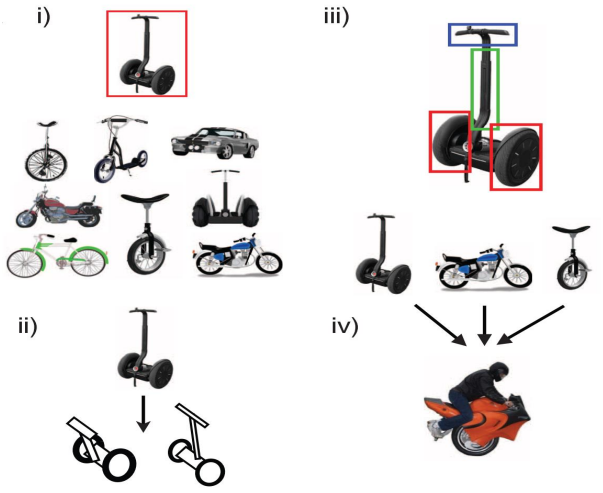

Abb. 1. Ein einzelnes Beispiel (rot) eines neuen Konzeptes liefert für den Menschen genug Information, um (i) neue Beispiele der gleichen Kategorie zuzuordnen und darüber hinaus (ii) neue Exemplare zu generieren (Lake et al., 2015).

Lernen zu ermöglichen. Heutige Systeme benötigen im Vergleich zum Menschen enorm viele Trainingsdaten, um Beispiele richtig zu kategorisieren. Unter anderem Tennenbaum et al. beschäftigen sich daher in ihrem Papier „Human-level concept learning through probabilistic program induction“ mit dem Spezialgebiet der One-Shot Klassifikation (Lake et al., 2015). Dabei soll ein System von einem oder sehr wenigen Trainingsbeispielen lernen. Der Ansatz ist sehr ambitioniert, da im Normalfall für eine komplexe Lernaufgabe mehr Trainingsdaten nötig sind. Das Papier zeigt allerdings, dass man im Fall handgeschriebener Buchstaben fähig ist, einige Aspekte des menschlichen Konzeptlernens zu simulieren. Ausgehend von nur wenigen handgeschriebenen Buchstaben kann das System die Generierung neuer Zeichen bewerkstelligen. Die Ergebnisse erreichen die Qualität realer Buchstabenkreationen von Testpersonen. Generell können in der jüngsten Vergangenheit vor allem Deep Learning Ansätze mit dem Menschen vergleichbare Leistungen erzielen. Diese benötigen wiederum extrem viele Trainingsbeispiele zum Lernen. Eine Auseinandersetzung mit der Schnittmenge von menschlicher Kognition auf der einen Seite und maschinellem Lernen auf der anderen ist daher aus ingenieurswissenschaftlicher Sicht erstrebenswert. Hier hat man das Ziel, funktionale und effiziente Algorithmen zu entwickeln.
Ich gehe mit meinem Artikel der Frage nach, zu welchen bekannten Mechanismen des menschlichen Konzeptlernens bereits korrespondierende Ansätze im maschinellen Lernen existieren. Der Fokus liegt hierbei insbesondere auf dem überwachten Lernen. Für ein generelles Verständnis wird zunächst darauf eingegangen, welche Theorien im Bezug auf die Kategorienbildung beim Menschen als mögliche Erklärungsversuche existieren. Darauf aufbauend folgt die angespro-

chene Gegenüberstellung. Im Anschluss daran werden Methoden und Werkzeuge vorgestellt mit deren Hilfe es möglich ist, Lernprozesse beim Menschen zu ergründen. Zudem wird anhand einer konkreten Forschungsarbeit ein Vorgehen skizziert, um effiziente Mechanismen des Menschen in ein lernendes System zu implementieren. Am Schluss werden offene Fragen und Probleme angesprochen.

2 Theoretische Grundlegung

Folgender Textauszug ist eine Übersicht der wichtigsten Theorien zum menschlichen Konzeptlernen. Zu Beginn befasst sich der Autor mit dem Inductive Bias – einer grundlegenden Voraussetzung für maschinelles und menschliches Lernen gleichermaßen.

Definition 1. *Ein Inductive Bias beschreibt alle Annahmen, die eine Person unterbewusst macht oder in einem System implementiert sind, sodass über bekannte Trainingsbeispiele hinaus das gelernte Konzept auf neue Beispiele übertragen werden kann (Mitchell, 1997).*

Zum Beispiel zeigt sich beim Spracherwerb ein besonders markanter Bias. Dieser erlaubt es, Kleinkindern die Vergangenheitsform von Wörtern zu bilden, nachdem sie wenige Beispiele von bereits konjugierten Wörtern gehört haben. Die dabei angewandten Heuristiken sind notwendig, um gegeben einer Lernerfahrung mit einer Teilmenge, Entscheidungen für folgende Exemplare treffen zu können. Ein Bias ist beim Menschen immer implizit. Durch produktive Bias ist es dem Menschen beim Konzeptlernen möglich, anhand einem oder sehr weniger Exemplare neue Beispiele zu klassifizieren (siehe Abb. 1).

Die kognitive Fähigkeit des Menschen Konzepte zu lernen ist erforderlich, damit wir mit unserem Umfeld effizient umgehen können. Kategorien stellen wichtige Voraussetzungen für unser Verständnis und logisches Denken dar. Sie sorgen dafür, dass wir gewisse Erwartungen an Objekte und Situationen haben. Ein Wissenstransfer bzw. nachhaltige Kommunikation zwischen Individuen wäre unvorstellbar, wenn diese nicht über einen ähnlichen Wissensstand an Kategorien verfügen würden. Somit wird es erst durch Kategorisierung möglich, miteinander über physisch nicht präsente Dinge zu kommunizieren.
In diesem Abschnitt behandle ich die wichtigsten Theorien zur Kategorienbildung beim Menschen. Jede einzelne versucht die Frage zu beantworten, wie Kategorien gebildet und im Bewusstsein repräsentiert sind. Es wird deutlich, dass in der Literatur verschiedene Ansichten vertreten sind und bisher keine allgemein akzeptierte Theorie vorherrscht.

2.1 Ähnlichkeitsbasierte Theorien

Die ähnlichkeitsbasierten Theorien sind am populärsten. Die wesentlichsten Anschauungen sind der regelbasierte Ansatz, die Prototypensicht, sowie die Exemplarsicht.

Der regelbasierte Ansatz – die klassische Sicht. Nach der klassischen Sicht besitzen Objekte harte, definierenden Merkmale, welche über die Kategorienzugehörigkeit entscheiden. Schwächen dieses frühen Ansatzes sind vor allem dadurch gegeben, dass die Experimentatoren bei der Untersuchung ausschließlich artifizielles Versuchsmaterial (z. B. geometrische Formen) für eine bessere Kontrolle in den Experimenten verwendeten. Die Versuche sind damit kaum realitätsnah. Letztendlich ist dieser Ansatz in seiner Grundform allein wenig plausibel. Grund hierfür ist unter anderem, dass die klaren Grenzen zur Unterscheidung der Kategorien in der Realität eher weicher verlaufen und schwerer auszumachen sind. Des Weiteren ist die klassische Sicht durch ihre Annahmen funktional limitiert. Kategorien sind von diesem Standpunkt aus betrachtet immer eine endliche Menge von Merkmalen, die sie charakterisieren (Müsseler & Rieger, 2017).

Die Prototypensicht. In einem Prototypen sind besonders typische Merkmale einer Kategorie vereint. Definitorische Merkmale sind im Vergleich mit der klassischen Sicht nicht zwingend für ein Exemplar erforderlich, um einer Kategorie anzugehören. Stattdessen genügt es, für ein Objekt eine ausreichend große Übereinstimmung mit dem Prototypen der Kategorie aufzuweisen. Für jede Kategorie wird zunächst die Repräsentation des Prototypen gebildet. Der Vergleich eines neuen Objektes mit den Prototypen der Kategorien liefert die notwendigen Informationen, um eine bestimmte Gruppe zuzuordnen. Dieser Ansatz berücksichtigt den Aspekt, dass gewisse Merkmale stärker eine Kategorie charakterisieren und andere wiederum schwächer. Zudem sind die Kategoriengrenzen weniger strikt als bei der klassischen Sicht.

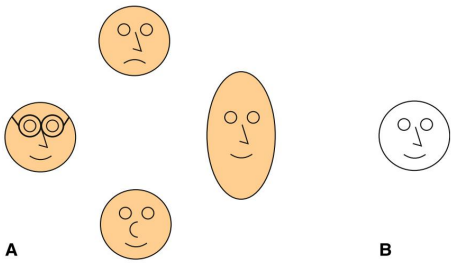

Abb. 2. Aus der Summe der charakterisierenden Merkmale bildet sich der Prototyp (B). Dieser besitzt kein Merkmal, dass von allen vier Beispielgesichtern (A) erfüllt wird. Beispielsweise besitzen nur drei die gleiche Form des Mundes. Im Vergleich zur klassischen Sicht ist eine Übereinstimmung nicht zwingend notwendig (Müsseler & Rieger, 2017).

Prototypen erweisen sich unter Laborbedingungen mit einfachen Kategorien, wie den Gesichtern in Abb. 2, als nachvollziehbare Strategie. Allerdings hat diese

Sichtweise ebenfalls Schwächen. Dem Ansatz mangelt es aufgrund der geschaffenen Versuchsbedingungen in den Experimenten an externer Validität. Andere Einwände sind, dass ein Prototyp in manchen Fällen nicht als das Mittel aller typischen Merkmale repräsentiert ist, sondern Idealtypen zum Vergleich herangezogen werden. Weiterhin findet die Interaktion zwischen den einzelnen Merkmalen oder die Abhängigkeit der Typizität eines Merkmales vom jeweiligen Kontext keine Erwähnung. Limitiert ist die Prototypensicht zudem dadurch, dass nur linear separierbare Kategorien für eine Kategorisierung infrage kommen. Hingegen kann der Mensch ebenso nicht linear trennbare Kategorien lernen (Müsseler & Rieger, 2017).

Die Exemplarsicht. Diese Perspektive beschreibt den Prozess der Kategoriezuordnung als einen Vergleich der Ähnlichkeiten zwischen dem unbekannten Objekt und bereits bekannten Exemplaren bestimmter Kategorien. Der Abgleich findet zum Zeitpunkt der Konfrontation mit einem neuen Objekt oder Situation statt. Demnach klassifizieren wir bspw. einen Stuhl als solchen, indem wir dessen (visuelle) Merkmale mit den Charakteristiken von bereits bekannten Exemplaren vergleichen. Wir erkennen das Objekt als Stuhl dadurch, dass wir die Kategorie mit der am höchsten berechneten Ähnlichkeit zuweisen. Über die abgerufenen Objekte und deren Kategorien findet keinerlei Generalisierung statt. Im Vergleich zur Prototypensicht behalten diese ihre ursprüngliche Repräsentation bei. Ihre Merkmale werden nicht zu einem Durchschnittsexemplar zusammengefasst. Trotz der unrealistischen Kapazitätsanforderungen an das menschliche Gedächtnis bleibt der Ansatz bestehen, da sich unter anderem einige Schwächen der Prototypensicht mit dem exemplarbasierten Ansatz erklären lassen. Beispielsweise kann die Interaktion zwischen Merkmalen mit dieser Theorie berücksichtigt werden oder die lineare Separierbarkeit als Bedingung für die Klassifikation entfallen.

Ich bringe bereits in den vorherigen Abschnitten einige spezifische Probleme der einzelnen Ansätze zur Sprache. Insgesamt ist bei den ähnlichkeitsbasierten Theorien die fehlende formale Beschreibung kritisch. Es existiert keine allgemeine Formel, um die Ähnlichkeit zwischen Objekten exakt zu bestimmen. Dieser Umstand macht es sehr schwierig, Charakteristiken zu finden, mithilfe derer die Kategorisierung von Objekten ohne Zweifel gelingt. Ein weiterer kritischer Punkt ist, dass die empfundene Ähnlichkeiten zwischen Objekten oder Situationen sehr kontextabhängig ist (Müsseler & Rieger, 2017).

2.2 Die Theoriensicht

Neben den ähnlichkeitsbasierten Theorien soll als Alternative die Theoriensicht kurz vorgestellt werden. Ein gleichartiger Ansatz aus dem maschinellen Lernen kann im weiteren Verlauf allerdings nicht identifiziert werden.
Die ähnlichkeitsbasierten Ansätze vertreten die Vorstellung, dass Kategorien basierend auf unabhängigen Merkmalen repräsentiert sind. Die Relationen zwischen den Merkmalen bleiben hierbei meist unbeachtet. Die Theoriensicht vertritt dagegen, dass Konzepte wissensbasiert und von Theorien über die Welt

beeinflusst werden. Dieser Anschauung nach ist Kategorisierung eher ein Prozess, der eine Beziehung zwischen Theorie und passender Exemplare herstellt. Dies ermöglicht den Menschen bspw. ein Konzept „Betrunkener“ auch auf eine Person anzuwenden, die auf einer Party in einen Pool springt. „In den Pool springen“ ist kein festes Merkmal eines Betrunkenen. Dennoch ist es durch die Verknüpfung mit dem theoretischen Vorwissen und dem beobachteten Verhalten der Person möglich, die Handlung als Folge aus übermäßigen Alkoholkonsum zu erkennen. In diesem Fall sind die Merkmale vielmehr konstruiert als objektiv erkennbar. Offen bei dieser Sichtweise bleibt, wie zuallererst das Vorwissen zustande kommt, das darauf für die Kategorisierung benutzt wird (Müsseler & Rieger, 2017).

2.3 Kategorisierung mithilfe von Analogien

Theorien zu Analogien können viele Aussagen betreffend der ähnlichkeitsbasierten Kategorisierung (vgl. 2.1) ebenso erklären.

Definition 2. *Eine Analogie beschreibt den Prozess des Verstehens einer neuen Situation in Bezug auf eine bereits bekannte Situation (Gentner & Holyoak, 1997).*

Eine Analogie funktioniert zum größten Teil auf Basis der Übereinstimmungen von Merkmalen zwischen Objekten oder Situationen. Der Vergleich von Ähnlichkeiten ist die Voraussetzung, um eine vertraute Kategorie zuzuordnen. Analogien zu erkennen, ist eine starke kognitive Fähigkeit. Menschen machen sich Analogien zunutze, wenn sie sich ausgehend von einer bereits bekannten Situation (Base Analog) eine neue Situation besser erklären (Target Analog). Beispielsweise bedienen sich viele der Analogie vom Wasser welches durch Leitungen fließt, um sich und anderen die Komponenten des Ohmschen Gesetzes verständlicher zu machen. Dabei dient uns die bekannte Situation als Modell, mithilfe dessen wir Schlussfolgerungen für die neue Situation ableiten und so Verständnislücken schließen (Gentner & Holyoak, 1997).
Die zentrale Annahme für die Entstehung einer Analogie ist, dass falls zwei Situationen in einigen Aspekten übereinstimmen, sie darüber hinaus noch weitere Gemeinsamkeiten haben müssen. Damit wir eine Analogie anwenden können, müssen beide Situationen in einigen Punkten übereinstimmen, jedoch in gewissem Maße auch unterschiedlich sein. Wären beide Situation identisch könnte kein Lernen (Inferenz) und in der Folge keine Kategorisierung mithilfe von Analogien stattfinden (Winston, 1980).

3 Parallelen zwischen dem menschlichen Konzeptlernen und überwachtem maschinellen Lernen

Im folgenden Kapitel werden die Gemeinsamkeiten von menschlichem Lernen und maschinellem Lernen analysiert. Es existieren einige Ansätze des maschinellen Lernens, die in ähnlicher Weise umsetzen, was auch als menschliche Strategie

angenommen wird.
Entscheidungsbäume sind ein prominentes Beispiel für das Lernen von Klassifikationen. Der erste Decision Tree Algorithmus wurde von der Arbeit der Kognitionspsychologen an regelbasierten Theorien inspiriert (Hunt, Stone, & Marin, 1966). Die Klassifikation von neuen Instanzen entscheidet sich auf Basis der gelernten Regeln, die aus dem Entscheidungsbaum ableitbar sind (Mitchell, 1997). Eine menschliche Strategie, die sich an der originalen Idee der klassischen Sicht orientiert ist unwahrscheinlich (vgl. 2.1). Neuronale Netze und Deep Learning sind ebenfalls von der Forschung am Menschen inspiriert. Hier dienten die neurophysiologischen Prozesse in Form des Informationsaustausches unter Neuronen als Inspiration. Davon abgesehen haben Neuronale Netze keine Gemeinsamkeiten mit den tatsächlichen neurophysiologischen und kognitiven Vorgängen beim Menschen. Deshalb widmen sich die folgenden Abschnitte den Ansätzen mit einem höheren Überschneidungsgrad.

3.1 Instanzbasiertes Lernen

Das instanzbasierte Lernen zeichnet sich vor allem durch den reaktiven Charakter einer *lazy* Lernmethode aus. Im Gegensatz zu *eager* Lernmethoden (z. B. Artificial Neural Networks, Hidden Markov Models, etc.) ist es möglich, inkrementell neue Trainingsbeispiele zu berücksichtigen. Eine wichtige Voraussetzung für ein lebenslanges Lernszenario, wie bspw. der Verlauf eines Menschenlebens. Der Ansatz ähnelt der exemplarbasierten Kategorisierung beim Menschen (vgl. 2.1).
Der k-nearest-neighbor Algorithmus, als ein Beispiel instanzbasierten Lernens benutzt das Wissen aus Beispieldaten und deren Kategorien. Darauf basierend entscheidet ein zuvor definiertes Distanz- bzw. Ähnlichkeitsmaß (z. B. euklidische Distanzen) über die Kategorie einer neuen Instanz. Der veränderbare Parameter k gibt hierbei an, wie viele der Nachbarn im Feaure-Raum bei der Kalkulation der Distanzen mit einbezogen werden (Mitchell, 1997).

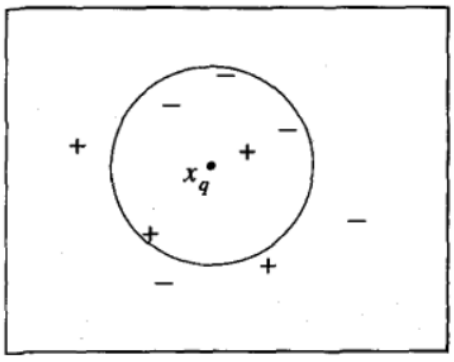

Abb. 3. k-nearest-neighbor. Die Kategorisierung der Instanz x_q als positiv oder negativ entscheidet sich anhand der k zu berücksichtigenden Nachbarn. Ein 1–nearest-neighbor Algorithmus klassifiziert die Instanz als positiv. Bei einem 5–nearest-neighbor fällt die Klassifikation negativ aus (Mitchell, 1997).

Analog für die Klassen „−“ und „+“ in Abb. 3 könnte es sich, wie beim verwendeten Beispiel zur Exemplarsicht (vgl. 2.1), genauso um Instanzen handeln, die das Konzept „Stuhl“ oder „Tisch“ repräsentieren. Der Algorithmus ordnet die Kategorie mit der größten Ähnlichkeit unter den k nächsten Nachbarn der neuen Instanz zu. K-nearest-neighbor macht implizit die Annahme, dass je ähnlicher (geringere Distanz im Feature-Raum) sich eine neue Instanz und Instanzen einer bestimmten Klasse sind, desto wahrscheinlicher wird das neue Exemplar dieser Kategorie angehören. Ein vergleichbarer Bias wirkt ebenso beim Menschen, wenn man sich im besonderen auf die Exemplarsicht bei den ähnlichkeitsbasierten Theorien bezieht.

3.2 Analytical Inductive Logic Programming

Mit ILP (Inductive Logic Programming) lernt man in aller Regel rekursive Programme. In diesem speziellen Forschungszweig des maschinellen Lernens setzt man, neben den hier thematisierten analytischen Ansätzen, auch Generate-Then-Test Algorithmen ein, wie sie bspw. im evolutionären Lernen zu finden sind. Generate-Then-Test hat im Gegensatz zum analytischen Ansatz wenig mit den Prozessen beim Menschen gemein.
Generell unterscheidet sich ILP stark von den häufig statistischen Standardlernverfahren wie Neuronale Netze oder Support Vector Machine. Im Unterschied zum klassischen Klassifikationslernen müssen die gelernten Programme die korrekte Lösung für das Lernproblem liefern und alle gestellten Beispiele korrekt erkennen. Man verfolgt an dieser Stelle keine Approximation einer Zielfunktion im Sinne einer höchstwahrscheinlich korrekten Lösung oder Ähnlichem. In anderen Domänen würde man das Vorgehen als Overfitting betrachten. Vergleichbar zu Entscheidungsbäumen ist ILP ein symbolischer White Box Ansatz. Das Gelernte kann also von Menschen verbalisiert, inspiziert und kommuniziert werden. Für die Synthese des Programmes durchsucht der Algorithmus einen Hypothesenraum bestehend aus Programmen. Das gelernte Programm besteht am Ende oft aus Anweisungen in einer deklarativen Programmiersprache (z. B. Haskell oder Prolog). Menschen setzen erwiesenermaßen bei Kategorien wie „Vorfahre“ oder „Primzahl“ rekursive Konzepte ein. Durch die Möglichkeit, Konzepte um Rekursion zu erweitern, kann eine Fähigkeit des Menschen erschlossen werden, die andere Klassifkationsverfahren nicht lernen können. Den beispielbasierten Ansätzen reichen zudem wenig Trainingsdaten aus (Schmid & Kitzelmann, 2011).

Das exemplarische System IGOR2 (Schmid & Kitzelmann, 2011) ist in der Lage die menschliche Fähigkeit nachzubilden, effizient über gegebene Strukturen zu verallgemeinern (vgl. visuell dargestellt in in Abb. 1). Das System kann bereits von drei positiven Beispielen die Rekursion der „Türme von Hanoi“ lernen. Dies ist damit ein gelungener Versuch einen ähnlich produktiven Bias des Menschen im maschinellen Lernen umzusetzen. Beeindruckend ist, dass das System mit Beispielen lernt, die mit einigen syntaktischen Vorkenntnissen der verwendeten Programmiersprache ebenso von einem Menschen für die Lösung des Problems interpretierbar sind.

IGOR2 als ein konkreter Ansatz hat weitreichendere Relevanz für das Feld der künstlichen Intelligenz in Bezug auf die Modellierung von Programmen mithilfe kognitiver Architekturen. In diesem Kontext kann IGOR2 als ein Modul gesehen werden, das den Inhalt des Arbeitsgedächtnisses verallgemeinert. Das nächste Kapitel greift kognitive Architekturen etwas detaillierter auf.

4 Methoden zur Erforschung menschlicher Lernmechanismen

In Experimenten studieren Forscher die menschlichen Lernmechanismen. Insbesondere die Untersuchung, bei welchen Gegebenheiten Fehlinterpretationen bzw. Fehlklassifikation beim Menschen auftreten hat eine wichtige Bedeutung für die Entdeckung menschlicher Bias. Zunächst stelle ich im folgenden Abschnitt kognitive Architekturen vor, die man u .a. bei der kognitiven Modellierung in Versuchen einsetzt. Insgesamt soll ein Eindruck darüber vermittelt werden, wie man menschliche Bias identifizieren und die Forschungsergebnisse zusätzlich mit Modellen verifizieren kann. Schließlich wird eine konkrete Methode zur Extraktion der visuellen Bias beim Menschen und deren anschließender Implementierung in ein System vorgestellt.
Kognitive Architekturen sind dafür geeignet, um mehr über die Prozesse menschlicher Kognition und unserer Lernmechanismen zu erfahren. Ziel der verschiedenen Architekturen, wie bspw. ACT-R, ist es, mit ihnen grundlegendes, intelligentes, menschliches Handeln für die Programmierung intelligenter Agenten oder Modelle zur Verfügung zu stellen. Das notwendige Wissen über die menschliche Kognition zum Bau einer kognitiven Architektur stammt aus hochvaliden Versuchsresultaten in der Psychologie. Architekturen stellen ein Framework bereit, mit dessen Hilfe es für Forscher möglich ist, Programme zu modellieren, die in bestimmten Situationen ihre Aktionen ähnlich einem menschlichen Akteur wählen. Um dies zu gewährleisten enthalten ausgereifte Architekturen z. B. eine Funktion für ein Kurzzeit- und Langzeitgedächtnis. Im Idealfall verfügt eine Architektur über Lernfähigkeit und ist damit imstande, Fähigkeiten mit der Zeit zu verbessern – vergleichbar mit einem Schachspieler, der mit zunehmender Anzahl an Spielen Situationen besser zu lösen weiß. Experimentatoren erhoffen sich von den Modellen, dass sie konsistente Ergebnisse im Vergleich zu Experimenten mit realen Testpersonen zurückliefern. Zeigen sich Übereinstimmungen, so erhöht dies die Validität der Forschungsergebnisse. Kognitive Architekturen leisten damit einen wichtigen Beitrag, um die Forschung an kognitiven Prozessen beim Menschen zu unterstützen (Langley, Laird, & Rogers, 2009).
Menschliche Bias zeigen sich unter anderem, wenn der Mensch Fehlkonzeptionen oder einer optischen Illusion aufliegt. Die Untersuchung der Bedingungen unter denen sich beim Menschen beispielsweise eine optische Täuschung provozieren lässt, kann viel darüber aussagen, auf welche Art und Weise die menschliche Kognition funktioniert. Es ist wichtig zu wissen, wann menschliche Klassifikationsmechanismen versagen, um mehr über die Prozesse im Hintergrund in Erfahrung zu bringen.

4.1 Transfer menschlicher Bias in Systeme

Vondrick et al. versuchen in ihrem Papier „Learning visual biases from human imagination“ mithilfe einer Methode aus der experimentellen Psychologie die Bias des visuellen Systems zu schätzen. Durch diese gelingt es uns, in unserer Umgebung sehr schnell Konzepte zu erkennen. Die Forscher versuchen, die visuellen Bias in ein System für maschinelles Sehen zu transferieren. Konkret besteht die Aufgabe darin, sich die Fähigkeit des Menschen zunutze zu machen, durch die er in der Lage ist visuelle Objekte von zufälligem Rauschen zu differenzieren. Die Ergebnisse legen nahe, dass der Einsatz von menschlichen Bias die Performance eines Klassifikationsalgorithmus zur Objekterkennung verbessert, wenn nur wenige Trainingsdaten zur Verfügung stehen.
Das eingesetzte Verfahren aus der Psychologie wurde ursprünglich zur Untersuchung der internen Vorlage des visuellen Systems beim Menschen entwickelt, die er für die Erkennung von Objekten nutzt. Im Originalverfahren werden Bilder mit zufälligem Rauschen verunreinigt. Anschließend soll von den Probanden beurteilt werden, ob sie eine bestimmte Kategorie (bspw. ein Auto oder die Gestalt einer Person) identifizieren können. Nach ausreichend vielen Versuchen schätzt man durch statistische Methoden eine Vorlage, wie die Zielkategorie intern repräsentiert sein könnte. Dies kann Aufschluss darüber geben, welche Muster bei der Diskriminierung von Kategorien eine Rolle spielen.
Der aktuelle Ansatz modifiziert das Originalverfahren dahingehen, dass die Vorlage in einem System zum maschinellen Sehen geschätzt wird. Die verfälschten Bilder erstellt man im Feature-Raum. Daraufhin nutzen sie Algorithmen zur Umwandlung der visuellen Merkmale in Bilder, welche anschließend von Menschen klassifiziert werden. Die menschlichen Bias in Form der approximierten Vorlage können im Anschluss in einen Algorithmus integriert werden (Vondrick, Pirsiavash, Oliva, & Torralba, 2015).

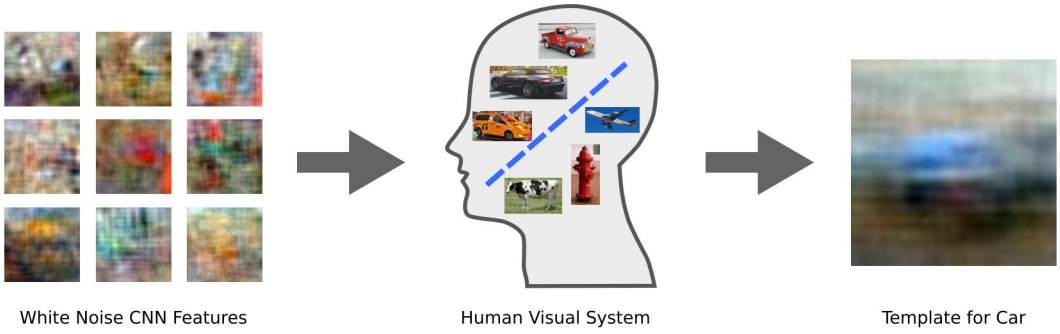

Abb. 4. Menschen identifizieren Objekte in Bildern aus Zufallsrauschen. Die Daten verwendet man anschließend, um eine Vorlage zu approximieren (Vondrick et al., 2015).

Eine weitere Anpassung des ursprünglichen Verfahrens besteht darin, dass man keine realen Bilder verfälscht, sondern die Bilder allein durch Zufallsrauschen im Feature-Raum generiert werden. Dadurch schließt man aus, dass die Probanden eventuell von den noch erkennbaren Strukturen des Originalbildes beeinflusst werden. Zudem erhöht sich die Wahrscheinlichkeit, die tatsächlichen Bias des Menschen zu erfassen und nicht solche, die von den verwendeten Daten verursacht werden.
Die geschätzte Repräsentation der Vorlage als Klassifikator (komplett ohne Trainingsdaten) erzielt Ergebnisse über Zufallswahrscheinlichkeit. In einem weiteren Versuch integrieren Vondrick et al. die Vorlage in eine Support Vector Machine. Dem Algorithmus stellte man wenige Trainingsdaten zur Verfügung. Die Hyperebene wurde dahingehend beeinflusst, dass deren Ausrichtung sich an dem zuvor geschätzten menschlichen Bias orientiert. Die Berücksichtigung des menschlichen Bias hatte eine positive Auswirkung auf die Performance des Algorithmus, wenn nur wenig Trainingsdaten verfügbar waren. Zusammengefasst demonstriert dies, dass ein Transfer menschlicher Bias in ein System wertvolle Signale geben kann, die dabei helfen Objekte in Bildern zu erkennen.

5 Schlussfolgerung und Ausblick

Der Artikel bietet eine Übersicht der grundlegenden Theorien zum menschlichen Konzeptlernen. Generell werden in der Wissenschaft auch Multi-Modelle und hybride Ansätze diskutiert. Aus neueren Erkenntnissen der Neurowissenschaften spricht viel für ein Zusammenwirken mehrerer Komponenten (Müsseler & Rieger, 2017). Bemerkenswert ist, dass betreffend der Akquisition von Kategorien wenig bekannt ist. Die vorhandenen Theorien setzen sich hauptsächlich mit der Frage auseinander, wie Kategorien mental repräsentiert werden. In Kapitel zwei werden Algorithmen aus dem Bereich des überwachten Lernens vorgestellt. Diese stehen für das Lernen auf Basis von Trainingsbeispielen. Auch Menschen verwenden Beispiele aus ihrer Erfahrung, um zu lernen sowie komplexe Anforderungen zu meistern. Durch unser Vorwissen sind wir in der Lage, Dinge vorauszuahnen, zu planen, sowie bereits bekannte Strategien bei neuen Aufgaben anzuwenden. Einige prägnante Parallelen z. B. zum instanzbasierten Lernen konnten identifiziert werden. Zahlreiche weitere Überschneidungen mit anderen Bereichen des maschinellen Lernens, die sich bspw. im Kontext von Reinforcement Learning mit lebenslangen Lernszenarien beschäftigen, sind anzunehmen. Weiterführende Arbeiten könnten dies aufgreifen und u. a. das Vergessen von Gelerntem für ein effizienteres Lernen behandeln.
Ich zeige in der Arbeit, dass bereits Versuche unternommen werden, um menschliche Bias in Systeme zu integrieren. Offen bleibt, inwiefern weiteres Know-how aus der Psychologie zur Entwicklung lernender Systeme von Nutzen sein kann. Beide Teildisziplinen haben das Potential bei der Erforschung künstlicher Intelligenz voneinander zu profitieren. Dadurch könnten sich zukünftig Chancen ergeben, weitere Bias des Menschen für eine Verbesserung bestehender und Innovation neuer Algorithmen zu reproduzieren.

Literatur

Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy. *American Psychologist*, *52*(1), 32–34. doi: 10.1037//0003-066X.52.1.32

Hunt, E. B., Stone, P. J., & Marin, J. (1966). *Experiments in induction*. New York.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science (New York, N.Y.)*, *350*(6266), 1332–1338. doi: 10.1126/science.aab3050

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160. doi: 10.1016/j.cogsys.2006.07.004

Mitchell, T. M. (1997). *Machine learning* (1st ed.). New York, NY, USA: McGraw-Hill, Inc.

Müsseler, J., & Rieger, M. (2017). *Allgemeine psychologie* (3. Auflage ed.). Berlin, Heidelberg: Springer.

Schmid, U., & Kitzelmann, E. (2011, September). Inductive rule learning on the knowledge level. *Cogn. Syst. Res.*, *12*(3-4), 237–248. doi: 10.1016/j.cogsys.2010.12.002

Vondrick, C., Pirsiavash, H., Oliva, A., & Torralba, A. (2015). Learning visual biases from human imagination. , 289–297. Retrieved from `http://papers.nips.cc/paper/5781-learning-visual-biases-from-human-imagination.pdf`

Winston, P. H. (1980). Learning and reasoning by analogy. *Communications of the ACM*, *23*(12), 689–703. doi: 10.1145/359038.359042

Erklärung

Ich erkläre hiermit, dass ich die Hausarbeit mit dem Titel „Parallelen zwischen dem menschlichen und maschinellen Lernen“ im Rahmen der Lehrveranstaltung „Seminar KI: gestern, heute, morgen“ im Wintersemester 2017/18 selbständig angefertigt, keine anderen Hilfsmittel als die im Quellen- und Literaturverzeichnis genannten benutzt und alle aus den Quellen und der Literatur wörtlich oder sinngemäß übernommenen Stellen als solche gekennzeichnet habe.

Bamberg, 13.03.2018