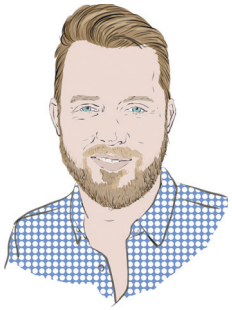
Jon Krohn · Grant Beyleveld · Aglaé Bassens

Deep Learning illustriert

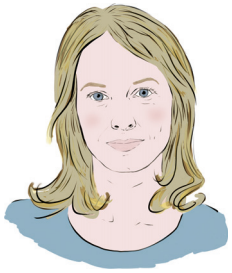
Eine anschauliche Einführung
in Machine Vision,
Natural Language Processing
und Bilderzeugung
für Programmierer und Datenanalysten

dpunkt.verlag

Jon Krohn ist Chief Data Scientist beim Machine-Learning-Unternehmen *untapt*. Er präsentiert eine viel gerühmte Serie aus Tutorials, die von Addison-Wesley herausgebracht wurden, darunter *Deep Learning with TensorFlow LiveLessons* und *Deep Learning for Natural Language Processing LiveLessons*. Jon unterrichtet Deep Learning an der *New York City Data Science Academy* und als Gastdozent an der Columbia University. Er besitzt einen Doktortitel in Neurowissenschaften von der Universität Oxford und veröffentlicht seit 2010 Artikel zum Thema Machine Learning in führenden Fachzeitschriften, darunter *Advances in Neural Information Processing Systems*.



Grant Beyleveld ist Data Scientist bei *untapt*, wo er auf dem Gebiet der Verarbeitung natürlicher Sprache mittels Deep Learning arbeitet. Er besitzt einen Doktortitel in biomedizinischer Wissenschaft von der *Icahn School of Medicine* am *Mount Sinai Hospital* in New York City, wo er die Beziehung zwischen Viren und ihren Wirten untersuchte. Er ist Gründungsmitglied von *deeplearning-studygroup.org*.



Aglaé Bassens ist eine in Paris lebende belgische Künstlerin. Sie studierte bildende Kunst an *The Ruskin School of Drawing and Fine Art* der Universität Oxford und an der *Slade School of Fine Arts des University College London*. Neben ihrer Arbeit als Illustratorin malt sie Stillleben und Wandbilder.

Jon Krohn · Grant Beyleveld · Aglaé Bassens

Deep Learning illustriert

Eine anschauliche Einführung in Machine Vision, Natural Language Processing und Bilderzeugung für Programmierer und Datenanalysten

Aus dem Englischen von Kathrin Lichtenberg

dpunkt.verlag

Jon Krohn · Grant Beyleveld · Aglaé Bassens

Lektorat: Gabriel Neumann
Übersetzung: Kathrin Lichtenberg, Ilmenau
Copy-Editing: Friederike Daenecke, Zülpich
Terminologie-Beratung: Marcus Fraaß
Satz: Birgit Bäuerlein
Herstellung: Stefanie Weidner
Umschlaggestaltung: Helmut Kraus, *www.exclam.de*
Druck und Bindung: mediaprint solutions GmbH, 33100 Paderborn

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über *http://dnb.d-nb.de* abrufbar.

ISBN:
Print 978-3-86490-663-3
PDF 978-3-96088-751-5
ePub 978-3-96088-752-2
mobi 978-3-96088-753-9

Hinweis:
Dieses Buch wurde auf PEFC-zertifiziertem Papier aus nachhaltiger Waldwirtschaft gedruckt. Der Umwelt zuliebe verzichten wir zusätzlich auf die Einschweißfolie.

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: *hallo@dpunkt.de*.

5 4 3 2 1 0

❖ *Für Gigi* ❖

Vorwort

Machine Learning gilt vielen Menschen als die Zukunft der Statistik und Computertechnik, da es völlig neue Akzente in Kundendienst, Design, Bankwesen, Medizin, Produktion und in vielen anderen Bereichen und Branchen setzt. Es ist kaum möglich, seinen Einfluss auf die Welt und jene Veränderungen, die Machine Learning in den kommenden Jahren und Jahrzehnten bringen wird, überzubewerten. Von der Vielzahl an Machine-Learning-Methoden, die von Experten eingesetzt werden, etwa *Penalized Regression*, *Random Forest* und *Boosted Trees*, ist *Deep Learning* vermutlich die aufregendste.

Deep Learning hat die Gebiete *Computer Vision* (maschinelles Sehen) und *Natural Language Processing* (Verarbeitung natürlicher Sprache) revolutioniert, und Forscher finden immer neue Bereiche, die sie mit der Macht neuronaler Netze verwandeln wollen. Seine größte und beeindruckendste Wirkung zeigt Deep Learning oft bei den Versuchen, das menschliche Erleben nachzuahmen, wie bei der erwähnten Seh- und Sprachverarbeitung sowie bei der Audiosynthese und bei Übersetzungen. Die Berechnungen und Konzepte, die dem Deep Learning zugrunde liegen, wirken möglicherweise abschreckend und hindern Menschen unnötigerweise daran, sich damit zu befassen.

Die Autoren von *Deep Learning illustriert* gehen diese traditionell wahrgenommenen Hürden an und vermitteln ihr Wissen ruhig und gelassen – und das entstandene Buch ist eine wahre Freude. Wie die anderen Bücher aus dieser Reihe – *R for Everyone*, *Pandas for Everyone*, *Programming Skills for Data Science* und *Machine Learning with Python for Everyone* – wendet sich dieses Buch an ein breites Publikum mit ganz unterschiedlichem Wissen und Können. Die mathematischen Notationen sind auf ein Minimum beschränkt, und falls dennoch Gleichungen erforderlich sind, werden sie von verständlichem Text begleitet. Die meisten Erkenntnisse werden durch Grafiken, Illustrationen und Keras-Code ergänzt, der in Form leicht nachzuvollziehender Jupyter-Notebooks zur Verfügung steht.

Jon Krohn unterrichtet schon seit vielen Jahren Deep Learning. Besonders denkwürdig war eine Präsentation beim *Open Statistical Programming Meetup* in New York – bei derselben Vereinigung, in der er seine *Deep Learning Study Group* startete. Seine Brillanz in diesem Thema zeigt sich an seinen Texten, die Lesern Bildung vermitteln und ihnen gleichzeitig zeigen, wie spannend und aufregend das Material ist. Für dieses Buch arbeitet er mit Grant Beyleveld und Aglaé Bassens zusammen, die ihr Wissen bei der Anwendung von Deep-Learning-Algorithmen und ihre gekonnten und witzigen Zeichnungen beisteuern.

Deep Learning illustriert kombiniert Theorie, Mathematik (dort, wo es nötig ist), Code und Visualisierungen zu einer umfassenden Behandlung des Themas Deep Learning. Das Buch behandelt die volle Breite des Themas, einschließlich vollständig verbundener Netzwerke, Convolutional Neural Networks, Recurrent Neural Networks, Generative Adversarial Networks und Reinforcement Learning sowie deren Anwendungen. Dadurch ist dieses Buch die ideale Wahl für jemanden, der neuronale Netze kennenlernen und gleichzeitig praktische Hinweise für deren Implementierung haben möchte. Jeder kann und sollte davon profitieren und außerdem seine Zeit beim Lesen mit Jon, Grant und Aglaé genießen.

Jared Lander
Herausgeber der Reihe

Einführung

Milliarden miteinander verbundener Neuronen, gemeinhin als Gehirn bezeichnet, bilden Ihr Nervensystem und erlauben es Ihnen, zu spüren, zu denken und zu handeln. Durch akribisches Einfärben und Untersuchen dünner Scheiben von Gehirnmasse konnte der spanische Arzt Santiago Cajal (Abbildung 1) als erster[1] Neuronen identifizieren (Abbildung 2). In der ersten Hälfte des 20. Jahrhunderts begannen Forscher zu verstehen, wie diese Zellen arbeiten. In den 1950er-Jahren experimentierten Wissenschaftler, die von unserem zunehmenden Verständnis für das Gehirn inspiriert waren, mit computerbasierten künstlichen Neuronen und verknüpften diese zu künstlichen neuronalen Netzen, die versuchten, die Funktionsweise ihres natürlichen Namensvetters nachzuahmen.

Gewappnet mit dieser kurzen Geschichte der Neuronen, können wir den Begriff Deep Learning täuschend leicht definieren: Deep Learning beinhaltet ein Netzwerk, in dem künstliche Neuronen – üblicherweise Tausende, Millionen oder noch mehr davon – wenigstens mehrere Schichten tief gestapelt sind. Die künstlichen Neuronen in der ersten Schicht übergeben Informationen an die zweite, die zweite Schicht reicht sie an die dritte und so weiter, bis die letzte Schicht irgendwelche Werte ausgibt. Wie wir allerdings im Laufe dieses Buches zeigen werden, kann diese simple Definition die bemerkenswerte Breite der Funktionalität des Deep Learning sowie seine außerordentlichen Zwischentöne nicht annähernd erfassen.

Wie wir in Kapitel 1 genauer ausführen werden, war die erste Welle des Deep-Learning-Tsunami, die metaphorisch gesprochen ans Ufer brandete, eine herausragende Leistung in einem wichtigen Machine-Vision-Wettbewerb im Jahre 2012. Sie wurde getrieben und unterstützt durch das Vorhandensein einigermaßen preiswerter Rechenleistung, ausreichend großer Datensätze und einer Handvoll wesentlicher theoretischer Fortschritte. Akademiker und Techniker merkten auf, und in den turbulenten Jahren seither hat das Deep Learning zahlreiche, mittlerweile alltägliche Anwendungen gefunden. Von Teslas Autopilot bis zur Stimmerkennung von Alexa, von Echtzeitübersetzungen zwischen Sprachen bis hin zu seiner Integration in Hunderte von Google-Produkten hat Deep Lear-

1. Cajal, S.-R. (1894). *Les Nouvelles Idées sur la Structure du Système Nerveux chez l'Homme et chez les Vertébrés*. Paris: C. Reinwald & Companie.

ning die Genauigkeit vieler durch Computer erledigter Aufgaben von 95 Prozent auf teils mehr als 99 Prozent verbessert – die entscheidenden Prozentpunkte, die dafür sorgen, dass ein automatisierter Dienst sich tatsächlich anfühlt, als würde er von Zauberhand ausgeführt werden. Auch wenn die in diesem Buch gelieferten interaktiven Codebeispiele die vorgebliche Magie entzaubern, verschafft das Deep Learning den Maschinen eine übermenschliche Fähigkeit bei komplexen Aufgaben, die so verschieden sind wie das Erkennen von Gesichtern, das Zusammenfassen von Texten und das Spielen schwieriger Brettspiele.[2] Angesichts dieser markanten Fortschritte überrascht es kaum, dass »Deep Learning« gleichgesetzt wird mit »künstlicher Intelligenz« – in der Presse, am Arbeitsplatz und zu Hause.

Abb. 1 *Santiago Cajal (1852–1934)*

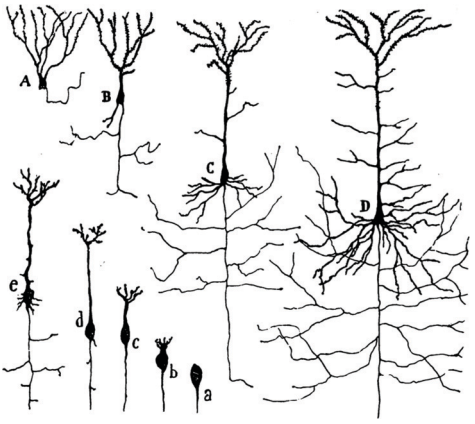


Abb. 2 *Ein handgezeichnetes Diagramm aus Cajals Veröffentlichung (1894) zeigt das Wachstum eines Neurons (a–e) und verschiedenartige Neuronen eines Frosches (A), einer Eidechse (B), einer Ratte (C) und eines Menschen (D)*

2. Unter *bit.ly/aiindex18* finden Sie einen Vergleich zwischen menschlicher und maschineller Leistungsfähigkeit.

Es sind aufregende Zeiten, weil – wie Sie in diesem Buch entdecken werden – vermutlich nur einmal im Leben ein einziges Konzept in so kurzer Zeit so umfassende Umstürze mit sich bringt. Wir sind hocherfreut, dass auch Sie Interesse an Deep Learning gefunden haben, und können es kaum erwarten, unseren Enthusiasmus für diese beispiellose Technik mit Ihnen zu teilen.

Wie Sie dieses Buch lesen sollten

Dieses Buch besteht aus vier Teilen. Teil I, »Deep Learning vorgestellt«, eignet sich für alle interessierten Leserinnen und Leser. Es ist ein allgemeiner Überblick, der uns verrät, was Deep Learning eigentlich ist, wie es sich entwickelt hat und wie es mit Konzepten wie KI, Machine Learning und Reinforcement Learning verwandt ist. Voller eigens geschaffener Illustrationen, eingängiger Analogien und auf das Wesentliche konzentrierter Beschreibungen, sollte Teil I für alle erhellend sein, also auch für diejenigen, die keine besondere Programmiererfahrung mitbringen.

Die Teile II bis IV wenden sich hingegen an Softwareentwickler, Data Scientists, Forscher, Analysten und andere, die gern lernen möchten, wie sich Deep-Learning-Techniken auf ihrem Gebiet einsetzen lassen. In diesen Teilen unseres Buches wird die wesentliche zugrunde liegende Theorie behandelt. Hierbei wird der Einsatz mathematischer Formeln auf das Mindestmaß reduziert und stattdessen auf intuitive visuelle Darstellungen und praktische Beispiele in Python gesetzt. Neben dieser Theorie vermitteln funktionierende Codeausschnitte, die in den begleitenden Jupyter-Notebooks[3] zur Verfügung stehen, ein praktisches Verständnis für die wichtigsten Familien der Deep-Learning-Ansätze und -Anwendungen: Maschinelles Sehen (Machine Vision) (Kapitel 10), Verarbeitung natürlicher Sprache (Natural Language Processing) (Kapitel 11), Bildherstellung (Kapitel 12) und Spiele (Kapitel 13). Damit er besser zu erkennen ist, geben wir Code immer in einer `solchen Nichtproportionalschrift (also in einer Schrift mit fester Breite)` an. Außerdem verwenden wir in den Codeausschnitten den üblichen Jupyter-Stil (Zahlen in Grün, Strings in Rot usw.).

Falls Sie sich nach detaillierteren Erklärungen der mathematischen und statistischen Grundlagen des Deep Learning sehnen, als wir in diesem Buch anbieten, könnten Sie sich unsere Tipps für weitere Studien anschauen:

1. Michael Nielsens E-Book *Neural Networks and Deep Learning*[4], das kurz ist, Konzepte mithilfe netter interaktiver Applets demonstriert und eine ähnliche mathematische Notation verwendet wie wir

3. *github.com/the-deep-learners/deep-learning-illustrated*
4. Nielsen, M. (2015). *Neural Networks and Deep Learning*. Determination Press. Kostenlos verfügbar unter: *neuralnetworksanddeeplearning.com*

2. Das Buch *Deep Learning*[5] von Ian Goodfellow (vorgestellt in Kapitel 3), Yoshua Bengio (Abbildung 1–10) und Aaron Courville, das ausführlich die mathematischen Grundlagen neuronaler Netzwerktechniken behandelt

Überall im Buch finden Sie freundliche Trilobiten, die Ihnen gern kleine Schnipsel nicht ganz so notwendiger Informationen anbieten möchten, die für Sie vielleicht dennoch interessant oder hilfreich sein könnten. Der *lesende Trilobit* (wie in Abbildung 3) ist ein Bücherwurm, der Freude daran hat, Ihr Wissen zu erweitern. Der Trilobit, der um Ihre Aufmerksamkeit bittet (wie in Abbildung 4), hat eine Textpassage bemerkt, die möglicherweise problematisch für Sie ist, und würde in dieser Situation gern helfen. Zusätzlich zu den Trilobiten, die die Kästen bevölkern, haben wir reichlich Gebrauch von Fußnoten gemacht. Diese müssen Sie nicht unbedingt lesen, aber sie enthalten kurze Erklärungen neuer Begriffe und Abkürzungen sowie Quellenangaben zu wichtigen Artikeln, Büchern und anderen Referenzen, die Sie bei Interesse bemühen können.

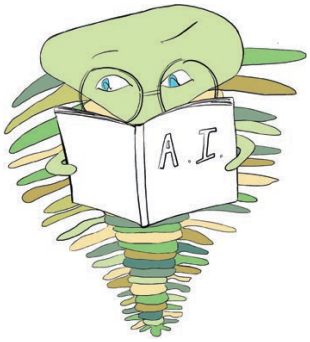


Abb. 3 *Der lesende Trilobit hat Freude daran, Ihr Wissen zu erweitern.*

Abb. 4 *Dieser Trilobit möchte Ihre Aufmerksamkeit auf eine schwierige Textpassage lenken. Achten Sie auf ihn!*

5. Goodfellow, I., et al. (2016). *Deep Learning*. MIT Press. Kostenlos verfügbar unter: *deeplearningbook.org*

Für einen Großteil des Inhalts dieses Buches gibt es begleitende Video-Tutorials in englischer Sprache. Dieses Buch bot uns die Möglichkeit, die theoretischen Konzepte gründlicher darzustellen, und die Videos erlauben es Ihnen, sich aus einer anderen Perspektive mit den Jupyter-Notebooks vertraut zu machen: Hier wird die Bedeutung der einzelnen Codezeilen bereits beim Eintippen beschrieben.[6] Die Serie der Video-Tutorials verteilt sich über drei Titel, die jeweils bestimmte Kapitel dieses Buches begleiten:

1. *Deep Learning with TensorFlow LiveLessons:*[7]
   Kapitel 1 und Kapitel 5 bis 10
2. *Deep Learning for Natural Language Processing LiveLessons:*[8]
   Kapitel 2 und 11
3. *Deep Reinforcement Learning and GANs LiveLessons:*[9]
   Kapitel 3, 4, 12 und 13

6. Viele der Jupyter-Notebooks, die in diesem Buch behandelt werden, sind direkt aus den Videos abgeleitet, die alle vor dem Schreiben des Buches aufgezeichnet wurden. An manchen Stellen haben wir entschieden, den Code für das Buch zu aktualisieren, sodass die Video-Version und die Buchversion eines Notebooks einander zwar ähnlich sind, aber nicht unbedingt völlig identisch sein müssen.
7. Krohn, J. (2017). *Deep Learning with TensorFlow LiveLessons: Applications of Deep Neural Networks to Machine Learning Tasks* (Videokurs). Boston: Addison-Wesley.
8. Krohn, J. (2017). *Deep Learning for Natural Language Processing LiveLessons: Applications of Deep Neural Networks to Machine Learning Tasks* (Videokurs). Boston: Addison-Wesley.
9. Krohn, J. (2018). *Deep Reinforcement Learning and GANs LiveLessons: Advanced Topics in Deep Learning* (Videokurs). Boston: Addison-Wesley.

Danksagungen

Wir danken dem Team bei *untapt*, vor allem Andrew Vlahutin, Sam Kenny und Vince Petaccio II, die uns unterstützten, während wir dieses Buch schrieben. Besonders erwähnen wollen wir Ed Donner, der neuronale Netze liebt und uns pausenlos ermutigte, unserer Leidenschaft auf dem Gebiet des Deep Learning zu folgen.

Außerdem danken wir den Mitgliedern der *Deep Learning Study Group*[1], die regelmäßig unsere stimulierenden und lebhaften Treffen im New Yorker Büro von untapt besuchen. Da dieses Buch aufgrund der Diskussionen unserer Study Group entstand, kann man sich kaum vorstellen, wie es ohne diese Treffen zustande gekommen wäre.

Dank geht an unsere technischen Gutachter für ihre wertvollen Ratschläge, die den Inhalt des Buches deutlich verbessert haben: Alex Lipatov, Andrew Vlahutin, Claudia Perlich, Dmitri Nesterenko, Jason Baik, Laura Graesser, Michael Griffiths, Paul Dix und Wah Loon Keng. Danke auch an die Lektoren und Hersteller des Buches – Chris Zahn, Betsy Hardinger, Anna Popick und Julie Nahil –, deren Sorgfalt und Aufmerksamkeit die Qualität, Klarheit und Gestaltung dieses Buches sicherstellten. Dank an Jared Lander, der die New Yorker Open-Statistical-Programming-Gemeinschaft leitet, die sowohl unsere Deep Learning Study Group begründete als auch ein Treffen mit Debra Williams Cauley in die Wege leitete. Ein besonderer Dank gilt Debra selbst, die unsere fantasievollen Publikationsideen von dem Tag an unterstützt hat, an dem wir sie kennenlernten, und die entscheidend an ihrer Umsetzung beteiligt war. Wir danken auch den Wissenschaftlern und Machine-Learning-Experten, die uns akademisch geleitet haben und uns weiterhin inspirieren, vor allem Jonathan Flint, Felix Agakov und Will Valdar.

Und schließlich geht ein unendlicher Dank an unsere Familien und Freunde, die nicht nur ertragen haben, dass wir auch im Urlaub und an den Wochenenden gearbeitet haben, sondern uns auch selbstlos motiviert haben, es zu tun.

1. *deeplearningstudygroup.org*

Inhaltsübersicht

Inhaltsverzeichnis

Teil I

Deep Learning vorgestellt

1 Biologisches und maschinelles Sehen

In diesem Kapitel und auch im Laufe dieses Buches dient uns das visuelle System biologischer Organismen als Analogie, um das Deep Learning »zum Leben zu erwecken«. Diese Analogie vermittelt nicht nur ein tiefgreifendes Verständnis für das, was Deep Learning ausmacht, sondern verdeutlicht auch, weshalb Deep-Learning-Ansätze so machtvoll und so überaus vielfältig einsetzbar sind.

1.1 Das biologische Sehen

Vor 550 Millionen Jahren, in der prähistorischen Periode des Kambrium, stieg die Anzahl der Arten auf unserem Planeten schlagartig an (Abbildung 1–1). Aus den Fossilienfunden lässt sich ablesen,[1] dass diese explosionsartige Zunahme (die auch tatsächlich als Kambrische Explosion bezeichnet wird) durch die Entwicklung von Lichtdetektoren bei Trilobiten gefördert wurde, einem kleinen Meereslebewesen, das mit den heutigen Krebsen verwandt ist (Abbildung 1–2). Ein visuelles System, selbst wenn es nur primitiv ausgebildet ist, bringt eine wunderbare Vielfalt neuer Fähigkeiten mit sich. Man kann beispielsweise bereits aus einiger Entfernung Nahrung, Feinde und freundlich aussehende Gefährten ausmachen. Auch andere Sinne, wie der Geruchssinn, erlauben es Tieren, diese Dinge wahrzunehmen, allerdings nicht mit der Genauigkeit und Schnelligkeit des Sehvermögens. Die Hypothese besagt, dass mit dem Sehvermögen der Trilobiten ein Wettrüsten einsetzte, dessen Ergebnis die Kambrische Explosion war: Die Beutetiere und auch die Feinde der Trilobiten mussten sich weiterentwickeln, um zu überleben.

1. Parker, A. (2004). *In the Blink of an Eye: How Vision Sparked the Big Bang of Evolution.* New York: Basic Books.

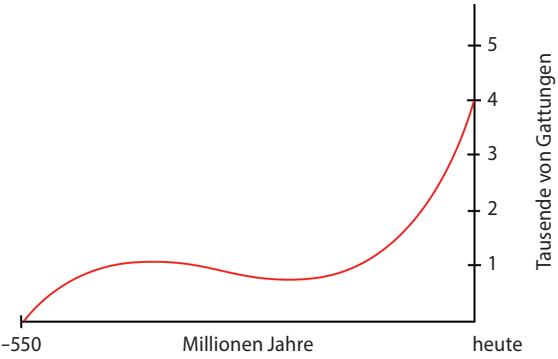


Abb. 1–1 *Die Anzahl der Arten auf unserem Planeten begann vor 550 Millionen Jahren, während der Periode des Kambrium, schlagartig anzusteigen. »Gattungen« sind Kategorien miteinander verwandter Arten.*

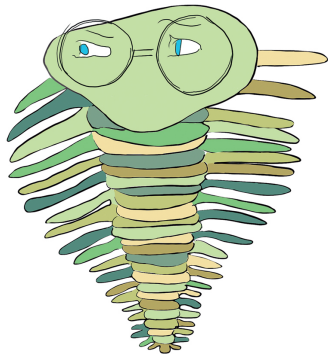


Abb. 1–2 *Ein Trilobit mit Brille*

In den mehr als eine halbe Milliarde Jahren, seit die Trilobiten das Sehen entwickelten, hat die Komplexität dieses Sinnes ungemein zugenommen. So ist bei heutigen Säugetieren ein Großteil der Großhirnrinde – das ist die äußere graue Masse des Gehirns – der visuellen Wahrnehmung vorbehalten.[2] Ende der 1950er-Jahre begannen die Physiologen David Hubel und Torsten Wiesel (Abbildung 1–3) an der John Hopkins University mit ihren bahnbrechenden Forschungen darüber, wie visuelle Informationen in der Großhirnrinde von Säugetieren verarbeitet werden,[3] für die sie später mit dem Nobelpreis[4] ausgezeichnet wurden. Wie in Abbildung 1–4 dargestellt wird, führten Hubel und Wiesel ihre Forschungen durch, indem sie narkotisierten Katzen Bilder zeigten, während sie gleichzeitig die Aktivität einzelner Neuronen aus dem primären visuellen Cortex aufzeichneten, also dem ersten Teil der Großhirnrinde, der visuellen Input von den Augen erhält.

Hubel und Wiesel zeigten den Katzen mithilfe von Dias, die sie auf eine Leinwand projizierten, einfache Formen, wie den Punkt aus Abbildung 1–4. Ihre ersten Ergebnisse waren entmutigend: Ihre Bemühungen lösten keine Reaktion der

Neuronen des primären visuellen Cortex aus. Sie waren frustriert, weil diese Zellen, die anatomisch das Eingangstor für die visuellen Informationen in die restliche Großhirnrinde zu sein schienen, nicht auf visuelle Stimuli reagierten. Verzweifelt versuchten Hubel und Wiesel vergeblich, die Neuronen anzuregen, indem sie vor der Katze auf und ab sprangen und mit den Armen fuchtelten. Nichts. Und dann, wie bei vielen der großen Entdeckungen, von Röntgen-Strahlen über das Penicillin bis zum Mikrowellenofen, machten Hubel und Wiesel eine unverhoffte Beobachtung: Als sie eines der Dias aus dem Projektor entfernten, löste dessen gerader Rahmen das unverkennbare Knistern ihres Aufzeichnungsgerätes aus, das damit signalisierte, dass ein Neuron des primären visuellen Cortex feuerte. Voller Freude feierten sie dies auf den Korridoren der Labors ihrer Universität.

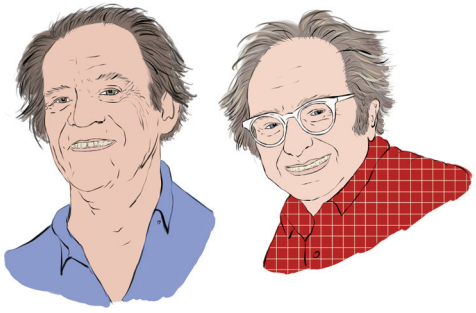


Abb. 1–3 *Die Neurophysiologen und Nobelpreis-Gewinner Torsten Wiesel (links) und David Hubel*

---

2. Ein paar interessante Fakten über die Großhirnrinde: Erstens: Sie gehört zu den neueren evolutionären Entwicklungen des Gehirns und ist verantwortlich für die Komplexität des Verhaltens von Säugetieren im Vergleich zum Verhalten älterer Klassen von Tieren, wie Reptilien und Amphibien. Zweitens: Auch wenn das Gehirn zwanglos als *graue Masse* bezeichnet wird, weil die Großhirnrinde die äußere Schicht des Gehirns bildet und dieses Gewebe grau ist, handelt es sich beim größten Teil des Gehirns um *weiße* Masse. Im Großen und Ganzen ist die weiße Masse verantwortlich für das Übertragen von Informationen über längere Distanzen als die graue Masse, weshalb ihre Neuronen eine weiße, fetthaltige Ummantelung haben, die die Signalübertragung beschleunigt. Man könnte sich die Neuronen in der weißen Masse quasi wie eine »Autobahn« vorstellen. Diese Schnellstraßen besitzen nur wenige Auf- oder Abfahrten, können ein Signal aber in Blitzesschnelle von einem Teil des Gehirns zum anderen befördern. Im Gegensatz dazu bieten die »Landstraßen« der grauen Masse eine Unzahl an Möglichkeiten für Verbindungen zwischen Neuronen, allerdings auf Kosten der Geschwindigkeit. Eine krasse Verallgemeinerung ist es daher, die Großhirnrinde – die graue Masse – als den Teil des Gehirns zu betrachten, in dem die komplexesten Berechnungen erfolgen, die den Tieren mit dem größten Anteil daran – wie den Säugetieren, speziell den Menschenaffen wie dem *Homo sapiens* – ihre komplexen Verhaltensweisen ermöglichen.

3. Hubel, D. H. und Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148*, 574–91.
4. Nobelpreis für Physiologie oder Medizin 1981, gemeinsam mit dem amerikanischen Neurobiologen Roger Sperry.

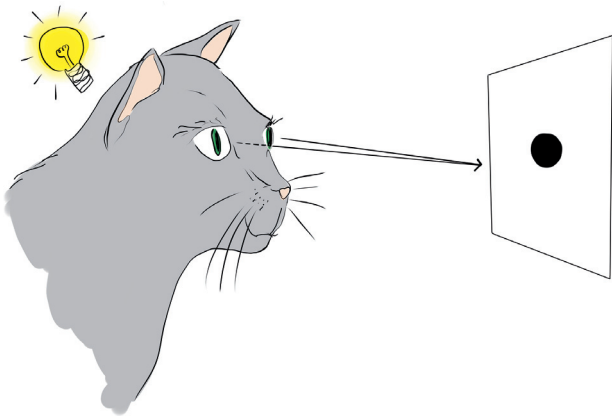


Abb. 1–4 *Hubel und Wiesel nutzten einen Lichtprojektor, um narkotisierten Katzen Dias zu zeigen, während sie die Aktivitäten im primären visuellen Cortex aufzeichneten. Für diese Experimente waren den Katzen elektrische Aufzeichnungsvorrichtungen in den Schädel implantiert worden. Wir schätzen, dass es angenehmer ist, die Aktivierung der Neuronen durch eine Glühlampe zu versinnbildlichen, statt die eigentliche Versuchsanordnung darzustellen. Gezeigt wird in diesem Bild ein Neuron aus dem primären visuellen Cortex, das zum Glück durch die gerade Kante eines Dias aktiviert wurde.*

Die glückliche Zufallsentdeckung des feuernden Neurons zeigte keine Anomalie. Durch weitere Experimente entdeckten Hubel und Wiesel, dass die Neuronen, die einen visuellen Input vom Auge empfingen, im Allgemeinen am empfänglichsten für einfache, gerade Kanten waren. Passenderweise nannten sie diese Zellen *einfache* Neuronen.

Wie Abbildung 1–5 zeigt, stellten Hubel und Wiesel fest, dass ein bestimmtes einfaches Neuron optimal auf eine Kante mit einer jeweils speziellen Ausrichtung reagiert. Eine große Gruppe aus Neuronen, die jeweils darauf spezialisiert sind, eine bestimmte Kantenausrichtung zu entdecken, kann gemeinsam die insgesamt möglichen 360 Grad an Ausrichtung darstellen. Diese einfachen Zellen für die Erkennung der Kantenausrichtung übergeben die Informationen dann weiter an eine große Zahl sogenannter *komplexer* Neuronen. Ein bestimmtes komplexes Neuron empfängt visuelle Informationen, die bereits durch mehrere einfache Zellen verarbeitet wurden, sodass es in der Lage ist, mehrere Linienausrichtungen zu einer komplexeren Form zu kombinieren, wie etwa zu einer Ecke oder einer Kurve.

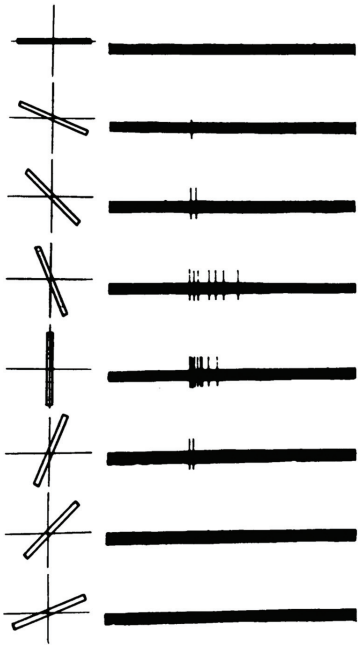


Abb. 1–5 *Eine einfache Zelle im primären visuellen Cortex einer Katze feuert in unterschiedlichen Raten, die von der Ausrichtung einer Linie abhängig sind, die der Katze gezeigt wird. Die Ausrichtung der Linie ist in der linken Spalte zu sehen, während die rechte Spalte das Feuern (die elektrische Aktivität) der Zelle über eine bestimmte Zeitspanne (eine Sekunde) zeigt. Eine senkrechte Linie (in der fünften Zeile von oben) verursacht die stärkste elektrische Aktivität für diese spezielle einfache Zelle. Linien, die nicht ganz senkrecht stehen (in den Zwischenzeilen) verursachen eine geringere Aktivität in der Zelle, während Linien, die nahezu waagerecht sind (in der obersten und untersten Zeile) kaum bis gar keine Aktivität auslösen.*

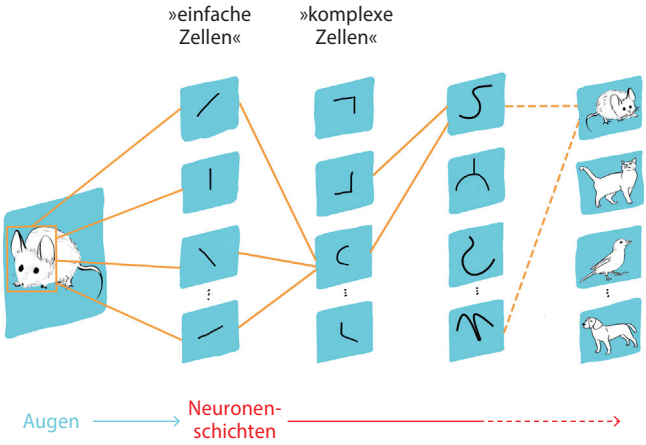


Abb. 1–6 *Diese Zeichnung zeigt, wie aufeinanderfolgende Ebenen aus biologischen Neuronen visuelle Informationen im Gehirn etwa einer Katze oder eines Menschen darstellen.*

Abbildung 1–6 illustriert, wie über viele hierarchisch organisierte Ebenen aus Neuronen, die Informationen an zunehmend übergeordnete Neuronen weiterreichen, schrittweise immer komplexere visuelle Stimuli durch das Gehirn dargestellt werden können. Die Augen sind auf das Bild eines Mäusekopfes gerichtet. Lichtphotonen stimulieren Neuronen in der Retina der einzelnen Augen. Diese visuellen Rohinformationen werden von den Augen in den primären visuellen Cortex des Gehirns übertragen. Die erste Schicht der Neuronen des primären visuellen Cortex, die diesen Input empfangen – Hubel und Wiesels *einfache Zellen* –, ist darauf spezialisiert, Kanten (gerade Linien) mit bestimmten Ausrichtungen zu erkennen. Es gibt viele Tausend solcher Neuronen; aus Gründen der Einfachheit zeigen wir in Abbildung 1–6 nur vier von ihnen. Diese einfachen Neuronen übermitteln Informationen über das Vorhandensein oder Fehlen von Linien einer bestimmten Ausrichtung an eine nachfolgende Ebene *komplexer Zellen*, die die Informationen aufnehmen und neu kombinieren, um auf diese Weise die Darstellung komplexerer visueller Stimuli, wie etwa der Wölbung des Mäusekopfes, zu ermöglichen. Während die Informationen mehrere aufeinanderfolgende Schichten durchlaufen, können die Darstellungen visueller Stimuli schrittweise immer komplexer und abstrakter werden. Wie durch die ganz rechte Schicht der Neuronen gezeigt wird, ist das Gehirn nach vielen Schichten dieser hierarchischen Verarbeitung (der gestrichelte Pfeil soll andeuten, dass viele weitere Verarbeitungsschichten vorhanden sind, aber nicht gezeigt werden) schließlich in der Lage, visuelle Konzepte darzustellen, die so komplex sind wie eine Maus, eine Katze, ein Vogel oder ein Hund.

Heute haben Neurowissenschaftler mithilfe zahlloser weiterer Aufzeichnungen aus den kortikalen Neuronen von Gehirnchirurgie-Patienten sowie aus nichtinvasiven Techniken wie der Magnetresonanztomographie (MRT)[5] eine ziemlich hoch aufgelöste Karte der Regionen zusammengestellt, die sich auf die Verarbeitung bestimmter visueller Stimuli spezialisiert haben, wie etwa Farbe, Bewegung und Gesichter (siehe Abbildung 1–7).

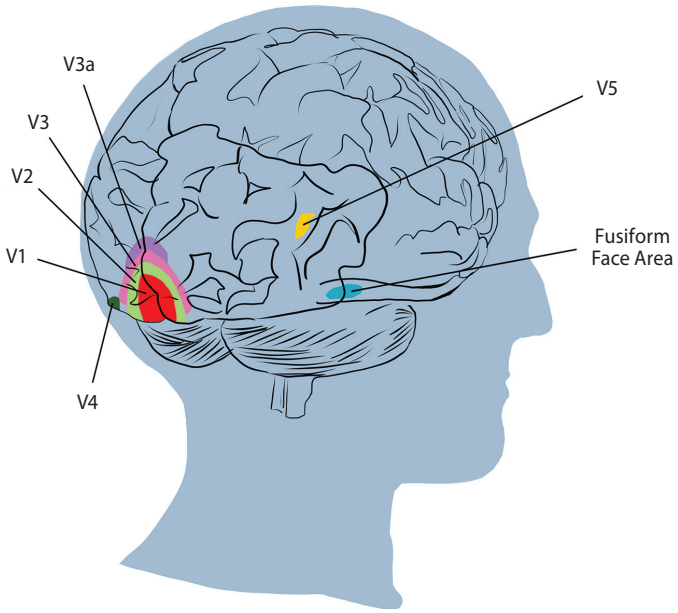


Abb. 1–7 *Regionen des visuellen Cortex. Die Region V1 empfängt Input von den Augen und enthält die einfachen Zellen, die die Kantenausrichtung erkennen. Durch die Neukombination von Informationen über eine Vielzahl nachfolgender Schichten aus Neuronen (unter anderem in den Regionen V2, V3 und V3a) werden zunehmend abstrakter werdende visuelle Stimuli dargestellt. Im menschlichen Gehirn (hier zu sehen) gibt es Regionen, die besonders viele Neuronen mit bestimmten Spezialisierungen enthalten, zum Beispiel für die Erkennung von Farbe (V4), Bewegung (V5) oder Gesichtern von Menschen (die »Fusiform Face Area« oder FFA).*

5. Besonders aus der funktionalen Magnetresonanztomographie, die Einblicke darüber erlaubt, welche Regionen der Großhirnrinde besonders aktiv oder inaktiv sind, wenn das Gehirn mit einer speziellen Aktivität befasst ist.

1.2 Maschinelles Sehen

Wir haben das biologische visuelle System nicht nur deswegen diskutiert, weil es interessant ist (obwohl Sie hoffentlich den vorangegangenen Abschnitt absolut faszinierend fanden), sondern weil es als Inspiration für die Deep-Learning-Ansätze des maschinellen Sehens (Machine Vision) dient, wie in diesem Abschnitt deutlich werden soll.

Abbildung 1–8 bietet einen kurzgefassten historischen Zeitstrahl des Sehens in biologischen Organismen sowie in Maschinen. Der obere, blaue Zeitstrahl hebt die Entwicklung des Sehens bei den Trilobiten sowie die Veröffentlichung von Hubel und Wiesel aus dem Jahre 1959 über das hierarchische Wesen des primären visuellen Cortex hervor, von dem im vorangegangenen Abschnitt die Rede war. Der Zeitstrahl zum maschinellen Sehen ist in zwei parallele Strömungen aufgeteilt, die zwei alternative Ansätze verkörpern. Der mittlere, rosa Zeitstrahl stellt den Deep-Learning-Ansatz dar, der in diesem Buch behandelt wird. Der untere, lila Zeitstrahl repräsentiert derweil den traditionellen Machine-Learning-Weg (ML) zum Sehen. Der Vergleich der beiden Vorgehensweisen verdeutlicht, wieso das Deep Learning so leistungsfähig und revolutionär ist.

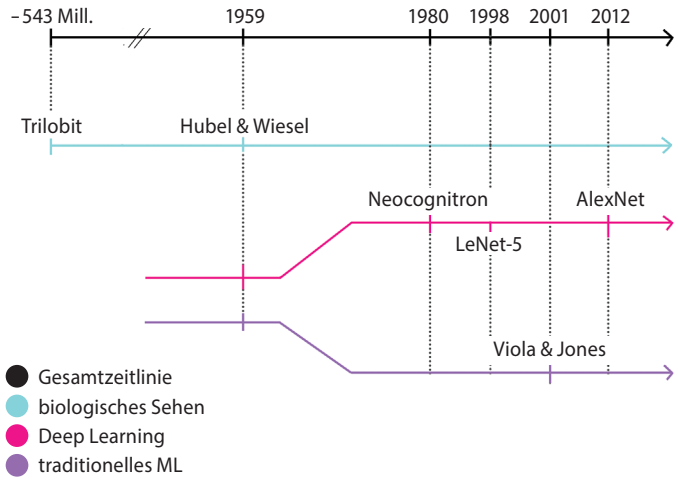


Abb. 1–8 *Verkürzte Zeitlinie für das biologische und maschinelle Sehen. Darin haben wir hervorgehoben, wann welche Ansätze für das Deep Learning sowie für das traditionelle Machine Learning aufkamen, auf die in diesem Abschnitt eingegangen wird.*

1.2.1 Das Neocognitron

Inspiriert durch Hubel und Wiesels Entdeckung der einfachen und komplexen Zellen, die die Hierarchie des primären visuellen Cortex bilden, schlug der japanische Elektroingenieur Kunihiko Fukushima Ende der 1970er-Jahre eine ana-

loge Architektur für das maschinelle Sehen vor, die er als *Neocognitron*[6] bezeichnete. Zwei Dinge sind besonders bemerkenswert:

1. Fukushima bezog sich in seinen Schriften explizit auf die Arbeit von Hubel und Wiesel. Im Speziellen verweist sein Artikel auf ihre entscheidenden Artikel zur Organisation des primären visuellen Cortex und nutzt ebenfalls die Terminologie der »einfachen« und »komplexen« Zellen, um die erste bzw. zweite Schicht seines Neocognitron zu beschreiben.
2. Wenn man künstliche Neuronen[7] auf diese hierarchische Weise anordnet, repräsentieren diese Neuronen – genau wie ihre biologische Inspiration aus Abbildung 1–6 – im Allgemeinen die Zeilenausrichtungen in den Zellen, die dem visuellen Rohbild am nächsten liegen, während die tiefer gelegenen Schichten zunehmend komplexer und abstrakter werdende Objekte darstellen. Um diese mächtige Eigenschaft des Neocognitron und seiner Deep-Learning-Abkömmlinge zu verdeutlichen, werden wir am Ende dieses Kapitels ein interaktives Beispiel zeigen, das sie demonstriert.[8]

1.2.2 LeNet-5

Abb. 1–9 *Der in Paris geborene Yann LeCun gehört zu den bedeutendsten Gestalten in der Forschung zu künstlichen neuronalen Netzen und Deep Learning. LeCun ist Gründungsdirektor des New Yorker »University Center for Data Science« sowie Leiter der KI-Forschung des sozialen Netzwerks Facebook.*

6. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.
7. Wir werden in Kapitel 7 genau definieren, was *künstliche Neuronen* sind. Im Moment reicht es, wenn Sie sich jedes künstliche Neuron als einen flinken kleinen Algorithmus vorstellen.
8. Insbesondere Abbildung 1–19 zeigt diese Hierarchie mit ihren zunehmend abstrakteren Darstellungen.

Abb. 1–10 *Yoshua Bengio ist eine weitere führende Person auf dem Gebiet der künstlichen neuronalen Netze und des Deep Learning. Geboren in Frankreich, arbeitet er jetzt als Informatikprofessor an der University of Montreal und gehört zu den Leitern des renommierten »Machines and Brains«-Programms am kanadischen »Institute for Advanced Research«.*

Während das Neocognitron zum Beispiel in der Lage war, handgeschriebene Zeichen zu identifizieren[9], stellte die Genauigkeit und Effizienz des *LeNet-5*-Modells[10] von Yann LeCun (Abbildung 1–9) und Yoshua Bengio (Abbildung 1–10) eine beeindruckende Weiterentwicklung dar. Die hierarchische Architektur von LeNet-5 (Abbildung 1–11) baute auf dem Modell von Fukushima und dessen biologischer Inspiration durch Hubel und Wiesel[11] auf. Darüber hinaus genossen LeCun und seine Kollegen den Vorteil besserer Daten zum Trainieren ihres Modells[12], einer schnelleren Verarbeitungsleistung und – was entscheidend war – des Backpropagation-Algorithmus.

Backpropagation (auch *Rückpropagierung* oder *Rückführung* genannt) ermöglicht ein effizientes Lernen durch die Schichten künstlicher Neuronen in einem Deep-Learning-Modell.[13] Die Daten der Forscher und die Verarbeitungsleistung sorgten dafür, dass LeNet-5 ausreichend zuverlässig für eine frühe kommerzielle Anwendung des Deep Learning wurde: Der *United States Postal Service* (USPS) nutzte es, um das Lesen der ZIP-Codes[14] auf Briefumschlägen zu automatisieren.

---

9. Fukushima, K. und Wake, N. (1991). Handwritten alphanumeric character recognition by the neocognitron. *IEEE Transactions on Neural Networks, 2*, 355–65.
10. LeCun, Y., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*, 2278–2324.
11. LeNet-5 war das erste *Convolutional Neural Network* (auf Deutsch in etwa: faltendes neuronales Netz), eine Deep-Learning-Variante, die im modernen maschinellen Sehen dominiert und die wir in Kapitel 10 genauer betrachten werden.
12. Ihr klassischer Datensatz, die handgeschriebenen MNIST-Ziffern, kommt umfassend in Teil II, »Die nötige Theorie«, zum Einsatz.
13. Wir untersuchen den Backpropagation-Algorithmus in Kapitel 7.

In Kapitel 10, wenn es um das maschinelle Sehen geht, werden Sie LeNet-5 aus erster Hand erleben, wenn Sie es selbst entwerfen und auf die Erkennung handgeschriebener Ziffern trainieren.

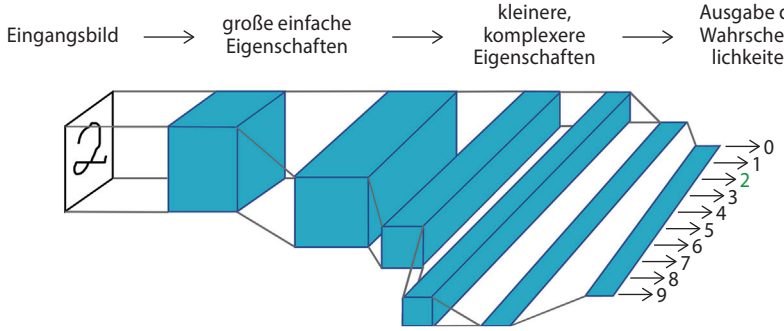


Abb. 1–11 *LeNet-5 behielt die hierarchische Architektur bei, die von Hubel und Wiesel im primären visuellen Cortex entdeckt und von Fukushima in seinem Neocognitron benutzt wurde. Wie in diesen anderen Systemen auch repräsentiert die ganz links gelegene Schicht einfache Kanten, während nachfolgende Schichten zunehmend komplexer werdende Eigenschaften darstellen. Durch eine solche Verarbeitung von Informationen sollte zum Beispiel eine handgeschriebene »2« korrekt als Zahl Zwei erkannt werden (in der Ausgabe ganz rechts grün hervorgehoben).*

In LeNet-5 hatten Yann LeCun und seine Kollegen einen Algorithmus, der korrekt die handgeschriebenen Ziffern vorhersagen konnte, ohne dass besondere Expertise über handgeschriebene Ziffern in dessen Code vorhanden sein musste. Entsprechend bietet LeNet-5 eine Gelegenheit, einen grundlegenden Unterschied zwischen Deep Learning und dem traditionellen Machine Learning vorzustellen. Wie in Abbildung 1–12 verdeutlicht wird, zeichnet sich der traditionelle Machine-Learning-Ansatz dadurch aus, dass seine Anwender den größten Teil ihrer Bemühungen in das Entwickeln von sogenannten *Features* (Eigenschaften, Merkmale) stecken. Dieses *Feature Engineering* ist die Anwendung ausgeklügelter und oft sehr aufwendiger Algorithmen auf Rohdaten, um eine Vorverarbeitung dieser Daten zu Eingabevariablen vorzunehmen, die dann leicht durch herkömmliche statistische Techniken modelliert werden können. Diese Techniken – wie *Regression*, *Random Forest* und *Support Vector Machine* – lassen sich auf nicht verarbeitete Daten nur selten effektiv anwenden, sodass die Entwicklung der Eingabedaten in der Vergangenheit der Hauptfokus der Machine-Learning-Forscher war.

Im Allgemeinen wenden Nutzer des traditionellen Machine Learning nur wenig Zeit für das Optimieren von ML-Modellen auf oder darauf, das effektivste Modell aus dem vorhandenen Angebot auszuwählen. Der Deep-Learning-Ansatz stellt diese Prioritäten auf den Kopf. *Ein Deep-Learning-Anwender verbringt*

14. Die Bezeichnung des USPS für Postleitzahlen.

üblicherweise kaum oder keine Zeit mit dem Entwickeln von Features, sondern verbringt sie damit, Daten mit verschiedenen Architekturen künstlicher neuronaler Netze zu modellieren, die die rohen Eingabedaten automatisch zu sinnvollen Features verarbeiten. Dieser Unterschied zwischen Deep Learning und dem traditionellen Machine Learning (TML) ist das entscheidende Thema dieses Buches. Im nächsten Abschnitt finden Sie ein klassisches Beispiel für das Feature Engineering, das diesen Unterschied genauer erläutern soll.

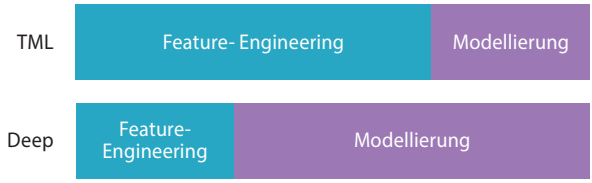


Abb. 1–12 *Feature Engineering – die Umwandlung von Rohdaten in sinnvoll gestaltete Inputvariablen – beherrscht oft den Einsatz traditioneller Machine-Learning-Algorithmen (TML). Im Gegensatz dazu kommt es bei der Anwendung von Deep Learning kaum zu Feature Engineering, sondern der größte Teil der Zeit wird mit dem Entwurf und der Anpassung der Modellarchitekturen zugebracht.*

1.2.3 Der traditionelle Machine-Learning-Ansatz

Im Anschluss an LeNet-5 kam die Forschung zu künstlichen neuronalen Netzen sowie zu Deep Learning gewissermaßen aus der Mode. Der Konsens lautete, dass die automatisierte Feature-Generierung dieser Methode nicht pragmatisch war – dass das feature-freie Vorgehen zwar ganz gut bei der Handschriftenerkennung funktionierte, aber ansonsten nur eingeschränkt einsetzbar sei.[15] Das traditionelle Machine Learning inklusive seines Feature Engineering schien vielversprechender zu sein und die Deep-Learning-Forschung verlor eine Menge Fördergelder.[16]

Um zu verdeutlichen, was Feature Engineering ist, sehen Sie in Abbildung 1–13 ein berühmtes Beispiel von Paul Viola und Michael Jones aus den frühen 2000er-Jahren.[17] Viola und Jones verwendeten rechteckige Filter, wie die senkrechten oder waagerechten schwarzweißen Balken, die in der Abbildung gezeigt werden.

---

15. Damals sah sich die Optimierung der Deep-Learning-Modelle vor Hindernisse gestellt, die inzwischen ausgeräumt wurden, darunter die schlechte Initialisierung der Gewichte (behandelt in Kapitel 9), Kovarianz-Verschiebung (ebenfalls Kapitel 9) und die Vorherrschaft der relativ ineffizienten Sigmoid-Aktivierungsfunktion (Kapitel 6).
16. Die Förderung durch die öffentliche Hand für die Forschung an künstlichen neuronalen Netzen nahm global ab. Eine Ausnahme bildete die anhaltende Unterstützung durch die kanadische Regierung, die es den Universitäten Montreal, Toronto und Alberta erlaubte, auf dem Gebiet zu führenden Kräften zu werden.
17. Viola, P. und Jones, M. (2001). Robust real-time face detection. *International Journal of Computer Vision, 57*, 137–54.

Features, die generiert werden, wenn man diese Filter über ein Bild führt, können in Machine-Learning-Algorithmen eingegeben werden, um zuverlässig zu erkennen, ob ein Gesicht vorhanden ist. Diese Arbeit ist deshalb so bemerkenswert, weil der Algorithmus effizient genug war, um die erste Echtzeit-Gesichtserkennung zu liefern, die nicht auf dem Gebiet der Biologie basierte.[18]

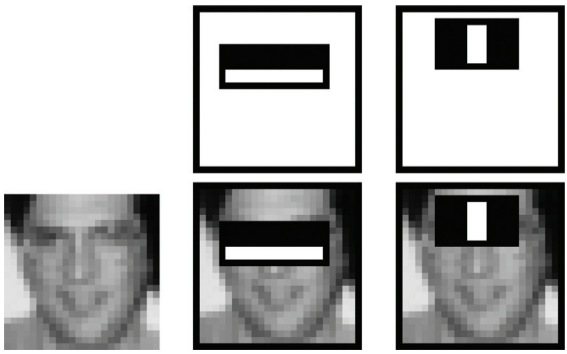


Abb. 1–13 *Beispiele für Feature Engineering durch Viola und Jones (2001) zum zuverlässigen Erkennen von Gesichtern. Ihr effizienter Algorithmus fand seinen Weg in Fujifilm-Kameras, die dann zum ersten Mal Echtzeit-Autofokus boten.*

Das Konstruieren der cleveren Gesichtserkennungsfilter zum Verarbeiten der Rohpixel zu Features, die als Input für das Machine-Learning-Modell dienen konnten, war das Ergebnis jahrelanger Forschung und Zusammenarbeit zum Thema »Eigenschaften von Gesichtern« Und natürlich beschränkt sich das Ganze auf das Erkennen von Gesichtern im Allgemeinen – es werden also keine speziellen Gesichter erkannt, wie zum Beispiel das von Angela Merkel oder Oprah Winfrey. Um Features zu entwickeln, die etwa Oprahs Gesicht oder andere Klassen von Objekten, die keine Gesichter sind, erkennen könnten (zum Beispiel Häuser, Autos oder Yorkshire-Terrier), müsste Kompetenz in diesen Kategorien aufgebaut werden, was wieder eine jahrelange Zusammenarbeit der wissenschaftlichen Gemeinschaft erfordern würde, um es effizient und akkurat zu schaffen. Hm, wenn es doch nur möglich wäre, sich die nötige Zeit und Mühe einfach irgendwie zu sparen!

1.2.4 ImageNet und die ILSVRC

Wie bereits erwähnt, bestand einer der Vorteile von LeNet-5 gegenüber dem Neocognitron in seinem größeren und höherwertigen Satz an Trainingsdaten. Der nächste Durchbruch bei neuronalen Netzen wurde ebenfalls durch einen qualita-

18. Einige Jahre später fand der Algorithmus seinen Weg in die digitalen Fujifilm-Kameras und erlaubte zum ersten Mal Autofokus auf Gesichter – eine mittlerweile alltägliche Funktion in Digitalkameras und Smartphones.

tiv hochwertigen öffentlichen Datensatz ermöglicht, der dieses Mal viel größer war. ImageNet, eine kategorisierte Bilderdatenbank, die von Fei-Fei Li (Abbildung 1–14) begründet wurde, gibt Machine-Learning-Forschern einen riesigen Katalog mit Trainingsdaten an die Hand.[19, 20] Zum Vergleich: Die handgeschriebenen Zifferndaten, die zum Trainieren von LeNet-5 benutzt wurden, enthielten Zehntausende von Bildern. ImageNet besteht aus mehr als 14 *Millionen* Fotos.

Die 14 Millionen Bilder in der ImageNet-Datenbank sind in über 22.000 Kategorien eingeteilt. Diese Kategorien enthalten so unterschiedliche Dinge wie Containerschiffe, Leoparden, Seesterne und Holunderbeeren. Seit 2010 veranstaltet Li jährlich einen offenen Wettstreit namens ILSVRC (*ImageNet Large Scale Visual Recognition Challenge*) auf einer Teilmenge der ImageNet-Daten, der mittlerweile das wichtigste Terrain zum Beurteilen der modernsten Algorithmen zum maschinellen Sehen darstellt. Die ILSVRC-Teilmenge besteht aus 1,4 Millionen Bildern aus 1.000 Kategorien. Dabei wird nicht nur ein breites Spektrum an Kategorien geboten; viele der ausgewählten Kategorien sind darüber hinaus Hunderassen, sodass nicht nur die Fähigkeit der Algorithmen abgeschätzt wird, deutlich verschiedene Bilder zu unterscheiden, sondern auch solche zu erkennen, die sich nur leicht voneinander abheben.[21]

Abb. 1–14 *Der gigantische ImageNet-Datensatz ist das geistige Kind der chinesisch-amerikanischen Informatikprofessorin Fei-Fei Li und ihrer Kollegen in Princeton und entstand 2009. Li, die mittlerweile an der Stanford University arbeitet, ist außerdem Chefwissenschaftlerin für KI/ML bei Googles Cloud-Plattform.*

19. *image-net.org*
20. Deng, J., et al. (2009), ImageNet: A large-scale hierarchical image database. *Proceedings of the Conference on Computer Vision and Pattern Recognition*.
21. Versuchen Sie einmal, Fotos von Yorkshire-Terriern von solchen mit Australian-Silky-Terriern zu unterscheiden. Es ist schwierig, aber Juroren der Westminster Dog Show können es, genau wie moderne Machine-Vision-Modelle. Im Übrigen sind diese hundelastigen Datensätze der Grund dafür, dass Deep-Learning-Modelle, die mit ImageNet trainiert wurden, dazu neigen, von Hunden zu »träumen« (siehe z.B. *deepdreamgenerator.com*).

1.2.5 AlexNet

Wie Sie in Abbildung 1–16 sehen, stammten in den ersten beiden Jahren des ILSVRC alle Algorithmen, die in den Wettbewerb eingereicht wurden, aus dem traditionellen Machine Learning, setzten also auf Feature Engineering. Im dritten Jahr waren alle Teilnehmer *mit Ausnahme eines einzigen* herkömmliche ML-Algorithmen. Wenn dieses Deep-Learning-Modell im Jahre 2012 nicht entwickelt worden wäre oder wenn seine Schöpfer nicht am ILSVRC teilgenommen hätten, wäre die Exaktheit der von Jahr zu Jahr zu beobachtenden Bildklassifizierung vernachlässigbar gewesen. Stattdessen zerschmetterten Alex Krizhevsky und Ilya Sutskever – beide von der University of Toronto, wo sie unter Leitung von Geoffrey Hinton (Abbildung 1–15) arbeiteten – mit ihrem Beitrag, der heute als AlexNet (Abbildung 1–17) bekannt ist, die vorhandenen Benchmarks.[22, 23]

Abb. 1–15 *Der überragende, britisch-kanadische Pionier auf dem Gebiet der künstlichen neuronalen Netze, Geoffrey Hinton, wird in der Presse oft als »Pate des Deep Learning« bezeichnet. Hinton ist emeritierter Professor an der University of Toronto und arbeitet außerdem für Google, wo er das »Brain Team« des Suchmaschinenriesen, eine Forschungsabteilung in Toronto, leitet. 2019 wurden Hinton, Yann LeCun (Abbildung 1–9) und Yoshua Bengio (Abbildung 1–10) gemeinsam für ihre Arbeit auf dem Gebiet des Deep Learning mit dem Turing Award geehrt – der höchsten Auszeichnung in der Informatik.*

Dies war ein Wendepunkt. Deep-Learning-Architekturen traten aus dem Schatten des Machine Learning heraus. Akademische und kommerzielle Anwender bemühten sich hastig, die Grundlagen der neuronalen Netze zu verstehen und Softwarebibliotheken herzustellen – von denen viele Open Source sind –, um mit den Deep-Learning-Modellen auf ihren eigenen Daten und Anwendungsfällen zu experimen-

22. Krizhevsky, A., Sutskever, I. und Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*.
23. Die Bilder in Abbildung 1–17 unten stammen aus Yosinski, J., et al. (2015). Understanding neural networks through deep visualization. *arXiv: 1506.06579*.

tieren, egal ob diese maschinelles Sehen oder anderes betrafen. In Abbildung 1–16 ist zu erkennen, dass seit 2012 alle Modelle, die im ILSVRC an der Spitze stehen, auf Deep Learning basieren.

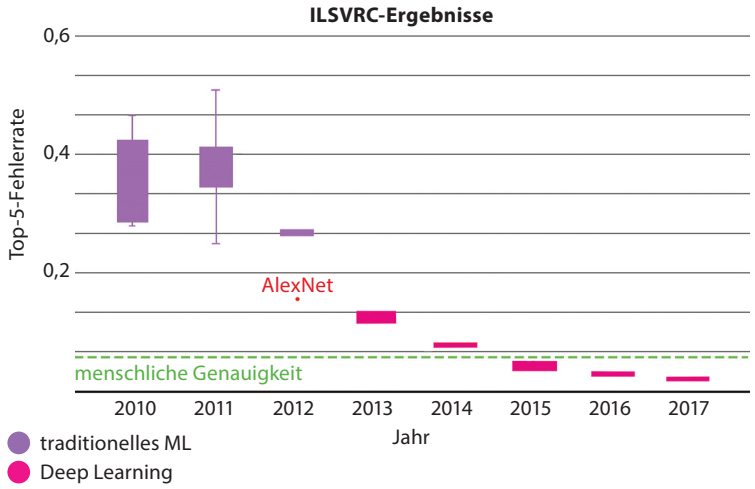


Abb. 1–16 *Leistung der besten Teilnehmer am ILSVRC in den einzelnen Jahren. AlexNet war 2012 der um Längen (um 40 %!) bessere Gewinner. Seitdem waren die besten Algorithmen immer Deep-Learning-Modelle. 2015 übertrafen die Maschinen dann die menschliche Genauigkeit.*

Auch wenn die hierarchische Architektur von AlexNet an LeNet-5 erinnert, gibt es drei wesentliche Faktoren, die dafür sorgten, dass AlexNet im Jahre 2012 der führende Algorithmus für das maschinelle Sehen wurde. Der erste Faktor waren die Trainingsdaten. Krizhevsky und seine Kollegen hatten nicht nur Zugriff auf die riesige ImageNet-Datenbank, sondern erweiterten die verfügbaren Daten auch noch künstlich, indem sie Transformationen auf die Trainingsbilder anwandten (Sie werden dies in Kapitel 10 ebenfalls tun). Der zweite Faktor ist die Verarbeitungsleistung. Zum einen war die Rechenleistung pro Kosteneinheit zwischen 1998 und 2012 drastisch angestiegen, zum anderen programmierten Krizhevsky, Hinton und Sutskever zwei GPUs[24], um ihre großen Datensätze mit bisher nie gesehener Effizienz zu trainieren. Der dritte Faktor waren die Fortschritte in der Architektur. AlexNet ist tiefer (besitzt mehr Schichten) als LeNet-5 und nutzt sowohl einen neuen Typ künstlicher Neuronen[25] als auch einen raffinierten Trick[26], der dabei hilft, Deep-Learning-Modelle über die Daten hinaus zu verallgemeinern, mit denen sie trainiert wurden. Genau wie LeNet-5 werden Sie AlexNet in Kapitel 10 selbst bauen und es nutzen, um Bilder zu klassifizieren.

24. Graphical Processing Units: Diese sind vor allem für das Darstellen von Videospielen gedacht, eignen sich aber genauso gut für das Durchführen von Matrixmultiplikationen, die es in Deep-Learning-Systemen in Form Hunderter paralleler Rechenoperationen zuhauf gibt.
25. Die *Rectified Linear Unit* (ReLU), die in Kapitel 6 eingeführt wird.
26. *Dropout*, vorgestellt in Kapitel 9.

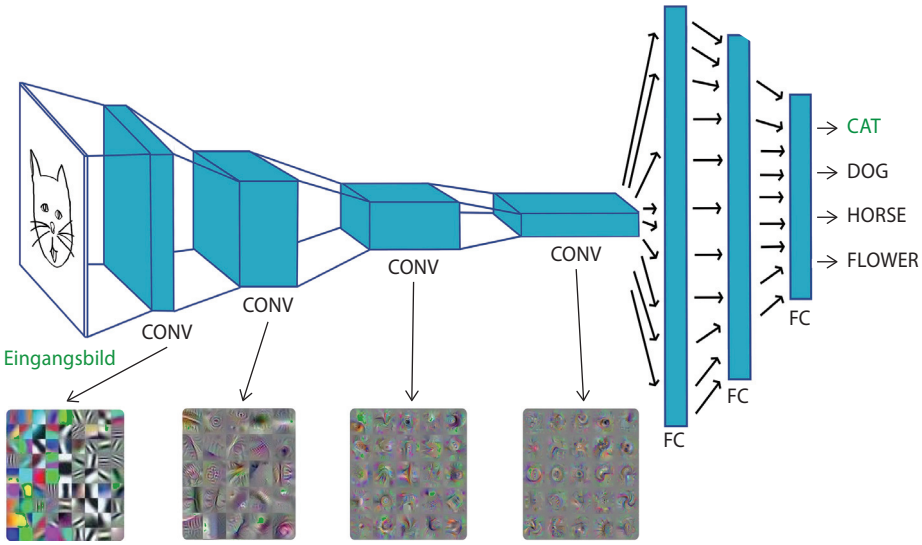


Abb. 1–17 *Die hierarchische Natur von AlexNet erinnert an LeNet-5. Die erste Schicht (ganz links) repräsentiert einfache visuelle Merkmale wie Kanten, während tiefer gelegene Schichten zunehmend komplexer werdende Merkmale und abstrakte Konzepte darstellen. Am unteren Rand sehen Sie Beispiele für Bilder, auf die die Neuronen in dieser Schicht eine maximale Reaktion zeigen. Dies erinnert an die Schichten des biologischen visuellen Systems aus Abbildung 1–6 und demonstriert die hierarchische Zunahme der visuellen Komplexität. Im hier gezeigten Beispiel wird das Bild einer Katze, das AlexNet präsentiert wurde, korrekt als solche erkannt (wie der grüne Ausgabetext »CAT« impliziert). »CONV« deutet an, dass ein sogenannter Convolutional Layer verwendet wird, »FC« ist eine vollständig verknüpfte Schicht. Wir werden diese Schichttypen in Kapitel 7 bzw. Kapitel 10 einführen.*

Unsere ILSVRC-Fallstudie unterstreicht, wieso Deep-Learning-Modelle wie AlexNet so ungemein nützlich und bahnbrechend in allen Branchen und Computeranwendungen sind: Sie reduzieren ganz drastisch das themenbezogene Fachwissen, das zum Herstellen hochgradig exakter Vorhersagemodelle erforderlich ist. Dieser Trend weg vom fachlich gestützten Feature Engineering und hin zu überraschend leistungsstarken Deep-Learning-Modellen mit automatischer Feature-Generierung wird nicht nur von Vision-Anwendungen getragen, sondern zum Beispiel auch von Computerspielen (sie sind Thema von Kapitel 4) und von der Verarbeitung natürlicher Sprache (Kapitel 2)[27]. Man muss kein Spezialist für die visuellen Attribute von Gesichtern sein, um einen Gesichtserkennungsalgorithmus herzustellen. Man benötigt kein umfassendes Verständnis mehr für die Strategie eines Spiels, um ein Programm zu schreiben, das es meistern kann. Man

27. Einen besonders unterhaltsamen Bericht über den Durchbruch im Feld des maschinellen Übersetzens lieferte Gideon Lewis-Kraus in seinem Artikel »The Great A. I. Awakening«, erschienen im *New York Times Magazine* am 14. Dezember 2016.

muss keine Autorität für die Struktur und Semantik aller betreffenden Sprachen sein, um ein Übersetzungswerkzeug zu schreiben. Für immer mehr Anwendungsfälle ist es wichtiger, Deep-Learning-Techniken anzuwenden, als Kenntnisse auf dem entsprechenden Gebiet zu haben. Während diese Kenntnisse früher wenigstens einen Doktorgrad oder vielleicht jahrelange Forschungen in diesem Bereich erfordert haben, kann ein hinreichendes Niveau auf dem Feld des Deep Learning relativ einfach erreicht werden – etwa, indem man dieses Buch durcharbeitet!

1.3 TensorFlow Playground

Wenn Sie auf nette und interaktive Weise das hierarchische Wesen des Deep Learning erkunden wollen, bei dem selbsttätig Features gefunden werden, sollten Sie einmal den *TensorFlow Playground* unter *bit.ly/TFplayground* besuchen. Das Netzwerk, das Sie hinter diesem eigens eingerichteten Link finden, sollte automatisch so ähnlich aussehen wie in Abbildung 1–18. In Teil II werden wir wieder dorthin zurückkehren und alle Begriffe definieren, die auf dem Bildschirm zu sehen sind; für die aktuelle Übung können Sie sie getrost ignorieren. Momentan reicht es zu wissen, dass dies ein Deep-Learning-Modell ist. Die Modellarchitektur besteht aus sechs Schichten künstlicher Neuronen: einer Eingabeschicht ganz links (unter der Überschrift »FEATURES«), vier »HIDDEN LAYERS« (verborgene Schichten; diese sind verantwortlich für das Lernen) und einer »OUTPUT«-Schicht (Ausgabeschicht; das Raster ganz rechts, das an beiden Achsen von -6 bis +6 verläuft). Ziel des Netzwerks ist es zu lernen, orange Punkte (negative Fälle) ausschließlich durch ihre Lage im Raster von blauen Punkten (positive Fälle) zu unterscheiden. Daher geben wir in der Eingabeschicht nur zwei Informationen über jeden Punkt ein: seine horizontale Position (X_1) und seine vertikale Position (X_2). Die als Trainingsdaten verwendeten Punkte sind standardmäßig im Raster zu sehen. Wenn Sie die Checkbox *Show test data* anklicken, können Sie außerdem die Lage der Punkte sehen, die benutzt werden, um die Leistung des Netzes beim Lernen abzuschätzen. Entscheidend ist, dass diese Testdaten dem Netz nicht zur Verfügung stehen, während es lernt, sodass wir sichergehen können, dass das Netz gut auf neue, ungesehene Daten verallgemeinert.

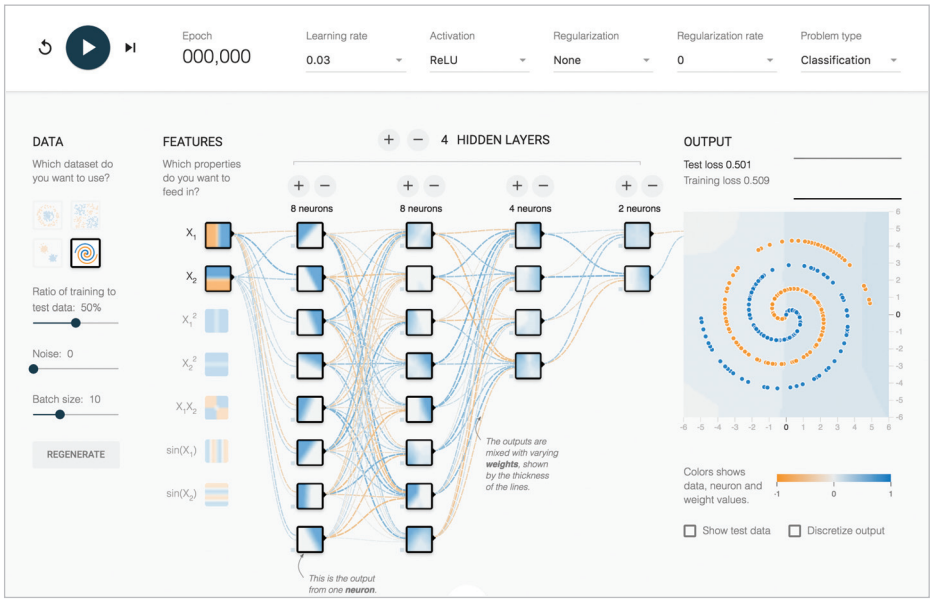


Abb. 1–18 *Dieses tiefe neuronale Netz ist bereit zu lernen, wie es eine Spirale aus orange Punkten (negative Fälle) von blauen Punkten (positive Fälle) unterscheiden kann. Als Grundlage hierzu dienen lediglich die Positionen der Punkte auf den Achsen X_1 und X_2 im Raster auf der rechten Seite.*

Klicken Sie auf den großen *Play*-Pfeil in der oberen linken Ecke. Erlauben Sie es dem Netzwerk zu trainieren, bis die Werte »Training loss« und »Test loss« in der oberen rechten Ecke jeweils fast auf Null zurückgegangen sind – sagen wir, auf weniger als 0.05. Wie lange das dauert, hängt von Ihrer Hardware ab, aber es werden hoffentlich nur ein paar Minuten sein.

Wie in Abbildung 1–19 festgehalten wurde, sollten Sie nun die künstlichen Neuronen des Netzwerks sehen, die die Eingabedaten repräsentieren. Je tiefer (also je weiter rechts) sie positioniert sind, umso komplexer und abstrakter sollte die Darstellung werden – wie im Neocognitron, im LeNet-5 (Abbildung 1–11) und im AlexNet (Abbildung 1–17). Bei jedem Durchlauf des Netzes löst das Netz das Problem der Spiralklassifikation auf Neuronen-Ebene anders, aber der allgemeine Ansatz bleibt gleich (wenn Sie dies nachprüfen wollen, laden Sie die Seite neu und trainieren Sie das Netz erneut). Die künstlichen Neuronen in der ganz linken verborgenen Schicht sind darauf spezialisiert, Kanten (gerade Linien) zu unterscheiden, und zwar jeweils in einer bestimmten Ausrichtung. Die Neuronen aus der ersten verborgenen Schicht übergeben die Informationen an Neuronen in der zweiten verborgenen Schicht, die wiederum die Kanten zu etwas komplexeren Merkmalen, wie etwa Kurven, neu kombinieren. Die Neuronen in jeder nachfolgenden Schicht setzen die Informationen der Neuronen aus den vorangegangenen Schichten neu zusammen, sodass sich die Komplexität und der Abstraktionsgrad

der Features, die die Neuronen repräsentieren können, schrittweise erhöhen. Wenn die letzte (ganz rechts gelegene) Schicht erreicht ist, sind die Neuronen geschult darin, die Feinheiten der Spiralform darzustellen, was es dem Netz erlaubt, anhand seiner Position im Raster (seiner X_1- und X_2-Koordinaten) akkurat vorherzusagen, ob ein Punkt orange (ein negativer Fall) oder blau (ein positiver Fall) ist. Halten Sie den Mauszeiger über ein Neuron, um es auf das äußerst rechts gelegene »OUTPUT«-Raster zu projizieren und seine jeweilige Spezialisierung im Detail zu untersuchen.

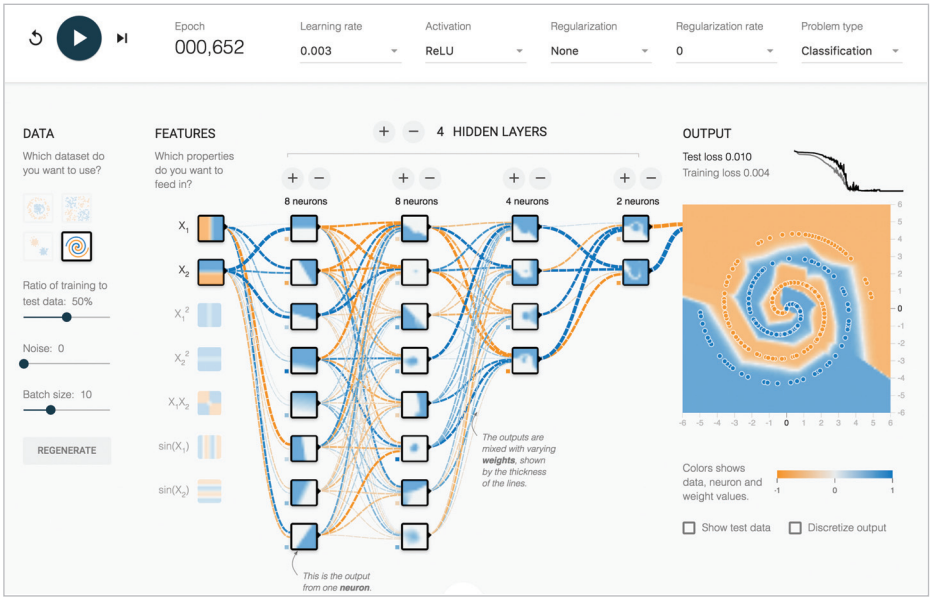


Abb. 1–19 *Das Netzwerk nach dem Trainieren*

1.4 Quick, Draw!

Um interaktiv zu erleben, wie ein Deep-Learning-Netzwerk in Echtzeit eine Aufgabe im maschinellen Sehen (Machine Vision) ausführt, spielen Sie einmal das *Quick, Draw!*-Spiel (oder auf Deutsch *Flugs gezeichnet!*) unter *quickdraw.withgoogle.com*. Klicken Sie auf *Und los!*, um das Spiel zu beginnen. Sie werden aufgefordert, ein Objekt zu zeichnen, und der Deep-Learning-Algorithmus versucht zu erraten, was es ist. Am Ende von Kapitel 10 werden wir die ganze Theorie und die praktischen Codebeispiele behandelt haben, die erforderlich sind, um einen Algorithmus zum maschinellen Sehen zu entwerfen, der diesem ganz ähnlich ist. Die Zeichnungen, die Sie herstellen, werden außerdem zu dem Datensatz hinzugefügt, den Sie in Kapitel 12 verwenden, um ein Deep-Learning-Modell herzustellen, das überzeugend Kritzeleien imitieren kann, die von Menschen gemacht werden. Schnallen Sie sich an! Wir gehen auf eine fantastische Reise.