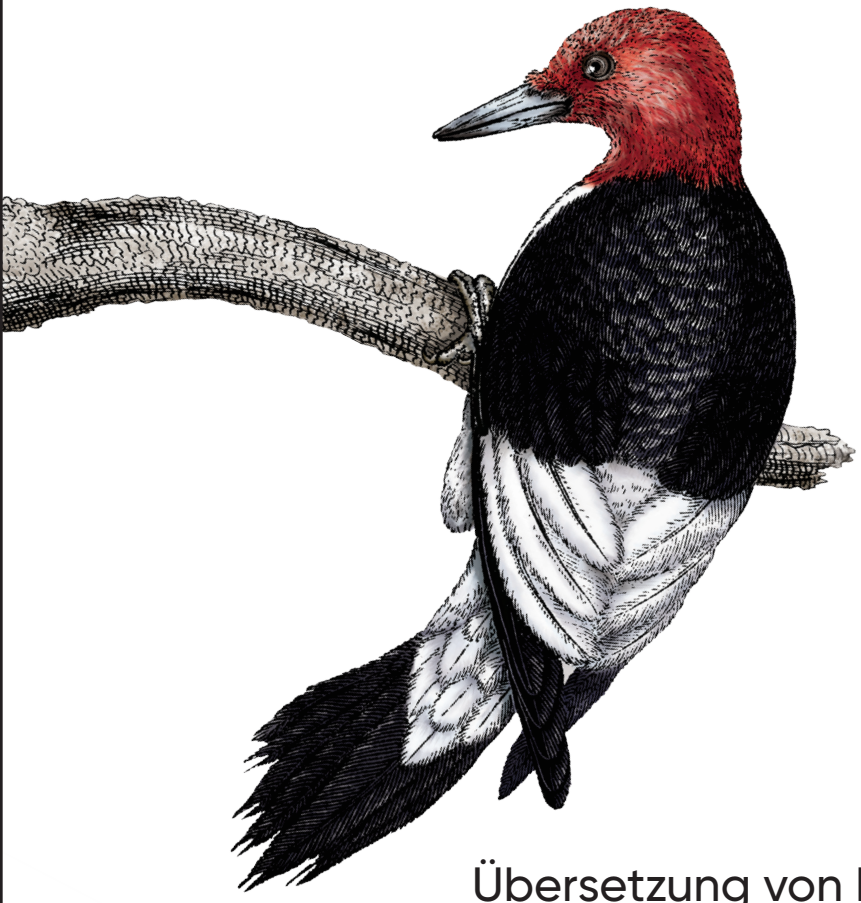
O'REILLY®

Deutsche Ausgabe

PyTorch für Deep Learning

Anwendungen für Bild-, Ton- und Textdaten entwickeln und deployen



Ian Pointer

Übersetzung von Marcus Fraaß

PyTorch für Deep Learning

Anwendungen für Bild-, Ton- und Textdaten entwickeln und deployen

Ian Pointer

Deutsche Übersetzung von Marcus Fraaß

O'REILLY®

Ian Pointer

Lektorat: Alexandra Follenius
Übersetzung: Marcus Fraaß
Korrektorat: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner und Frank Heidt
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.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 978-3-96009-134-9
PDF 978-3-96010-399-8
ePub 978-3-96010-400-1
mobi 978-3-96010-401-8

1. Auflage 2021

Wieblinger Weg 17
69123 Heidelberg

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«. O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

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: kommentar@oreilly.de.

5 4 3 2 1 0

Inhalt

Vorwort

Deep Learning in der heutigen Zeit

Hallo und herzlich willkommen! Dieses Buch wird Sie in das Thema Deep Learning mit PyTorch, einer Open-Source-Bibliothek, die im Jahr 2017 von Facebook öffentlich bereitgestellt wurde, einführen. Wenn Sie in den letzten Jahren nicht wie ein Vogel Strauß mit dem Kopf im Sand gesteckt haben, werden Sie nicht umhingekommen sein, zu bemerken, dass neuronale Netzwerke heutzutage überall zugegen sind. Sie sind von dem »wirklich coolen Teil der Informatik, über den die Menschen lernen und danach nichts damit anfangen können« dazu gereift, jeden Tag in unseren Smartphones mit uns herumgetragen zu werden, um unsere Bilder zu verbessern oder unseren Sprachbefehlen zuzuhören. Unsere E-Mail-Programme lesen unsere E-Mails und produzieren kontextbezogene Antworten, unsere Lautsprecher hören uns zu, Autos fahren von selbst, und der Computer hat die besten Spieler bei Go inzwischen besiegt. Wir sehen auch, dass die Technologie in autoritären Ländern für ruchlosere Zwecke eingesetzt wird, etwa wenn Wächter mithilfe von neuronalen Netzwerken Gesichter aus der Menge herauspicken und darüber entscheiden, ob die Menschen verhaftet werden sollen.

Auch wenn es einem so vorkommen mag, dass dies alles sehr rasch voranging, so reichen die zugrunde liegenden Konzepte der neuronalen Netze und des Deep Learning dennoch weit zurück. Der Beweis, dass ein solches Netz in der Lage ist, *jede* mathematische Funktion in einer approximativen Weise zu ersetzen – was die Idee untermauert, dass neuronale Netze für viele verschiedene Aufgaben trainiert werden können –, geht zurück ins Jahr 1989.[1] Neuronale Konvolutionsnetze wurden bereits in den späten 1990er-Jahren zur Erkennung von Ziffern auf Schecks verwendet. In dieser Zeit hat sich ein solides Fundament aufgebaut. Warum also fühlt es sich so an, als hätte es in den letzten zehn Jahren eine explosionsartige Entwicklung in diesem Bereich gegeben?

1 Siehe »Approximation by Superpositions of Sigmoidal Functions« (*https://oreil.ly/BQ8-9*) von George Cybenko (1989).

Hierfür gibt es zahlreiche Gründe, wobei der wichtigste wohl der Anstieg der Leistungsfähigkeit von *Grafikprozessoren* (GPUs) und deren zunehmende Erschwinglichkeit ist. Ursprünglich für Spiele entwickelt, sind GPUs darauf ausgelegt, unzählige Millionen von Matrixoperationen pro Sekunde auszuführen, um alle Polygone für das Fahr- oder Ballerspiel, das Sie auf Ihrer Konsole oder Ihrem PC spielen, zu rendern – Operationen, für die ein gewöhnlicher Hauptprozessor (CPU) einfach nicht optimiert ist. Der im Jahr 2009 erschienene Forschungsbeitrag »Large-Scale Deep Unsupervised Learning Using Graphics Processors« von Rajat Raina et al.[2] wies darauf hin, dass das Training neuronaler Netze ebenfalls auf der Ausführung zahlreicher Matrixoperationen basiert, sodass zusätzliche Grafikkarten das Training beschleunigen und dadurch auch erstmals noch größere, *tiefere* neuronale Netzwerkarchitekturen eingesetzt werden können.

Andere wichtige Methoden wie *Dropout* (die wir in Kapitel 3 betrachten werden) wurden ebenfalls in den letzten zehn Jahren eingeführt, um das Training nicht nur zu beschleunigen, sondern auch besser zu *verallgemeinern* (damit das Netzwerk nicht einfach nur die Trainingsdaten auswendig lernt – ein Problem namens *Überanpassung*, auf das wir im nächsten Kapitel stoßen werden). In den letzten Jahren haben Unternehmen diesen GPU-basierten Ansatz auf die nächste Ebene gehoben. Dabei hat Google etwas entwickelt, das es als *Tensorprozessoren* (TPUs) bezeichnet. Es handelt sich dabei um anwendungsspezifische Chips, die speziell dafür entwickelt wurden, so schnell wie möglich Deep-Learning-Anwendungen durchzuführen. Darüber hinaus stehen sie sogar der allgemeinen Öffentlichkeit als Teil des Google-Cloud-Ökosystems zur Verfügung.

Eine weitere Möglichkeit, den Fortschritt des Deep Learning in den letzten zehn Jahren zu skizzieren, bietet der ImageNet-Wettbewerb. ImageNet ist eine riesige Datenbank mit über 14 Millionen Bildern, die zu Zwecken des maschinellen Lernens für 20.000 Kategorien manuell gelabelt sind. Seit dem Jahr 2010 wird jährlich im Rahmen des Wettbewerbs *ImageNet Large Scale Visual Recognition Challenge* versucht, die Anwendungen aller Teilnehmer anhand eines Auszugs aus der Datenbank mit 1.000 Kategorien zu testen. Bis 2012 lagen die Fehlerquoten für die Bewältigung der Aufgaben bei etwa 25%. In jenem Jahr gewann jedoch ein tiefes neuronales Konvolutionsnetz den Wettbewerb mit einer Fehlerquote von 16%, womit es alle anderen Teilnehmer deutlich übertraf. In den folgenden Jahren wurde diese Fehlerquote immer weiter nach unten gedrückt, bis die ResNet-Architektur im Jahr 2015 ein Ergebnis von 3,6% erreichte, das unterhalb der durchschnittlichen menschlichen Leistung (5%) lag. Wir wurden also überholt.

2 *http://www.robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf*

Aber was genau ist Deep Learning, und brauche ich einen Doktortitel, um es verstehen zu können?

Die Definitionen von Deep Learning sind oft eher verwirrend als erhellend. Eine Definitionsmöglichkeit besteht darin, Deep Learning als eine Methode des maschinellen Lernens zu bezeichnen, die mehrere und zahlreiche Schichten nicht linearer Transformationen verwendet, um nach und nach Merkmale aus der ursprünglichen Eingabe zu extrahieren. Das ist zwar richtig, aber es hilft nicht wirklich weiter, oder? Ich ziehe es vor, es als eine Methode zu beschreiben, die Aufgaben löst, bei denen Sie dem Computer die Eingaben und die gewünschten Ausgaben liefern und er die Lösung findet, im Regelfall unter Verwendung eines neuronalen Netzes.

Auf viele Menschen abschreckend wirkt im Zusammenhang mit Deep Learning die dahinterstehende Mathematik. Wenn Sie sich irgendeinem Forschungspapier in diesem Bereich widmen, werden Sie fast überall auf undurchdringliche Mengen an Notationen in Form von griechischen Buchstaben stoßen und wahrscheinlich schreiend im Dreieck springen. Die Sache ist die: Man muss in den meisten Fällen kein Mathegenie sein, um die Methoden des Deep Learning anwenden zu können. Tatsächlich muss man für die meisten alltäglichen, grundlegenden Anwendungen der Technologie überhaupt nicht viel wissen. Um wirklich zu verstehen, was vor sich geht (wie Sie in Kapitel 2 sehen werden), muss man sich nur ein wenig anstrengen, um die Konzepte nachvollziehen zu können, die man wahrscheinlich bereits in der Oberstufe gelernt hat. Lassen Sie sich also nicht von der Mathematik verunsichern. Am Ende von Kapitel 3 werden Sie in der Lage sein, einen Bildklassifikator zusammenzustellen, der mit nur wenigen Zeilen Code mit dem konkurriert, was die besten Köpfe im Jahr 2015 anbieten konnten.

PyTorch

Wie bereits zu Beginn erwähnt, ist PyTorch eine Open-Source-Bibliothek von Facebook, die das Programmieren von Deep-Learning-Code in Python erleichtert. Der Ursprung der Bibliothek fußt wiederum auf zwei weiteren Bibliotheken. Sie leitet, und angesichts ihres Namens vielleicht nicht ganz überraschend, viele Funktionen und Konzepte von *Torch* ab, einer Lua-basierten Bibliothek für neuronale Netzwerke, die auf das Jahr 2002 zurückgeht. Der zweite wichtige Vorläufer ist die Bibliothek *Chainer*, die im Jahr 2015 in Japan entwickelt wurde. Chainer war eine der ersten Bibliotheken für neuronale Netzwerke, die einen eifrigen bzw. interaktiven (engl. Eager Execution) Ansatz zur Differenzierung anstelle der Definition statischer Graphen bot. Das ermöglichte eine größere Flexibilität bei der Erstellung, dem Training und der Handhabung von Netzwerken. Die Kombination aus Elementen von Torch und Ideen von Chainer hat PyTorch in den letzten Jahren populär gemacht.[3]

3 Beachten Sie, dass PyTorch Ideen von Chainer entlehnt, aber nicht den eigentlichen Code.

Die Bibliothek umfasst auch hilfreiche Module zur Text-, Bild- und Tondatenmanipulation (`torchtext`, `torchvision` und `torchaudio`) sowie integrierte Varianten populärer Architekturen wie ResNet (mit Gewichten, die heruntergeladen werden können, um bei Methoden wie *Transfer Learning*, das Sie in Kapitel 4 kennenlernen werden, Unterstützung zu bieten).

Auch über Facebook hinaus gewann PyTorch in der Industrie schnell an Akzeptanz. Unternehmen wie Twitter, Salesforce, Uber und NVIDIA nutzen es auf verschiedene Weise für ihre Deep-Learning-Anwendungen. Ich ahne aber bereits die nächste Frage ...

Warum nicht TensorFlow?

Genau, richten wir unseren Blick auf das ziemlich große von Google konzipierte Schwergewicht in der gegenüberliegenden Ecke. Was bietet PyTorch, was TensorFlow nicht bieten kann? Warum sollten Sie stattdessen PyTorch lernen?

Die Antwort ist, dass das traditionelle TensorFlow anders funktioniert als PyTorch, was erhebliche Auswirkungen auf den Code und die Fehlersuche mit sich bringt. In TensorFlow verwenden Sie die Bibliothek, um eine Graphendarstellung der Architektur des neuronalen Netzwerks aufzubauen, und führen dann Operationen auf diesem Graphen aus, was innerhalb der TensorFlow-Bibliothek geschieht. Diese Methode der deklarativen Programmierung steht im Widerspruch zum imperativen Paradigma von Python, was bedeutet, dass TensorFlow-Programme in Python etwas seltsam und schwer verständlich aussehen und sich auch so anfühlen können. Das andere Problem ist, dass im Vergleich zum Ansatz mit PyTorch die statische Graphendeklaration die dynamische Änderung der Architektur während der Trainings- und Inferenzphase sehr viel komplizierter macht und mit Boilerplate-Code vollstopfen kann.

Aus diesen Gründen ist PyTorch in den forschungsorientierten Fachkreisen populär geworden. Die Anzahl der bei der »International Conference on Learning Representations« eingereichten Beiträge, in denen *PyTorch* erwähnt wird, ist im letzten Jahr um 200% gestiegen. Die Anzahl der Beiträge, in denen *TensorFlow* erwähnt wird, hat fast genauso stark zugenommen. PyTorch wird mit Sicherheit auch in Zukunft eine Rolle spielen.

Mit dem Erscheinen neuerer Versionen von TensorFlow änderten sich die Dinge jedoch ein wenig. Eine neue Funktion namens *Eager Execution*, die es ermöglicht, ähnlich wie mit PyTorch zu arbeiten, wurde unlängst der Bibliothek hinzugefügt. Das wird auch das Paradigma sein, das in TensorFlow 2.0 vorangetrieben wird. Da jedoch nur wenige neue Informationsmaterialien außerhalb von Google zur Verfügung stehen, die Ihnen helfen, diese neue Arbeitsweise mit TensorFlow zu erlernen, sind die Möglichkeiten sehr begrenzt. Sie müssten jahrelang daran arbeiten, um das andere Paradigma zu verstehen, damit Sie das Beste aus der Bibliothek herausholen können.

Aber nichts davon sollte Ihnen ein schlechtes Bild von TensorFlow vermitteln; es bleibt eine industrieerprobte Bibliothek, die durch eines der größten Unternehmen der Welt gefördert wird. PyTorch (natürlich ebenfalls gestützt durch eine andere »größte Firma der Welt«) ist, würde ich sagen, ein rationalisierter und fokussierterer Ansatz für Deep Learning und differenzielle Programmierung. Da keine älteren, eingestaubten APIs mehr unterstützt werden müssen, ist es leichter, in PyTorch zu lehren und produktiv zu werden als in TensorFlow.

Wie hängt Keras damit zusammen? So viele gute Fragen! Keras ist eine übergeordnete Deep-Learning-Bibliothek, die ursprünglich Theano und TensorFlow und nun auch andere Frameworks wie Apache MXNet unterstützt. Sie bietet bestimmte Funktionen wie Trainings-, Validierungs- und Testroutinen, deren Erstellung die untergeordneten Frameworks sonst dem Entwickler überlassen, sowie einfache Methoden zum Aufbau von neuronalen Netzwerkarchitekturen. Sie hat enorm zur Verbreitung von TensorFlow beigetragen, ist jetzt Teil von TensorFlow selbst (als `tf.keras`) und wird auch weiterhin ein eigenständiges Projekt bleiben. Im Vergleich dazu ist PyTorch so etwas wie der Mittelweg zwischen reinem TensorFlow und Keras; wir werden unsere eigenen Trainings- und Vorhersageroutinen schreiben müssen, aber das Erstellen von neuronalen Netzwerken ist fast genauso einfach (und meiner Meinung nach ist der Ansatz von PyTorch zur Erstellung und Wiederverwendung von Architekturen für einen Python-Entwickler bedeutend logischer als der bei Keras).

Wie Sie in diesem Buch noch sehen werden, ist PyTorch seit der Einführung von Version 1.0, obwohl es in eher forschungsorientierten Kreisen verbreitet ist, perfekt für Anwendungsfälle im Produktiveinsatz geeignet.

Ziel und Ansatz

Das vorliegende Buch führt Sie in die Grundlagen von PyTorch ein und zeigt Ihnen, wie Sie selbst Ihre eigenen Deep-Learning-Anwendungen Schritt für Schritt entwickeln, debuggen und in den Produktivbetrieb überführen können. Sie lernen, wie man neuronale Netzwerke in PyTorch erstellt, schrittweise komplexere Modellelemente integriert und schließlich auch Modelle nutzt, die den neuesten Stand der Technik abbilden. Sie erfahren, wie Sie Ihren eigenen Datensatz aufbauen und erweitern können, und auch, wie Sie sich Transfer Learning zunutze machen, wenn Sie nur auf wenige Daten zurückgreifen können oder die Leistungsfähigkeit Ihres Modells weiter verbessern möchten.

Darüber hinaus lernen Sie, wie Sie Deep-Learning-Modelle erfolgreich überwachen und debuggen. Sie werden auch sehen, wie Sie die Entscheidungsfindung eines Modells mithilfe von Visualisierungen offenlegen können. Sie erfahren, wie sich Ihre Anwendung als skalierbarer Webdienst unter Nutzung von Docker, Kubernetes und Google Cloud in den Produktivbetrieb überführen lässt. Zudem sehen wir uns verschiedene Ansätze dazu an, wie wir bereits trainierte Modelle einerseits

komprimieren und andererseits auch das Modelltraining, beispielsweise durch die direkte oder indirekte Nutzung der Programmiersprache C++ oder einer GPU, beschleunigen können.

Wir trainieren unsere Modelle mit Daten aus drei verschiedenen Domänen: Bilder, Texte und Töne. Dabei lernen Sie, in welche Form Sie diese Datentypen zur weiteren Verarbeitung bringen müssen, welche Unterschiede Sie hinsichtlich der Modellwahl beachten sollten, welche Modellarchitekturen sich für welche Domäne besonders anbieten, und sogar, wann es sich lohnt, in eine andere Domäne mithilfe einer Transformation der Daten zu wechseln. Dabei erfahren Sie, wie Sie in PyTorch sowohl integrierte Modelle und Klassen nutzen als auch eigenständig Erweiterungen oder Modifizierungen umsetzen können.

Im Verlauf des Buchs begegnen Ihnen zahlreiche hilfreiche Tipps und Tricks, die Ihnen das Leben bei der täglichen Anwendung von PyTorch erleichtern sollen. Das Buch soll Ihnen nicht nur einen Leitfaden an die Hand geben, sondern Sie auch zu selbstständigem Lernen anregen und Ihnen aufzeigen, was PyTorchs Ökosystem sonst noch für Sie bereithält. Scheuen Sie sich also nicht, zu gegebener Zeit einen Blick in die Dokumentationen oder in den Quellcode der genutzten Bibliotheken und Pakete zu werfen – es lohnt sich ungemein!

Voraussetzungen

Die Hauptzielgruppe dieses Buchs sind Entwickler und Data Scientists, die gegebenenfalls noch wenige Kenntnisse im Bereich Deep Learning haben, aber mehr darüber lernen möchten. Auch wenn Vorkenntnisse wie z.B. ein grundlegendes Verständnis für die Funktionsweise von neuronalen Netzwerken und deren Optimierungsroutinen hilfreich sind, so sind sie nicht unbedingt erforderlich.

Darüber hinaus richtet sich das Buch an diejenigen, die bereits Erfahrungen in der Programmierung im Bereich Deep Learning haben und auf PyTorch umsteigen oder es noch eingehender lernen möchten. Sie sollten über eine gewisse Erfahrung in der Programmierung mit Python oder anderen Programmiersprachen (z.B. Java, Scala, Ruby, R usw.) verfügen und sich bereits ein wenig mit der Handhabung der Kommandozeile auskennen.

Wenn Sie mit Python noch nicht vertraut sind, finden Sie einen guten Einstieg auf *https://www.learnpython.org*. Darüber hinaus gibt es unzählige kostenlose Onlineressourcen, die es Ihnen ermöglichen, sich genügend Python-Wissen anzueignen, um mit den Beispielen in diesem Buch arbeiten zu können.

Der Fokus des Buchs liegt auf der eigenständigen Entwicklung und Umsetzung von Deep-Learning-Modellen in PyTorch. Dabei wird weitestgehend auf mathematische Notationen und komplexe Formeln verzichtet. Sie müssen dementsprechend auch nicht über einen nennenswerten mathematischen Hintergrund verfügen, um dem Buch folgen zu können.

Sie benötigen eine Rechnerumgebung, in der Sie die verwendeten Codebeispiele, die Sie auch in dem GitHub-Repository (*https://oreil.ly/pytorch-github*) des Buchs finden, ausführen können. Das Repository können Sie mit dem Befehl `git clone https://github.com/falloutdurham/beginners-pytorch-deep-learning.git` herunterladen (oder besser, Sie forken zunächst das Repository und laden es dann von Ihrem eigenen Account herunter) und die Beispiele in Ihrer gewünschten Entwicklungsumgebung ausführen (ggf. müssen Sie noch die Pfadangaben individuell an Ihre eigene Verzeichnisstruktur anpassen). In Kapitel 1 werden Sie noch ausführlich erfahren, wie Sie Ihre Entwicklungsumgebung einrichten können.

Wegweiser durch dieses Buch

In Kapitel 1, *Einstieg in PyTorch*, richten wir alles ein, um mit der Programmierung in PyTorch loslegen zu können. Hierbei zeige ich Ihnen, wie Sie PyTorch sowohl lokal als auch in der Cloud nutzen können, und ich gehe auch kurz auf die grundlegendsten Elemente und Befehle der Bibliothek ein.

In Kapitel 2, *Bildklassifizierung mit PyTorch*, lernen Sie die Grundlagen neuronaler Netze kennen und erfahren, wie Sie ein neuronales Netzwerk in PyTorch selbst programmieren, das zur Bildverarbeitung bereitstehende Paket `torchvision` verwenden und Ihr Modell zur Bildklassifizierung nutzen können. Hierfür entwickeln wir zunächst ein relativ einfaches, vollständig verbundenes neuronales Netz, einen eigenen Datensatz mit Bildern aus dem Internet und eine Trainingsroutine, mit der wir das neuronale Netz auf der GPU trainieren. Zudem sehen wir uns an, wie Sie Vorhersagen für Bilder vornehmen und wie Sie Modelle auf und von der Festplatte speichern bzw. wiederherstellen können. Bis einschließlich Kapitel 4 entwickeln wir davon ausgehend unterschiedliche Netzwerke, mit denen wir Bilder klassifizieren und feststellen möchten, ob sich auf einem Bild eine Katze oder ein Fisch befindet. Dabei lernen Sie die wichtigsten Bestandteile und Parameter neuronaler Netzwerke kennen, wie Aktivierungs- und Verlustfunktionen, Lernrate sowie Optimierungsalgorithmen.

Im weiteren Verlauf erhöhen wir kontinuierlich die Leistungsfähigkeit unserer Netzwerke. In Kapitel 3 lernen Sie mit den neuronalen Konvolutionsnetzen (engl. Convolutional Neural Networks, CNNs) eine weitere Architektur kennen, und Sie werden beobachten, dass sich CNNs schneller trainieren lassen und auch genauer sind. In diesem Zusammenhang zeige ich Ihnen verschiedene neuronale Schichttypen und deren Funktionsweise. Zudem gebe ich Ihnen einen kurzen geschichtlichen Einblick in die Entwicklung der bemerkenswertesten Netzwerkarchitekturen im Bereich des bildbasierten Deep Learning. Anschließend werfen wir einen Blick darauf, wie man bereits zuvor trainierte Modelle in PyTorch nutzen kann.

In Kapitel 4, *Transfer Learning und andere Kniffe*, vertiefen wir das Thema Bildverarbeitung noch weiter und greifen dabei auf eine unglaublich effektive Technik im Deep Learning namens Transfer Learning zurück. Hierbei wird ein Modell auf eine

Aufgabe (vor-)trainiert und im Anschluss daran auf eine andere Aufgabe übertragen, weshalb es sich sehr schnell und äußerst effektiv in Anwendung bringen lässt. In diesem Zusammenhang greifen wir auf ein großes, auf einen anderen Datensatz vortrainiertes Netzwerk namens ResNet zurück und nehmen darauf basierend eine Feinabstimmung auf unseren Datensatz vor, um unsere Katze-oder-Fisch-Klassifizierung noch akkurater zu gestalten. Hierbei zeige ich Ihnen auch, wie Sie die optimale Lernrate finden, differenzielle Lernraten im Rahmen der Feinabstimmung heranziehen und auch Datenaugmentationstechniken sowie Ensemble-Modelle zur weiteren Verbesserung Ihres Modells nutzen können.

In Kapitel 5, *Textklassifizierung*, lassen wir die Bilddomäne vorerst hinter uns und widmen uns textbasierten Ansätzen des Deep Learning zur Klassifizierung. Das Kapitel bietet einen Schnelleinstieg in die Thematik rund um rekurrente neuronale Netzwerke und deren Erweiterungen. Wir sehen uns an, wie Textdaten in Form von Einbettungen (engl. Embeddings) codiert und für das Deep Learning nutzbar gemacht werden. Außerdem erkunden wir das `torchtext`-Paket sowie die Erstellung eines Twitter-Datensatzes und wie man das Paket im Rahmen der Textverarbeitung für die Entwicklung eines LSTM-basierten Modells zur Klassifizierung von Tweets nutzt. Darüber hinaus sehen wir uns ebenso für Textdaten Strategien an, die Ihnen helfen, Ihren Datensatz zu erweitern.

Kapitel 6, *Eine Reise in die Welt der Klänge*, bietet Ihnen einen Einstieg in die Verarbeitung von Ton- bzw. Audiodaten. Wir sehen uns zwei sehr unterschiedliche Ansätze zur Klassifizierung an: einen, der auf der reinen Wellenform der Audiodaten und einem neuronalen Konvolutionsnetz beruht, und einen anderen, bei dem wir die Audiodaten zuvor in den Frequenzbereich überführen und uns erneut Transfer Learning zunutze machen. Dabei unternehmen wir einen kurzen Einstieg in das `torchaudio`-Paket und sehen uns an, wie sich Transformationen von Datensätzen effektiv berechnen lassen. Außerdem widmen wir uns abschließend mehreren Ansätzen zur Datenaugmentation von Audiodaten, die die Leistungsfähigkeit der Modelle und ihre Fähigkeit, zu verallgemeinern, verbessern sollen.

In Kapitel 7, *PyTorch-Modelle debuggen*, kommen wir darauf zu sprechen, wie man Modelle debuggen kann, wenn das Training nicht korrekt oder nicht schnell genug vonstattengeht. Wir werfen einen Blick darauf, wie wir mit sogenanntem Class Activation Mapping unter Nutzung von PyTorch-Hooks Deep-Learning-Modelle nachvollziehbarer gestalten können und wie man PyTorch zu Debugging-Zwecken mit Googles TensorBoard verbindet. Zudem erfahren Sie, wie Flammendiagramme eingesetzt werden können, um rechenintensive Funktionsaufrufe zu identifizieren und zu beseitigen, speziell in Bezug auf mögliche Verzögerungen bei Transformationen. Schließlich sehen wir uns an, wie Sie die Auslastung Ihrer GPU im Blick behalten, den Speicherbedarf reduzieren und wie Sie insbesondere bei der Verwendung größerer Modelle mithilfe von Checkpoints mehr Speicherkapazität auf Ihrer GPU erlangen können.

In Kapitel 8, *PyTorch im Produktiveinsatz*, kommen wir darauf zu sprechen, wie Sie PyTorch-Anwendungen in eine Produktionsumgebung überführen, das Deployment des Modells gestalten und Ihr Modell weiter optimieren können. Wir entwickeln einen skalierbaren Webdienst, mit dem wir in PyTorch erstellte Vorhersagemodelle über HTTP und gRPC bereitstellen. Dabei packen wir unsere Anwendung in einen Docker-Container, stellen sie in einem Kubernetes-Cluster über Google Cloud bereit und führen uns vor Augen, wie wir die Anwendung mittels Telemetrie und Logging überwachen können. Im zweiten Teil beschäftigen wir uns mit Optimierungsmöglichkeiten unseres Codes. Wir erkunden in diesem Zusammenhang TorchScript, das uns aufgrund seiner statischen Typisierung und graphenbasierten Darstellung ermöglicht, Python-basierte Modelle weiter zu optimieren. Ich zeige Ihnen, wie Sie mithilfe von Just-in-Time-(JIT-)Tracing und Scripting optimierte und Python-kompatible TorchScript-Modelle erstellen können, die überdies mithilfe von libTorch in C++ ausgeführt und dadurch noch weiter optimiert werden können. Abschließend werfen wir auch einen kurzen Blick darauf, wie große Modelle mittels Quantisierung komprimiert werden können.

In Kapitel 9, *Praxiserprobte PyTorch-Modelle in Aktion*, sehen wir uns an, wie PyTorch von Personen und Unternehmen rund um den Globus genutzt wird und wie Sie selber diese hochaktuellen Modelle mit bereits zur Verfügung stehenden Bibliotheken für Ihre Anwendung nutzen können. Dabei lernen Sie einige neue Ansätze und Modelle kennen. Zuerst beschäftigen wir uns damit, wie Sie Ihr Modell mithilfe weiterer Datenaugmentationstechniken noch verbessern und einer Überanpassung vorbeugen können. Anschließend skizziere ich Ihnen den Aufbau von Super-Resolution-Modellen, die auf Generative Adversarial Networks (GANs) basieren und mit denen wir Bilder hochauflösend hochskalieren können. Im Anschluss daran zeige ich Ihnen, wie Sie relativ schnell mit dem Konvolutionsnetzwerk Faster R-CNN Objekte auf Bildern erkennen und wie Sie Bilder erstellen können, die sogar dazu in der Lage sind, neuronale Netzwerke zu täuschen – sogenannte Adversarial Samples. Wir schauen uns die Grundstruktur von Transformer-basierten Architekturen an und wie diese insbesondere im Bereich des Natural Language Processing als Sprachmodelle zum Tragen kommen. Hierbei zeige ich Ihnen, wie Sie Ihre Feinabstimmung mit einem Transformer-basierten Modell vornehmen (FastBERT), synthetische Tweets mit GPT-2 erzeugen und ein vortrainiertes ULMFiT-Modell sehr schnell mithilfe der fast.ai-Bibliothek zur Klassifizierung von Tweets heranziehen können, das zwar nicht auf dieser Architektur basiert, aber dennoch zufriedenstellende Ergebnisse liefert. Zu guter Letzt greifen wir noch auf PyTorch Lightning zurück, und ich zeige Ihnen verschiedene Anwendungsfälle des selbstüberwachten Lernens im Rahmen des Vorabtrainings, mit dem sich die Modellleistung für verschiedene Aufgabenstellungen weiter verbessern lässt. Hier lernen Sie einerseits den Ansatz kennen, Eingaben zu verändern und diese mithilfe von Selbstüberwachung wiederherzustellen, und andererseits den vielversprechenden Ansatz, Labels für Ihre Daten ohne jeglichen menschlichen Annotationsaufwand zu erstellen.

Veränderungen gegenüber der englischsprachigen Ausgabe

Die vorliegende deutsche Übersetzung unterscheidet sich an wenigen Stellen gegenüber dem englischsprachigen Original. Zum einen wurde der abgedruckte Code aktualisiert, um die Kompatibilität zu neueren Versionen der genutzten Bibliotheken zu gewährleisten, und Errata wurden ausgebessert. Zum anderen wurden der deutschsprachigen Ausgabe in enger Zusammenarbeit mit dem Autor Ian Pointer neue Abschnitte hinzugefügt oder überarbeitet. Dies betrifft den Abschnitt zur Quantisierung von Modellen zum Ende des Kapitels 8 und den Abschnitt zum selbstüberwachten Lernen zum Abschluss des Kapitels 9. Der Abschnitt zu den Transformer-basierten Modellen (ULMFiT, GPT-2 und FastBERT) wurde überarbeitet, und die Codebeispiele wurden komplett in PyTorch überführt. Darüber hinaus wurde der Abschnitt »Alles in einem« in Kapitel 2 inhaltlich ergänzt.

Softwareversionen

Der Code in diesem Buch wurde mit den folgenden Versionen auf Funktionsfähigkeit getestet: Python 3.6.9, PyTorch (torch) 1.6.0 (bis auf den Code im Abschnitt »Faster R-CNN und Mask R-CNN« auf Seite 203, der PyTorch-Version 1.0 nutzt), torchvision 0.7.0, torchtext 0.7.0 und torchaudio 0.6.0. Auf der GitHub-Seite des Buchs finden Sie Dateien mit einer vollständigen Auflistung aller genutzten Versionen. Diese können Sie zur Versionskontrolle nutzen, insbesondere dann, wenn Sie die Installationen lokal vornehmen möchten und in einer eigens dafür vorgesehenen virtuellen Umgebung arbeiten.

Sollten Sie mit dem Paketverwaltungsprogramm `pip` arbeiten, empfehle ich Ihnen, mit dem Befehl `python3 -m venv .venv` von Ihrer Kommandozeile aus (wenn Sie sich im Zielordner befinden) eine virtuelle Umgebung einzurichten (*https://docs.python.org/3/library/venv.html#module-venv*). Nach Aktivierung der virtuellen Umgebung können Sie auf Basis der Datei *requirements.txt* alle notwendigen Bibliotheken und Pakete mit dem Befehl `pip install -r requirement.txt` installieren. Wenn Sie sich für `conda` entscheiden, steht Ihnen eine Datei namens *environment.yml* zur Verfügung, mit der Sie die Installationen mit dem Befehl `conda env create` durchführen können. Auf gleiche Weise können Sie auch die Version bei der Installation einzelner Pakete kontrollieren. Beispielsweise könnten Sie die Befehle `pip install torch==1.6.0` oder `conda install torch==1.6.0` ausführen, um die Installation von PyTorch-Version 1.6.0 zu gewährleisten.

Es ist nicht auszuschließen, dass Veränderungen an den zugrunde liegenden Bibliotheken oder Paketen vorgenommen werden oder dass sich Schnittstellendefinitionen verändern. Sollten Sie einmal feststellen, dass der Code nicht wie gewünscht läuft, zögern Sie nicht, im Repository auf GitHub (*https://oreil.ly/pytorch-github*) nachzusehen, ob neuere Codeversionen verfügbar sind (oder gesonderte Branches) oder ob es offene oder geschlossene Issues gibt, die Ihnen bei Ihrem Problem weiterhelfen könnten.

Weitere Ressourcen

- *Deep Learning – Grundlagen und Implementierung: Neuronale Netze mit Python und PyTorch programmieren* von Seth Weidman, O'Reilly 2020
- *Natural Language Processing mit PyTorch: Intelligente Sprachanwendungen mit Deep Learning erstellen* von Delip Rao und Brian McMahan, O'Reilly 2019
- *GANs mit PyTorch selbst programmieren: Ein verständlicher Einstieg in Generative Adversarial Networks* von Tariq Rashid, O'Reilly 2020

In diesem Buch verwendete Konventionen

Die folgenden typografischen Konventionen werden in diesem Buch eingesetzt:

Kursiv
: Kennzeichnet neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateiendungen.

`Konstante Zeichenbreite`
: Wird für Programmlistings und für Programmelemente in Textabschnitten wie Namen von Variablen und Funktionen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter verwendet.

`Konstante Zeichenbreite, fett`
: Kennzeichnet Befehle oder anderen Text, den der Nutzer wörtlich eingeben sollte.

`Konstante Zeichenbreite, kursiv`
: Kennzeichnet Text, den der Nutzer je nach Kontext durch entsprechende Werte ersetzen sollte.

Dieses Symbol steht für einen Tipp oder eine Empfehlung.

Dieses Symbol steht für einen allgemeinen Hinweis.

Dieses Symbol warnt oder mahnt zur Vorsicht.

Verwenden von Codebeispielen

Zusätzliche Materialien (Codebeispiele, Übungen und so weiter) können Sie von der Adresse *https://oreil.ly/pytorch-github* herunterladen.

Dieses Buch dient dazu, Ihnen beim Erledigen Ihrer Arbeit zu helfen. Im Allgemeinen dürfen Sie die Codebeispiele aus diesem Buch in Ihren eigenen Programmen und der dazugehörigen Dokumentation verwenden. Sie müssen uns dazu nicht um Erlaubnis bitten, solange Sie nicht einen beträchtlichen Teil des Codes reproduzieren. Beispielsweise benötigen Sie keine Erlaubnis, um ein Programm zu schreiben, in dem mehrere Codefragmente aus diesem Buch vorkommen. Wollen Sie dagegen eine CD-ROM mit Beispielen aus Büchern von O'Reilly verkaufen oder verteilen, benötigen Sie eine Erlaubnis. Eine Frage zu beantworten, indem Sie aus diesem Buch zitieren und ein Codebeispiel wiedergeben, benötigt keine Erlaubnis. Eine beträchtliche Menge Beispielcode aus diesem Buch in die Dokumentation Ihres Produkts aufzunehmen, bedarf hingegen unserer ausdrücklichen Zustimmung.

Wir freuen uns über Zitate, verlangen diese aber nicht. Ein Zitat enthält Titel, Autor, Verlag und ISBN. Beispiel: »*PyTorch für Deep Learning* von Ian Pointer (O'Reilly). Copyright 2021 dpunkt.verlag, ISBN 978-3-96009-134-9.«

Wenn Sie glauben, dass Ihre Nutzung von Codebeispielen über den übliche Rahmen hinausgeht oder außerhalb der oben vorgestellten Nutzungsbedingungen liegt, kontaktieren Sie uns bitte unter *kommentar@oreilly.de*.

Danksagungen

Ein herzliches Dankeschön an meine Lektorin, Melissa Potter, meine Familie und Tammy Edlund für all ihre Hilfe bei der Umsetzung dieses Buchs.

Vielen Dank auch an die technischen Gutachter, die während des gesamten Schreibprozesses wertvolles Feedback gaben, darunter Phil Rhodes, David Mertz, Charles Givre, Dominic Monn, Ankur Patel und Sarah Nagy.

KAPITEL 1

Einstieg in PyTorch

In diesem Kapitel richten wir alle erforderlichen Elemente ein, die wir für die Arbeit mit PyTorch benötigen. Jedes folgende Kapitel wird anschließend auf diesem anfänglichen Grundgerüst aufbauen, daher ist es wichtig, dass wir es richtig machen. Dies führt zu unserer ersten grundlegenden Frage: Sollten Sie besser einen maßgeschneiderten Computer für Deep Learning zusammenstellen oder einfach eine der vielen verfügbaren Cloud-basierten Ressourcen verwenden?

Zusammenbau eines maßgeschneiderten Deep-Learning-Rechners

Wenn man sich ins Deep Learning stürzt, verspürt man den Drang, sich ein Rechenmonstrum zusammenzustellen. Sie können Tage damit verbringen, sich verschiedene Arten von Grafikkarten anzusehen, die verschiedenen Speichertypen in Erfahrung zu bringen, die mit Ihrer CPU kompatibel sind, die beste Art von Speicher zu kaufen und zu durchdenken, wie groß ein SSD-Laufwerk im Rahmen Ihres Budgets sein mag, damit Sie so schnell wie möglich auf Ihre Festplatte zugreifen können. Auch ich war davor nicht gefeit; ich habe vor ein paar Jahren einen Monat damit verbracht, eine Liste mit Komponenten zu erstellen und einen neuen Computer auf meinem Esstisch zusammenzubauen.

Mein Rat – vor allem wenn Sie neu in die Welt des Deep Learning einsteigen – lautet: Tun Sie es nicht. Sie können leicht mehrere Tausend Euro in einen Rechner investieren, den Sie möglicherweise gar nicht so oft benutzen. Stattdessen empfehle ich Ihnen, dieses Buch mithilfe von Cloud-Ressourcen (entweder in Amazon Web Services, Google Cloud oder Microsoft Azure) durchzuarbeiten und erst dann über den Bau eines eigenen Rechners nachzudenken, wenn Sie das Gefühl haben, dass Sie einen eigenen Rechner für den 24/7-Betrieb benötigen. Sie müssen *in keiner Weise* große Investitionen in Hardware tätigen, um den Code in diesem Buch ausführen zu können.

Unter Umständen werden Sie nie in die Lage kommen, einen maßgeschneiderten Rechner für sich selbst bauen zu müssen. Es gibt sicherlich Fälle, in denen es billi-

ger sein kann, einen maßgeschneiderten Rechner zu bauen, insbesondere wenn Sie wissen, dass sich Ihre Berechnungen immer auf einen einzigen Rechner (mit bestenfalls einer Handvoll GPUs) beschränken werden. Wenn Ihre Berechnungen jedoch allmählich mehrere Rechner und GPUs erfordern, erscheint die Cloud-Lösung attraktiver. Angesichts der Kosten für die Zusammenstellung eines maßgeschneiderten Rechners würde ich vor einem möglichen Kauf lange und intensiv nachdenken.

Falls es mir nicht gelungen ist, Sie vom Bau Ihres eigenen Rechners abzubringen, finden Sie in den folgenden Absätzen Vorschläge dazu, was Sie dafür benötigen würden.

Grafikprozessor (GPU)

Das Herzstück jedes Deep-Learning-Rechners ist der Grafikprozessor bzw. die GPU. Sie ist die Grundlage für die Vielzahl an Berechnungen in PyTorch und wird die wahrscheinlich teuerste Komponente in Ihrem Computer sein. In den letzten Jahren sind die Preise für GPUs aufgrund ihres Einsatzes beim Mining von Krypto-Währungen wie dem Bitcoin gestiegen und die Vorräte spürbar gesunken. Glücklicherweise scheint diese Blase allmählich zu verschwinden, und das Angebot an GPUs ist wieder etwas größer geworden.

Zum Zeitpunkt des Verfassens dieses Buchs empfehle ich die Beschaffung der NVIDIA GeForce RTX 2080 Ti. Eine billigere Variante ist die GTX 1080 Ti (wenn Sie jedoch die Entscheidung für die 1080 Ti aus Budgetgründen abwägen, schlage ich erneut vor, stattdessen die Cloud-Optionen in Betracht zu ziehen). Obwohl es auch Grafikkarten von AMD gibt, ist ihre Unterstützung in PyTorch derzeit nicht gut genug, um etwas anderes als eine NVIDIA-Karte zu empfehlen. Aber behalten Sie ihre ROCm-Technologie im Auge, die sie allmählich zu einer ernst zu nehmenden Alternative im GPU-Bereich machen sollte.

Hauptprozessor (CPU) und Motherboard

Wahrscheinlich werden Sie sich für ein Motherboard der Z370-Serie entscheiden wollen. Viele Menschen werden Ihnen sagen, dass die CPU für das Deep Learning keine Rolle spielt und dass Sie mit einer CPU mit einer relativ geringen Rechengeschwindigkeit auskommen können, solange Sie eine leistungsstarke GPU haben. Meiner Erfahrung nach werden Sie überrascht sein, wie oft die CPU zum Engpass werden kann, insbesondere bei der Arbeit mit augmentierten Daten.

Arbeitsspeicher (RAM)

Mehr Arbeitsspeicher ist immer ratsam, da Sie dadurch mehr Daten im Speicher halten können, ohne auf den viel langsameren Festplattenspeicher zurückgreifen

zu müssen (besonders wichtig während des Trainings). Sie sollten mindestens 64 GB DDR4-Speicher für Ihren Rechner in Betracht ziehen.

Speicher

Der Speicher für ein individuelles System sollte zwei Arten umfassen: zum einen eine Festplatte mit M2-Schnittstelle (SSD) – so groß wie möglich – für Ihre »heißen« Daten, damit Sie so schnell wie möglich auf diese zugreifen können, wenn Sie aktiv an einem Projekt arbeiten. Als zweites Speichermedium fügen Sie ein 4 TB (Terabyte) großes Serial-ATA-(SATA-)Laufwerk für Daten hinzu, an denen Sie nicht aktiv arbeiten; wechseln Sie je nach Bedarf zwischen »heißem« und »kaltem« Speicher.

Ich empfehle Ihnen, einen Blick auf die Webseite PCPartPicker (*https://pcpartpicker.com*) zu werfen, um einen Eindruck von den Deep-Learning-Rechnern anderer Leute zu bekommen. (Sie können sich auch all die schrägen und verrückten Ideen ansehen!) Sie erhalten ein Gefühl für die Auswahl an möglichen Computerteilen und den dazugehörigen Preisen, die insbesondere bei Grafikkarten sehr stark schwanken können.

Nachdem Sie nun die Möglichkeiten für Ihren lokalen physischen Rechner gesehen haben, ist es an der Zeit, auf die Cloud-Alternativen zu sprechen zu kommen.

Deep Learning in der Cloud

Okay, nun könnten Sie fragen, warum die Cloud-Variante besser sein sollte – besonders dann, wenn Sie sich das Preisschema von Amazon Web Services (AWS) angeschaut und herausgefunden haben, dass sich der Eigenbau eines leistungsfähigen Rechners innerhalb von sechs Monaten bezahlt macht. Denken Sie darüber nach: Wenn Sie gerade erst anfangen, werden Sie diesen Rechner in den besagten sechs Monaten nicht rund um die Uhr nutzen. Sie werden es schlichtweg nicht tun. Das bedeutet, dass Sie den Cloud-Rechner abschalten und in der Zwischenzeit lediglich ein paar Cents für die gespeicherten Daten bezahlen dürften.

Sie brauchen nicht gleich zu Beginn eine von NVIDIAs Profikarten wie die Tesla V100 zu verwenden, die ebenfalls in Ihrer Cloud-Instanz verfügbar sind. Sie können mit einer der wesentlich preisgünstigeren (manchmal sogar kostenlosen) K80-basierten Instanz beginnen und zu einer leistungsstärkeren Grafikkarte wechseln, sobald diese benötigt wird. Das ist ein wenig günstiger als der Kauf einer einfachen GPU-Karte und die Aufrüstung auf eine RTX 2080 Ti für Ihre eigene Rechnerumgebung. Wenn Sie acht V100-Karten zu einer einzelnen Instanz hinzufügen möchten, können Sie dies mit nur wenigen Klicks erledigen. Versuchen Sie das einmal mit Ihrer eigenen Hardware.

Ein weiterer Punkt ist die Wartung. Wenn Sie es sich zur Gewohnheit machen, Ihre Cloud-Instanzen regelmäßig neu zu erstellen (idealerweise jedes Mal, wenn Sie

wieder an Ihren Projekten arbeiten), werden Sie fast immer auf einem Rechner arbeiten, der auf dem neuesten Stand ist. Gehört der Rechner Ihnen, obliegt Ihnen auch die Aktualisierung. An dieser Stelle muss ich gestehen, dass ich meinen eigenen, speziell konzipierten Deep-Learning-Rechner habe und die Ubuntu-Installation darauf so lange ignoriert habe, bis sie nicht mehr unterstützt wurde. Dies führte dazu, dass ich schließlich einen Tag damit verbrachte, das System erneut an einen Punkt zu bringen, an dem es wieder Updates erhalten konnte. Peinlich.

Wie dem auch sei, nehmen wir an, Sie haben die Entscheidung getroffen, die Cloud zu nutzen. Hurra! Nun zur nächsten Frage: Welchen Anbieter sollten Sie wählen?

Google Colaboratory

Doch warten Sie, bevor wir uns die Anbieter ansehen: Was ist, wenn Sie überhaupt keine Lust haben, einen nennenswerten Aufwand zu treiben? Keine Lust auf das lästige Aufbauen eines Rechners oder die ganze Mühe, Instanzen in der Cloud einzurichten? Was wäre denn tatsächlich die bequemste Wahl? Dafür hat Google das Richtige für Sie. *Colaboratory* (oder *Colab*) (*https://colab.research.google.com*) ist eine weitgehend kostenlose, installationsfreie, benutzerdefinierte Jupyter-Notebook-Umgebung, für die Sie ein Google-Konto benötigen, um Ihre eigenen Notebooks einzurichten. Abbildung 1-1 zeigt einen Screenshot eines in Colab erstellten Notebooks.

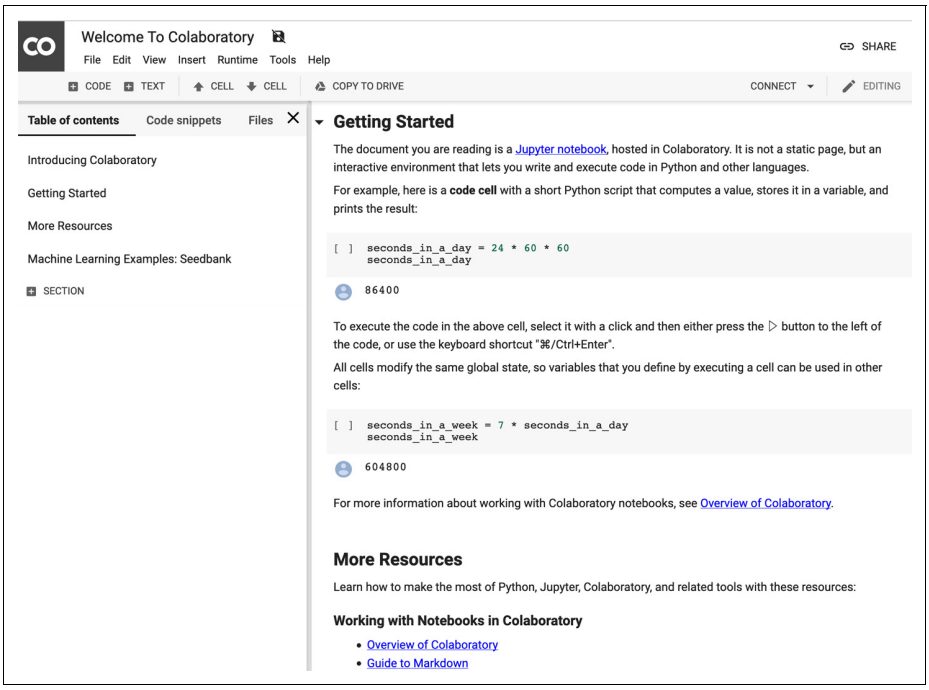


Abbildung 1-1: Google Colab(oratory)

Colab ist ein großartiges Tool, um ins Deep Learning einzutauchen, da es bereits vorinstallierte Versionen von TensorFlow und PyTorch bereithält. Dementsprechend müssen Sie nichts weiter eingeben als den `import torch`-Befehl und auch keine weiteren Einstellungen vornehmen. Jeder Benutzer erhält einen freien Zugang zu einer NVIDIA-T4-Grafikkarte mit bis zu zwölf Stunden ununterbrochener Laufzeit. Kostenlos. Um das einzuordnen: Empirische Untersuchungen haben ergeben, dass Sie etwa die Hälfte der Geschwindigkeit einer 1080 Ti für Ihr Training erhalten, aber zusätzlich noch 5 GB Speicher, sodass Sie größere Modelle speichern können. Es besteht zudem die kostenpflichtige Möglichkeit, eine Verbindung zu neueren GPUs oder Googles eigener TPU-Hardware herzustellen. Aber Sie können so ziemlich jedes Beispiel aus diesem Buch mit Colab ausführen, ohne dass irgendwelche Kosten für Sie anfallen. Aus diesem Grund empfehle ich, Colab zunächst parallel zu diesem Buch zu verwenden, um dann bei Bedarf zu entscheiden, ob Sie auf dedizierte Cloud-Instanzen und/oder Ihren eigenen persönlichen Deep-Learning-Server ausweichen möchten.

Wenn Sie zum ersten Mal mit Google Colab arbeiten, werden Sie feststellen, dass es verschiedene Möglichkeiten gibt, die Dateien des GitHub-Repositorys dieses Buchs einzubinden. Eine gute Einführung bietet Ihnen das Willkommens-Notebook, das sich beim Aufruf von Google Colab im Hintergrund lädt, und hier insbesondere der Abschnitt »Weitere Ressourcen«.

Colab ist der Ansatz, mit dem am wenigsten Aufwand verbunden ist. Sie möchten aber vielleicht ein wenig mehr Kontrolle darüber haben, wie Dinge installiert werden oder wie Sie einen Secure-Shell-(SSH-)Zugang zu Ihrer Instanz in der Cloud erhalten. Werfen wir am besten einen Blick auf das Angebot der wichtigsten Cloud-Anbieter.

Cloud-Anbieter

Jeder der drei großen Cloud-Anbieter (Amazon Web Services, Google Cloud Platform und Microsofts Azure) bietet GPU-basierte Instanzen (auch als *virtuelle Maschinen* oder *VMs* bezeichnet) und offizielle Images, die auf diesen Instanzen bereitgestellt werden. Sie haben alles, was Sie brauchen, um loszulegen, ohne selbst Treiber oder Python-Bibliotheken installieren zu müssen. Gehen wir die Angebote der einzelnen Anbieter einmal durch.

Amazon Web Services

AWS, das Schwergewicht auf dem Cloud-Markt, erfüllt nur zu gern Ihre GPU-Anforderungen und bietet P2- und P3-Instanzarten an, um Sie zu unterstützen. (Der G3-Instanztyp wird eher in tatsächlich grafikbasierten Anwendungen wie der Videocodierung verwendet, daher werden wir hier nicht weiter darauf eingehen). Die P2-Instanzen verwenden die älteren NVIDIA-K80-Karten (maximal 16 können an eine Instanz angeschlossen werden), und die P3-Instanzen setzen die extrem

schnellen NVIDIA-V100-Karten ein (Sie können acht davon auf eine Instanz legen, wenn Sie sich trauen).

Sollten Sie sich für AWS entscheiden, empfehle ich, für dieses Buch die Klasse `p2.xlarge` zu verwenden. Das kostet Sie zum Zeitpunkt des Verfassens dieses Buchs rund einen US-Dollar pro Stunde und bietet Ihnen ausreichend Leistung, um die Beispiele problemlos durchzuarbeiten. Sollten Sie mit diversen anspruchsvollen Kaggle-Wettbewerben beginnen, bietet es sich gegebenenfalls an, auf P3-Instanzen aufzustocken.

Es ist unglaublich einfach, eine Deep-Learning-Umgebung auf AWS zu erstellen:

1. Melden Sie sich an der AWS-Konsole an.
2. Wählen Sie *EC2* und klicken Sie auf *Launch a virtual machine*.
3. Suchen Sie nach der Option *Deep Learning AMI (Ubuntu)* und wählen Sie diese aus.
4. Wählen Sie `p2.xlarge` als Ihren Instanztyp.
5. Starten Sie die Instanz, indem Sie entweder ein neues Schlüsselpaar erstellen oder ein vorhandenes Schlüsselpaar wiederverwenden.
6. Stellen Sie mittels Ihrer Kommandozeile eine Verbindung zur Instanz her, indem Sie SSH verwenden und Port 8888 auf Ihrem lokalen Rechner auf die Instanz umleiten:

```
ssh -L localhost:8888:localhost:8888 \
-i /path/my-key-pair.pem my-instance-user-name@my-instance-public-dns-name
```

7. Starten Sie Jupyter Notebook, indem Sie **`jupyter notebook`** eingeben. Kopieren Sie die erzeugte URL und fügen Sie sie in Ihren Browser ein, um auf Jupyter zuzugreifen.

Vergessen Sie nicht, Ihre Instanz abzuschalten, wenn Sie sie nicht benutzen! Sie können das tun, indem Sie mit der rechten Maustaste auf die Instanz in der Weboberfläche klicken und unter *Instance State* die Option *Stop* wählen. Dadurch wird die Instanz heruntergefahren, und es entstehen Ihnen keine Kosten für die Instanz, solange sie nicht läuft. Allerdings wird Ihnen der Speicherplatz, den Sie der Instanz zugewiesen haben, auch dann in Rechnung gestellt, wenn die Instanz ausgeschaltet ist; seien Sie sich dessen bewusst. Um die Instanz und den Speicherplatz vollständig zu löschen, wählen Sie stattdessen die Option *Terminate*.

Azure

Wie AWS bietet auch Azure eine Mischung aus günstigen K80-basierten Instanzen und teureren Tesla-V100-Instanzen an. Azure stellt außerdem Instanzen zur Verfügung, die auf der älteren P100-Hardware basieren und als Kompromiss zwischen den beiden anderen Instanzen gelten. Für dieses Buch empfehle ich, als Instanztyp eine einzelne K80 (NC6) zu verwenden, die 90 Cent pro Stunde kostet. Wechseln