Fakultät Informatik/Mathematik

Wissenschaftliches Seminar

im Wintersemester 2017/2018

Eine Einführung in TensorFlow

---

bearbeitet von: Bierschneider Christian 3118760
Maximilian Poeschl 3121342
Benjamin Maiwald 3097528
Studiengang: Informatik
Schwerpunkt: Software Engineering

Betreuer: Prof. Dr. Jan Dünnweber
OTH Regensburg

Regensburg, 31. Dezember 2017

Abstract

Abkürzungsverzeichnis

KI	Künstliche Intelligenz.
ML	Machine Learning, bzw. Maschinelles Lernen.
Neuronales Netz	Neuronales Netz, bzw. Neural Network.
TF	TensorFlow.

Inhaltsverzeichnis

Kapitel 1

1 Eine Einführung in Maschinelles Lernen

Da Vorwissen in den Bereichen Künstliche Intelligenz (KI) und Machine Learning (ML) selbst bei Studierenden der Informatik nicht generell vorausgesetzt werden kann, gibt dieses Kapitel eine kurze Einführung in das Thema. Es wird über die Grundlagen im Bereich des Maschinellen Lernens mit dem Schwerpunkt auf Neuronale Netze (NN) informiert, die zum Verständnis der Arbeit benötigt werden. Sollte sich der Leser bereits mit dem Thema auseinander gesetzt haben und Begriffe wie „Label“, „Schichten (Layers)“ und „Gewichte“ schon bekannt sein, kann dieses Kapitel auch übersprungen werden.

Für große Teile des geschichtlichen Abrisses diente das Buch „Künstliche Intelligenz, ein moderner Ansatz“ von Stuart Russell und Peter Norvig als Quelle, welches wohl eines der bekanntesten Werke zum Thema KI sein dürfte [1]. Generell kann ich dieses Buch allen Interessierten empfehlen, gleichwohl aufgrund des doch sehr großen Umfangs je nach Situation meist nur einzelne Kapitel hilfreich sein werden.

1.1 Meilensteine des Maschinellen Lernens

In den letzten Jahren (seit ca. 2010) hat sich das Thema KI zu einem regelrechten Hype-Thema entwickelt — und das nicht ganz zu unrecht. Denn gerade durch Anwendungen in den Bereichen Bildverarbeitung, Spracherkennung, sogenannter „Recommender Systems“ oder auch des automatisierten Fahrens, gab es enorme Fortschritte. Diese machten KI für die breite Masse salonfähig und ermöglichten die Entwicklung von Produkten wie Siri, den Skype Translator, Filmempfehlungen auf Netflix oder die Fahrerassistenzsysteme im Tesla Model S, die heute von Millionen von Menschen täglich genutzt werden.

Allen voraus liegt das Hauptaugenmerk vieler Informatiker und Forscher gerade auf den sogenannten „Künstlichen Neuronalen Netzen“ (Artificial Neural Networks). Manche dieser Neuronalen Netze wurden sogar zu echten Superstars in der Szene, wie zum Beispiel „AlphaGo“, das von Google entwickelt wurde und 2016 den damaligen Vize-Weltmeister Lee Sedol in vier von fünf Runden im Brettspiel „Go“ besiegte. Aufgrund der unglaublichen

Komplexität[1] des Spiels galt der Sieg einer Maschine über einen realen Meister lange Zeit als unmöglich.

Dabei sind die meisten Grundlagen auf diesem Gebiet bereits Jahrzehnte alt. Schon in den 1940er Jahren und somit unmittelbar nach der Erfindung des modernen Computers, begannen die ersten Forscher damit, ein Modell für Künstliche Neuronale Netze zu entwickeln und behaupteten sogar bereits, dass entsprechend definierte Netze auch lernfähig seien. In demselben Artikel, in dem Alan Turing 1950 die Idee des weltbekannten Turing-Tests — in [2] „The Imitation Game“ genannt — veröffentlichte, schrieb er außerdem zum ersten Mal über „lernende Maschinen“ und philosophierte darüber, wie man einer Maschine beibringen könne, im Imitation Game zu bestehen. In dieser Niederschrift formulierte Turing auch die Grundideen zum heute als „Reinforcement Learning“ bezeichneten Lernen durch Bestrafung und Belohnung.

1956 war schließlich offiziell das Geburtsjahr der Künstlichen Intelligenz. Am Dartmouth College (Hanover, New Hampshire) veranstalteten McCarthy, Minsky, Shannon und Rochester (allesamt Größen in der Entwicklung der KI) ein „Summer Research Project on Artificial Intelligence“ zusammen mit weiteren Forschern aus ganz Amerika. Hier wurde der Begriff „Artificial Intelligence“ zum ersten Mal überhaupt benutzt. Ziel des Workshops war es, in zwei Monaten einen signifikanten Fortschritt bei der Entwicklung einer intelligenten Maschine zu erreichen. Dieses Ziel konnte zwar nicht erfüllt werden, jedoch sorgte das Treffen dafür, dass sich die wichtigsten Personen kennenlernten, die in den darauffolgenden 20 Jahren die größten Neuerungen auf diesem Gebiet entwickelten.

Wie auch heute gab es schon einmal in den 1980er Jahren einen großen Boom in der KI-Industrie. Die Investitionen stiegen von einigen Millionen Dollar im Jahr 1980 auf mehrere Milliarden Dollar im Jahr 1988. Viele der KI-Firmen konnten ihre Versprechen jedoch nicht halten, weshalb der Markt in den 90er Jahren zusammenbrach. In Folge dessen ging auch die Forschung auf dem Gebiet zurück und es kam zu keinen nennenswerten Erkenntnisgewinnen in den 90er Jahren. Deshalb wird dieses Jahrzehnt auch als „KI-Winter“ bezeichnet.

Wenn alle diese Entwicklungen im Bereich der KI aber bereits so lange zurück liegen, warum hat es dann bis heute gedauert, dass es die ersten Anwendungen zum Endkunden schaffen?

Wir leben heute in einer spannenden Zeit, denn einige wichtige Faktoren, die für den Erfolg der KI wichtig sind, wurden nahezu zeitgleich verfügbar.

[1] Ein Go-Brett besteht aus einem Raster von 19 x 19 Plätzen zum Setzen. Für den ersten Zug existieren also 361 Möglichkeiten, für den zweiten 360 und so weiter. Dies ergibt bereits für die ersten drei Züge mehr als 46 Millionen mögliche Spielabläufe.

Zum einen stieg die Leistung von Computern seit deren Erfindung stetig an. Für die meisten Berechnungen im Bereich des Maschinellen Lernens wird eine sehr hohe Rechenleistung benötigt. Vor allem die Verwendung von GPUs zur parallelen Berechnung von allgemeinen Aufgaben beschleunigt das Trainieren von Neuronalen Netzen um ein Vielfaches. Rechenvorgänge, die noch vor zehn Jahren große Rechencluster für viele Monate beanspruchten, können heute in Stunden, maximal aber wenigen Tagen abgeschlossen werden – und das auf kompakten Rechnern, die sogar für Privatpersonen erschwinglich sind. Dies gibt den Forschern und Softwareingenieuren die Möglichkeit, schon nach kurzer Zeit ein Feedback zu erhalten, ob die von Ihnen gewählten Ansätze richtig sind und falls nicht, Anpassungen vorzunehmen.

Zum anderen leben wir im Zeitalter von „Big Data“. Für ML ist es extrem wichtig, dass große Datenmengen zur Verfügung stehen, die für den Trainingsprozess verwendet werden können. Firmen wie Google, Facebook, Apple oder Microsoft (und natürlich vielen anderen) steht ein schier unerschöpflicher Pool an Informationen zur Verfügung. Diese maschinell generierten Daten eignen sich aufgrund ihres großen Umfangs perfekt dazu, neuronale Netze zu trainieren und das wird von diesen Firmen natürlich auch genutzt, um neue Geschäftsideen zu entwickeln und die angebotenen Dienste für ihre Kunden kontinuierlich zu verbessern.

Natürlich birgt die Entwicklung der Künstlichen Intelligenz auch Gefahren. Diese können ganz real sein, wie der Wegfall tausender Arbeitsplätze [3] oder aber spekulativ und in der Zukunft liegend, wie die mögliche Gefährdung der Menschheit durch eine künstliche Superintelligenz [4]. In jedem Fall wird die Weiterentwicklung und Forschung auf dem Gebiet auch die nächsten Jahre ein extrem vielfältiges und spannendes Thema bleiben.

1.2 Grundlagen von Neuronalen Netzen

Kapitel 2

2 Neuronale Netze in Tensorflow

Neuronale Netze sind das am häufigsten benutzte Werkzeug in Tensorflow. Im Folgenden werden die Grundlagen für Neuronale Netze vorgestellt, sowie deren Umsetzung in Tensorflow. Diverse Codeauschnitte in diesem und alle nachfolgenden Kapitel sind Implementierungen in der Programmiersprache Python.

2.1 Feedforwardnetzwerk

Die häufigste in Tensorflow implementierte Version Neuronaler Netze ist das Feed Forward Netz. Diese werden auch am häufigsten für Deep Learning verwendet.[1] Ein Feed Forward Netz besteht aus einer Eingabeschicht, einer Ausgabeschicht und n versteckten Schichten dazwischen. Hierbei ist jedes Neuron mit jedem Neuron der nachfolgenden Schicht durch

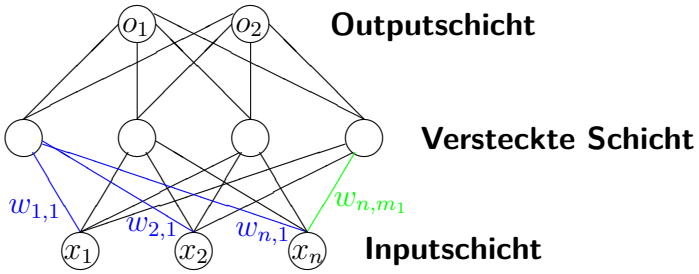

Abbildung 2.1: Feed Forward Netz

Gewichte verbunden.
Die Abbildung zeigt beispielhaft ein Feed Forward Netzwerk mit 3 Input-Neuronen in der Eingabeschicht, einer versteckten Schicht mit 4 versteckten Neuronen und einer Ausgabeschicht mit 2 Ausgabe-Neuronen.[2]
Ein Feed Forward Netzwerk kann beliebig viele versteckten Schichten enthalten, aber nur eine Eingabe und eine Ausgabeschicht. Ebenso kann die Anzahl der Neuronen in den versteckten Schichten frei gewählt werden. Eine sinnvolle Anzahl von Neuronen in den versteckten Schichten lässt sich nur experimentell bestimmen.[3] Man sollte darauf achten

[1] [5] S.168
[2] [6]s.117
[3] [7] S.385

nicht zu wenige Neuronen zu verwenden da sonst die Lernkapazität möglicherweiße zu eingeschränkt ist. Auch zu viele Neuronen können problematisch sein, da es sehr lang dauern kann jede der vielen Neuronen zu trainieren und die Effizienz des Netzwerks darunter leidet.[4]
Man kann bereits mit einer versteckten Schicht und einer ausreichend großen Anzahl Neuronen jedes Problem simulieren, tendenziell ist es aber besser die Anzahl der Schichten zu erhöhen anstatt die Anzahl der Neuronen pro Schicht.[5]

2.2 Der Input Tensor

Vektoren und Matrizen werden in Tensorflow als Tensors bezeichnet. Der Input eines Neuronalen Netzes ist ein n-dimensionaler Vektor der für jedes Input-Neuron einen Wert enthält. Die Anzahl der Input-Neuronen richtet sich nach dem untersuchten Problem. Das Einführungsbeispiel für Tensorflow das in etwa dem "Hello World" für Programmiersprachen entspricht, behandelt ein Klassifikationsproblem.[6] In diesem Problem sollen mithilfe der MNIST Datenbank die tausende von handgeschriebenen 28×28 Bilder der Zahlen von 1-9 enthält, das Neuronale Netz lernen handgeschriebene Ziffern zu unterscheiden. Für jedes dieser Bilder hat man entsprechend eine 28×28 Matrix mit Grauwerten. Mit dem Befehl `reshape(-1,..)` [7] lässt es sich in einen eindimensionalen Vektor mit $28 * 28 = 784$ Input Neuronen verwandeln.
Input-Tensore werden in Tensorflow mit Platzhaltern angelegt, da während des Lernprozesses der Inputtensor für jedes zu lernende Beispiel aktualisiert wird. Um den Platzhalter anzulegen verwendet man den Befehl

```
1 input= tf.placeholder("float",[None,Input_Neuronen])
```

[8] Mit float wird der Datentyp des inputs für die einzelnen Neuronen gewählt. Die Variable `Input_Neuronen` gibt die Anzahl der Input Neuronen an. Das None steht für die Anzahl der Trainingsdaten, die später noch dynamisch eingefügt wird. So muss man sich zunächst nicht festlegen wie viele Trainingsdaten man benutzen will.[9]

[4] [8] S.341
[5] [7] S.384
[6] [7] S.124
[7] [7] S.99
[8] [9] S.218
[9] [7] S.377

2.3 Gewichte

Man kann alle Gewichte zwischen Inputschicht und der ersten versteckten Schicht als Gewichtsmatrix

$$W^{(1)} := \begin{pmatrix} {w_{1,1}}^{(1)} & {w_{1,2}}^{(1)} & \cdots & {w_{1,m_1}}^{(1)} \\ {w_{2,1}}^{(1)} & {w_{2,2}}^{(1)} & \cdots & . \\ {w_{3,1}}^{(1)} & {w_{3,2}}^{(1)} & \cdots & . \\ \vdots & \vdots & \vdots & \vdots \\ {w_{n,1}}^{(1)} & {w_{n,2}}^{(1)} & \cdots & {w_{n,m_1}}^{(1)} \end{pmatrix} \tag{2.1}$$

auffassen.

Die Zeilen von $W^{(1)}$ entsprechen allen ausgehenden Verbindungen für ein Neuron aus der Input-Schicht. Die Spalten entsprechen den eingehenden Verbindungen für ein Neuron aus der versteckten Schicht.

Der Tensorflow Befehl um die Gewichte zwischen zwei Schicht zu Initialisieren lautet:

```
1 init = tf.truncated_normal(n_inputs, n_neurons)
2 W =tf.Variable(init,name="weights")
```

[10] Damit werden die Gewichte mit zufälligen Gewichten belegt und es wird eine Matrix bzw Tensor der Dimension `(input_neuronen×versteckte_neuronen)` erzeugt.
Um den Output des Neuronalen Netzes zu berechnen führt man den Befehl `tf.matmul(input,W)` eine Matrix Multiplikation zwischen Input-Tensor und Gewichtsmatrix durch.[11] Auf diese Weise bekommt man einen weiteren Tensor der für jedes versteckte Neuron einen Wert hat. Auf diese Werte wendet man nun eine Aktivierungsfunktion an, das Ergebnis nennt man **Aktivität**[12] der versteckten Schicht. Dieses Verfahren kann man nun für beliebig viele versteckte Schichten wiederholen, dabei nutzt man die Aktivität und die Gewichtsmatrix für die zweite versteckte Schicht um eine Matrix Multiplikation durchzuführen und die Aktivität der nächsten Schicht zu bekommen.

[10] [7] S.377
[11] [7] S.377
[12] [10] S.247

2.4 Aktivierungsfunktionen

Aktivierungsfunktionen werden verwendet um den Wertebereich den die Neuronen annehmen können einzugrenzen. So liegen manche Aktivierungsfunktionen nur zwischen den Werten -1 und 1 oder bilden alle negative Werte auf Null ab. Exemplarisch werden 3 der beliebtesten Aktivierungsfunktionen beschrieben.

2.4.1 Rectified linear unit function

Eine beliebte Möglichkeit ist die Rectified linear unit function, kurz relu.
Für relu gilt in der parametrisierten Form[13]

$$\sigma(x) = \begin{cases} 0 & \text{falls} \quad x <= 0 \\ ax & \text{sonst} \end{cases}$$

a ist dabei frei wählbar und kann an das jeweilige Beispiel angepasst werden. Obwohl die Relufunktion eine sehr einfache fast lineare Funktion ist, erweist sie sich als sehr Leistungsfähig. Sie dient als Standardaktivierungsfunktion, die für die meisten FeedForward Netzwerke empfohlen wird.[14] In Tensorflow ist die Relu Funktion auf verschiedene Arten implementiert, die oben beschriebene Standard Relu Funktion allerdings nur für den Parameterwert a=1. Sie wird mit

```
1 tf.nn.relu(features, name=None)
```

[15] eingebunden. Eine andere in tensorflow verwendete Versionen der relu Funktion ist

```
1 tf.nn.relu6(features, name=None)
```

[9] Für diese gilt:

$$\sigma(x) = \begin{cases} 0 & \text{falls} \quad x <= 0 \\ x & \text{falls,} \quad 0 < x < 6 \\ 6 & \text{sonst} \end{cases}$$

Sie kann schneller berechet werden und hat den Vorteil das weder Werte nahe der Null verschwinden, noch die Werte zu großwerden können.[9]

[13] [5]S.174
[14] [5]S.175
[15] [9]S.58f.

2.4.2 Sigmoidfunktion

Eine andere der Standardaktivierungsfunktionen ist die Sigmoid Funktion. Sie eignet sich gut für das lernen mittels Backpropagation, welches ein sehr häufig eingesetztes Lernverfahren ist.[16]
Die Sigmoidfunktion hat folgende Form:[11]

$$\sigma_c(x) = \frac{1}{1 + e^{-cx}} \tag{2.2}$$

In Tensorflow ist sie nur mit dem Parameterwert $c = 1$ enthalten.

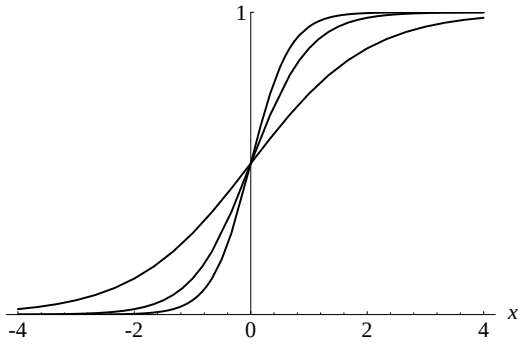

Abbildung 2.2: Sigmoidfunktion für c=1 [11]

Der Wertebereich der Sigmoidfunktion liegt zwischen 0 und 1, was sich sehr gut eignet falls die Ausgabewerte des Netzwerks ebenfalls in diesem Bereich liegen. In Tensorflow ist sie mit dem Befehl `tf.sigmoid(x, name=None)`[17] definiert.

2.4.3 Tangenshyperbolicus

Der Tangenshyperbolicus ist definiert als[18]

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{2.3}$$

Die Ableitungen der Sigmoidfunktion und des Tangenshyperbolicus sind leicht zu berechnen, weshalb sie sich gut zur Berechnung des Gradienten der Fehlerfunktion eignen.
Wie in der Abbildung ersichtlich wird, befindet sich der Wertebereich des Tangenshyperbolicus im Intervall von [-1,1]. Es wird mit `f.tanh(x, name=None)`[19] aufgerufen und eignet

[16] [11] S.151f.
[17] [12]S.175
[18] [6]S.127
[19] [12]S.175

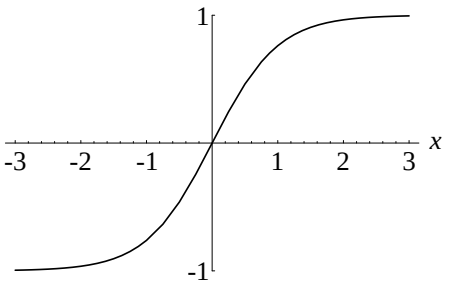

Abbildung 2.3: Graph des Tangenshyperbolicus [11]

sich besonders gut, wenn man in den Lernbeispielen auch negative Zahlen berücksichtigen will.

2.5 Kostenfunktionen

Kosten- oder Fehlerfunktionen sind ein Maßdafür wie gut ein Neuronales Netz lernt. Es stellt eine berechenbare Formel bereit, die den durch das Netz berechneten Output für ein Trainingsbeispiel mit dem zu lernenden Output vergleicht. Außerdem werden sie für das lernen des Neuronales Netz benötigt, da aus den Ableitungen der jeweiligen Kostenfunktion für das zugehörige Gewicht die neuen Gewichte gebildet werden.[20]

2.5.1 MSE

Eine mögliche Kostenfunktion ist der Mean Squared Error, kurz MSE.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{o}_i - o_i)^2. \tag{2.4}$$

[21] $\hat{o}$ steht für den gewünschten Output und o für den vom Netz für den zugehörigen Input Vektor generierten Output.
Der MSE wird mit `Kosten=tf.losses.mean_squared_error(labels,predictions)`[22] eingebunden und summiert die Quadrate der Abweichungen auf, d.h. je geringer der MSE wird, desto genauer hat das Netz die Trainingsdaten gelernt.

[20] [5]S.177f.
[21] [11]S.156
[22] [9]S.94f.

2.5.2 cross entropy

Eine andere Möglichkeit ist die sogenannte Cross Entropy Funktion.[23]

$$CE = -\frac{1}{N}\sum_{i=1}^{N} \hat{o}_i \ln o_i + (1-\hat{o}_i)\ln(1-o_i) \tag{2.5}$$

Der Vorteil der cross entropy Funktion besteht daran, dass sie je schneller lernt je größer der anfängliche Fehler ist. Fängt man also bei sehr ungünstig gewählten zufälligen Gewichten mit dem Lernprozess an, wird cross entropy schneller bessere Ergebnisse liefern als der MSE.[24] Welche man letztendlich verwendet hängt vom zugrunde liegenden Problem ab.
Cross entropy wird in Tensorflow mit dem Befehl

```
1 Kosten=tf.nn.softmax_cross_entropy_with_logits()
```

verwendet.[25]

2.6 Lernprozess

Damit das Netz lernt müssen Stück für Stück die Gewichte angepasst werden. Zu diesem Zweck berechnet man die Ableitung der Kostenfunktion und benutzt sie um die Gewichte abzuändern. Dieses Verfahren wird "Backpropagation of Error" mittels "Gradient Descent" - dem Gradientabstiegsverfahren genannt.
Das Update der Gewichte funktioniert nach folgender Regel:[26]

$$w_{i,j} = w_{i,j} - \frac{\partial Kosten}{\partial w_{i,j}}\eta, \tag{2.6}$$

wobei η die Lernrate darstellt. Diese ist frei wählbar, üblicherweise liegt sie im Bereich von 0.01 und 0.5.
Die optimale Lernrate für das gegebene Problem muss experimentell bestimmt werden, da man dazu im Vorfeld keine genauen Vorhersagen machen kann. Trainiert wird das Netzwerk letztendlich mit dem Befehl `tf.train.GradientDescentOptimizer(Lernrate ).minimize(Kosten)`[27]

[23] [13]
[24] [13]
[25] [9]S.88
[26] [11]S.157
[27] [12]S.156

Kapitel 3

3 Das Framework TensorFlow

TensorFlow (TF) ist eine plattformunabhängige Bibliothek für ML, die für große und variable Architekturen entwickelt [14] und Ende 2015 veröffentlicht wurde [15]. Um die einzelnen Verarbeitungsschritte der Daten darzustellen, werden von TF sogenannte Datenfluss Graphen verwendet. Diese bieten auch die Möglichkeit für alle Operationen festzulegen, von welcher Hardware sie berechnet werden sollen. TF unterstützt dabei CPUs, GPGPUs[1] und eigens für ML entwickelte Hardware. Besonders umfangreich unterstützt das Framework Arbeiten im Bereich der (tiefen) Neuronales Netz. Schnittstellen zu den Hochsprachen Python und C++ sollen den Einstieg erleichtern und sicherstellen, dass die verfügbare Hardware immer bestmöglich genutzt werden kann. Mit dem sogenannte Tensorboard bringt TF außerdem eine Weboberfläche mit, die ohne großen Aufwand für den Entwickler viele relevante Informationen ausgibt und teilweise auch grafisch aufbereitet [16].

3.1 Eine Einführung zu TensorFlow

3.2 Die Entwicklung von Tensorflow

Bereits 2011 begann das Google Brain Projekt damit, den Nutzen von sehr großen tiefen Neuronales Netz zu erforschen. Einen Teil davon bildete der Aufbau von „DistBelief“, ein System, das Training und Vorhersage im Bereich des ML verteilt und skalierbar ermöglichte. Neben einigen Forschungsprojekten wurde DistBelief auch bereits produktive in einigen Google Produkten, wie Google Search, Google Tanslate oder Youtube eingesetzt [16]. 2012 wurde von Google Mitarbeitern ein Paper veröffentlicht, das über die Nutzung von zentausenden CPU-Kernen durch DistBelief berichtete, wodurch auch sehr große Modelle in absehbarer Zeit trainiert werden konnten [17]. Auf Basis der Erfahrungen beim Einsatz von DistBelief arbeitete Google dann am System der zweiten Generation für groß-skalierende ML Modelle und veröffentlichte im November 2015 TF [15]. TF ist seit dem auf github unter der Apache License 2.0 verfügbar [18]. Mit der Veröffentlichung

[1] general-purpose graphics processing units

von Version 1.0 im Februar 2017 wurde schließlich eine verlässliche API eingeführt, die auch in Zukunft sicher stellen soll, dass der geschriebene Code mit neuen Versionen von TF kompatibel ist [19]. Des weiteren wurde die Leistung weiter verbessert und die Einführung eines neuen Moduls ermöglicht seither die Nutzung von TF mit Keras[2].

3.3 Angesprochene Zielgruppe

Im Gegensatz zu DistBelief, welches viele Forschungsgebiete zu unflexibel war, ist TF sowohl für die Forschung als auch für den Produktiven Einsatz in großen Softwareprojekten geeignet. Abbildung 3.1 wurde beim TensorFlow Dev Summit 2017 [20] gezeigt und veranschaulicht die Abdeckung der Zielgruppen beider Systeme.

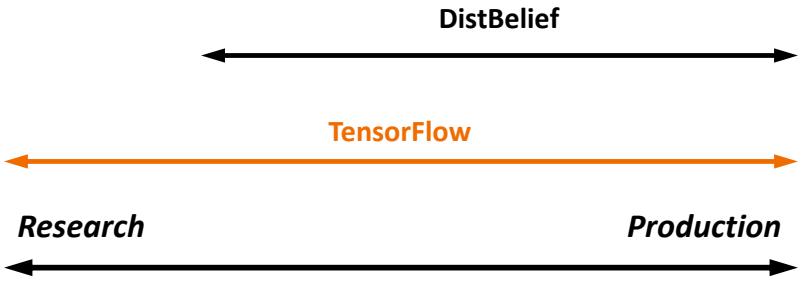

Abbildung 3.1: Zielgruppenabdeckung von DistBelief und TF [20]

TF bietet zum einen genug Flexibilität für Forschungsprojekte, um mit neuen Modellen zu experimentieren, ist gleichzeitig aber hochperformant und robust, was beim Einsatz in Produktivsoftware wichtig ist. Modelle können somit also oft aus der Forschung direkt in Produktivumgebungen übernommen werden [16].

3.4 Hard- und Software Anforderungen

3.4.1 Hardware Anforderungen

3.4.2 Software Anforderungen

3.5 Softwarearchitektur von TensorFlow

[2]Keras ist eine Deep Learning Library, die auf TF, CNTK oder Theano aufsetzt.

Kapitel 4

4 Der Allgemeine Workflow in TensorFlow

Zu Beginn dieses Kapitels wird die Methodik der Aufteilung der Datensätze in Trainings-, Validierungs- und Testdaten erläutert, welches für die Beurteilung der Ergebnisse eines Modells unerlässlich ist. Anschließend erfolgt eine Einführung in das Visualisierungstool TensorBoard, mit dessen Hilfe sich die hier vorgestellten und noch weitere Vorgänge anschaulich abbilden lassen. Am Schluss werden noch die einzelnen Graphelemente näher betrachtet, wodurch sich umfangreiche Graphabbildungen erzeugen lassen.

4.1 Vorgehensweise beim Trainingsprozess

Eines der entscheidenden Kriterien über die Qualität des Modells ist die Prognose der Zielgröße auf zuvor ungesehenen Daten. Die Genauigkeit auf den Daten, mit denen das Modell trainiert wurde, geben keine grundlegende Aussage über die zu erwartende Genauigkeit auf zukünftige Daten. Deshalb ist es essentiell wichtig ungenutzte Testdaten zur Beurteilung mit einzubeziehen. Häufig müssen Entscheidungen über Metaparameter wie Neuronenanzahl, Stärke der Regularisierung, Art des Kernels getroffen werden. Daher unterteilt man die Datensätze in Trainings-, Validierungs- und Testdaten.

- **Training**

  Trainingsdaten bilden die Grundlage für das überwachte Lernen. Hierbei werden durch eine Reihe von Beispielen Parameter wie die Gewichte der Verbindungen zwischen den Neuronen angepasst. Häufig besteht der Trainingsdatensatz aus Paaren von Eingangsvektoren und der dazugehörigen Antwortvektoren.

- **Validierung**

  Validierungsdaten dienen der Festlegung der optimalen Struktur des Modells, insbesondere zur Einstellung der optimalen Parameter des Lernalgorithmus wie die

Lernrate oder die Anzahl der Trainingsepochen. Validierungsdatensätze werden auch für die Regularisierung von "early stopping“ und bei der Zunahme des Fehlers im Validierungsdatensatz, da dies ein Zeichen für "overfitting“ ist, angewandt.

- **Testdaten**

  Testdaten werden verwendet, um eine Bewertung eines endgültigen Modells zu ermöglichen, welche auf ungesehene Daten angewandt wurde.

Cross-Validation

Die Unterteilung des Datensatzes in einen festen Trainingssatz und einen festen Testsatz kann problematisch sein, wenn der Testsatz zu klein ist. Dies impliziert statistische Unsicherheit im Bereich des geschätzten Testfehlers. Abhilfe schafft die k-fache Kreuzvalidierung (Cross-Validation), welche die Daten in k disjunkte Teilmengen unterteilt, auf denen k Tests durchgeführt werden, wobei jeweils k-1 Teilmengen für das Training und die verbliebene Teilmenge zum Testen verwendet wird.

4.2 Die Visualisierung mit TensorBoard

TensorBoard ist eine in TensorFlow enthaltene Webanwendung zur Visualisierung der Abläufe in einem neuronalen Netz. TensorBoard stellt eine Vielzahl an graphischen Elementen zur Verfügung, welche dem Entwickler vorallem das Debuggen oder die Optimierung eines erstellten Modells erleichtern. Ebenso können damit insbesondere komplexe Datenstrukturen zum besseren Verständnis anschaulich visualisiert werden.

Bevor mit TensorBoard gearbeitet werden kann, muss folgender Befehl 4.2 in die Kommandozeile eingegeben werden:

```
1 C:\> tensorboard --logdir=path/to/log_directory
```

wobei vorher im spezifizierten Log Ordner die gewünschten Event Daten mit der Klasse

```
1 tf.summary.FileWriter('path/to/log_directory')
```

abgespeichert werden müssen. Nachdem TensorBoard gestart wurde, navigiert man mit dem Browser zu folgender Seite:

```
1 http://localhost:6006
```

Falls keine Fehlermeldungen aufgetreten sind, sollte folgender Bildschirminhalt Abbildung 4.1 angezeigt werden. Um die Visualisierungen in den einzelnen Reiter zu erhalten, müssen diese vorher im Programm mit speziellen TensorFlow Klassen abgespeichert werden. In dem nachfolgenden Kapitel werden die einzelnen Reiter und der dazugehörigen Klasse vorgestellt.

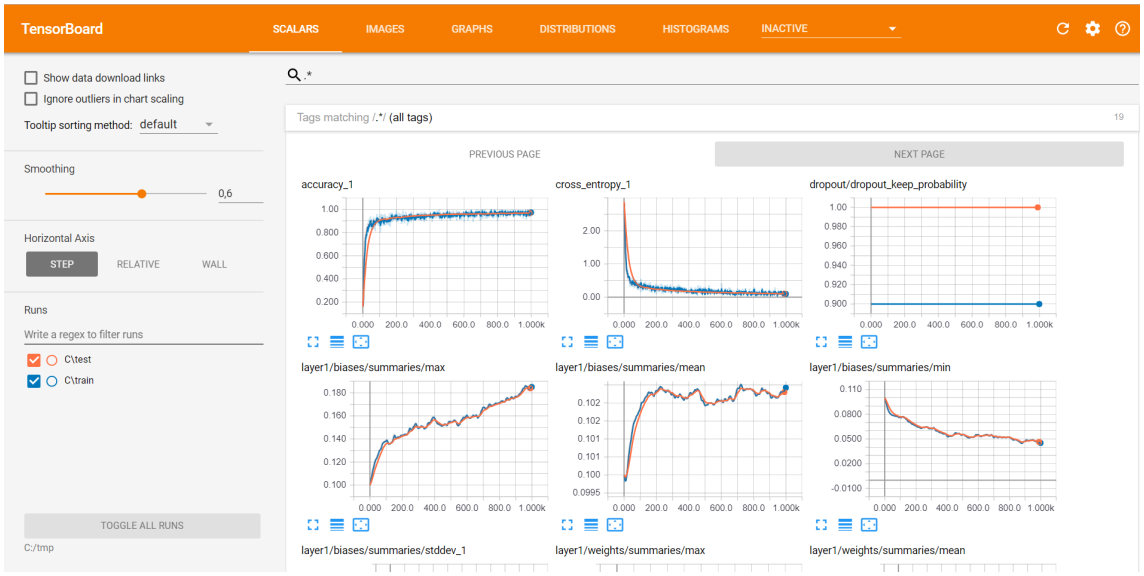

Abbildung 4.1: Die Startseite von TensorBoard

4.3 Die einzelnen Visualisierungsmöglichkeiten im Detail

4.3.1 Skalare

Unter dem Reiter Skalare, welche mit der Klasse

```
1 tf.summary.scalar(name, tensor, collections=None, family=None)
```

abgespeichert werden, können verschiedenste Statistiken während eines Trainingsprozesses visualisiert werden. Dies könnten zum Beispiel die Genauigkeit oder Cross-Entropie sein, welche in Abbildung 4.2 dargestellt sind. Hierbei wird die Genauigkeit über den einzelnen

Trainingsschritten aufgetragen. Wählt man mit der Maus einen bestimmten Datenpunkt aus, so werden zahlreiche weitere Informationen angezeigt. Ebenso ist hierbei ersichtlich, dass auch Trainings- und Testdaten gleichzeitig angezeigt und miteinander verglichen werden können.

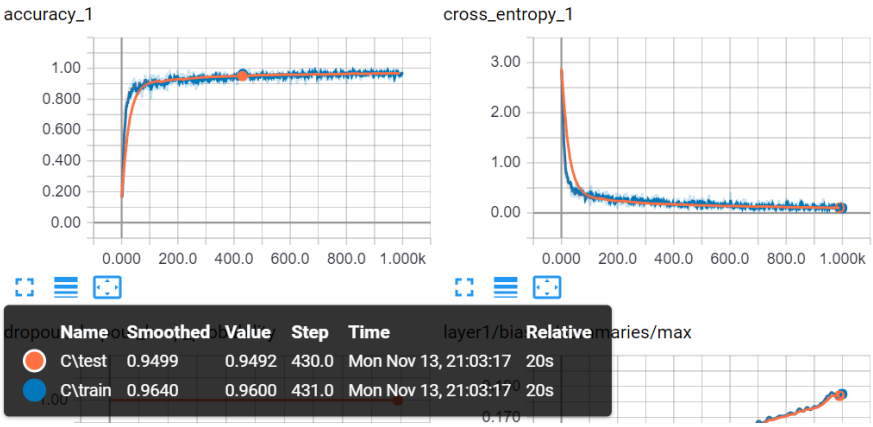

Abbildung 4.2: Visualisierung der 'Accuracy' und 'Cross_entropy' über die einzelnen Trainingschritte

4.3.2 Bilder

Innerhalb des Reiters Bilder, welche mit der Klasse

```
1 tf.summary.image(name, tensor, max_outputs=3, collections=None, family=
      None)
```

abgespeichert werden, können zur genaueren Analyse die Test- und Trainingsbilder eingesehen werden. Über den Bildern ist eine Scrollbar vorhanden mit dieser können einzelne Test- und Trainingsschritte ausgewählt werden, wodurch genau ersichtlich wird, welches Bild zum aktuellen Durchlauf gehört.

Abbildung 4.3: Bild des Testschrittes 490

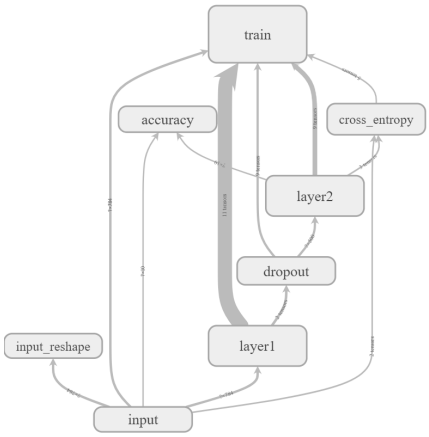

Abbildung 4.4: TensorFlow Graph mit definierten Name scopes

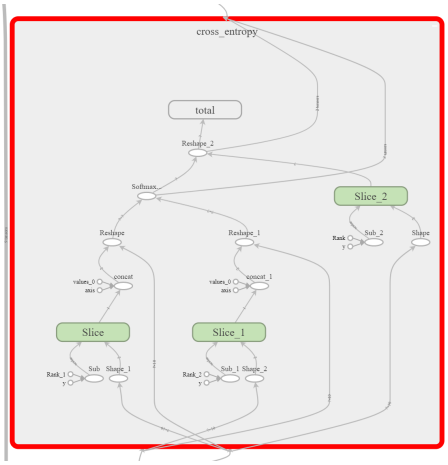

Abbildung 4.5: Name scope 'cross_entropy' unterteilt in weiteren Operationen

4.3.3 Graphen

Unter dem Reiter Graphen befindet sich das komplette TensorFlow Model als Graph wie in Abbildung 4.4 zu sehen ist. Um einen Graphen in Tensorboard zu erhalten, müssen im Programm die gewünschten Operationen als Name scope erstellt werden. Mit nachfolgendem Befehl **??** erhält man einen 'cross_entropy' Name scope:

```
1 with tf.name_scope('cross_entropy'):
```

Der Name scope *cross_entropy* Abbildung 4.5 kann natürlich wiederum in mehrere Untergruppierungen aufgeteilt werden. So erhält man eine übersichtliche Visualisierung komplexer TensorFlow Modelle.

4.3.4 Histogramme

Unter dem Reiter Histogramme, welche mit der Klasse

```
1 tf.summary.histogram(name, values, collections=None, familiy=None)
```

abgespeichert werden, wird die statistische Verteilung eines Tensors über der Zeit dargestellt. Im Histogramm sind zeitliche "Slices“ der Daten visualisiert, wobei jeder einzelne Slice ein Histogramm des Tensors in einem einzelnen Schritt darstellt. In Abbildung 4.9 ist ein einzelner Slice schwarz markiert.

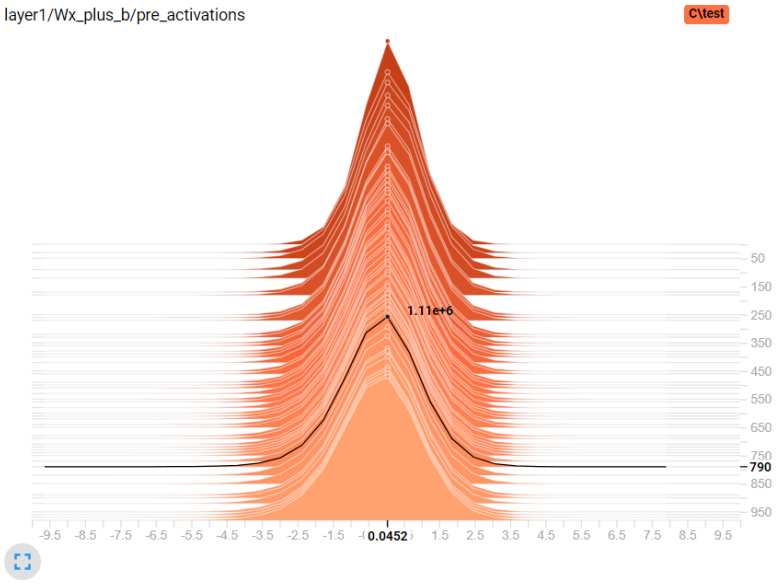

Abbildung 4.6: Visualisierung eines Histogramms über die einzelnen Trainingschritte

4.3.5 Verteilungen

Unter dem Reiter Verteilungen, welche ebenfalls mit der Klasse

```
1 tf.summary.histogram(name, values, collections=None, familiy=None)
```

abgespeichert werden, befindet sich eine weitere Möglichkeit der Visualisierung der statistischen Verteilung. Hierbei repräsentiert die oberste Linie den über die Zeit veränderten maximalen Wert, die unterste Linie den minimalen Wert und die mittlere Linie den veränderten Median über der Zeit.

4.3.6 Projektor

Die Auswahl des Reiters Projektors ermöglicht die höherdimensionale Visualisierung der Eingabedaten. Nachfolgend eine mögliche Konfiguration des Projektors:

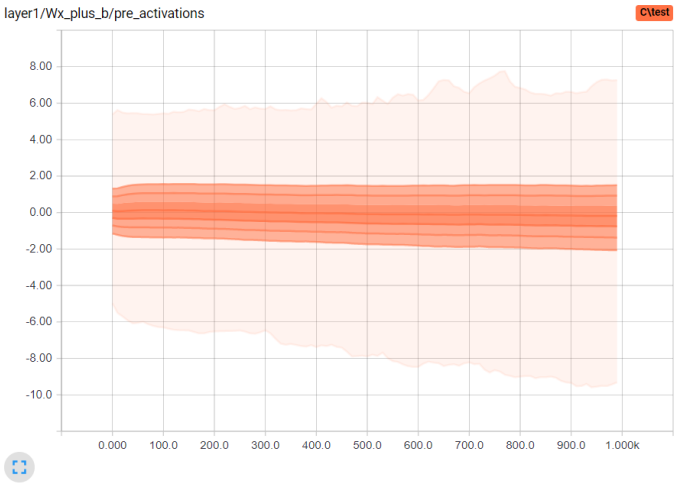

Abbildung 4.7: Eine weitere Möglichkeit der Visualisierung der statistischen Verteilung

Listing 4.1: Konfiguration eines Projektors in TensorFlow

```
1  from tensorflow.contrib.tensorboard.plugins import projector
2
3  # Use the same LOG_DIR where you stored your checkpoint.
4  summary_writer = tf.train.SummaryWriter(LOG_DIR)
5
6  # Format: tensorflow/contrib/tensorboard/plugins/projector/
       projector_config.proto
7  config = projector.ProjectorConfig()
8
9  # You can add multiple embeddings. Here we add only one.
10 embedding = config.embeddings.add()
11 embedding.tensor_name = embedding_var.name
12
13 # Link this tensor to its metadata file (e.g. labels).
14 embedding.metadata_path = os.path.join(LOG_DIR, 'metadata.tsv')
15
16 # Saves a configuration file that TensorBoard will read during startup.
17 projector.visualize_embeddings(summary_writer, config)
```

Hierbei werden zwei wesentliche Darstellungen unterschieden.

- PCA (Prinicipal Component Analysis)

Ein häufiges Problem bei multivariaten Daten ist, dass diese nicht im zweidimensionalem Raum dargestellt werden können. Hierbei werden bei der Hauptkomponentenanalyse (PCA) die Daten so auf eine zweidimensionale (bzw. dreidimensionaler) Ebene projiziert, mit der Erwartung, dass diese neue Darstellung eventuell vorhandenes Rauschen herausfiltert und versteckte Strukturen zum Vorschein bringt.

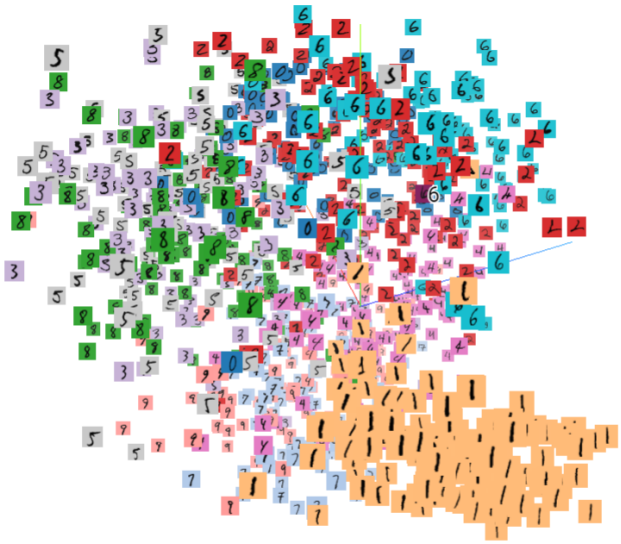

Abbildung 4.8: Visualisierung des PCA

- T-SNE (t-distributed stochastic neighbor embedding)

  Diese Technik der Dimensionsreduktion eignet sich besonders gut, um hochdimensionale Daten in einen Raum von zwei oder drei Dimensionen zu projizieren. Hierbei wird jedes hochdimensionale Objekt durch einen zwei- oder dreidimensionalen Punkt modelliert, sodass ähnliche Objekte durch nahe gelegene Punkte und ungleiche Objekte durch entfernte Punkte modelliert werden.

4.3.7 Audio und Text

Unter dem Reiter Audio können mit der Klasse

```
1 tf.summary.audio(name, tensor, sample_rate, max_outputs=3, collections=
      None, family=None)
```

abspielbare Audio-Widgets eingebettet werden.

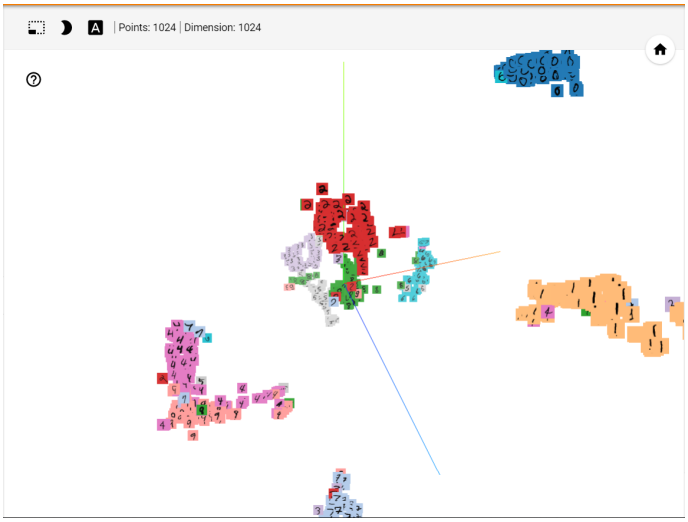

Abbildung 4.9: Visualisierung des T-SNE

Ebenso können unter dem Reiter Text mit der Klasse

```
1 tf.summary.text(name, tensor, collections=None)
```

Textausschnitte abgespeichert werden. Zusätzliche Funktionen wie Hyperlinks, Listen und Tabellen werden unterstützt.

4.4 Die Graphelemente im Datenfluss

Dafür wurden zahlreiche unterschiedliche Elemente definiert, welche nachfolgend kurz erläutert werden.

Namespace

High-level Knoten repräsentiert einen definierten Namensbereich.

Unconnected series

Nummerierte Knoten, die nicht miteinander verbunden sind.

Connected series

Nummerierte Knoten, die miteinander verbunden sind.

Operation node

Ein Knoten der eine einzelne Operation darstellt.

Constant

Repräsentiert eine Konstante im Programm.

Summary node

Dieser Knoten stellt eine Zusammenfassung dar.

Dataflow edge

Durchgezogener Pfeil zeigt den Datenfluss zwischen den Operationen an.

Control edge

Gepunkteter Pfeil zeigt die Steuerungsabhängigkeit zwischen den Operationen an.

Reference edge

Gelber Pfeil bedeutet, dass die ausgehende Operation die ankommende mutieren kann

Kapitel 5

5 Ausblick

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

Literaturverzeichnis

[1] Russell, Stuart ; Norvig, Peter ; Kirchner, Frank: *Künstliche Intelligenz: Ein moderner Ansatz*. 3., aktualisierte Aufl. München : Pearson Higher Education, 2012 (Always learning). – ISBN 978–3–86894–098–5

[2] TURING, A. M.: I.—COMPUTING MACHINERY AND INTELLIGENCE. In: *Mind* LIX (1950), Nr. 236, S. 433–460. http://dx.doi.org/10.1093/mind/LIX.236.433. – DOI 10.1093/mind/LIX.236.433. – ISSN 0026–4423

[3] Russell, Stuart ; Dewey, Daniel ; Tegmark, Max: Research Priorities for Robust and Beneficial Artificial Intelligence. (2015), jan. http://futureoflife.org/data/documents/research_priorities.pdf

[4] Barrat, James: *Our Final Invention: Artificial Intelligence and the End of the Human Era*. THOMAS DUNNE BOOKS, 2013 http://www.ebook.de/de/product/20253628/james_barrat_our_final_invention_artificial_intelligence_and_the_end_of_the_human_era.html. – ISBN 0312622376

[5] Goodfellow, Ian ; Bengio, Yoshua ; Courville, Aaron: *Deep Learning*. MIT Press, 2016. – http://www.deeplearningbook.org

[6] Bishop, Christopher M.: *Neural Networks for Pattern Recognition*. Oxford : Clarendon Press, 1995. – ISBN 978–0–198–53864–6

[7] Géron, A.: *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017 https://books.google.de/books?id=bRpYDgAAQBAJ. – ISBN 9781491962244

[8] Rashid, Tariq: *Make Your Own Neural Network*. 1st. USA : CreateSpace Independent Publishing Platform, 2016. – ISBN 1530826608, 9781530826605

[9] McClure, N.: *TensorFlow Machine Learning Cookbook*. Packt Publishing, 2017 https://books.google.de/books?id=LVQoDwAAQBAJ. – ISBN 9781786466303

[10] Ertel, Wolfgang: *Grundkurs Künstliche Intelligenz - Eine praxisorientierte Einführung*. Berlin Heidelberg New York : Springer-Verlag, 2013. – ISBN 978–3–834–82157–7

[11] Rojas, Raul ; Varga, Peter: *Neural Networks - A Systematic Introduction*. Berlin Heidelberg : Springer Science, Business Media, 1996. – ISBN 978–3–540–60505–8

[12] Bonnin, R.: *Building Machine Learning Projects with TensorFlow*. Packt Publishing, 2016 https://books.google.de/books?id=pZ3cDgAAQBAJ. – ISBN 9781786466822

[13] Michael, Nielsen: Improving the way neural networks learn. (2017), Aug. http://neuralnetworksanddeeplearning.com/chap3.html

[14] Abadi, Martín ; Agarwal, Ashish ; Barham, Paul ; Brevdo, Eugene ; Chen, Zhifeng ; Citro, Craig ; Corrado, Greg S. ; Davis, Andy ; Dean, Jeffrey ; Devin, Matthieu ; Ghemawat, Sanjay ; Goodfellow, Ian ; Harp, Andrew ; Irving, Geoffrey ; Isard, Michael ; Jia, Yangqing ; Jozefowicz, Rafal ; Kaiser, Lukasz ; Kudlur, Manjunath ; Levenberg, Josh ; Mané, Dan ; Monga, Rajat ; Moore, Sherry ; Murray, Derek ; Olah, Chris ; Schuster, Mike ; Shlens, Jonathon ; Steiner, Benoit ; Sutskever, Ilya ; Talwar, Kunal ; Tucker, Paul ; Vanhoucke, Vincent ; Vasudevan, Vijay ; Viégas, Fernanda ; Vinyals, Oriol ; Warden, Pete ; Wattenberg, Martin ; Wicke, Martin ; Yu, Yuan ; Zheng, Xiaoqiang: *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/. Version: 2015. – Software available from tensorflow.org

[15] Dean, Jeff ; Monga, Rajat: *TensorFlow - Google's latest machine learning system, open sourced for everyone*. https://research.googleblog.com/2015/11/tensorflow-googles-latest-machine_9.html, . – [letzter Zugriff: 28. Nov 2017]

[16] Abadi, Martín ; Barham, Paul ; Chen, Jianmin ; Chen, Zhifeng ; Davis, Andy ; Dean, Jeffrey ; Devin, Matthieu ; Ghemawat, Sanjay ; Irving, Geoffrey ; Isard, Michael ; Kudlur, Manjunath ; Levenberg, Josh ; Monga, Rajat ; Moore, Sherry ; Murray, Derek G. ; Steiner, Benoit ; Tucker, Paul ; Vasudevan, Vijay ; Warden, Pete ; Wicke, Martin ; Yu, Yuan ; Zheng, Xiaoqiang: *TensorFlow: A system for large-scale machine learning*. https://www.tensorflow.org/. Version: 2016. – Software available from tensorflow.org

[17] Dean, Jeffrey ; Corrado, Greg ; Monga, Rajat ; Chen, Kai ; Devin, Matthieu ; Mao, Mark ; Ranzato, Marc'aurelio ; Senior, Andrew ; Tucker, Paul ; Yang, Ke ; Le, Quoc V. ; Ng, Andrew Y.: Large Scale Distributed Deep Networks. Version: 2012. http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf. In: Pereira, F. (Hrsg.) ; Burges, C. J. C. (Hrsg.) ; Bottou, L. (Hrsg.) ; Weinberger, K. Q.

(Hrsg.): *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, 1223–1231

[18] *tensorflow, Computation using data flow graphs for scalable machine learning*. https://github.com/tensorflow/tensorflow, . – [letzter Zugriff: 28. Nov 2017]

[19] Sandjideh, Amy M.: *Announcing TensorFlow 1.0*. https://developers.googleblog.com/2017/02/announcing-tensorflow-10.html, . – [letzter Zugriff: 29. Nov 2017]

[20] Jeff Dean, Rajat M. ; Kacholia, Megan: *Keynote (TensorFlow Dev Summit 2017)*. https://www.youtube.com/watch?v=4n1AHvDvVvw, . – [letzter Zugriff: 01. Dez 2017]

[21] Jeffrey Dean, Rajat Monga Kai Chen Matthieu Devin Quoc V. Le Mark Z. Mao Marc'Aurelio Ranzato Andrew Senior Paul Tucker Ke Yang Andrew Y. N. Greg S. Corrado C. Greg S. Corrado: *Large Scale Distributed Deep Networks* . https://static.googleusercontent.com/media/research.google.com/de//pubs/archive/40565.pdf, . – [letzter Zugriff: 28. Nov 2017]

[22] *TensorBoard*. https://github.com/tensorflow/tensorboard. – Zuletzt besucht am 25.11.2017