

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 Dännweber
 OTH Regensburg

Regensburg, 21. Dezember 2017

Abstract

Abkürzungsverzeichnis

KI	Künstliche Intelligenz.
ML	Machine Learning, bzw. Maschinelles Lernen.
NN	Neuronale Netze.
TF	TensorFlow.
TPU	Tensor Processing Unit.

Inhaltsverzeichnis

Abkürzungsverzeichnis	iii
1 Eine Einführung in Maschinelles Lernen	1
1.1 Meilensteine des Maschinellen Lernens	1
1.2 Grundlagen von Neuronalen Netzen	3
2 Neuronale Netze in Tensorflow	4
2.1 Feedforwardnetzwerk	4
2.2 Der Input Tensor	5
2.3 Gewichte	6
2.4 Aktivierungsfunktionen	7
2.4.1 Rectified linear unit function	7
2.4.2 Sigmoidfunktion	8
2.4.3 Tangenshyperbolicus	8
2.5 Kostenfunktionen	9
2.5.1 MSE	9
2.5.2 cross entropy	10
2.6 Lernprozess	10
3 Das Framework TensorFlow	11
3.1 Eine Einführung zu TensorFlow	11
3.2 Die Entwicklung von Tensorflow	11
3.3 Angesprochene Zielgruppe	12
3.4 Hard- und Software Anforderungen	12
3.4.1 Betriebssysteme	12
3.4.2 Hardware Unterstützung	13
3.4.3 Software Anforderungen	14
3.5 Softwarearchitektur von TensorFlow	14
4 Der Allgemeine Workflow in TensorFlow	15
4.1 Die Tooling Pipeline	15

4.2	Vorgehensweise beim Trainingsprozess	15
4.3	Die Visualisierung mit TensorBoard	16
4.3.1	Die einzelnen Visualisierungsmöglichkeiten im Detail	17
4.3.2	Die Graphenelemente im Datenfluss	19
5	Ausblick	21
	Abbildungsverzeichnis	22
	Tabellenverzeichnis	23
	Anhang	24
	Literaturverzeichnis	25

1

Kapitel 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, geben dieses und das nächste 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, können dieses Kapitel auch übersprungen 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 Komplexität¹ des Spiels galt der Sieg einer Maschine über einen realen Meister lange Zeit als unmöglich [1].

¹Ein 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.

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 [3].

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.

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. Rechengvorgänge, 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.

1.2 Grundlagen von Neuronalen Netzen

2 Kapitel 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.

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

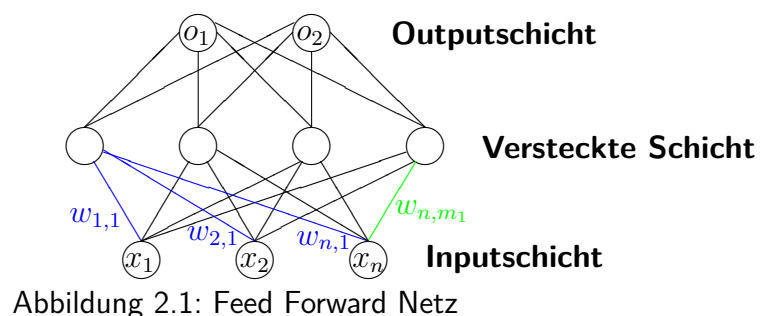


Abbildung 2.1: Feed Forward Netz

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[4] S.168

²[5]s.117

³[6] S.385

nicht zu wenige Neuronen zu verwenden da sonst die Lernkapazität möglicherweise 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-Tensoren 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[7] S.341

⁵[6] S.384

⁶[6] S.124

⁷[6] S.99

⁸[8] S.218

⁹[6] S.377

2.3 Gewichte

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

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

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[6] S.377

¹¹[6] S.377

¹²[9] 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 Form¹³

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

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)
```

[8] Für diese gilt:

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

Sie kann schneller berechnet werden und hat den Vorteil dass weder Werte nahe der Null verschwinden, noch die Werte zu groß werden können.[8]

¹³[4]S.174

¹⁴[4]S.175

¹⁵[8]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:[10]

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

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

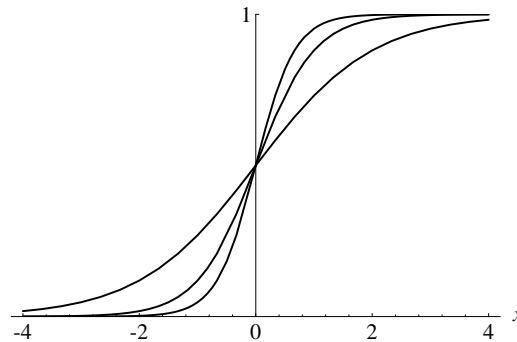


Abbildung 2.2: Sigmoidfunktion für $c=1$ [10]

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 als¹⁸

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

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[10] S.151f.

¹⁷[11]S.175

¹⁸[5]S.127

¹⁹[11]S.175

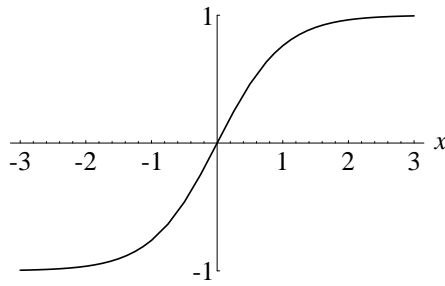


Abbildung 2.3: Graph des Tangenshyperbolicus [10]

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

2.5 Kostenfunktionen

Kosten- oder Fehlerfunktionen sind ein Maß dafür wie gut ein NN 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 NN 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.

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

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[4]S.177f.

²¹[10]S.156

²²[8]S.94f.

2.5.2 cross entropy

Eine andere Möglichkeit ist die sogenannte Cross Entropy Funktion.²³

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

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 Gradientenabstiegsverfahren genannt.

Das Update der Gewichte funktioniert nach folgender Regel:²⁶

$$w_{i,j} = w_{i,j} - \frac{\partial \text{Kosten}}{\partial w_{i,j}} \eta, \quad (2.6)$$

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[12]

²⁴[12]

²⁵[8]S.88

²⁶[10]S.157

²⁷[11]S.156

3 Kapitel 3

Das Framework TensorFlow

TensorFlow (TF) ist eine plattformunabhängige Bibliothek für ML, die für große und variable Architekturen entwickelt [13] und Ende 2015 veröffentlicht wurde [14]. 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, GPGPUs¹ und eigens für ML entwickelte Hardware. Besonders umfangreich unterstützt das Framework Arbeiten im Bereich der (tiefen) NN. Schnittstellen zu den Hochsprachen Python und C++ sollen den Einstieg erleichtern und sicherstellen, dass die verfügbare Hardware immer bestmöglich genutzt werden kann. Mit dem sogenannten Tensorboard bringt TF außerdem eine Weboberfläche mit, die ohne großen Aufwand für den Entwickler viele relevante Informationen ausgibt und teilweise auch grafisch aufbereitet [15].

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 NN 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 Translate oder Youtube eingesetzt [15]. 2012 wurde von Google Mitarbeitern ein Paper veröffentlicht, das über die Nutzung von zehntausenden CPU-Kernen durch DistBelief berichtete, wodurch auch sehr große Modelle in absehbarer Zeit trainiert werden konnten [16]. 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 [14]. TF ist seit dem auf github unter der Apache License 2.0 verfügbar [17]. Mit der Veröffentlichung von Version 1.0 im Februar 2017 wurde schließlich eine verlässliche API eingeführt, die

¹general-purpose graphics processing units

auch in Zukunft sicher stellen soll, dass der geschriebene Code mit neuen Versionen von TF kompatibel ist [18]. Des weiteren wurde die Leistung weiter verbessert und die Einführung eines neuen Moduls ermöglicht seither die Nutzung von TF mit Keras².

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 [19] gezeigt und veranschaulicht die Abdeckung der Zielgruppen beider Systeme.

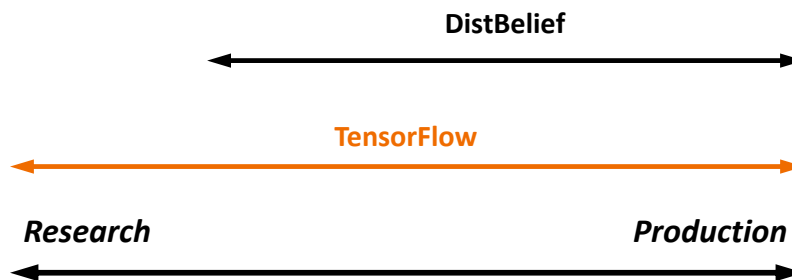


Abbildung 3.1: Zielgruppenabdeckung von DistBelief und TF [19]

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 [15].

3.4 Hard- und Software Anforderungen

Tensorflow wurde von Grund auf so gestaltet, dass es eine möglichst großen Bandbreite an Hardware zur Berechnung der NN nutzen kann und auch größtenteils unabhängig von dem System ist, auf dem es ausgeführt wird.

3.4.1 Betriebssysteme

TF ist seit dem Anfang auf Linux und MacOS lauffähig. Seit Version 0.12 wird außerdem Windows unterstützt [20]. Die reine Ausführung eines bereits trainierten Netzes (Inference) ist außerdem auch auf Android und iOS möglich [13][15].

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

3.4.2 Hardware Unterstützung

Lernalgorithmen basieren oft auf aufwändigen Rechenoperationen, die zwar hoch parallelisierbar sind, meist jedoch auch große Abhängigkeiten beim Datenzugriff aufweisen. Ein Beispiel hierfür sind Matrix Multiplikationen. Mit dem Aufkommen von so genannten „General Purpose GPUs“ die eine große Anzahl an Recheneinheiten und schnellen lokalen Speicher bieten, konnten NN stark beschleunigt werden.

TF bietet hier den Vorteil, dass es unabhängig von der vorhandenen Hardware eingesetzt werden kann. Es werden sowohl CPUs als auch eine Vielzahl von NVIDIA GPUs unterstützt. Im Moment können GPUs anderer Hersteller noch nicht genutzt werden, da glsTF die CUDA-API von NVIDIA nutzt. Diese „CUDA Compute Capability“ muss von der Hardware auch mindestens in Version 3.0 unterstützt werden. Eine Liste mit allen unterstützten Karten kann auf der [NVIDIA Homepage](#) gefunden werden [21].

Neben diesen allgemein gehaltenen Recheneinheiten kann TF auch spezielle Hardware für die Berechnungen nutzen. So gibt es von Google selbst sogenannte Tensor Processing Units (TPUs), die eine enorme Beschleunigung von Inference ermöglichen und dabei vor allem wesentlich energieeffizienter sind, als CPUs oder auch GPUs [22]. Da es sich hierbei jedoch um proprietäre Hardware handelt, können diese TPUs im Moment nur als Service in der Google Cloud genutzt werden [23]. Außerdem kann der sogenannte „Neural Compute Stick“ von Movidius (inzwischen Intel) [24] verwendet werden, der es auch auf leistungsschwacher Hardware wie zum Beispiel einem Raspberry Pi möglich macht, Inference von NN durchzuführen.

In Tensorflow kann genau definiert werden, welche Schritte im Programmcode von welcher Recheneinheit ausgeführt werden sollen. So wird in Zeile zwei von Listing 3.1 angegeben, dass die dritte GPU im System verwendet werden soll³ [25].

Listing 3.1: Festlegung des Geräts, auf dem die Berechnung getätigt werden soll

```
1 # Creates a graph.
2 with tf.device('/device:GPU:2'):
3     a = tf.constant([1.0, 2.0, 3.0, 4.0, 5.0, 6.0], shape=[2, 3], name='a')
4     b = tf.constant([1.0, 2.0, 3.0, 4.0, 5.0, 6.0], shape=[3, 2], name='b')
5     c = tf.matmul(a, b)
6 # Creates a session with log_device_placement set to True.
7 sess = tf.Session(config=tf.ConfigProto(log_device_placement=True))
8 # Runs the op.
9 print(sess.run(c))
```

³TF beginnt bei null zu zählen

In den Zeilen drei und vier werden zwei konstante Arrays definiert und in Zeile fünf wird die durchzuführende Rechenoperation festgelegt. In Zeile sieben wird eine TensorFlow Session erstellt und mit „sess.run“ in Zeile neun wird schließlich die Berechnung angestoßen.

3.4.3 Software Anforderungen

TF ist sowohl für Python 2.7 verfügbar, als auch für Python 3 ab Version 3.4 [21]. Alternativ sind aber auch Docker Images verfügbar.

Für die Verwendung von NVIDIA GPUs muss außerdem das CUDA® Toolkit in Version 8.0 sowie cuDNN v6.0 installiert werden [21].

3.5 Softwarearchitektur von TensorFlow

4 Kapitel 4 **Der Allgemeine Workflow in TensorFlow**

4.1 Die Tooling Pipeline

4.2 Vorgehensweise beim Trainingsprozess

4.3 Die Visualisierung mit TensorBoard

TensorBoard ist eine in der Installation von 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 4.1 in die Kommandozeile eingegeben werden:

```
tensorboard --logdir = path/to/log_directory (4.1)
```

wobei vorher im spezifizierten Log Ordner die gewünschten Event Daten mit der Klasse `tf.summary.FileWriter('path/to/log-directory')` abgespeichert werden müssen. Nachdem TensorBoard gestartet wurde, navigiert man mit dem Browser zu folgender Seite:

```
http : //localhost : 6006 (4.2)
```

Falls keine Fehlermeldungen aufgetreten sind, sollte folgender Bildschirminhalt angezeigt werden:

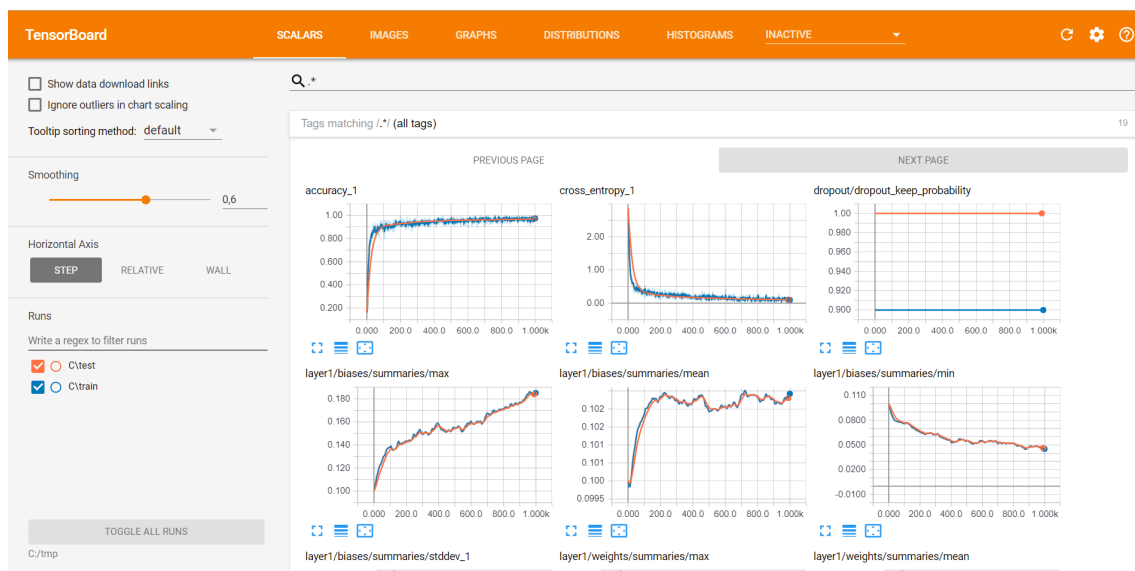


Abbildung 4.1: Die Startseite von TensorBoard

4.3.1 Die einzelnen Visualisierungsmöglichkeiten im Detail

Skalare

Unter dem Reiter Skalare, welche mit der Klasse "tf.summary.scalar" 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.

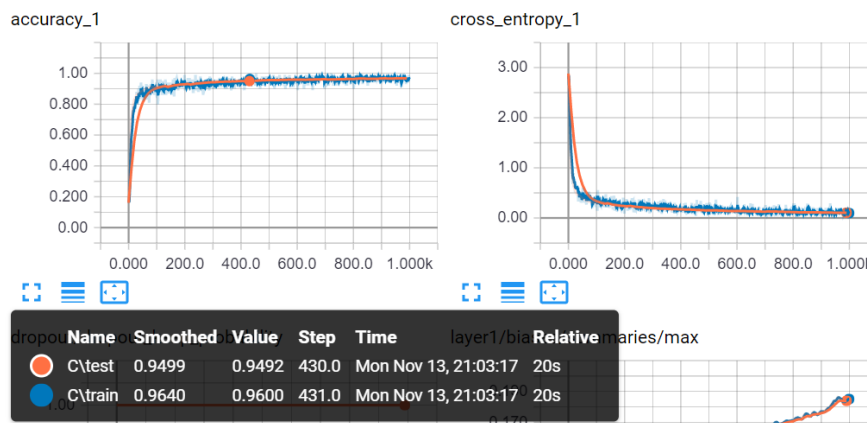


Abbildung 4.2: Visualisierung der 'Accuracy' und 'Cross_entropy' über die einzelnen Trainingschritte

Bilder

Innerhalb des Reiters Bilder, welche mit der Klasse "tf.summary.image" 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

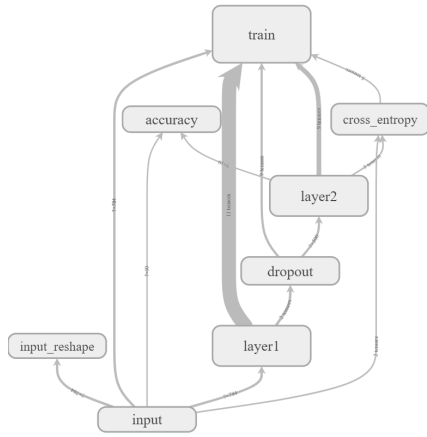


Abbildung 4.4: TensorFlow Graph mit definierten Name scopes

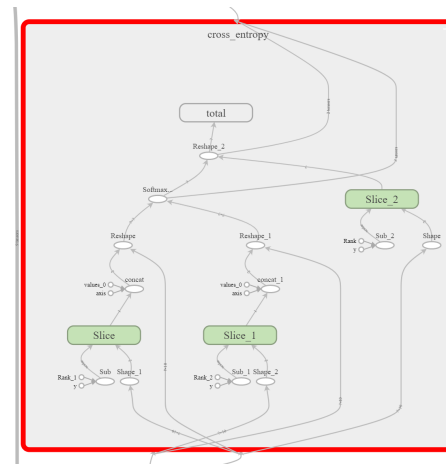


Abbildung 4.5: Name scope 'cross_entropy' unterteilt in weiteren Operationen

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 4.3 erhält man einen 'cross_entropy' Name scope:

with tf.name_scope('cross_entropy') : (4.3)

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.

Histogramme

Unter dem Reiter Histogramme, welche mit der Klasse "tf.summary.histogram" 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 4.6 ist ein einzelner Slice schwarz markiert.

Verteilungen

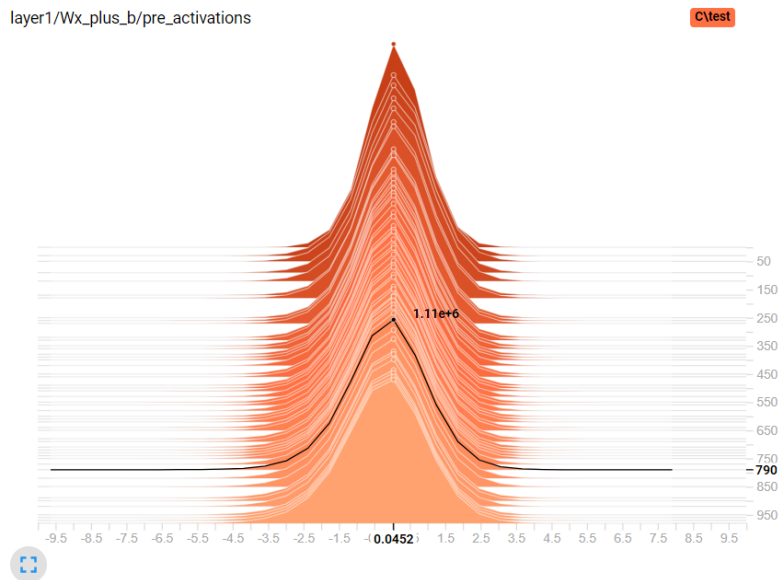
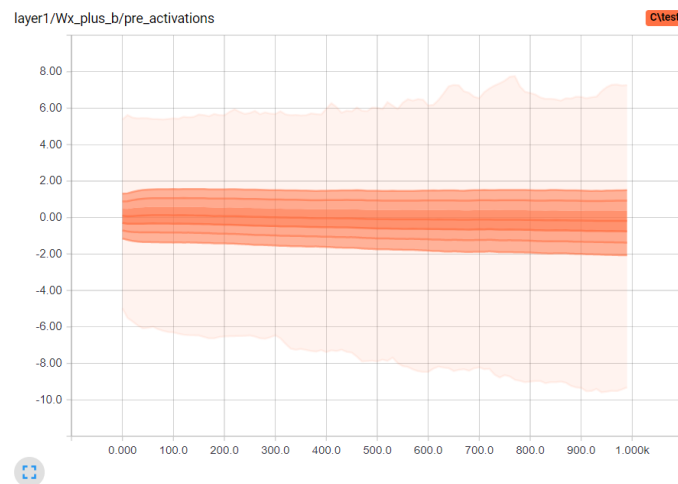


Abbildung 4.6: Visualisierung eines Histogramms über die einzelnen Trainingschritte



Audio

Projektor

Text

4.3.2 Die Graphenelemente 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

5

Kapitel 5

Ausblick

Abbildungsverzeichnis

2.1	Feed Forward Netz	4
2.2	Sigmoidfunktion für $c=1$ [10]	8
2.3	Graph des Tangenshyperbolicus [10]	9
3.1	Zielgruppenabdeckung von DistBelief und TF [19]	12
4.1	Die Startseite von TensorBoard	16
4.2	Visualisierung der 'Accuracy' und 'Cross_entropy' über die einzelnen Trainingsschritte	17
4.3	Bild des Testschrittes 490	17
4.4	TensorFlow Graph mit definierten Name scopes	18
4.5	Name scope 'cross_entropy' unterteilt in weiteren Operationen	18
4.6	Visualisierung eines Histogramms über die einzelnen Trainingsschritte	19

Tabellenverzeichnis

Anhang

Literaturverzeichnis

- [1] LEE, C. S. ; WANG, M. H. ; YEN, S. J. ; WEI, T. H. ; WU, I. C. ; CHOU, P. C. ; CHOU, C. H. ; WANG, M. W. ; YAN, T. H.: Human vs. Computer Go: Review and Prospect [Discussion Forum]. In: *IEEE Computational Intelligence Magazine* 11 (2016), Aug, Nr. 3, S. 67–72. <http://dx.doi.org/10.1109/MCI.2016.2572559>. – DOI 10.1109/MCI.2016.2572559. – ISSN 1556–603X
- [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 ; 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
- [4] GOODFELLOW, Ian ; BENGIO, Yoshua ; COURVILLE, Aaron: *Deep Learning*. MIT Press, 2016. – <http://www.deeplearningbook.org>
- [5] BISHOP, Christopher M.: *Neural Networks for Pattern Recognition*. Oxford : Clarendon Press, 1995. – ISBN 978–0–198–53864–6
- [6] 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
- [7] RASHID, Tariq: *Make Your Own Neural Network*. 1st. USA : CreateSpace Independent Publishing Platform, 2016. – ISBN 1530826608, 9781530826605
- [8] MCCLURE, N.: *TensorFlow Machine Learning Cookbook*. Packt Publishing, 2017 <https://books.google.de/books?id=LVQoDwAAQBAJ>. – ISBN 9781786466303
- [9] ERTEL, Wolfgang: *Grundkurs Künstliche Intelligenz - Eine praxisorientierte Einführung*. Berlin Heidelberg New York : Springer-Verlag, 2013. – ISBN 978–3–834–82157–7
- [10] ROJAS, Raul ; VARGA, Peter: *Neural Networks - A Systematic Introduction*. Berlin Heidelberg : Springer Science, Business Media, 1996. – ISBN 978–3–540–60505–8

- [11] BONNIN, R.: *Building Machine Learning Projects with TensorFlow*. Packt Publishing, 2016 <https://books.google.de/books?id=pZ3cDgAAQBAJ>. – ISBN 9781786466822
- [12] MICHAEL, Nielsen: Improving the way neural networks learn. (2017), Aug. <http://neuralnetworksanddeeplearning.com/chap3.html>
- [13] 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
- [14] 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]
- [15] 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
- [16] 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.) ; BURGESS, C. J. C. (Hrsg.) ; BOTTOU, L. (Hrsg.) ; WEINBERGER, K. Q. (Hrsg.): *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, 1223–1231

- [17] *tensorflow*, *Computation using data flow graphs for scalable machine learning*. <https://github.com/tensorflow/tensorflow>, . – [letzter Zugriff: 28. Nov 2017]
- [18] SANDJIDEH, Amy M.: *Announcing TensorFlow 1.0*. <https://developers.googleblog.com/2017/02/announcing-tensorflow-10.html>, . – [letzter Zugriff: 29. Nov 2017]
- [19] JEFF DEAN, Rajat M. ; KACHOLIA, Megan: *Keynote (TensorFlow Dev Summit 2017)*. <https://www.youtube.com/watch?v=4n1AHvDvVvw>, . – [letzter Zugriff: 01. Dez 2017]
- [20] *TensorFlow v0.12.0 RC0*. <https://github.com/tensorflow/tensorflow/releases/tag/0.12.0-rc0>, . – [letzter Zugriff: 19. Dez 2017]
- [21] *Installing TensorFlow on Ubuntu*. https://www.tensorflow.org/install/install_linux, . – [letzter Zugriff: 19. Dez 2017]
- [22] JOUPPI, Norman P. ; YOUNG, Cliff ; PATIL, Nishant ; PATTERSON, David ; AGRAWAL, Gaurav ; BAJWA, Raminder ; BATES, Sarah ; BHATIA, Suresh ; BODEN, Nan ; BORCHERS, Al ; BOYLE, Rick ; CANTIN, Pierre-luc ; CHAO, Clifford ; CLARK, Chris ; CORIELL, Jeremy ; DALEY, Mike ; DAU, Matt ; DEAN, Jeffrey ; GELB, Ben ; GHAEMMAGHAMI, Tara V. ; GOTTIPATI, Rajendra ; GULLAND, William ; HAGMANN, Robert ; HO, Richard C. ; HOGBERG, Doug ; HU, John ; HUNDT, Robert ; HURT, Dan ; IBARZ, Julian ; JAFFEY, Aaron ; JAWORSKI, Alek ; KAPLAN, Alexander ; KHAITAN, Harshit ; KOCH, Andy ; KUMAR, Naveen ; LACY, Steve ; LAUDON, James ; LAW, James ; LE, Diemthu ; LEARY, Chris ; LIU, Zhuyuan ; LUCKE, Kyle ; LUNDIN, Alan ; MACKEAN, Gordon ; MAGGIORE, Adriana ; MAHONY, Maire ; MILLER, Kieran ; NAGARAJAN, Rahul ; NARAYANASWAMI, Ravi ; NI, Ray ; NIX, Kathy ; NORRIE, Thomas ; OMERNICK, Mark ; PENUKONDA, Narayana ; PHELPS, Andy ; ROSS, Jonathan ; SALEK, Amir ; SAMADIANI, Emad ; SEVERN, Chris ; SIZIKOV, Gregory ; SNELHAM, Matthew ; SOUTER, Jed ; STEINBERG, Dan ; SWING, Andy ; TAN, Mercedes ; THORSON, Gregory ; TIAN, Bo ; TOMA, Horia ; TUTTLE, Erick ; VASUDEVAN, Vijay ; WALTER, Richard ; WANG, Walter ; WILCOX, Eric ; YOON, Doe H.: *In-Datacenter Performance Analysis of a Tensor Processing Unit*. In: *CoRR* abs/1704.04760 (2017). <http://arxiv.org/abs/1704.04760>
- [23] JEFF DEAN, Urs H.: *Installing TensorFlow on Ubuntu*. <https://blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/>, . – [letzter Zugriff: 21. Dez 2017]

- [24] *Introducing: TensorFlow™ Support for Neural Compute Stick.* <https://www.movidius.com/news/introducing-tensorflow-support-for-neural-compute-stick>, . – [letzter Zugriff: 19. Dez 2017]
- [25] *Using GPUs / TensorFlow.* https://www.tensorflow.org/tutorials/using_gpu, . – [letzter Zugriff: 21. Dez 2017]
- [26] 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
- [27] 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]
- [28] RUSSELL, Stuart ; DEWEY, Daniel ; TEGMARK, Max: *Research Priorities for Robust and Beneficial Artificial Intelligence.* (2015), jan. http://futureoflife.org/data/documents/research_priorities.pdf
- [29] *TensorBoard.* <https://github.com/tensorflow/tensorboard>. – Zuletzt besucht am 25.11.2017