

Distributed Tensorflow

Michael Zipperle

259564

Michael.Zipperle@hs-furtwangen.de

Abstract—

I. EINFÜHRUNG

Eine Künstliche Intelligenz (KI) wird heutzutage immer mehr eingesetzt, um die Interaktion zwischen Mensch und Maschine zu bessern. Zusätzliche ermöglicht eine KI vollkommen neue Möglichkeiten in Bereichen wie Robotik, Daten Analyse, Gesundheitswesen, Autonomes Fahren und vielen mehr. Doch was steckt hinter einer künstlichen Intelligenz? Eine künstliche Intelligenz versucht durch maschinelles Lernen die menschliche Wahrnehmung und das menschliche Handeln durch Maschinen nachzubilden. Bei der KI-Forschung gibt es viele Verbindung zur Neurologie und Psychologie, den das menschliche Denken muss erforscht und verstanden werden, bevor Maschinen dies nachahmen können. Bis heute ist dies noch nicht annähernd gelungen, die meisten KI beschränken sich auf einen bestimmten Teilbereich und sind somit optimiert für einen bestimmten Anwendungsfall [1].

Bei einer KI kommt meist maschinelles Lernen zum Einsatz, wobei von erhobenen Daten gelernt wird, um dann Entscheidung treffen zu können. Es gibt eine Reihe an vordefinierten Algorithmen die an sogenannte Entscheidungsbäume erinnern. Diese Algorithmen sind meist nicht dynamisch genug, um mehrere Variablen verarbeiten zu können. Mit der Einführung von Deep Learning als Zweig des maschinellen Lernens ist dieses Problem größtenteils behoben. Durch Deep Learning kann ein kontinuierlicher Lernprozess umgesetzt werden, der sich stetig an neue Situation anpasst. Dabei basiert Deep Learning auf der Analyse von Big Data und gräbt sich durch riesige Mengen an Daten aus Datenbank, Internetquellen und mehr. Es nutzt neuronale Netze, um der Denkweise eines Menschen möglichst gut nachzuahmen. Deep Learning war bis vor wenigen Jahren nur von wenig Unternehmen genutzt, da der Einsatz viel Ressourcen benötigt und somit relativ teuer war. Durch den Fortschritt der Technik und der besseren Unterstützung der Graphics Processing Unit (GPU) hat sich Deep Learning durchgesetzt und kommt in heutigen Durchbrüchen wie autonomes Fahren, maschinelle Übersetzung, Spracherkennung und vieles mehr zum Einsatz [2].

Aktuell stellt der Lernprozess in Echtzeit eine große Herausforderung dar, oftmals benötigt der Lernprozess mehrere Stunden bzw. Tage. Für einige Anwendungsfälle reicht dies nicht aus und es wird ein Lernprozess nahe zu in Echtzeit benötigt. Es gibt bereits zahlreiche Deep Learning Frameworks, welche

in diesem Artikel aufgezeigt werden. Sehr beliebt ist aktuell das Framework Tensorflow von Google. Dieses Framework wird im Laufe dieses Artikel genauer untersucht und es wird anhand eines Tutorials die Einrichtung einer verteilten Tensorflow Umgebung aufgezeigt. Durch diese verteilte Umgebung kann die Performance des Lernprozess verbessert werden.

II. RELATED WORK

Bevor auf Tensorflow genauer eingegangen wird, soll dieses Kapitel eine Übersicht über folgende Deep Learning Frameworks geben:

- Torch
- Caffe
- Caffe2
- Keras
- Chainer
- CNTK
- Apache MXNet
- Amazon DSSTNE

A. Torch

Torch wurde ursprünglich von Facebook entwickelt und 2017 als Open-Source Projekt veröffentlicht. Das Framework bietet zwei Hauptfunktionen: Erstens, mathematische Berechnung unter starker Unterstützung der GPU. Dabei wird es oft als Ersatz für das bekannte Python Framework Numpy eingesetzt, da es durch die GPU-Unterstützung eine wesentlich bessere Performance bietet. Zweitens, zur Bildung von neuronalen Netze für Deep Learning, dabei wirbt das Framework mit maximaler Flexibilität und Geschwindigkeit [4]. Laut DL4J zeichnet sich das Framework durch viele Modulare Funktionen aus, die sich einfach kombinieren lassen. Des Weiteren lassen sich einfach neue Layer definieren und auf der GPU ausführen. Jedoch bietet Torch keinen kommerziellen Einsatz und die Dokumentation soll nicht vollständig sein [3].

B. Caffe

Caffe ist ein Open-Source Projekt, das von Yangqing Jia im Rahmen seiner Doktorarbeit beim Berkley AI Research (BAIR) initiiert wurde. Aktuell wird es vom BAIR und Community Entwicklern weiterentwickelt. Mit nur wenigen Zeilen Code lassen sich Modelle erstellen, dabei lässt sich konfigurieren ob die Berechnungen auf der CPU oder GPU durchgeführt werden. Caffe wird hauptsächlich zur Verarbeitung von Bildern eingesetzt und kann mit nur einer GPU über 60 Millionen Bilder pro Tag verarbeiten [5]. Aktuell bietet Caffe noch keine Möglichkeit die Berechnung verteilt von mehreren GPU durchzuführen [3].

C. Caffe2

Der Erfinder von Caffe Yangqing Jia arbeitet jetzt bei Facebook und entwickelte dabei an einer Erweiterung von Caffe. Diese wurde unter dem Name Caffe2 veröffentlicht und bietet mehr Skalierbarkeit und Leichtgewichtigkeit gegenüber Caffe. Zudem erlaubt Caffe2 das Verteilen von Aufgaben bzw. Berechnung auf mehrere Instanzen [6].

D. Keras

Keras ist ein High-Level Application Programming Interface (API) für neuronale Netze und basiert auf Tensorflow, welches in Kapitel III genauer erläutert wird. Die Entwicklung von Keras folgte dem Motto: "Die Fähigkeit, mit möglichst wenig Verzögerung von der Idee zum Ergebnis zu kommen, ist der Schlüssel für eine gute Forschung" [7]. Somit ermöglicht Keras das schnell aufsetzen und testen von Neuronalen Netzen. Elephas ist eine Erweiterung für Keras, die das Verteilen einer Anwendung über mehrerer Instanzen erlaubt [8].

E. Chainer

Chainer ist auch ein Open-Source Deep Learning Framework und bietet ein flexibles, intuitives und leistungsstarkes Mittel zur Implementierung einer ganzen Reihe von Deep-Learning-Modellen, einschließlich State-of-the-Art-Modellen wie z. B. rekurrenten neuronalen Netzen und variationalen Autoencodern. Mit Chainer können Anwendungen auf mehrere Instanzen verteilt werden und dadurch wie in Abbildung 1 zu sehen im Bereich Performance andere Frameworks klar hinter sich lassen. Die optimale Performance wurde beim Verteilen einer Anwendung auf 128 GPUs erreicht [9].

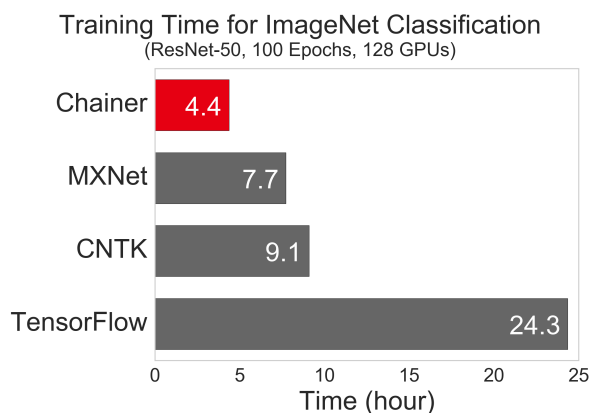


Fig. 1: Deep Learning Framework Vergleich [9]

F. Microsoft Cognitive Toolkit (CNTK)

Microsoft Cognitive Toolkit (CNTK) ist ein Open-Source Framework für das Verteilen einer Deep-Learning Anwendung im kommerziellen Bereich. Es beschreibt neuronale Netze als eine Reihe von Rechenschritten über einen gerichteten

Graphen. CNTK erlaubt dem Benutzer, populäre Modelltypen wie feed-forward DNNs, konvolutionelle neuronale Netze (CNNs) und wiederkehrende neuronale Netze (RNNs / LSTMs) leicht zu verwirklichen und zu kombinieren. CNTK implementiert stochastisches Gradientenabstiegsverfahren (SGD, Error Backpropagation) mit automatischer Differenzierung und Parallelisierung über mehrere GPUs und Server hinweg [10].

G. Apache MXNet

Apache MXNet ist ein modernes Open-Source Deep Learning Framework, mit dem neuronale Netze trainiert und implementiert werden können. Es ist skalierbar, ermöglicht schnelles Modelltraining und unterstützt ein flexibles Programmiermodell und mehrere Programmiersprachen. Die MXNet-Bibliothek ist portabel und kann auf mehreren GPUs und mehreren Instanzen skaliert werden. MXNet wird von den wichtigsten Public Cloud-Anbietern wie Amazon Web Services (AWS) und Azure unterstützt [11]. Amazon und Microsoft arbeiteten zusammen an einer API für Apache MXNet und veröffentlichten im Oktober 2017 Gluon, eine neue Open-Source Deep-Learning Schnittstelle, die es Entwicklern ermöglicht, einfach und schnell Machine-Learning-Modelle zu erstellen, ohne dabei die Leistung zu beeinträchtigen [13].

H. Amazon DSSTNE

Amazon Deep Scalable Sparse Tensor Network Engine (DSSTNE) ist eine Open-Source Deep Learning Framework zum Entwickeln von Empfehlungsmodellen. DSSTNE wurde bei Amazon eingesetzt, um personalisierte Produktempfehlungen für ihre Kunden zu erstellen. Es ist für den produktiven Einsatz von realen Anwendungen ausgelegt, bei denen Geschwindigkeit und Skalierbarkeit gegenüber experimenteller Flexibilität im Vordergrund stehen müssen. Training und Vorhersagen werden beide skaliert, wobei Berechnung und Speicherung für jede Schicht modellparallel verteilt werden [14].

III. TENSORFLOW

IV. TUTORIAL: DISTRIBUTED TENSORFLOW

V. EVALUIERUNG

VI. ZUSAMMENFASSUNG

REFERENCES

- [1] "Planet Wissen - Künstliche Intelligenz" https://www.planet-wissen.de/technik/computer_und_roboter/kuenstliche_intelligenz/ Accessed 20.06.2018
- [2] "Big Data Insider - Maschinelles Lernen und Deep Learning" <https://www.bigdata-insider.de/ki-maschinelles-lernen-und-deep-learning-das-sind-die-unterschiede-a-588067/> Accessed 20.06.2018
- [3] "DL4J - Deep Learning Frameworks Comparison" <https://deeplearning4j.org/compare-dl4j-tensorflow-pytorch> Accessed 20.06.2018
- [4] "Torch" <https://pytorch.org/about/> Accessed 20.06.2018
- [5] "Caffe" <http://caffe.berkeleyvision.org/> Accessed 21.06.2018
- [6] "Caffe2" <https://caffe2.ai/> Accessed 21.06.2018
- [7] "Keras" <https://keras.io/> Accessed 24.06.2018
- [8] "Elephas" <https://github.com/maxpumperla/elephas> Accessed 24.06.2018

- [9] "Chainer" <https://chainer.org/blog/> Accessed 01.07.2018
- [10] "Microsoft Cognitive Toolkit (CNTK)" <https://docs.microsoft.com/de-de/cognitive-toolkit/> Accessed 01.07.2018
- [11] "Wikipedia - Apache MXNet" https://en.wikipedia.org/wiki/Apache_MXNet Accessed 01.07.2018
- [12] "AWS Apache MXNet" <https://aws.amazon.com/de/mxnet/> Accessed 01.07.2018
- [13] "AWS Introducing Gluon" <https://aws.amazon.com/de/blogs/aws/introducing-gluon-a-new-library-for-machine-learning-from-aws-and-microsoft/> Accessed 01.07.2018
- [14] "Amazon DSSTNE" <https://aws.amazon.com/de/blogs/aws/introducing-gluon-a-new-library-for-machine-learning-from-aws-and-microsoft/> Accessed 01.07.2018