Neue Einsatzmöglichkeit von Hardwarebeschleunigern für nachhaltigere KI-Modelle: Entwicklung und Evaluation der Boltzmann Maschinen auf einem physikinspirierten Hardwarebeschleuniger

Bachelorarbeit

vorgelegt am 16. Februar 2024

Fakultät Wirtschaft und Gesundheit

Studiengang Wirtschaftsinformatik

Kurs WWI2021F

von

Simon Spitzer

Betreuer in der Ausbildungsstätte:

⟨ Name des Unternehmens ⟩
⟨ Titel, Vorname und Nachname der Betreuerin ⟩
⟨ Funktion der Betreuerin/des Betreuers ⟩

Unterschrift der Betreuerin/des Betreuers

DHBW Stuttgart:

⟨ Titel, Vorname und Nachname ⟩
⟨ der/des wissenschaftlichen Betreuerin/Prüferin ⟩

Inhaltsverzeichnis

Abkürzungsverzeichnis

Ein Abkürzungsverzeichnis ist optional. Das Paket `acronym` kann weit mehr, als hier gezeigt.[1] Beachten Sie allerdings, dass Sie die Einträge selbst in sortierter Reihenfolge angeben müssen.

CRM Customer Relationship Management

Ergänzende Bemerkung: Eine im Text verwendete Abkürzung sollte bei ihrer ersten Verwendung erklärt werden. Falls Sie sich nicht selbst darum kümmern möchten, kann das das Paket `acronym` übernehmen und auch automatisch Links zum Abkürzungsverzeichnis hinzufügen. Dazu ist an allen Stellen, an denen die Abkürzung vorkommt, `\ac{ITIL}` zu schreiben.

Das Ergebnis sieht wie folgt aus:

- erstmalige Verwendung von `\ac{ITIL}` ergibt: **ITIL!** (**ITIL!**),
- weitere Verwendung von `\ac{ITIL}` ergibt: **ITIL!**

Wo benötigt, kann man mit dem Befehl `\acl{ITIL}` wieder die Langfassung ausgeben lassen: **ITIL!**.

Falls man die Abkürzungen durchgängig so handhabt, kann man durch Paket-Optionen (in `_dhbw_praeambel.tex`) erreichen, dass im Abkürzungsverzeichnis nur die tatsächlich verwendeten Quellen aufgeführt werden (Option: `printonlyused`) und zu jedem Eintrag die Seite der ersten Verwendung angegeben wird (Option: `withpage`).

[1] siehe `http://ctan.org/pkg/acronym`

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

1.2 Problemstellung

In der Forschung und Entwicklung von Generativen KI-Modellen rückt die Rechengeschwindigkeit und Energieeffizienz zunehmend in den Fokus[2] Die Autor*innen von Open AI bestätigen, dass die Wachstumsrate von Machine-Learning-Modellen die Effizienzrate von Computerchips schon längst übertroffen hat. So verdoppeln sich jede 3-4 Monate der Rechenbedarf dieser Modelle jedoch verdoppeln sich nach Moore's Law die Leistung der Computerchips nur jede 2 Jahre.[3] Angesichts der Probleme des steigenden Energieverbrauchs von Rechenzentren und den damit verbundenen Treibhausgasemissionen dieser, ist die Suche nach effizienteren Lösungen essenziell für die Zukunft. Weltweit steigern Datenzentren ihren Energieverbrauch jährlich um 20-40%, wodurch sie 2022 etwa 1,3% des globalen Energieverbrauchs und 1% der energiebedingten globalen Treibhausgasemissionen verursacht haben.[4] Jedoch ist hier nicht zu erkennen, wie groß dabei der KI-Anteil zur Grundgesamtheit beiträgt.

Ein bereits bekannter Ansatz ist die Benutzung von KI-Beschleunigern basierend auf ASICs (Application-specific Integrated Circuits) - also Schaltungen, die anwendungsspezifisch verwendet werden, wie zum Beispiel Google TPUs (Tensor Processing Unit).[5] Dies ist auch sinnvoll, da die Verwendung von Mehrzweckmodellen für diskriminierende Aufgaben im Vergleich zu aufgabenspezifischen Modellen energieintensiver ist.[6] Ein alternatives vielversprechendes Konzept in der Forschung ist die Verwendung von physikinspirierten Hardwarebeschleunigern, die primär bei Optimierungsalgorithmen eingesetzt werden aufgrund ihrer Fähigkeit Probleme schneller und effizienter als GPUs lösen zu können.[7] Ein skalierbarer physikinspirierter Hardwarebeschleuniger (auch Ising-Maschine genannt), der die Leistung bestehender Standard-Digitalrechner übertrifft, könnte einen großen Einfluss auf praktische Anwendungen für eine Vielzahl von Optimierungsproblemen haben.[8]

Solche physikinspirierten Hardwarebeschleuniger bieten durch ihre besondere Berechnungsweise Potenzial für eine effizientere Verarbeitung von rechenintensiven Aufgaben. Konkret wird die Beschleunigung, anders als es bei digitalen Computern der Fall ist, durch die Berechnung rechenintensiver Aufgaben mit analogen Signalen erreicht. Die Implementierung auf dedizierter

[2] Vgl. Luccioni/Jernite/Strubell 2023, S. 1
[3] Vgl. Dario Amodei/Danny Hernandez 2024, S. 1
[4] Vgl. Hintemann/Hinterholzer 2022, S. 1
[5] Vgl. Wittpahl 2019, S. 39
[6] Vgl. Luccioni/Jernite/Strubell 2023, S. 5
[7] Vgl. Mohseni/McMahon/Byrnes 2022, S. 1
[8] Vgl. Mohseni/McMahon/Byrnes 2022, S. 1

Hardware bietet darüber hinaus die Möglichkeit, die Parallelisierung von digitalen Hardwarebeschleunigern und analogem Rechnen auszunutzen.[9]

Interessanterweise zeigen die Energiefunktionen von Hardwarebeschleunigern, die in Ising-Maschinen verwendet werden, große Parallelen zu denen in Boltzmann Maschinen, trotz ihrer unterschiedlichen Anwendungen, daher liegt es nahe, dass Ising Maschinen auch für KI gut funktionieren.[10] Ising-Maschinen zielen darauf ab, ihre Energie zu minimieren, wobei sie Energie als eine paarweise Interaktion von binären Variablen „Spins“ definieren.[11] BoltzmannMaschinen hingegen sind energiebasierte neuronale Netzwerke, die Klassifizierungen durchführen, indem sie jeder Konfiguration der Variablen eine skalare Energie zuordnen. Die Netzwerkenergie zu minimieren ist hierbei vergleichbar mit der Lösung des Optimierungsproblems.[12] Aktuelle Probleme mit Boltzmann-Maschinen umfassen die hohe Komplexität und Anforderungen an die All-to-All-Kommunikation zwischen Verarbeitungseinheiten, was ihre Implementierung auf herkömmlichen digitalen Computern ineffizient macht, sowie eine inhärent langsame Konvergenz in bestimmten Prozessen wie Simulated Annealing.[13] Diese Herausforderungen erschweren das Training und die Anwendung von Boltzmann-Maschinen insbesondere für große Datenmengen und komplexe Optimierungsaufgaben.[14] Nichtsdestotrotz impliziert die Ähnlichkeit der beiden, dass Ising-Maschinen in der Lage sein könnten, dieses spezielle KI-Modell, energieeffizienter und mit höherer Rechengeschwindigkeit auszuführen. Aktuell existieren nur wenige Konzepte eine Implementierung von Boltzmann Maschinen auf Ising-Maschinen zu erreichen. Das Paper der Autoren Mahdi Nazm BojnordiEngin und Engin Ipek ist hier ein vielversprechender Ansatz, jedoch konnte nicht gezeigt werden, wie es auf einem richtigen Beschleunigerchip funktionieren würde.

Vor diesem Hintergrund ergeben sich folgende zentrale Forschungsfragen:

1. Können Boltzmann Maschinen auf physikinspiriertenHardwarebeschleunigern durch analoge Rauschinjektion effizient implementiert werden?

    - Wie ist die Genauigkeit des KI-Modells im Hardwarebeschleuniger? Metrik: Prediction Accuracy

    - ergleichen mit anderen Hardwarebeschleuniger, FPGA, GPU oderCPU aus der Literatur (gute und schlechte) in Bezug auf Energieeffizienz und Rechengeschwindigkeit – Metriken: Troughput(Samples/Sec), Energieverbrauch (Energy/Operation)

Daher gilt es zu testen, ob dieses generative KI-Modell mit Ising Maschinen kompatibel ist und ob diese Lösung effizient ist oder nicht.

[9] Vgl. Mohseni/McMahon/Byrnes 2022, S. 4
[10] Vgl. Cai u. a. 2019, S. 10
[11] Vgl. Wang/Roychowdhury 2017, S. 1
[12] Vgl. Nazm Bojnordi/Ipek 2016, S. 2
[13] Vgl. Nazm Bojnordi/Ipek 2016, S. 1
[14] Vgl. Nazm Bojnordi/Ipek 2016, S. 2

1.3 Zielsetzung(ohne gneaue Metriken nennen, generell halten)

Das primäre Ziel dieser Bachelorarbeit ist die Erforschung und Erweiterung eines bestehenden physikinspirierten Hardwarebeschleunigers (ISING Maschine) zur Implementierung und Evaluation von Boltzmann Maschinen, einem energiebasierten KI-Modell. Dabei sollen die aufgestellten Forschungsfragen beantwortet werden.

Hierzu ist es zu Beginn nötig eine Simulator Pipeline zu konstruieren mit der Boltzmann Maschinen auf dem Hardwarebeschleuniger übersetzt werden. Die Simulator Pipeline besteht dabei aus einer bestehender KI-Bibliothek und bestehenden Hardwarebeschleuniger, die miteinander verbunden werden. Mit der Simulator Pipeline soll gezeigt werden, dass der Hardwaresimulator die Boltzmann Maschinen umsetzen kann. Aus der Simulator Pipeline heraus werden die Aktivierungswahrscheinlichkeiten der einzelnen Neuronen auf der simulierten Hardware gemessen und bei Erfolg bis zu einem vollständigen Neuronalen Netzwerk erweitert. Finaler Schritt ist, dass der Hardwarebeschleuniger für Training und Interferenz genutzt werden kann und dabei vergleichbar mit herkömmlichen MLLibraries ist. Diese Phase umfasst die sorgfältige Anpassung und möglicherweise Erweiterung des bestehenden Beschleunigers, um die spezifischen Anforderungen der Boltzmann Maschinen zu erfüllen.

Wenn die Simulator Pipeline validiert werden kann, wird ein Workload auf ein Standarddatenset zur Handschrifterkennung getestet. Dabei werden die Prediction Accuracy, Troughput (Samples/Sec) und der Energieverbrauch (Energy/Operation) der Boltzmann Maschinen auf dem ISING Hardwarebeschleuniger untersucht und dadurch die aufgestellten Forschungsfragen beantwortet.

1.4 Forschungsmethodik

Design Science Research

1. **Problemorientierung:** DSR fokussiert auf die Lösung praktischer Probleme, wie die Forschung zur Steigerung der Effizienz und Rechengeschwindigkeit in KI-Modellen.
2. **Artefakt Entwicklung:** Zentral in DSR ist die Entwicklung innovativer Artefakte. Die Arbeit zielt darauf ab, ein solches Artefakt in Form des physikinspirierten Hardwarebeschleunigers weiterzuentwickeln und für KI-Modelle einzusetzen.
3. **Iterative Evaluation:** Durch die iterative Vorgehensweise in DSR kann die Ausarbeitung der Lösung fortlaufend verbessert und angepasst werden, was für die Entwicklung und Optimierung von KI-Systemen entscheidend ist (ebenfalls das Konzept).
4. **Beitrag zur Wissensbasis und Praxisrelevanz:** DSR unterstützt die Generierung neuer Erkenntnisse und stellt sicher, dass Forschungsergebnisse sowohl theoretisch fundiert als

auch praktisch anwendbar sind, was mit den Zielen Ihres Projekts im Einklang steht. Untermethodik könnte hierbei eine Simulation sein. Variabel, je nach Verlauf der Forschung.

1.5 Aufbau der Arbeit

2 Aktueller Stand der Forschung und Praxis (generell auch wiedergeben von aktuell existierenden Lösungsmustern)

2.1 Ressourcenverbrauch bei KI-Modellen

2.1.1 Ressourcenverbrauch bei KI-Modellen

Nachhaltigkeit

Stromverbrauch

Rechenleistung begrenzt, KI-Modelle wachsen schneller als verfügbare Leistung

2.2 Deep Neural Network - Boltzmann Maschinen (Erstmal DNN erklären generell)

2.2.1 Konzept und Anwendung des Modells

Eine Boltzmann Maschine ist ein symmetrisches Netzwerk bestehend aus Neronen.[15]

2.2.2 Energiefunktion

2.2.3 Training von BMs

Markov-Chain-Monte-Carlo-Verfahren

Metropolis Hastings Conrtrastive Divergence

[15] Vgl. Amari/Kurata/Nagaoka 1992, S. 260

2.2.4 Aktuelle Probleme mit RBM/BM

2.3 Hardwarebeschleuniger

2.3.1 Aktuelle Ansätze im Bereich KI und weitere Lösungen

Asics

Quantencomputing

2.3.2 ISING Maschine/ Physikinspirierter Hardwarebeschleuniger

Konzept (mit Energiefunktion), Probleme der Digitalrechner bzw. Unterschied zu Digitalrechner

Aktuelle Anwendung

Potentielle Einsatzgebiete für KI-Modelle

Parallelen Energiefunktion BM und ISING Maschine

2.4 Memristor Hopfield Network

2.4.1 Memristor

2.4.2 Hopfield Network

2.4.3 Crossbar

2.4.4 Output Hopfield Networtk

2.4.5 Noisy HNN

3 Zielspezifikation und Darlegung der Forschungsmethodik

3.1 Zielspezifikation (genauer als in Einleitung, Metriken erwähnen, Erfolg meiner Methode bewerten, Welcher Teil der Forschungsfrage wird beantwortet?)

3.2 Design Science Research

3.3 Zielsetzung(ohne gneaue Metriken nennen, generell halten)

3.4 Laborexperiment für die Umsetzung

4 Implementierung/Laborexperiment der Simulator Pipeline

4.1 Zielsetzung und Forschungsmethodik

4.2 Aufbau der Simulator Pipeline

4.3 KI-Bibliothek Scikit-Learn

5 Evaluation der BM auf dem physikinspiriertem Hardwarebeschleuniger

5.1 Zielsetzung und Forschungsmethodik

5.1.1 Prediction Accuracy

5.1.2 Troughput (Samples/Sec)

5.1.3 Energieverbrauch (Energy/Operation)

5.2 Vergleichen mit anderen Hardwarebeschleuniger, FPGA, GPU oder CPU aus der Literatur

6 Kritische Reflexion und Ausblick

6.1 Evaluation der Erkenntnisse in Bezug auf die Zielsetzung der Arbeit

6.2 Kritische Reflexion der Ergebnisse und Methodik

6.3 Zielsetzung(ohne gneaue Metriken nennen, generell halten)

6.4 Ergebnisextration für Theorie und Praxis (evtl. mit 6.4 Zusammenlegen)

6.5 Ausblick

Anhang

Anhangverzeichnis

Anhang 1: So funktioniert's

Um den Anforderungen der Zitierrichtlinien nachzukommen, wird das Paket `tocloft` verwendet. Jeder Anhang wird mit dem (neu definierten) Befehl **anhang**{Bezeichnung} begonnen, der insbesondere dafür sorgt, dass ein Eintrag im Anhangsverzeichnis erzeugt wird. Manchmal ist es wünschenswert, auch einen Anhang noch weiter zu unterteilen. Hierfür wurde der Befehl **anhangteil**{Bezeichnung} definiert.

In Anhang 1/1 finden Sie eine bekannte Abbildung und etwas Source Code in **??**.

Anhang 1/1: Wieder mal eine Abbildung

Abb. 1: Mal wieder das DHBW-Logo.

Literaturverzeichnis

Amari, S./Kurata, K./Nagaoka, H. (1992): Information Geometry of Boltzmann Machines. In: *IEEE Transactions on Neural Networks* 3.2, S. 260–271. ISSN: 1941-0093. DOI: 10.1109/72.125867. URL: https://ieeexplore.ieee.org/abstract/document/125867 (Abruf: 16.02.2024).

Cai, F./Kumar, S./Van Vaerenbergh, T./Liu, R./Li, C./Yu, S./Xia, Q./Yang, J. J./Beausoleil, R./Lu, W./Strachan, J. P. (2019): Harnessing Intrinsic Noise in Memristor Hopfield Neural Networks for Combinatorial Optimization. DOI: 10.48550/arXiv.1903.11194. arXiv: 1903.11194 [cs]. URL: http://arxiv.org/abs/1903.11194 (Abruf: 15.02.2024). preprint.

Dario Amodei/Danny Hernandez (2024): AI and Compute. URL: https://openai.com/research/ai-and-compute (Abruf: 15.02.2024).

Hintemann, R./Hinterholzer, S. (2022): Data Centers 2021: Data Center Boom in Germany Continues - Cloud Computing Drives the Growth of the Data Center Industry and Its Energy Consumption. DOI: 10.13140/RG.2.2.31826.43207.

Luccioni, A. S./Jernite, Y./Strubell, E. (2023): Power Hungry Processing: Watts Driving the Cost of AI Deployment? DOI: 10.48550/arXiv.2311.16863. arXiv: 2311.16863 [cs]. URL: http://arxiv.org/abs/2311.16863 (Abruf: 15.02.2024). preprint.

Mohseni, N./McMahon, P. L./Byrnes, T. (2022): Ising Machines as Hardware Solvers of Combinatorial Optimization Problems. DOI: 10.48550/arXiv.2204.00276. arXiv: 2204.00276 [physics, physics:quant-ph]. URL: http://arxiv.org/abs/2204.00276 (Abruf: 15.02.2024). preprint.

Nazm Bojnordi, M./Ipek, E. (2016): Memristive Boltzmann Machine: A Hardware Accelerator for Combinatorial Optimization and Deep Learning, S. 13. 1 S. DOI: 10.1109/HPCA.2016.7446049.

Wang, T./Roychowdhury, J. (2017): Oscillator-Based Ising Machine. DOI: 10.48550/arXiv.1709.08102. arXiv: 1709.08102 [physics]. URL: http://arxiv.org/abs/1709.08102 (Abruf: 15.02.2024). preprint.

Wittpahl, V., Hrsg. **(2019)**: Künstliche Intelligenz: Technologie | Anwendung | Gesellschaft. Berlin, Heidelberg: Springer. ISBN: 978-3-662-58041-7 978-3-662-58042-4. DOI: 10.1007/978-3-662-58042-4. URL: http://link.springer.com/10.1007/978-3-662-58042-4 (Abruf: 15.02.2024).

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Thema: *Mein Titel* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

(Ort, Datum) (Unterschrift)