

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

Bachelorarbeit

submitted on March 22, 2024

Fakultät Wirtschaft und Gesundheit

Studiengang Wirtschaftsinformatik

Kurs WWI2021F

von

SIMON SPITZER

Betreuer in der Ausbildungsstätte:

⟨ Hewlett Packard GmbH ⟩
⟨ Dr. Fabian Böhm ⟩
⟨ Research Scientist at Hewlett Packard Labs ⟩

DHBW Stuttgart:

⟨ Prof. Dr., Kai Holzweißig ⟩
⟨ der/des wissenschaftlichen Betreuerin/Prüferin ⟩

Unterschrift der Betreuerin/des Betreuers

Contents

List of abbreviations	IV
List of figures	V
List of tables	VI
1 Einleitung	1
1.1 Motivation	1
1.2 Problem statement	1
1.3 Objective	2
1.4 Research method	3
1.5 Aufbau der Arbeit	4
2 Aktueller Stand der Forschung und Praxis	5
2.1 Ressourcenverbrauch bei KI-Modellen	5
2.1.1 Ressourcenverbrauch bei KI-Modellen	5
2.2 Neural Networks - Boltzmann Machines	5
2.2.1 Energy-based models	7
2.2.2 concept of Boltzmann Maschines	9
2.2.3 Restricted Boltzmann Machines	11
2.2.4 Current Problems with BMs and RBMs	16
2.3 Hardware accelerators	17
2.3.1 Current approaches in the field of AI and other solutions	17
2.3.2 GPU	18
2.3.3 Field programmable gate arrays	19
2.3.4 Application specific integrated circuit	20
2.4 Memristor Hopfield Neural Network	20
2.4.1 Hopfield Network	22
2.4.2 Memristor Crossbar Array	24
2.4.3 Output Hopfield Network	26
2.4.4 Noisy HNN	27
3 Zielspezifikation und Darlegung der Forschungsmethodik	28
3.1 Zielspezifikation (genauer als in Einleitung, Metriken erwähnen, Erfolg meiner Methode bewerten, Welcher Teil der Forschungsfrage wird beantwortet?)	28
3.2 Design Science Research	28
3.3 Zielsetzung(ohne genaue Metriken nennen, generell halten)	28
3.4 Laborexperiment für die Umsetzung	28
4 Implementierung/Laborexperiment der Simulator Pipeline	29
4.1 Zielsetzung und Forschungsmethodik	29
4.2 Aufbau der Simulator Pipeline	29
4.3 KI-Bibliothek Scikit-Learn	29
5 Evaluation der BM auf dem physikinspiertem Hardwarebeschleuniger	30
5.1 Zielsetzung und Forschungsmethodik	30

5.1.1	Prediction Accuracy	30
5.1.2	Troughput (Samples/Sec)	30
5.1.3	Energieverbrauch (Energy/Operation)	30
5.2	Vergleichen mit anderen Hardwarebeschleuniger, FPGA, GPU oder CPU aus der Literatur	30
6	Kritische Reflexion und Ausblick	31
6.1	Evaluation der Erkenntnisse in Bezug auf die Zielsetzung der Arbeit	31
6.2	Kritische Reflexion der Ergebnisse und Methodik	31
6.3	Zielsetzung(ohne gneaue Metriken nennen, generell halten)	31
6.4	Ergebnisextraction für Theorie und Praxis (evtl. mit 6.4 Zusammenlegen)	31
6.5	Ausblick	31
	Appendix	32
	List of references	34

List of abbreviations

BM	Boltzmann Maschine
RBM	Restricted Boltzmann Maschine
DNN	Deep Neural Network
EBM	Energy Based Model
MCMC	Markov chain Monte Carlo
DNN	Deep Neural Networks
CPU	Central Processing Unit
GPU	Graphics Processing Unit
ASIC	Application Specific Integrated Circuit
FPGA	Field Programmable Gate Array
TPU	Tensor Processing Unit
mem-HNN	memristor-Hopfield Neural Network

List of Figures

1	figure of a neural network	6
2	figure of a neural network	6
3	Figure of a simplified energy landscape	8
4	figure of a general Boltzmann Machine	10
5	Figure of a Restricted Boltzmann Maschine (RBM)	12
6	Figure of a logistic sigmoid function RBM	14
7	Flip Probability Function in Metropolis-Hastings Algorithm	16
8	Physical and microscopic view of memristor-Hopfield Neural Network (mem-HNN) .	22
9	Figure of a hopfield network	22
10	Modell of the mem-HNN	24
11	3D modell of the memristor crossbar array	25
12	Memristor-based learning	26
13	figure of a general Boltzmann Machine	27
14	Mal wieder das DHBW-Logo.	33

List of Tables

1 Einleitung

1.1 Motivation

In the research and development of generative AI models, the computing speed and energy efficiency are increasingly becoming the focus¹ The authors of Open AI confirm, that the growth rate of machine learning models surpassed the growth of efficiency within computerchips. The required computing power of the models double each 3-4 months but the power of computerchips, after Moore's Law double only every 2 years.² Focusing on current problems like rising energy consumption of datacenters and the associated greenhouse gas emissions, the search for more efficient solutions is essential for the future. Worldwide energy consumption of datacenters increase yearly around 20-40%, which means that in 2022 about 1,3% of the total global energy consumption and about 1% of energy-related global greenhouse gas emissions was caused by them.³ However, it is not clear how large the AI share contributed to the total numbers.

1.2 Problem statement

One approach that is already well known is the use of AI accelerators based on ASICs (application-specific integrated circuits) - i.e. circuits that are used application specific, such as Google TPUs (Tensor Processing Unit).⁴ This is useful because the usage of multimodels for discriminating tasks compared to task specific models are more energy intense.⁵ One promising alternative concept in research is the usage on physics-inspired hardware accelerators, that are primarily used for optimization problems because of their ability to solve problems faster and more efficient than GPUs.⁶ A scalable physics-inspired hardware accelerator (also called Ising-machine), that surpasses the power of existing standard digital computers, could have a large influence on practical applications for a variety of optimization problems.⁷

Such physics-inspired hardware accelerator offer, due to their special calculation method, potential for efficient processing of computationally intensive tasks. Specifically, the acceleration in contrast to digital computers is achieved by calculating the computationally intense tasks with analog signals. On top of that the implementation on dedicated hardware offers the possibility to exploit the parallelization of digital hardware accelerators and analog computation.⁸

¹Vgl. Luccioni/Jernite/Strubell 2023, p. 1

²Vgl. Dario Amodei/Danny Hernandez 2024, p. 1

³Vgl. Hintemann/Hinterholzer 2022, p. 1

⁴Vgl. Wittpahl 2019, p. 39

⁵Vgl. Luccioni/Jernite/Strubell 2023, p. 5

⁶Vgl. Mohseni/McMahon/Byrnes 2022a, p. 1

⁷Vgl. Mohseni/McMahon/Byrnes 2022a, p. 1

⁸Vgl. Mohseni/McMahon/Byrnes 2022a, p. 4

Interesting enough, despite their different applications, the energy function of the hardware accelerator that is used in Ising-machines shows big parallels to those used in Boltzmann Maschine (BM), therefore it can be suggested that Ising-machines could work well for AI.⁹ Ising-machines strive to minimize their energy, which is defined by the pairwise interaction of binary variables (Spins).¹⁰ In contrast, BMs are energy-based neuronal networks that are used for classification tasks by allocating a scalar energy for each configuration of variables. Minimizing the total network energy is therefore equal with the solution of a optimization problem.¹¹ Current problems with BMs are the high complexity and high requirements for the all-to-all communication between the processing units, which causes the implementation on conventional digital computers to be inefficient, but also an inherently slow convergence in certain processes such as simulated annealing.¹² These challenges complicate the training and the usage of BMs, especially for large data volumes and complex optimization tasks.¹³ Nevertheless the similarities of both models implicate, that Ising-machines could be able to execute this specific AI-model with higher energy efficiency and with higher computing speed. Currently there are only few concepts that exist on how to achieve a implementation of a BM within on a Ising-machine. The paper of the authors Mahdi, Nazm, Bojnordi and Ipek is a promising approach, However it could not be shown how a implementation on a real accelerator chip could function.

With the given background, the following central research question for this thesis arises:

1. Can Boltzmann Machines be efficiently implemented on physics-inspired Hardware accelerators by analog noise injection?
 - What is the accuracy of the AI-model on the hardware accelerator?
 - Metric: Prediction accuracy and negative Likelihood
 - Comparison between other hardware accelerators, FPGA, GPU, or CPU within the literature in terms of energy efficiency and computing speed.
 - Metrics: Throughput (Samples/Sec), Energy usage (Energy/Operation)

It is therefore necessary to test whether this generative AI model is compatible with Ising machines and whether this solution is efficient or not.

1.3 Objective

The primary objective of this bachelor thesis is the research and extension of a existing physics-inspired hardware accelerator (Ising machine) for the implementation and evaluation of BMs as an energy based AI-model. The aim is to answer the posed research question. In addition to

⁹Vgl. Cai/Kumar, Suhas/Van Vaerenbergh/Liu, R., et al. 2019, p. 10

¹⁰Vgl. Wang, T./Roychowdhury 2017, p. 1

¹¹Vgl. Nazm Bojnordi/Ipek 2016, p. 2

¹²Vgl. Nazm Bojnordi/Ipek 2016, p. 1

¹³Vgl. Nazm Bojnordi/Ipek 2016, p. 2

that, it would be beneficial if rules for the influence of hyperparameters could be established since there is no data available for this new method.

To initially accomplish this objective it is necessary to establish a simulator pipeline to the hardware accelerator that translates BM on top of it. The simulator pipeline consists of an existing machine learning library and an existing hardware accelerator that are connected to each other. With the simulator pipeline it needs to be shown that it is possible for the hardware accelerator to realize BMs.

Within the simulator pipeline, the activation probabilities of individual neurons are measured on the simulated hardware. If this process proves successful, it is then expanded to simulate a complete neuronal network. The final step is that the hardware accelerator can be used for training and inference and is comparable to conventional machine learning libraries. This phase includes a careful adjustment and possibly extension of the existing accelerator to be compliant with the specific requirements of BMs.

If the simulator pipeline can be validated a workload consisting of a standard data set to recognize handwritten digits will be tested. The prediction accuracy, throughput (samples/sec) and the energy consumption (energy/operation) of the Boltzmann machines on the Ising hardware accelerator will be investigated as metrics in order to thereby answer the second part of the posed research questions.

1.4 Research method

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

2.1 Ressourcenverbrauch bei KI-Modellen

2.1.1 Ressourcenverbrauch bei KI-Modellen

substantial challenges in high consumption of computational, memory, energy, and financial resources, especially in environments with limited resource capabilities¹⁴

Nachhaltigkeit

Stromverbrauch

Rechenleistung begrenzt, KI-Modelle wachsen schneller als verfügbare Leistung

2.2 Neural Networks - Boltzmann Machines

Over the past few years, the emergence of artificial neural networks has transformed the field of computer vision and extended its influence to other areas. These include natural language processing, game strategy development and execution (with examples in playing Atari and Go), and optimization of navigation tasks, such as determining the most efficient routes on maps.¹⁵ Therefore, it is fair to say that neural networks are part of various important applications.¹⁶ Particularly in the last two years, artificial intelligence has also garnered widespread interest from the public, especially regarding chatbots like ChatGPT and Google Bard.¹⁷ An important feature of a neural network-based system that are inspired by our brain, is that they can learn and adapt to data.¹⁸

Internally, neural networks are computational models that consist of many simple processing units, called neurons that work together in parallel often structured within interconnected layers.¹⁹ They consist out of a network architecture, which describes the layout and how the neurons are wired. Secondly, they have a optimization function which specifies the goals persued in the learning process.²⁰ Lastly, there is a training algorithm that varies all of the hyperparameters,

¹⁴cf. Bai et al. 2024, pp. 1–2

¹⁵cf. Cichy/Kaiser 2019, p. 305

¹⁶cf. Gawlikowski et al. 2023, p. 1513

¹⁷cf. Singh/Kumar, Shubham/Mehra 2023, pp. 1–2

¹⁸cf. Cichy/Kaiser 2019, p. 305

¹⁹cf. Cichy/Kaiser 2019, p. 305

²⁰cf. Durstewitz/Koppe/Meyer-Lindenberg 2019, p. 1583

like connection strengths between neurons, training iterations, the learning rate, etc..²¹ The following figure 1 shows a typical neural network that consists out of a input layer, a hidden layer and an output layer with dots representing the neurons within the network.

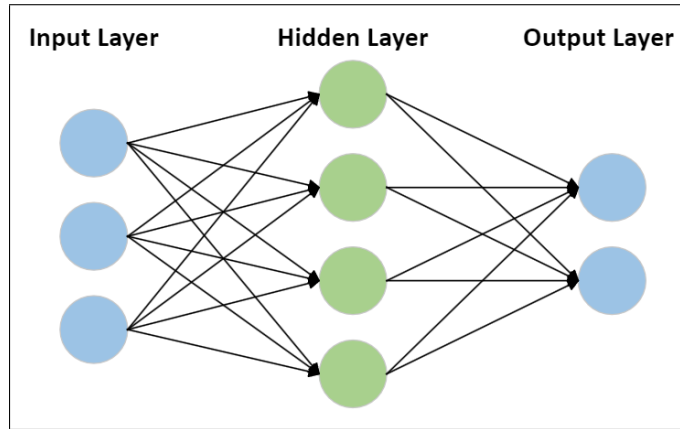


Fig. 1: figure of a neural network

Although, when these interconnected layers are stacked on top of each other, so multiple hidden layers are stacked on top of each other, the network is called deep.²² In general, deep learning methods can be seen as subset of machine learning methods and are today's fundament of artificial intelligence allowing to solve more complex tasks.²³ Deep Neural Networks (DNN)s are constantly growing and currently have around 1200 interconnected layers that equal to more than 16 million neurons inside a network .²⁴ An example of a deep neural network is presented in figure 2 which shows the stacked layers in the middle of the network.

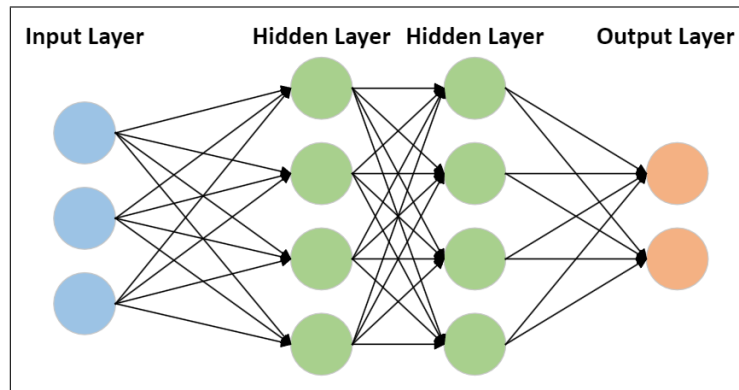


Fig. 2: figure of a neural network

Some examples for regression tasks within a DNN in the field of acomputer vision include object detection, medical image registration, head- and body-pose estimation, age estimation and visual

²¹cf. Durstewitz/Koppe/Meyer-Lindenberg 2019, p. 1583

²²cf. Cichy/Kaiser 2019, p. 305

²³cf. Durstewitz/Koppe/Meyer-Lindenberg 2019, p. 1583

²⁴cf. Mall et al. 2023, p. 2

tracking.²⁵ Nowadays, verly large neural networks with millions of parameters can be created due to the research achievements made in the field of neural networks and deep learning leading to highly performing models.²⁶ Nonetheless, such models often have a negative effect on the environment in terms of unnecessary energy consumption and a limitation to their deployment on low-resource devices because they are excessively oversized and redundant.²⁷

2.2.1 Energy-based models

An Energy Based Model (EBM) is a type of neural network that has special characteristics. One characteristic is that an EBM is a statistical model.²⁸ This probabilistic approach willingly uses uncertainty into the model calculations to draw the models inputs randomly from its underlying distribution.²⁹ This is done because the conventional deterministic method of backpropagation is known to potentially convert to local minimas, and requires long computation time.³⁰ As a result with conventional backpropagation more frequently incorrect classification would take place. The second characteristic is that an EBM is determined by an energy function that needs to be minimized in order to find the solution of the optimization problem.³¹ Since 1982, those statistical neural network models have been continuously emerging in the machine learning field when J.J. Hopfield introduced the Hopfield Network.³² Current developments include their use in reinforcement learning, potential replacements for discriminators in generative adversarial networks and for quantum EBMs.³³ In addition to that, Open AI showed that EBMs are useful models across a wide variety of tasks like achieving state-of-the-art out-of-distribution classification and continual online class learning to name a few.³⁴ The underlying idea behind EBMs is to establish a probabilistic physical system that is able to learn and memorize patterns but most importantly generalize it.³⁵ Especially, it involes learning an energy function $E_{\theta}(x) \in \mathbb{R}$, with x representing the configuration of the network, and assigning the low energy to observed data x_i and high energy to other values x .³⁶

²⁵cf. Gustafsson et al. 2020, pp. 325–326

²⁶cf. Marinó et al. 2023, p. 152

²⁷cf. Marinó et al. 2023, p. 152

²⁸cf. Huembeli et al. 2022, p. 2

²⁹cf. Uusitalo et al. 2015, pp. 25–27

³⁰cf. Specht 1990, p. 109

³¹cf. Huembeli et al. 2022, p. 2; cf. Ranzato, a. et al. 2006, p. 1

³²cf. Hopfield 1982

³³cf. Verdon et al. 2019, p. 1; cf. Du/Lin/Mordatch 2021, p. 1

³⁴cf. Du/Mordatch 2020, pp. 1–2

³⁵cf. Huembeli et al. 2022, p. 2

³⁶cf. Gustafsson et al. 2020, p. 330

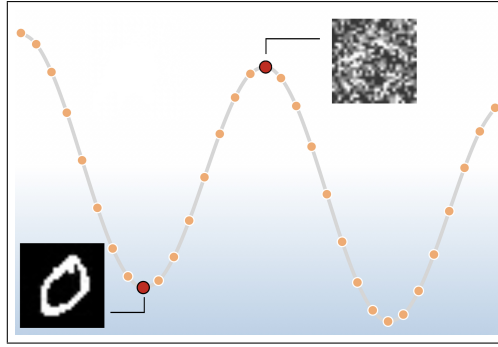


Fig. 3: Figure of a simplified energy landscape

In this figure 1 a simplified energy landscape is shown where the local minima corresponds to states that encode an MNIST digit.³⁷ It is visible that observed data settles in the local minimum of the energy landscape, in this case a clear 0. On the other hand close to the local maxima of the energy landscape the 0 is only barely recognizable and therefore got a higher energy value assigned to it. The assumption of the underlying distribution function $P(x)$ represents the probability distribution over the input data x , indicating how likely different configurations of x are under the models learned patterns:

$$P(x) = \frac{1}{Z} \exp\left(-\frac{E(x)}{T}\right), \quad (2.1)$$

where Z is the partition function to ensure that the density function normalizes to a total probability of 1 and T is interpreted as the temperature.³⁸ The partition function Z used in 2.1 is given by summing over all possible pairs of visible and hidden vectors³⁹:

$$Z = \sum_x \exp\left(-\frac{E(x)}{T}\right) \quad (2.2)$$

The aim of the training in an EBM is to match the true probability distribution P_{data} as closely as possible with the internal probability distribution P_{model} learned by the model. What this means is that the specific aim is to adjust its parameters such that P_{model} becomes as close to P_{data} as possible, which shows the model has learned the distribution of the real world data. A practical method to achieve this goal is to use the KL divergence. KL divergence is a mathematical measure that helps to measure how close the predictions are by comparing the model's learned distribution to the true distribution of the data:

$$G = \sum_x P_{\text{data}}(x) \ln\left(\frac{P_{\text{data}}(x)}{P_{\text{model}}(x)}\right) \quad (2.3)$$

Here, P_{data} is the probability distribution when the network receives a specific data input from the environment, while P_{model} represents the internal network running freely, also referred to as

³⁷cf. Huembeli et al. 2022, p. 6

³⁸cf. Huembeli et al. 2022, pp. 2–3

³⁹cf. Hinton, G. E. 2012b, p. 4

“dreaming”.⁴⁰ In the training process the asymmetric divergence G needs to be minimized and therefore the probabilities of the model converge close to the ones in reality. To optimise the KL divergence the energy is adjusted, whereby data is assigned to low energy states (according to 2.1) and the training data receives high energy and therefore low probabilities.⁴¹

NACH UNTEN As a side note it is worth mentioning that using the maximum likelihood estimator for Z is intractable due to the requirement of summing over all possible states, which leads to an exponential increase in the number of states for larger systems.⁴²

2.2.2 concept of Boltzmann Machines

A BM is a type of symmetrical EBM consisting of binary neurons $\{0, 1\}$.⁴³ The neurons of the network can be split into two functional groups, a set of visible neurons and a set of hidden neurons.⁴⁴ Therefore, the BM is a two-layer model with a visible layer (“v”) and a hidden layer (“h”).⁴⁵ The visible layer is the interface between the network and the environment. It receives data inputs during training and sets the state of a neuron to either $\{0, 1\}$ which represents activated or not activated. On the other hand, the hidden units are not connected to the environment and can be used to explain underlying constraints in the internal model of input vectors and they cannot be represented by pairwise constraints.⁴⁶ The connection between the individual neurons is referred to as bidirectional, as each neuron communicates with each other in both directions.⁴⁷ As early as 1985, one of the founding fathers of artificial intelligence, Geoffrey Hinton, was aware that an BM is able to learn its underlying features by looking at data from a domain and developing a generative internal model.⁴⁸

Most machine learning models can be categorized in either generative or discriminative models. Both are strategies to estimate a probability that an specific object can be assigned to a category.⁴⁹ Discriminative models estimate the probability distribution based on category labels that are given to specific objects.⁵⁰ On the other hand, a generative model differ as follows. They generate a probabilistic model of the underlying probability distribution for each category, which is assumed as the basis of the data, and in a following step they use Baye’s rule to identify which category is very likely to have established the object.⁵¹ An real world example would be the following: to predict if a movie will be a hit, you could analyze past box office successes to model characteristics shared by hits (generative approach), or assess immediate audience reactions to

⁴⁰cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, pp. 154–155

⁴¹cf. Zhai et al. 2016, pp. 2–3

⁴²cf. Zhai et al. 2016, pp. 2–3

⁴³cf. Amari/Kurata/Nagaoka 1992, p. 260

⁴⁴cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, p. 154

⁴⁵cf. Salakhutdinov/Hinton, G. 2009, p. 448

⁴⁶cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, p. 154

⁴⁷cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, p. 149

⁴⁸cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, p. 148

⁴⁹cf. Hsu/Griffiths 2010, p. 1

⁵⁰cf. Gm et al. 2020, p. 2

⁵¹cf. Hsu/Griffiths 2010, p. 1

movie trailers and reviews to predict success without modeling historical data (discriminative approach). Therefore it can be said that BMs and EBMs are generative models. In the following figure 4, a general BM is depicted, where the upper layer embodies a vector of stochastic binary 'hidden' features, while the lower layer embodies a vector of stochastic binary 'visible' variables.⁵²

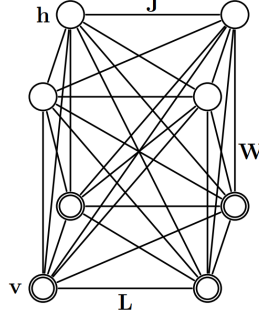


Fig. 4: figure of a general Boltzmann Machine

The model contains a set of visible units $v \in \{0, 1\}$, and a set of hidden units $h \in \{0, 1\}$ (see Fig. 1). The energy function of the BM with the states $\{v, h\}$ is defined as:

$$E(v, h; \theta) = -\frac{1}{2}v^T L v - \frac{1}{2}h^T J h - v^T W h, \quad (2.4)$$

where $\theta = \{W, L, J\}$ are the model parameters.⁵³ W, L, J represent visible-to-hidden, visible-to-visible and hidden-to-hidden weights. In BM each neurons works towards minimizing the global energy by entering a particular neuron configuration representing a input to the machine and the system will find the minimum energy configuration that is compatible with the given input.⁵⁴ A simple method to find a local energy minimum involves to switch into whichever of the two states (on or off) of a neuron result in a lower energy given the current state of the other neurons.⁵⁵ Integrating the function 2.4 into the earlier introduced KL-divergence 2.2 and doing gradient descent a learning rule to update the weights and biases appears.⁵⁶ The gradient descent algorithm is commonly used in machine learning and is an iterative technique that adjusts the model parameters (weights and biases).⁵⁷ It progressively acquires the gradient of the energy function, methodically advancing towards the optimal solution and ultimately achieves the minimum loss function along with adjusted parameters.⁵⁸ Consequently, this leads to the specific learning rule⁵⁹:

$$\Delta w_{ij} = \epsilon(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}) \quad (2.5)$$

⁵²cf. Salakhutdinov/Hinton, G. 2009, p. 449

⁵³cf. Salakhutdinov/Hinton, G. 2009, p. 448

⁵⁴cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, p. 150

⁵⁵cf. Fahlman/Hinton, G./Sejnowski, T. 1983, p. 110

⁵⁶cf. Hinton, G. E. 2012b, p. 5

⁵⁷cf. Wang, X./Yan/Zhang, Q. 2021, p. 11

⁵⁸cf. Wang, X./Yan/Zhang, Q. 2021, p. 11

⁵⁹cf. Hinton, G. E. 2012b, p. 5

The network can now update the weights “W” that exist between the neurons through the training rule based on the observations that served as input.⁶⁰ In this case, the square brackets represent expected values, as the training is based on the activation probability. In addition to that, the step sizes of updates to the weights are influenced by the learning rate ϵ within the iterative training process.

Performing exact training in this model is intractable because exact computation of the data predictions and the model predictions takes a time that is exponential in the number of hidden units.⁶¹ When the number of hidden units is large compared to the number of visible units it is impossible to achieve a perfect model because of the totally connected network and the resulting 2^n possibilities.⁶² Hereby, n represents the number of neurons in the network with each neuron being in one of the two states, the total sum of possibilities are 2^n . This leads back to the briefly mentioned constraint of equation 2.3, that is needed to calculate an activation probability of a neuron, which is required to update a weight in the training process shown in 2.5.

A specific example to demonstrate why it is intractable to calculate an activation of a BM is the following. A fictional BM has 80 visible nodes and 120 hidden nodes and therefore the possibilities of states of neurons are 2^{200} , which is 1.61×10^{60} . To put this into perspective, the total atoms that exist on earth are only estimated to be around 1.33×10^{50} .⁶³ That means even if it would be possible to store one information per atom it would just not be enough.

As a result, instead of directly trying to train the model sampling methods are used that are able to estimate these activation probabilities. This enables the training of BMs and RBMs.

2.2.3 Restricted Boltzmann Machines

As a simplification of the training problem Hinton and Sejnowski proposed Gibbs sampling as an algorithm to approximate both expectations.⁶⁴ Furthermore, the intralayer connections of the model got removed and the result is the so called RBM. To transform an BM into a RBM the diagonal elements L and J introduced earlier, are set to 0 and as a result the well-known model of a RBM establishes shown in fig.5.⁶⁵

⁶⁰cf. Barra et al. 2012, pp. 1–2

⁶¹cf. Salakhutdinov/Hinton, G. 2009, p. 449

⁶²cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, p. 154

⁶³cf. Helmenstine 2022, p. 478–480; cf. Schlamming 2014, p. 1

⁶⁴cf. Ackley/Hinton, G. E./Sejnowski, T. J. 1985, pp. 158–165

⁶⁵cf. Salakhutdinov/Hinton, G. 2009, p. 449

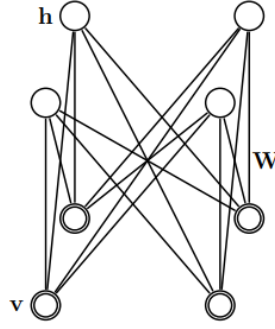


Fig. 5: Figure of a RBM

What can be recognized that no more visible-to-visible and hidden-to-hidden connections can be found in the model. The configuration of the visible and hidden units (v, h) therefore has also an updated energy function (Hopfield, 1982) given by:

$$E(v, h) = - \sum_{i \in \text{visible}} a_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}, \quad (2.6)$$

where v_i, h_j are the binary states of a visible unit i and hidden unit j , a_i and b_j are their biases and w_{ij} is the weight between them.⁶⁶ Despite, compared to the fully connected BM, the RBM is less complex but the advantages of training surpasses the loss in expressivity possibilities.⁶⁷ The RBM has recently been drawing attention in the machine learning community because of its adaption and extension for various tasks such as representational learning, document modeling, image recognition and for serving as foundational components for deep networks including Deep Boltzmann Machines, Deep Belief Networks and hybrid models with CNNs.⁶⁸

Training of RBMs

The training of RBMs can be established with the use of sampling methods that estimate the activation probabilities, which are needed to update the weights. There are currently two methods that can be chosen from: contrastive divergence and the Metropolis-Hastings algorithm. The goal of the techniques is to create a sequence of correlated steps from a random walk that, after enough iterations, makes it possible to sample a desired target probability distribution.⁶⁹ In the following part both methods will be explained in depth. Especially, they are interesting since they serve as baselines to compare against the new sampling method of a hopfield network that is to be achieved in the practical part of the thesis.

⁶⁶cf. Hinton, G. E. 2012a, pp. 3–4

⁶⁷cf. Huembeli et al. 2022, p. 4

⁶⁸cf. Zhang, N. et al. 2018, p. 1186

⁶⁹cf. Patrón et al. 2024, p. 1

Contrastive Divergence: Contrastive divergence is a special Gibbs Sampling training method developed by Geoffrey Hinton for the efficient training of RBMs.⁷⁰ In traditional, Gibbs sampling would have to generate a long chain of samples, until independent samples are obtained from the observed data distribution of the model.⁷¹ The samples are needed for each iteration of the gradient ascent on the log-likelihood resulting in large computational costs.⁷² To solve this issue contrastive divergence minimizes an approximation of the Kullback-Leibler divergence between the empirical distribution of the training data and the distribution generated by the model.⁷³ They way to achieve this is by initializing the Markov chain with the samples from the data distributon.⁷⁴ The outcome has been shown to heavily decrease the training time while only adding a small bias.⁷⁵ This allows to calculate the probabilities of equation 2.5. This entails initializing the visible units using an actual data input, such as an MNIST sample, and then commencing the subsequent steps with the hidden states. Often the process can be stopped after only sampling a very small number of steps.⁷⁶

1. Forward Pass (positive phase)

During the forward pass using the Gibbs Sampling method, the visible units are set to a completely random state. Next up the hidden units are computed. The computation of the hidden units involves calculating their activation probabilities and performing an actual sampling with their calculated activation probabilities. With the RBM it is now easy to get an analytical calculated sample of $(\mathbf{v}_i \mathbf{h}_j)_{data}$.⁷⁷ Given an input data out of the training images, v , the binary state, j , of each hidden unit, h_j , is set to 1 with following probability:

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{ij}), \quad (2.7)$$

where $\sigma(x)$ is the logistic sigmoid function with an unbiased sample. The sigmoid function is defined as $\sigma(x) = \frac{1}{1+\exp(-x)}$ and is needed because it is the underlying activation function of each neuron. A visual representation is shown in figure 6:

⁷⁰cf. Hinton, G. E. 2012b, pp. 4–5

⁷¹cf. Huembeli et al. 2022, pp. 5–6

⁷²cf. Upadhya/Sastry 2019, pp. 7–8

⁷³cf. Mocanu et al. 2016, p. 246

⁷⁴cf. Upadhya/Sastry 2019, pp. 7–8

⁷⁵cf. Larochelle/Bengio 2008, p. 537

⁷⁶cf. Larochelle/Mandel, et al. 2012, p. 646

⁷⁷cf. Hinton, G. E. 2012b, p. 5

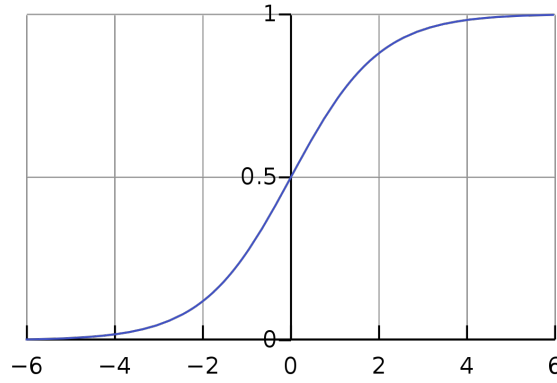


Fig. 6: Figure of a logistic sigmoid function RBM

The result is a set of probabilities that reflect how likely it is for each hidden unit to be on, which stands for 1, or, which stands for 0, given the input data.⁷⁸ The sampling part of the positive phase uses the just calculated activation probability of each hidden unit and performs a random experiment with it. That random experiment generates a uniform random number between 1 and 0 and if the random number is greater than the just calculated activation probability the hidden unit is set to activated. Afterwards, the hidden unit is either activated or not activated and the training process continues with the new state of the hidden units.

2. Reconstruction (negative phase)

In this phase, the sampled hidden states are used to reconstruct the visible units. This is essentially a prediction of the input, which is how the model sees the input based on the just updated hidden units and is calculated with following probability:⁷⁹

$$p(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_j h_j w_{ij}) \quad (2.8)$$

The sampling part of the negative phase uses the just calculated activation probability of each visible unit and performs a random experiment, like in the positive phase. Now, the result is a prediction of the input in the visible nodes. Afterwards, a half forward pass is made to calculate the activation probability of a hidden unit again based on the activated or not activated visible units.

3. Updating the weights

Now, all the requirements to update the weights are satisfied and can be used within the equation 2.5. The delta that results is summed to the current weight and the internal model gets closer to predicting the observed data. In total, one training iteration consists out of 1 Forward Pass, 1 Reconstruction and 0.5 Forward Pass again is accomplished. Repeating this training steps N

⁷⁸cf. Huembeli et al. 2022, p. 6

⁷⁹cf. Hinton, G. E. 2012b, p. 6

times for a suitable chosen N the model learns better, since more steps of alternating Gibbs sampling were performed.⁸⁰

Metropolis-Hastings: The Metropolis-Hastings algorithm, often only called Metropolis algorithm, is a technique out of Markov chain Monte Carlo (MCMC) class techniques.⁸¹ The Metropolis-Hastings method was invented by Metropolis et al. in 1953 when they noticed, that for an intractable distribution with too many states it can be seen as a limiting distribution of Markov chains.⁸² The intractable distribution to handle with the Metropolis-Hastings technique in the case of RBMs is equation 2.3. An Interpretation of the method can be expressed as: "A visitor to a museum that is forced by a general blackout to watch a painting with a small torch. Due to the narrow beam of the torch, the person cannot get a global view of the painting but can proceed along this painting until all parts have been seen."⁸³ The version already adjusted for RBMs incorporates the following functionality of the Metropolis technique:

First, select a random or given configuration x_{old} of a RBM that holds the states of all visible and hidden neurons.⁸⁴ Secondly, the energy of the configuration, noted as E_{old} , must be calculated using Equation 2.6, as previously introduced. Subsequently, this energy value is stored. Thirdly, the configuration gets updated by picking one random neuron and changing the state of it from 0 to 1 or vice versa.⁸⁵ This new configuration is stored as x_{new} . Following that the energy of the new configuration E_{new} is calculated and stored. Now the two energy values are compared and if $E_{new} \leq E_{old}$ the new configuration will be accepted and $x_{old} = x_{new}$.⁸⁶ If $E_{new} > E_{old}$ then there are some extra steps to be followed:

The flip probability is calculated as $p = \exp\left(-\frac{E_{new}-E_{old}}{kT}\right)$. KT is interpreted as the temperature in the network and with higher temperature it increases the activation probability leading to an faster exploration through the landscape but with less details.⁸⁷ For RBMs KT is assumed to be 1.⁸⁸ In the following figure 7 the resulting probability function is shown.

⁸⁰cf. Huembeli et al. 2022, p. 6

⁸¹cf. Patrón et al. 2024, p. 1

⁸²cf. Metropolis et al. 1953, pp. 1087–1092

⁸³cf. Robert 2016, p. 2

⁸⁴cf. Beichl/Sullivan 2000, p. 65

⁸⁵cf. Rosenthal 2009, p. 1

⁸⁶cf. Patrón et al. 2024, pp. 1–2

⁸⁷cf. Li et al. 2016, pp. 1–9

⁸⁸cf. Hinton, G. 2014, p. 3

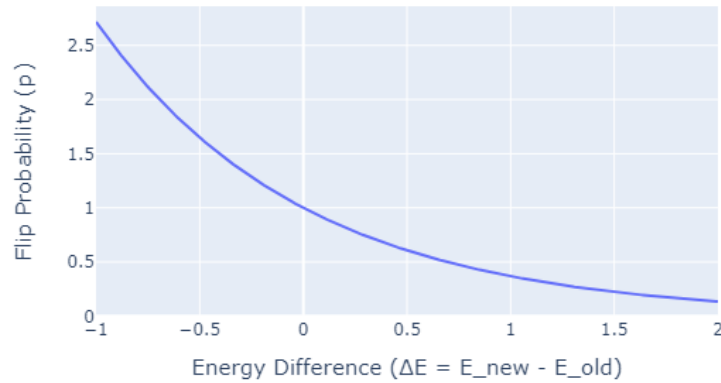


Fig. 7: Flip Probability Function in Metropolis-Hastings Algorithm

In the next step a uniform random number r between 0 and 1 is generated. After generating r the configuration will be accepted if $r \leq p$ (i.e., $x_{\text{old}} = x_{\text{new}}$).⁸⁹ Otherwise, a rejection takes place if $r > p$ (i.e., $x_{\text{old}} = x_{\text{old}}$).

Finally, the configuration x_{old} can be stored and the process repeats beginning from step 2 on.⁹⁰ After repeating enough times the activation probability for each neuron can be calculated by summing over all samples ($x_1 + x_2 + x_3 + \dots$) and the result is divided by the total number of samples.

2.2.4 Current Problems with BMs and RBMs

One general problem that occurs in the learning process of a BM is that it is both time-consuming and difficult.⁹¹ This is because sampling from an undirected graphical model is not straightforward and therefore RBMs make use of MCMC proposed methods like Contrastive Divergence and Metropolis Hastings.⁹² In addition to that, the selection of hyperparameters can be difficult since for the training of a practical model a large hyper-parameter space needs to be explored.⁹³ Hyperparameters are: the learning rate, size of the hidden layer, number of training iterations iteration count per bias (sampling step size), initializing the weight sizes in the beginning but also the method for calculating activation probabilities (Contrastive Divergence, Metropolis Hastings, etc.). As a result, establishing a RBM with perfect hyperparameters is time consuming and can be seen as art. Furthermore, training can become unstable and predictions become inaccurate

⁸⁹cf. Patrón et al. 2024, pp. 2–3

⁹⁰cf. Patrón et al. 2024, p. 17

⁹¹cf. Fischer/Igel 2012, pp. 1–2

⁹²cf. Fischer/Igel 2012, p. 2

⁹³cf. Larochelle/Bengio 2008, p. 536

due to an incompatible selected temperature.⁹⁴ A lower temperature reduces the system's possibility to explore the energy landscape thoroughly, leading to the false selection of local minima instead of finding the global minimum. Vice versa a too high temperature can cause that the energy landscape is not explored enough and has gaps between it missing some minima or skipping the global maxima. Luckily, the temperature for RBMs is expected to be 1 and only for specific use cases it makes sense to adjust internal temperature.

To accelerate the training process of a RBM, it is crucial to address the most computationally demanding aspect: the matrix-vector multiplication involved in the sampling process. A possibility of achieving this is using dedicated hardware, so called hardware accelerators for this problem. They are designed to tackle a specific task very efficiently, like matrix vector multiplications, which are widely used within most of the neural networks.⁹⁵ That is the reason why they are significant for the acceleration of this thesis and an interesting technology to look at.

2.3 Hardware accelerators

2.3.1 Current approaches in the field of AI and other solutions

Since Neural Networks and DNNs are growing their parameters at rapid rates they constantly achieve better and better results and are able to solve even more complex tasks.⁹⁶ This upcoming trend of growing network sizes exponentially also brings a dark side with it: An excessive increase in computational effort and memory size.⁹⁷ As a result Central Processing Unit (CPU)s can barely satisfy the required performance and specialized hardware accelerators are used to increase the performance of these Neural Networks.⁹⁸ In addition to that for many use cases, like autonomous driving, there are high energy, latency, and runtime predictability constraints CPUs are not able to meet.⁹⁹

The concept of developing hardware accelerators is not new. However, their limited adoption and the fast obsolescence of even the most outstanding accelerators have made the investment in them uneconomical compared to general-purpose processors that surpassed them.¹⁰⁰ At present, however they are seen as promising driving force of computer architecture since they are the optimal solution to satisfy the growing computation-hungry demands of businesses, especially within machine learning workloads.¹⁰¹ An hardware accelerator can be defined as "a separate architectural substructure that is architected using a different set of objectives than the base processor, where these objectives are derived from the needs of a special class of applications".¹⁰² Broken

⁹⁴cf. Huembeli et al. 2022, pp. 3–4

⁹⁵cf. Lehnert et al. 2023, pp. 3881–3882

⁹⁶cf. Baischer/Wess/TaheriNejad 2021, p. 1

⁹⁷cf. Baischer/Wess/TaheriNejad 2021, pp. 1–2

⁹⁸cf. Zhou et al. 2022, p. 1-2; cf. Baischer/Wess/TaheriNejad 2021, p. 2

⁹⁹cf. Ahmad/Pasha 2020, p. 2692

¹⁰⁰cf. Peccerillo et al. 2022, p. 2

¹⁰¹cf. Peccerillo et al. 2022, pp. 2–3

¹⁰²cf. Peccerillo et al. 2022, p. 2

down, they are specialized hardware, expertly optimized for the unique demands of certain application categories, freeing them from the restrictions usually placed on general-purpose processors. Moreover, a hardware accelerator doesn't replace the conventional processor, which is still used for the operating system, it rather enables specific workloads to be executed on it very efficiently.¹⁰³ There are different approaches such as Graphics Processing Unit (GPU)s, Application Specific Integrated Circuit (ASIC)s, Field Programmable Gate Array (FPGA)s, but also new approaches like Quantum Computations or Photonic matrix multiplication are researched.¹⁰⁴ All of these methods have different use cases and get more and more application specific. The list of the sequence, sorted by application-unspecific to specific for established approaches looks the following:

$$\text{CPU} \xrightarrow{\text{less flexible}} \text{GPU} \xrightarrow{\text{specialized}} \text{FPGA} \xrightarrow{\text{more customized}} \text{ASIC}$$

Currently the approaches can be segmented into three categories: **Firstly**, the design of data-driven digital circuits. It consists the shift from general-purpose GPUs to specialized dataflow architectures like systolic arrays, which are used in Google's Tensor Processing Units (TPUs). These architectures are noted for their efficiency in performing deep learning operations by reducing control hardware and keeping data movement local.¹⁰⁵ **Secondly**, network structure optimizations. Hereby modifications to the neural networks themselves are made to improve hardware efficiency. One method is quantization, which simplifies arithmetic operations and reduces memory needs by using fixed-point representations of data and weights instead of using for example 32-bit floating points. The other one is pruning, which involves setting certain weights to zero to reduce the complexity of operations.¹⁰⁶ **Thirdly**, technology-driven designs. Current research into using novel circuitry and memory cells include memristive memory cells and silicon photonics, to further enhance performance and energy efficiency. They work by storing the network weights and calculating the vector multiplications with analog signals with technologies like crossbar arrays. While these technologies promise significant advantages, their practical application is still being explored.¹⁰⁷ The following three accelerator approaches can be categorized into the first category:

2.3.2 GPU

The GPU, is by far the most common accelerator in the market with a focus at computational-complex workloads. Also GPUs enabled great achievements in DNNs due to their massive parallelization potential of computations and their high throughputs compared to FPGAs and ASICs.¹⁰⁸ A GPU is a manycore unit that features up to hundreds of multi-processors that consist of in-

¹⁰³cf. Peccerillo et al. 2022, pp. 2–3

¹⁰⁴Zhou et al. 2022, p. 1-2; cf. Baischer/Wess/TaheriNejad 2021, p. 2

¹⁰⁵cf. Lehnert et al. 2023, p. 3883

¹⁰⁶cf. Lehnert et al. 2023, p. 3883

¹⁰⁷cf. Lehnert et al. 2023, p. 3883

¹⁰⁸cf. Baischer/Wess/TaheriNejad 2021, p. 16

order cores which are able to exploit massive threadlevel parallelism.¹⁰⁹ They excel at performing numerous floating-point arithmetic operations for vector processing on large datasets with high degrees of data-parallelism. In practice this works by breaking down workloads into small tasks that can be processed by the enormous amount of cores in parallel.¹¹⁰ With development of real-time graphics GPUs became programmable too. The combination of programmability and floating-point performance make them very attractive for machine learning workloads and is the reason for their dominance in the market.¹¹¹ On top of that, the widespread adoption of GPUs has led to extensive support across numerous frameworks and high-level APIs commonly used in Machine Learning.¹¹² Well-known frameworks would be PyTorch or TensorFlow. However, compared to the other two accelerator approaches the GPU is not as flexible and has higher latency and energy consumption.¹¹³

2.3.3 Field programmable gate arrays

In contrast, FPGAs have also demonstrated enormous parallelization capabilities due to their fast digital signal processors and on-chip memory which result in lower energy cost than GPUs.¹¹⁴ They work by using reconfigurable logic blocks that can be interconnected using routing tracks with configurable switches at the intersections.¹¹⁵ This combined with the use of many digital signal processors and local storage of data in the hardware enables the development of the exact hardware needed for a workload.¹¹⁶ As a side note it is worth mentioning that the most energy-consuming task of a workload is the data transfer and not the computation itself. In the context of FPGAs, they use their on-chip memory to reduce the data transfer significantly and therefore achieve a sweet spot between computation speed and energy efficiency.¹¹⁷ Hence, they are utilized to design a specialized processor tailored for executing specific workloads, like machine learning, effectively.¹¹⁸ Furthermore, due to their reprogrammable nature they have a lower engineering cost and faster time-to-market compared to ASICs.¹¹⁹ With FPGAs the implementation time could only be a matter of weeks and also allows to support continuous upgrades and bug fixes even after the deployment which is not possible within ASICs¹²⁰ Even though the FPGA possesses all these advantages with their high flexibility, latency and low energy consumption, they are sometimes inferior in throughput compared to a GPU.¹²¹

¹⁰⁹cf. Peccerillo et al. 2022, p. 2

¹¹⁰cf. Hu/Liu, Y./Liu, Z. 2022, p. 101

¹¹¹cf. Dally/Keckler/Kirk 2021, p. 42

¹¹²cf. Baischer/Wess/TaheriNejad 2021, p. 16

¹¹³cf. Hu/Liu, Y./Liu, Z. 2022, p. 100

¹¹⁴cf. Ahmad/Pasha 2020, p. 2693

¹¹⁵cf. Babu/Parthasarathy 2021, p. 144

¹¹⁶cf. Baischer/Wess/TaheriNejad 2021, p. 19

¹¹⁷cf. Hu/Liu, Y./Liu, Z. 2022, pp. 101–102

¹¹⁸cf. Sipola et al. 2022, p. 322

¹¹⁹cf. Boutros/Betz 2021, p. 4

¹²⁰cf. Boutros/Betz 2021, p. 4

¹²¹cf. Hu/Liu, Y./Liu, Z. 2022, p. 100

2.3.4 Application specific integrated circuit

ASICs can be distinguished from FPGAs because they are not programmable. In addition to that, they offer the highest degree of customization and are designed to execute a specific application with the utmost efficiency.¹²² So the attributes of ASICs are the focus on specific tasks and the strong design freedom. Nowadays, the most common type of ASICs are Tensor Processing Unit (TPU)s because of their matrix vector multiplications abilities that are needed within machine learning. But conventional ASICs work by mapping neurons directly to the hardware.¹²³ Their design architecture enables them to outperform GPUs and FPGAs in terms of their small size, greater computation speed and high power efficiency.¹²⁴ Specifically when compared to corresponding FPGA circuits ASICs are 35x smaller and 4x faster. A more current study could show that this gap is reduced but they still are 9x smaller and also faster.¹²⁵

Nonetheless, developing an ASIC requires expert knowledge in chip design but also within neural networks and takes a lot of time.¹²⁶ Out of the three approaches, they provide the most efficient solution, yet it comes at the expense of lacking reconfigurability, no programmability, and incurring high engineering costs.¹²⁷ With a sustainability and climate-change aspect in mind, they are the by far the best option since they represent the most power efficient approach with also the best computation speed.

2.4 Memristor Hopfield Neural Network

The so called mem-HNN is a hardware accelerator that uses an emerging approach of combining analog signals and electronical signals to solve complex optimization problems.¹²⁸ It can be categorized into the ASIC family of hardware accelerators and its specific purpose is to solve Ising problems. In 1925 Ernst Ising, a German physicist, invented the Ising model which explained the interaction between ferromagnets.¹²⁹ This statistical model focuses on the state of a spin s_i (up and down; +1 and -1) representing their magnetic behaviour. With the so called Hamiltonian function the Ising model calculates the total energy of a system through the following energy equation:

$$E(\mathbf{s}) = \sum_i h_i s_i + \sum_{i,j} J_{ij} s_i s_j, \quad s_i = \pm 1, \quad (2.9)$$

where i is the label of the spins s_i , with h_i representing the external magnetic field interacting with each spin, and J_{ij} is the interaction strength between pairs of spins that are connected

¹²²cf. Baischer/Wess/TaheriNejad 2021, p. 17

¹²³cf. Hu/Liu, Y./Liu, Z. 2022, p. 104

¹²⁴cf. Baischer/Wess/TaheriNejad 2021, p. 17

¹²⁵cf. Boutros/Betz 2021, p. 5

¹²⁶cf. Baischer/Wess/TaheriNejad 2021, p. 17

¹²⁷cf. Peccerillo et al. 2022, p. 4; cf. Hu/Liu, Y./Liu, Z. 2022, p. 100

¹²⁸cf. Cai/Kumar, Suhas/Van Vaerenbergh/Sheng, et al. 2020, p. 410

¹²⁹cf. Ising 1925, pp. 253–258

by an edge (ij) .¹³⁰ Both values h_i and J_{ij} are real-valued allowing for a wide range of possible magnetic field intensities and vary in interaction strength.¹³¹ Nowadays, the Ising model is also attractive in other fields to describe the energy of a system and to transform them into an Ising problem.¹³² Ising problems are equal to finding the minimum energy state of a system. Hence, in practice transforming optimization problems into Ising problems, the optimal solution is equal to the minima of the Ising energy function. This transformation works by mapping each variable of the problem to Ising spins and designing an Ising model whose ground state represents the solution.¹³³

Equivalence between the energy function of a mem-HNN and the energy function of a RBM can be shown here. The energy function of the mem-HNN works by using the binary states of $+1$ and -1 while the RBM uses $+1$ and 0 but otherwise they are completely equal. To transform the RBM into the binary states of the mem-HNN its energy function from 2.6 needs to be modified using $\frac{x+1}{2}$ with x representing the energy function. The fact, that both energy functions are equal implies that the neural network of a RBM could possibly be trained on this Ising machine. Moreover, other artificial intelligence models could be compatible but currently there are no approaches on how to transform them.

Coming back to the Ising machine itself, the background for the development is the current slow down or failure of Moore's law which causes slow improving computation speed, energy efficiency and computation latency of conventional semiconductor electronic technology.¹³⁴ Since the mem-HNN is engineered to solve Ising problems, therefore also called Ising machine, it can tackle various problems that fall under the category of Ising problems.¹³⁵ Originally the mem-HNN was experimentally tested by the team of researchers to solve nondeterministic polynomial-time hard, or NP-hard for short, Ising problems directly on the hardware.¹³⁶ NP-hard problems are among the toughest problems to solve and have an exponential- or even factorial time to solve (2^n , $n!$) with no efficient solution, slow to solve and to verify.¹³⁷ Well-known examples would be the salesman problem, the maximum clique problem or the calculation of the partition function of the BM with 2^n possibilities.

In simple terms, the Ising machine is able to efficiently find the global minima of energy functions that are the underlying structure of an Ising problem. The name mem-HNN already indicates that the Ising machine is based on the concept of a Hopfield Network. All this is possible because of the update formula of the Hopfield Network, which is implemented directly on the hardware of the accelerator. A photograph mem-HNN of the physical accelerator (left side) and a microscopic view (right side) can be seen in following figure:

¹³⁰cf. Tanahashi et al. 2019, p. 2

¹³¹cf. Wang, T./Roychowdhury 2017, pp. 1-2

¹³²cf. Tanahashi et al. 2019, pp. 2-3

¹³³cf. Lucas 2014, pp. 2-3

¹³⁴cf. Cai/Kumar, Suhas/Van Vaerenbergh/Liu, R., et al. 2019, p. 1

¹³⁵cf. Mohseni/McMahon/Byrnes 2022b, p. 363

¹³⁶cf. Cai/Kumar, Suhas/Van Vaerenbergh/Sheng, et al. 2020, p. 410

¹³⁷cf. Izadkhah 2022, pp. 497-500

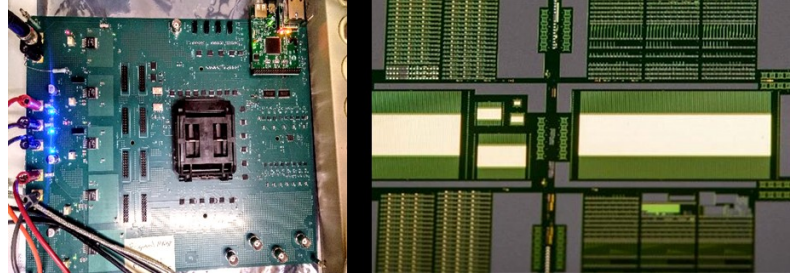


Fig. 8: Physical and microscopic view of mem-HNN

2.4.1 Hopfield Network

A Hopfield Network is an EBM and belongs to the field of recurrent neural networks.¹³⁸ Because the Hopfield Network is the underlying structure of the physical mem-HNN accelerator this model is explained first. The structure of the network consists of only one single layer with binary valued neurons inside.¹³⁹ Therefore, the neurons state can either be $\{1, 0\}$ or $\{1, -1\}$. The connections between the neurons are symmetrical, which means that the weights of the connections are the same in either direction.¹⁴⁰ Initially, the primary applications of this type of network were to serve as storage for associative patterns and to facilitate pattern retrieval.¹⁴¹ In practice given a query pattern a Hopfield Network can retrieve a pattern that is most similar or an is an average of similar patterns.¹⁴² In this paper the Hopfield Network's update function interests us because it possibly could be used to sample the intractable training of a RBM mentioned earlier. Surprisingly, since Hopfield networks were introduced by J.J Hopfield in 1982 the storage capacity got increased over time but the fundamentals stayed the same.¹⁴³ In following figure 6 an example of a Hopfield Network can be seen.¹⁴⁴

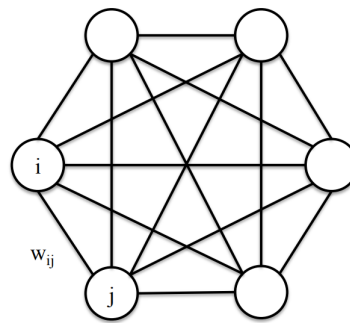


Fig. 9: Figure of a hopfield network

¹³⁸cf. Dramsch 2020, p. 35

¹³⁹cf. Ahad/Qadir/Ahsan 2016, p. 7

¹⁴⁰cf. Ahad/Qadir/Ahsan 2016, p. 7

¹⁴¹cf. Ramsauer et al. 2021, p. 2

¹⁴²cf. Ramsauer et al. 2021, p. 2

¹⁴³cf. Hopfield 1982, p. 2554-2558; cf. Ramsauer et al. 2021, p. 2

¹⁴⁴cf. Yao/Gripon/Rabbat 2013, pp. 1-2

The exemplary network has 6 neurons and bidirectional weights W_{ij} between the neurons. In addition to that, a Hopfield network has no input or output layer.¹⁴⁵ The main goal is to find the values for each neuron in the network given a specific input that minimizes the total energy of the system.¹⁴⁶ The minimum energy is then equal to the state where the network is able to perform as a memory item.¹⁴⁷ This energy state can be calculated with the following energy equation¹⁴⁸:

$$E = -\frac{1}{2} \sum_{i \neq j} T_{ij} V_i V_j. \quad (2.10)$$

This energy function invented by Hopfield has big similarities with a BM when comparing to the equation 2.4. This is one of the reasons why the execution on the mem-HNN could work out. When comparing a Hopfield Network, they seek to achieve the effect of changing node activation on the overall energy of the network but BMs replace this with the probability of a certain node being activated on the network energy.¹⁴⁹ The second important reason to research the hopfield networks is for their updating process. Approximately, the activation rule for each neuron is to update its state as if it were a single neuron with the threshold activation function.¹⁵⁰

$$s_i \leftarrow \begin{cases} +1 & \text{if } \sum_j w_{ij} s_j + b \geq \theta_i, \\ -1 & \text{otherwise.} \end{cases}$$

The state of the neuron will be updated to 1 if the sum over all weights multiplied with the states $\{1, -1\}$ added to a bias b is greater than the threshold θ_i . In the case of our accelerator the threshold is set to 0 but in theory can be used as an hyperparameter.

Since every neuron's output is an input to all the other neurons the order of the updates need to be specified.¹⁵¹ There is the possibility to update all neurons synchronous or asynchronous. There is no study that shows what update method leads to better results. Therefore, this paper follows the asynchronous option first and ensures to do enough iterations, so that every neuron has at least updated once before moving on. In addition to that, the idea of the updating method of the accelerator is slightly different. The idea behind this is to inject noise into the system so that the activation function could work together with the activation function that a RBM needs to perform. In detail the idea is to add a normal gaussian distribution $g(x)$ on top of the activation

¹⁴⁵cf. Yao/Gripon/Rabbat 2013, p. 3

¹⁴⁶cf. Ahad/Qadir/Ahsan 2016, p. 7

¹⁴⁷cf. Ahad/Qadir/Ahsan 2016, p. 7

¹⁴⁸cf. Hopfield 1982, p. 2556

¹⁴⁹cf. Ahad/Qadir/Ahsan 2016, p. 7

¹⁵⁰cf. MacKay 2003, p. 506

¹⁵¹cf. MacKay 2003, p. 506

function.¹⁵² As a result the new statistical updating function looks like the following:

$$s_i \leftarrow \begin{cases} +1 & \text{if } \sum_j w_{ij} + b + g(x) \geq \theta_i, \\ -1 & \text{otherwise.} \end{cases}$$

Now the system could potentially be used to update the states of the neurons within a RBM. Since the success of this method is not guaranteed or tested in literature yet the practical part first needs to validate if this concept is feasible.

2.4.2 Memristor Crossbar Array

Having set the foundational knowledge about the function of a Hopfield Network the mode of operation can be tackled. Since the mem-HNN saw the light of day in 2021, a number of improvements have been made to it and at the end of 2023 the individual components can be seen in following figure¹⁵³:

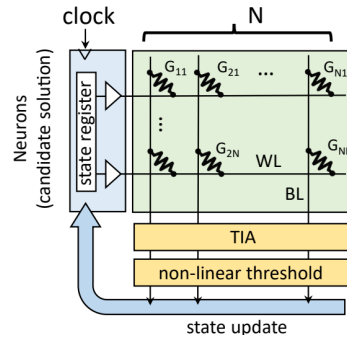


Fig. 10: Modell of the mem-HNN

The green square symbolizes the memristor crossbar array, which has the task of performing the in-memory-vector matrix multiplication. This is the core part of the architecture that is built around the update formula of the Hopfield Network. The memristor crossbar array gets his name because it consists of **memristors** that connect orthogonal **electric tracks** with each other. The G stands for conductance and represents the inverse of the resistance R of the memristors since $G = \frac{1}{R}$. On the other hand, BL (Bitline) and WL (Wordline) represent the electrical tracks. All the other parts of the model are handled in subchapter 2.4.3, so for now the spotlight belongs to the memristor crossbar array. A better perspective of the memristor crossbar array gives following 3D model¹⁵⁴:

¹⁵²cf. Böhm et al. 2022, pp. 4–5

¹⁵³cf. Hizzani et al. 2023, p. 2

¹⁵⁴cf. Cai/Kumar, Suhas/Van Vaerenbergh/Sheng, et al. 2020, p. 410

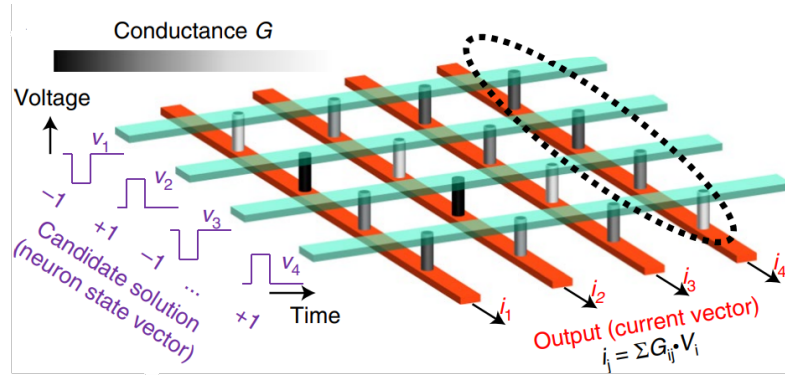


Fig. 11: 3D modell of the memristor crossbar array

To understand how this crossbar array implements the Hopfield network, it helps to have another look at the update formula.

$$s_i \leftarrow \begin{cases} +1 & \text{if } \sum_j w_{ij}s_j + b \geq \theta_i, \\ -1 & \text{otherwise.} \end{cases}$$

First of all, on the green Wordline a voltage which either is +1 or −1 that represent the state of the neurons in the hopfield network is applied. Then, the current is flowing towards the memristors, which is an electronic device that functions as a resistor, but with the ability to dynamically change its resistanceas while the device is being used.¹⁵⁵ The memristors are suitable for matrix vector multiplication because they can use their dynamical resistance to act as weigths.¹⁵⁶ In addition to that, this is the reason why this kind of memory has the potential of high switching speeds, high energy efficiency and high endurance.¹⁵⁷

In eletrical terminology, if the resistance of a memristor is higher, the conductance is low since the current can barely flow into the lower bitlane. For each of the intersections, when a Wordline meets the crossbar it represents the multiplication of w_{ij} and s_j within the update formula. After the matrix vector multiplication the single currents flow in the same direction on the Bitline and are summed up together within their current strength.¹⁵⁸ The reason for this phenomenon is the 1. Kirchhoff'sche rule, which is not further explained here since it would be out of scope. To make an example for each output current i_1 has for intput voltages V_1, V_2, V_3, V_4 . The current flowing into the lower Bitlane is now according to Ohm's rule $i_{out} = \frac{1}{R_{memristor}} * V_{in}$. Since $G = \frac{1}{R}$ and Kirchhoff'sche rule sums up each current i_1 is $i_1 = G_{11} * V_1 + G_{12} * V_2 + \dots + G_{14} * V_4$ As a result the first first part $\sum_j w_{ij}s_j$ of the updating formula of the Hopfield Network is established. Adding the bias b to the sum is achieved by simply adding a current, which is worth the amount of the bias, to the the total Bitline current.

¹⁵⁵cf Sung/Hwang/Yoo 2018, p. 124

¹⁵⁶cf Sung/Hwang/Yoo 2018, p. 124

¹⁵⁷cf. Amirsoleimani et al. 2020, p. 3

¹⁵⁸cf. Amirsoleimani et al. 2020, p. 3

Presently, the research on the application of memristor crossbar arrays in supervised learning is sparse, indicating a clear necessity for further investigative studies to understand their potential and usability.¹⁵⁹ In the practical part the mem-HNN model exactly this is put to the test and training data of the RBM should be generated to give clearance about the feasibility of this idea. Each i is the output of the current vector and equal to a neuron within the Hopfield Network with all the memristors on this lane symbolizing the weights that all the other neurons connect to it.

Memristor-based learning

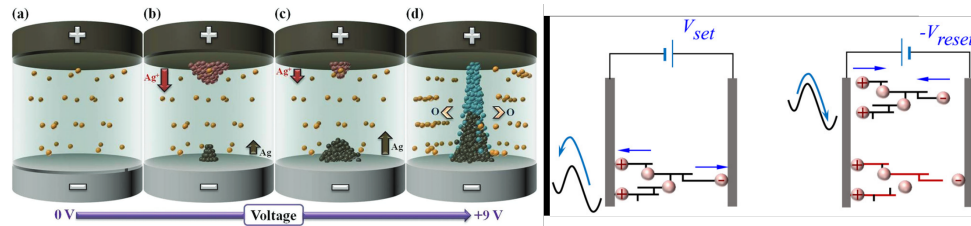


Fig. 12: Memristor-based learning

Eher Synapsen interessant -Wie programmiere ich den elektrischen Widerstand - dieses Konzept kurz erläutern -Filament formation, wie Plattenkondensator, Wasserleitung mit Ventil, dass man einstellen kann wie viel Wasser durchfließt - Wie kann man schaffen, dass dieser Widerstand geändert wird? - Bei hoher Spannung - V_{reset} -> Wenn hoch genug kann man anfangen Memristor zu verändern - 2 Operationsmodi, Programmierphase ermöglicht Widerstandseinstellen, andere Phase -> Gewichtsmatrix, Strom durchschicken, also geringe Spannung Inferenz. -Filament sind Ionen, geladen Teilchen die Strom leiten, bilden gemeinsam Filament (Equivalent wie ein Faden) -Spannung wird beim V_{reset} umgedreht, dadurch bricht Verbindung ab. Keine Verbindung mehr da, denn Spannung umgedreht. Filament kann auch größer und kleiner werden (Mehr Ionen zusammen), Mehrere Verbindungen möglich, -Lege Spannung für bestimmte Zeit an und dann lege ich eine Spannung ab und Filament ist gebildet -Non volatile Memory - im Betrieb, permanenter Speicher, Trainingsschritt trainieren mit V_{set} , dann $-V_{reset}$ und V_{set} für neue Trainingsiteration mit neuen Gewichten Memristor ist ein nicht volatiler Speicher

Kleine Graue Widerstände sind Memristoren, Das ganze ist das Crossbar Array

Alles zusammen ist das Memristor Hopfield Neural Network

2.4.3 Output Hopfield Network

Alle anderen Bauteile

¹⁵⁹cf.Amirsoleimani et al. 2020, p. 8; cf.Sung/Hwang/Yoo 2018, p. 124

TIA, kein Multiplexer mehr = Strom in Spannung umwandeln

non-linear threshold = Komperator, ist es größer oder kleiner gleich

Erklärung der Restlichen Bauteile außerhalb des Crossbar arrays (Alles noch Analog)

Hysteretic threshold ist gleich Komperator

Input/output buffer and driver = State register -> , Buffer = Speicher und Driver = Sitzt am Anfang des Crossbar arrays als Verstärker

—

Zusätzlicher Strom auf leiterbahn ist equivalent zum Bias.

Größer gleich ist mit dem Komperator initialisiert.

Anfangen mit der Update Funktion -> Dadurch wird Energiefunktion auch implementiert, mit jedem Update wird Energie gesenkt. Ising Modell ist ne Energiefunktion. Update funktion sorgt dafür, dass lokale minimas der Energiefunktion gefunden werden. Für Optimierungsprobleme ist die Lösung dadurch gut genug.

Ich benutze Sampling mit Rauschen, sodass ich auch mehrere Minimas erkennen kann. Wir sind an mehreren Minimas interessiert. System starten, iterativ herunterlaufen, in nächstes Minima springen durch rauschen, dadurch Sampling möglicih

Gewichte werden digital geupdated und ein Eingangssignal mitgegeben

—

2.4.4 Noisy HNN

Zwischen Bitline(BL) und TIA wird zusätzlich das rauschen noch eingefügt.

Picture is out of following paper.¹⁶⁰

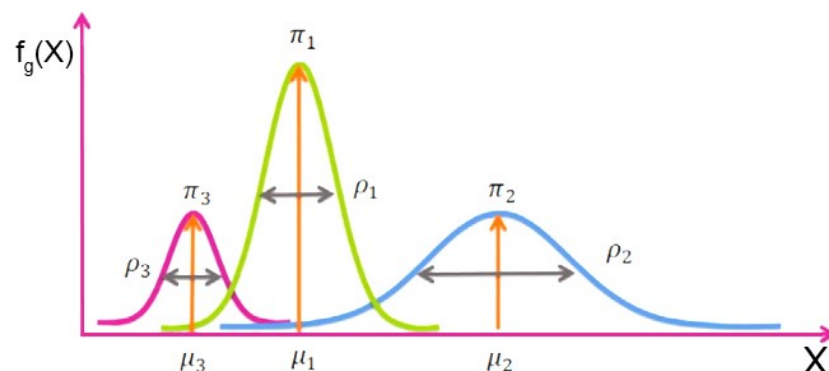


Fig. 13: figure of a general Boltzmann Machine

¹⁶⁰cf. Gm et al. 2020, p. 3

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 genaue Metriken nennen, generell halten)

3.4 Laborexperiment für die Umsetzung

4 Implementierung/Laborexperiment der Simulator Pipeline

Hopfield Netzwerk aktivierungsfunktion der Updating methode

-> Konzeptionell Art des Updates mit keiner Temperatur wie bei MCMC Unterschied von MCMC zu Hopfield Netzwerk -> Zufällige Konfiguration und minimale Energie finden. Jedoch hat ein Hopfield Netzwerk keine Temperatur

-> Starte zufällige Konfiguration -> Wähle ein Neuron aus und Berechne Summe und addiere mit Bias, -> Update wenn threshold überschritten 1 und dann auf 0 -> Speichern der neuen Konfiguration -> Starte iteration von gespeicherter Konfiguration -> Am Ende habe ich 10000 Vektoren (Die Konfigurationen) -> V1 Neuron wurde so und so oft aktiviert und ich muss average über das neuron und habe dadurch die Aktivierungswarscheinlichkeit.

-Aktivierungsfunktion einfügen (Binary Step und verfleich zu sigmoid von Abb.4)

-Testen der Aktivierungsfunktion, wenn ich ein Neuron trainiere und dann Mitteln - Von vornerein auf Netzwerk Basis arbeiten mit mehreren Neuron, jedoch für 1 Neuron testen

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 genaue Metriken nennen, generell halten)

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

6.5 Ausblick

Appendix

List of appendices

Anhang 1	So funktioniert's	33
Anhang 1/1	Wieder mal eine Abbildung	33

Appendix 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



Fig. 14: Mal wieder das DHBW-Logo.

List of references

- Ackley, D. H./Hinton, G. E./Sejnowski, T. J. (1985): A Learning Algorithm for Boltzmann Machines. In: *Cognitive Science* 9.1, pp. 147–169. ISSN: 0364-0213. DOI: 10.1016/S0364-0213(85)80012-4. URL: <https://www.sciencedirect.com/science/article/pii/S0364021385800124> (retrieval: 02/16/2024).
- Ahad, N./Qadir, J./Ahsan, N. (2016): Neural Networks in Wireless Networks: Techniques, Applications and Guidelines. In: *Journal of Network and Computer Applications* 68, pp. 1–27. ISSN: 1084-8045. DOI: 10.1016/j.jnca.2016.04.006. URL: <https://www.sciencedirect.com/science/article/pii/S1084804516300492> (retrieval: 02/28/2024).
- Ahmad, A./Pasha, M. A. (2020): Optimizing Hardware Accelerated General Matrix-Matrix Multiplication for CNNs on FPGAs. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 67.11, pp. 2692–2696. ISSN: 1558-3791. DOI: 10.1109/TCSII.2020.2965154. URL: https://ieeexplore.ieee.org/abstract/document/8954788?casa_token=Ff1Z-99s30MAAAAA:l-4hcDRIsH2x9gRN3bGMy8BAo1nbQbrJEhqZpdRnAR5IJSe2naviSLmKFiAuYV_yuWVlAPPPdb (retrieval: 03/18/2024).
- Amari, S./Kurata, K./Nagaoka, H. (1992): Information Geometry of Boltzmann Machines. In: *IEEE Transactions on Neural Networks* 3.2, pp. 260–271. ISSN: 1941-0093. DOI: 10.1109/72.125867. URL: <https://ieeexplore.ieee.org/abstract/document/125867> (retrieval: 02/16/2024).
- Amirsoleimani, A./Alibart, F./Yon, V./Xu, J./Pazhouhandeh, M. R./Ecoffey, S./Beilliard, Y./Genov, R./Drouin, D. (2020): In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives. In: *Advanced Intelligent Systems* 2.11, p. 2000115. ISSN: 2640-4567. DOI: 10.1002/aisy.202000115. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202000115> (retrieval: 02/15/2024).
- Babu, P./Parthasarathy, E. (2021): Reconfigurable FPGA Architectures: A Survey and Applications. In: *Journal of The Institution of Engineers (India): Series B* 102.1, pp. 143–156. ISSN: 2250-2114. DOI: 10.1007/s40031-020-00508-y. URL: <https://doi.org/10.1007/s40031-020-00508-y> (retrieval: 03/20/2024).
- Bai, G./Chai, Z./Ling, C./Wang, S./Lu, J./Zhang, N./Shi, T./Yu, Z./Zhu, M./Zhang, Y./Yang, C./Cheng, Y./Zhao, L. (2024): Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models. DOI: 10.48550/arXiv.2401.00625. arXiv: 2401.00625 [cs]. URL: <http://arxiv.org/abs/2401.00625> (retrieval: 02/23/2024). preprint.
- Baischer, L./Wess, M./TaheriNejad, N. (2021): Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. DOI: 10.48550/arXiv.2104.09252. arXiv: 2104.09252 [cs]. URL: <http://arxiv.org/abs/2104.09252> (retrieval: 03/18/2024). preprint.
- Barra, A./Bernacchia, A./Santucci, E./Contucci, P. (2012): On the Equivalence of Hopfield Networks and Boltzmann Machines. In: *Neural Networks* 34, pp. 1–9. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.06.003. URL: <https://www.sciencedirect.com/science/article/pii/S0893608012001608> (retrieval: 02/16/2024).

- Beichl, I./Sullivan, F. (2000):** The Metropolis Algorithm. In: *Computing in Science & Engineering* 2.1, pp. 65–69. ISSN: 1558-366X. DOI: 10.1109/5992.814660. URL: <https://ieeexplore.ieee.org/document/814660> (retrieval: 02/27/2024).
- Böhm, F./Alonso-Urquijo, D./Verschaffelt, G./Van der Sande, G. (2022):** Noise-Injected Analog Ising Machines Enable Ultrafast Statistical Sampling and Machine Learning. In: *Nature Communications* 13.1 (1), p. 5847. ISSN: 2041-1723. DOI: 10.1038/s41467-022-33441-3. URL: <https://www.nature.com/articles/s41467-022-33441-3> (retrieval: 02/15/2024).
- Boutros, A./Betz, V. (2021):** FPGA Architecture: Principles and Progression. In: *IEEE Circuits and Systems Magazine* 21.2, pp. 4–29. ISSN: 1558-0830. DOI: 10.1109/MCAS.2021.3071607. URL: <https://ieeexplore.ieee.org/abstract/document/9439568> (retrieval: 03/20/2024).
- Cai, F./Kumar, Suhas/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> (retrieval: 02/15/2024). preprint.
- Cai, F./Kumar, Suhas/Van Vaerenbergh, T./Sheng, X./Liu, R./Li, C./Liu, Z./Foltin, M./Yu, S./Xia, Q./Yang, J. J./Beausoleil, R./Lu, W. D./Strachan, J. P. (2020):** Power-Efficient Combinatorial Optimization Using Intrinsic Noise in Memristor Hopfield Neural Networks. In: *Nature Electronics* 3.7, pp. 409–418. ISSN: 2520-1131. DOI: 10.1038/s41928-020-0436-6. URL: <https://www.nature.com/articles/s41928-020-0436-6> (retrieval: 03/21/2024).
- Cichy, R. M./Kaiser, D. (2019):** Deep Neural Networks as Scientific Models. In: *Trends in Cognitive Sciences* 23.4, pp. 305–317. ISSN: 1364-6613, 1879-307X. DOI: 10.1016/j.tics.2019.01.009. pmid: 30795896. URL: [https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613\(19\)30034-8](https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(19)30034-8) (retrieval: 02/23/2024).
- Dally, W. J./Keckler, S. W./Kirk, D. B. (2021):** Evolution of the Graphics Processing Unit (GPU). In: *IEEE Micro* 41.6, pp. 42–51. ISSN: 1937-4143. DOI: 10.1109/MM.2021.3113475. URL: <https://ieeexplore.ieee.org/abstract/document/9623445> (retrieval: 03/20/2024).
- Dario Amodei/Danny Hernandez (2024):** AI and Compute. URL: <https://openai.com/research/ai-and-compute> (retrieval: 02/15/2024).
- Dramsch, J. S. (2020):** ‘Chapter One - 70 Years of Machine Learning in Geoscience in Review’. In: *Advances in Geophysics*. Ed. by Ben Moseley/Lion Krischer. Vol. 61. Machine Learning in Geosciences. Elsevier, pp. 1–55. DOI: 10.1016/bs.agph.2020.08.002. URL: <https://www.sciencedirect.com/science/article/pii/S0065268720300054> (retrieval: 02/28/2024).
- Du, Y./Lin, T./Mordatch, I. (2021):** Model Based Planning with Energy Based Models. DOI: 10.48550/arXiv.1909.06878. arXiv: 1909.06878 [cs, stat]. URL: <http://arxiv.org/abs/1909.06878> (retrieval: 02/19/2024). preprint.

- Du, Y./Mordatch, I. (2020):** Implicit Generation and Generalization in Energy-Based Models. DOI: 10.48550/arXiv.1903.08689. arXiv: 1903.08689 [cs, stat]. URL: <http://arxiv.org/abs/1903.08689> (retrieval: 02/23/2024). preprint.
- Durstewitz, D./Koppe, G./Meyer-Lindenberg, A. (2019):** Deep Neural Networks in Psychiatry. In: *Molecular Psychiatry* 24.11 (11), pp. 1583–1598. ISSN: 1476-5578. DOI: 10.1038/s41380-019-0365-9. URL: <https://www.nature.com/articles/s41380-019-0365-9> (retrieval: 02/23/2024).
- Fahlman, S./Hinton, G./Sejnowski, T. (1983):** Massively Parallel Architectures for AI: NETL, Thistle, and Boltzmann Machines., p. 113. 109 pp.
- Fischer, A./Igel, C. (2012):** An Introduction to Restricted Boltzmann Machines. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. by Luis Alvarez/Marta Mejail/Luis Gomez/Julio Jacobo. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 14–36. ISBN: 978-3-642-33275-3. DOI: 10.1007/978-3-642-33275-3_2.
- Gawlikowski, J./Tassi, C. R. N./Ali, M./Lee, J./Humt, M./Feng, J./Kruspe, A./Triebel, R./Jung, P./Roscher, R./Shahzad, M./Yang, W./Bamler, R./Zhu, X. X. (2023):** A Survey of Uncertainty in Deep Neural Networks. In: *Artificial Intelligence Review* 56.1, pp. 1513–1589. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10562-9. URL: <https://doi.org/10.1007/s10462-023-10562-9> (retrieval: 02/23/2024).
- Gm, H./Gourisaria, M. K./Pandey, M./Rautaray, S. S. (2020):** A Comprehensive Survey and Analysis of Generative Models in Machine Learning. In: *Computer Science Review* 38, p. 100285. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2020.100285. URL: <https://www.sciencedirect.com/science/article/pii/S1574013720303853> (retrieval: 03/08/2024).
- Gustafsson, F. K./Danelljan, M./Bhat, G./Schön, T. B. (2020):** Energy-Based Models for Deep Probabilistic Regression. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi/Horst Bischof/Thomas Brox/Jan-Michael Frahm. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 325–343. ISBN: 978-3-030-58565-5. DOI: 10.1007/978-3-030-58565-5_20.
- Helmenstine, A. (2022):** How Many Atoms Are in the World? Science Notes/Projects. URL: <https://sciencenotes.org/how-many-atoms-are-in-the-world/> (retrieval: 02/21/2024).
- 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.
- Hinton, G. (2014):** ‘Boltzmann Machines’. In: *Encyclopedia of Machine Learning and Data Mining*. Ed. by Claude Sammut/Geoffrey I. Webb. Boston, MA: Springer US, pp. 1–7. ISBN: 978-1-4899-7502-7. DOI: 10.1007/978-1-4899-7502-7_31-1. URL: https://link.springer.com/10.1007/978-1-4899-7502-7_31-1 (retrieval: 03/16/2024).
- Hinton, G. E. (2012a):** ‘A Practical Guide to Training Restricted Boltzmann Machines’. In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon/Geneviève B. Orr/Klaus-Robert Müller. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer,

- pp. 599–619. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_32. URL: https://doi.org/10.1007/978-3-642-35289-8_32 (retrieval: 02/15/2024).
- Hinton, G. E. (2012b)**: ‘A Practical Guide to Training Restricted Boltzmann Machines’. In: *Neural Networks: Tricks of the Trade*. Ed. by Grégoire Montavon/Geneviève B. Orr/Klaus-Robert Müller. Vol. 7700. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 599–619. ISBN: 978-3-642-35288-1 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_32. URL: http://link.springer.com/10.1007/978-3-642-35289-8_32 (retrieval: 02/15/2024).
- Hizzani, M./Heittmann, A./Hutchinson, G./Dobrynin, D./Van Vaerenbergh, T./Bhattacharya, T./Renaudineau, A./Strukov, D./Strachan, J. P. (2023)**: Memristor-Based Hardware and Algorithms for Higher-Order Hopfield Optimization Solver Outperforming Quadratic Ising Machines. DOI: 10.48550/arXiv.2311.01171. arXiv: 2311.01171 [cs]. URL: <http://arxiv.org/abs/2311.01171> (retrieval: 02/15/2024). preprint.
- Hopfield, J. J. (1982)**: Neural Networks and Physical Systems with Emergent Collective Computational Abilities. In: *Proceedings of the National Academy of Sciences* 79.8, pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554. URL: <https://www.pnas.org/doi/10.1073/pnas.79.8.2554> (retrieval: 02/19/2024).
- Hsu, A./Griffiths, T. (2010)**: Effects of Generative and Discriminative Learning on Use of Category Variability. In: *Proceedings of 32nd Annual Conference of the Cognitive Science Society*.
- Hu, Y./Liu, Y./Liu, Z. (2022)**: A Survey on Convolutional Neural Network Accelerators: GPU, FPGA and ASIC. In: *2022 14th International Conference on Computer Research and Development (ICCRD)*. 2022 14th International Conference on Computer Research and Development (ICCRD), pp. 100–107. DOI: 10.1109/ICCRD54409.2022.9730377. URL: <https://ieeexplore.ieee.org/abstract/document/9730377> (retrieval: 03/19/2024).
- Huembeli, P./Arrazola, J. M./Killoran, N./Mohseni, M./Wittek, P. (2022)**: The Physics of Energy-Based Models. In: *Quantum Machine Intelligence* 4.1, p. 1. ISSN: 2524-4914. DOI: 10.1007/s42484-021-00057-7. URL: <https://doi.org/10.1007/s42484-021-00057-7> (retrieval: 02/19/2024).
- Ising, E. (1925)**: Beitrag zur Theorie des Ferromagnetismus. In: *Zeitschrift für Physik* 31.1, pp. 253–258. ISSN: 0044-3328. DOI: 10.1007/BF02980577. URL: <https://doi.org/10.1007/BF02980577> (retrieval: 03/21/2024).
- Izadkhah, H. (2022)**: ‘P, NP, NP-Complete, and NP-Hard Problems’. In: *Problems on Algorithms: A Comprehensive Exercise Book for Students in Software Engineering*. Ed. by Habib Izadkhah. Cham: Springer International Publishing, pp. 497–511. ISBN: 978-3-031-17043-0. DOI: 10.1007/978-3-031-17043-0_15. URL: https://doi.org/10.1007/978-3-031-17043-0_15 (retrieval: 03/21/2024).
- Larochelle, H./Bengio, Y. (2008)**: Classification Using Discriminative Restricted Boltzmann Machines. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. New York, NY, USA: Association for Computing Machinery, pp. 536–543. ISBN: 978-1-60558-205-4. DOI: 10.1145/1390156.1390224. URL: <https://dl.acm.org/doi/10.1145/1390156.1390224> (retrieval: 02/22/2024).

- Larochelle, H./Mandel, M./Pascanu, R./Bengio, Y. (2012):** Learning Algorithms for the Classification Restricted Boltzmann Machine. In: *The Journal of Machine Learning Research* 13, pp. 643–669.
- Lehnert, A./Holzinger, P./Pfenning, S./Müller, R./Reichenbach, M. (2023):** Most Resource Efficient Matrix Vector Multiplication on FPGAs. In: *IEEE Access* 11, pp. 3881–3898. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3234622. URL: <https://ieeexplore.ieee.org/document/10007836?denied=> (retrieval: 03/16/2024).
- Li, G./Deng, L./Xu, Y./Wen, C./Wang, W./Pei, J./Shi, L. (2016):** Temperature Based Restricted Boltzmann Machines. In: *Scientific Reports* 6.1 (1), p. 19133. ISSN: 2045-2322. DOI: 10.1038/srep19133. URL: <https://www.nature.com/articles/srep19133> (retrieval: 02/27/2024).
- Lucas, A. (2014):** Ising Formulations of Many NP Problems. In: *Frontiers in Physics* 2. ISSN: 2296-424X. DOI: 10.3389/fphy.2014.00005. URL: <https://www.frontiersin.org/articles/10.3389/fphy.2014.00005> (retrieval: 03/22/2024).
- 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> (retrieval: 02/15/2024). preprint.
- MacKay, D. J. C. (2003):** Information Theory, Inference and Learning Algorithms. Cambridge University Press. 694 pp. ISBN: 978-0-521-64298-9. Google Books: AKuMj4PN EMC.
- Mall, P. K./Singh, P. K./Srivastav, S./Narayan, V./Paprzycki, M./Jaworska, T./Ganzha, M. (2023):** A Comprehensive Review of Deep Neural Networks for Medical Image Processing: Recent Developments and Future Opportunities. In: *Healthcare Analytics* 4, p. 100216. ISSN: 2772-4425. DOI: 10.1016/j.health.2023.100216. URL: <https://www.sciencedirect.com/science/article/pii/S2772442523000837> (retrieval: 02/23/2024).
- Marinó, G. C./Petrini, A./Malchiodi, D./Frasca, M. (2023):** Deep Neural Networks Compression: A Comparative Survey and Choice Recommendations. In: *Neurocomputing* 520, pp. 152–170. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2022.11.072. URL: <https://www.sciencedirect.com/science/article/pii/S0925231222014643> (retrieval: 02/23/2024).
- Metropolis, N./Rosenbluth, A. W./Rosenbluth, M. N./Teller, A. H./Teller, E. (1953):** Equation of State Calculations by Fast Computing Machines. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. ISSN: 0021-9606. DOI: 10.1063/1.1699114. URL: <https://doi.org/10.1063/1.1699114> (retrieval: 02/27/2024).
- Mocanu, D. C./Mocanu, E./Nguyen, P. H./Gibescu, M./Liotta, A. (2016):** A Topological Insight into Restricted Boltzmann Machines. In: *Machine Learning* 104.2, pp. 243–270. ISSN: 1573-0565. DOI: 10.1007/s10994-016-5570-z. URL: <https://doi.org/10.1007/s10994-016-5570-z> (retrieval: 02/22/2024).
- Mohseni, N./McMahon, P. L./Byrnes, T. (2022a):** 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> (retrieval: 02/15/2024). preprint.

- Mohseni, N./McMahon, P. L./Byrnes, T. (2022b):** Ising Machines as Hardware Solvers of Combinatorial Optimization Problems. In: *Nature Reviews Physics* 4.6, pp. 363–379. ISSN: 2522-5820. DOI: 10.1038/s42254-022-00440-8. URL: <https://www.nature.com/articles/s42254-022-00440-8> (retrieval: 03/22/2024).
- Nazm Bojnordi, M./Ipek, E. (2016):** Memristive Boltzmann Machine: A Hardware Accelerator for Combinatorial Optimization and Deep Learning, p. 13. 1 p. DOI: 10.1109/HPCA.2016.7446049.
- Patrón, A./Chepelianskii, A. D./Prados, A./Trizac, E. (2024):** On the Optimal Relaxation Rate for the Metropolis Algorithm in One Dimension. DOI: 10.48550/arXiv.2402.11267. arXiv: 2402.11267 [cond-mat, physics:math-ph]. URL: <http://arxiv.org/abs/2402.11267> (retrieval: 02/27/2024). preprint.
- Peccerillo, B./Mannino, M./Mondelli, A./Bartolini, S. (2022):** A Survey on Hardware Accelerators: Taxonomy, Trends, Challenges, and Perspectives. In: *Journal of Systems Architecture* 129, p. 102561. ISSN: 1383-7621. DOI: 10.1016/j.sysarc.2022.102561. URL: <https://www.sciencedirect.com/science/article/pii/S1383762122001138> (retrieval: 03/19/2024).
- Ramsauer, H./Schäfl, B./Lehner, J./Seidl, P./Widrich, M./Adler, T./Gruber, L./Holzleitner, M./Pavlović, M./Sandve, G. K./Greiff, V./Kreil, D./Kopp, M./Klambauer, G./Brandstetter, J./Hochreiter, S. (2021):** Hopfield Networks Is All You Need. DOI: 10.48550/arXiv.2008.02217. arXiv: 2008.02217 [cs, stat]. URL: <http://arxiv.org/abs/2008.02217> (retrieval: 02/28/2024). preprint.
- Ranzato, M. a./Poultney, C./Chopra, S./Cun, Y. (2006):** Efficient Learning of Sparse Representations with an Energy-Based Model. In: *Advances in Neural Information Processing Systems*. Vol. 19. MIT Press. URL: <https://proceedings.neurips.cc/paper/2006/hash/87f4d79e36d68c3031ccf6c55e9bbd39-Abstract.html> (retrieval: 03/07/2024).
- Robert, C. P. (2016):** The Metropolis-Hastings Algorithm. DOI: 10.48550/arXiv.1504.01896. arXiv: 1504.01896 [stat]. URL: <http://arxiv.org/abs/1504.01896> (retrieval: 02/27/2024). preprint.
- Rosenthal, S. (2009):** Optimal Proposal Distributions and Adaptive MCMC. In: *Handbook of Markov Chain Monte Carlo*.
- Salakhutdinov, R./Hinton, G. (2009):** Deep Boltzmann Machines. In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 448–455. URL: <https://proceedings.mlr.press/v5/salakhutdinov09a.html> (retrieval: 02/16/2024).
- Schlamming, S. (2014):** A Cool Way to Measure Big G. In: *Nature* 510.7506 (7506), pp. 478–480. ISSN: 1476-4687. DOI: 10.1038/nature13507. URL: <https://www.nature.com/articles/nature13507> (retrieval: 02/21/2024).
- Singh, S. K./Kumar, Shubham/Mehra, P. S. (2023):** Chat GPT & Google Bard AI: A Review. In: *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*. 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), pp. 1–6. DOI: 10.1109/ICICAT57735.2023.10263706. URL: <https://doi.org/10.1109/ICICAT57735.2023.10263706>

- [//ieeexplore.ieee.org/abstract/document/10263706?casa_token=JMHwBzQgxnwAAAAA:700nfYs5ECetZhuq8D_F3QXyua1Xu65rLOa_Ywve3mch00UAeSs0yVjhWCUVdUBpMX83NAbpUpM](https://ieeexplore.ieee.org/abstract/document/10263706?casa_token=JMHwBzQgxnwAAAAA:700nfYs5ECetZhuq8D_F3QXyua1Xu65rLOa_Ywve3mch00UAeSs0yVjhWCUVdUBpMX83NAbpUpM) (retrieval: 02/23/2024).
- Sipola, T./Alatalo, J./Kokkonen, T./Rantonen, M. (2022):** Artificial Intelligence in the IoT Era: A Review of Edge AI Hardware and Software. In: *2022 31st Conference of Open Innovations Association (FRUCT)*. 2022 31st Conference of Open Innovations Association (FRUCT), pp. 320–331. DOI: 10.23919/FRUCT54823.2022.9770931. URL: <https://ieeexplore.ieee.org/abstract/document/9770931> (retrieval: 03/18/2024).
- Specht, D. F. (1990):** Probabilistic Neural Networks. In: *Neural Networks* 3.1, pp. 109–118. ISSN: 0893-6080. DOI: 10.1016/0893-6080(90)90049-Q. URL: <https://www.sciencedirect.com/science/article/pii/089360809090049Q> (retrieval: 03/07/2024).
- Sung, C./Hwang, H./Yoo, I. K. (2018):** Perspective: A Review on Memristive Hardware for Neuromorphic Computation. In: *Journal of Applied Physics* 124.15, p. 151903. ISSN: 0021-8979. DOI: 10.1063/1.5037835. URL: <https://doi.org/10.1063/1.5037835> (retrieval: 02/15/2024).
- Tanahashi, K./Takayanagi, S./Motohashi, T./Tanaka, S. (2019):** Application of Ising Machines and a Software Development for Ising Machines. In: *Journal of the Physical Society of Japan* 88.6, p. 061010. ISSN: 0031-9015. DOI: 10.7566/JPSJ.88.061010. URL: <https://journals.jps.jp/doi/full/10.7566/JPSJ.88.061010> (retrieval: 03/21/2024).
- Upadhyaya, V./Sastry, P. (2019):** An Overview of Restricted Boltzmann Machines. In: *Journal of the Indian Institute of Science* 99. DOI: 10.1007/s41745-019-0102-z.
- Uusitalo, L./Lehikoinen, A./Helle, I./Myrberg, K. (2015):** An Overview of Methods to Evaluate Uncertainty of Deterministic Models in Decision Support. In: *Environmental Modelling & Software* 63, pp. 24–31. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2014.09.017. URL: <https://www.sciencedirect.com/science/article/pii/S1364815214002813> (retrieval: 03/07/2024).
- Verdon, G./Marks, J./Nanda, S./Leichenauer, S./Hidary, J. (2019):** Quantum Hamiltonian-Based Models and the Variational Quantum Thermalizer Algorithm. arXiv: 1910.02071 [quant-ph]. URL: <http://arxiv.org/abs/1910.02071> (retrieval: 02/19/2024). preprint.
- 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> (retrieval: 02/15/2024). preprint.
- Wang, X./Yan, L./Zhang, Q. (2021):** Research on the Application of Gradient Descent Algorithm in Machine Learning. In: *2021 International Conference on Computer Network, Electronic and Automation (ICCNEA)*. 2021 International Conference on Computer Network, Electronic and Automation (ICCNEA), pp. 11–15. DOI: 10.1109/ICCNEA53019.2021.00014. URL: <https://ieeexplore.ieee.org/document/9603742> (retrieval: 03/08/2024).
- Wittpahl, V., ed. (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> (retrieval: 02/15/2024).

- Yao, Z./Gripon, V./Rabbat, M. (2013):** A Massively Parallel Associative Memory Based on Sparse Neural Networks. In.
- Zhai, S./Cheng, Y./Lu, W./Zhang, Z. (2016):** Deep Structured Energy Based Models for Anomaly Detection. In: *Proceedings of The 33rd International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 1100–1109. URL: <https://proceedings.mlr.press/v48/zhai16.html> (retrieval: 02/19/2024).
- Zhang, N./Ding, S./Zhang, J./Xue, Y. (2018):** An Overview on Restricted Boltzmann Machines. In: *Neurocomputing* 275, pp. 1186–1199. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2017.09.065. URL: <https://www.sciencedirect.com/science/article/pii/S0925231217315849> (retrieval: 02/15/2024).
- Zhou, H./Dong, J./Cheng, J./Dong, W./Huang, C./Shen, Y./Zhang, Q./Gu, M./Qian, C./Chen, H./Ruan, Z./Zhang, X. (2022):** Photonic Matrix Multiplication Lights up Photonic Accelerator and Beyond. In: *Light: Science & Applications* 11.1, p. 30. ISSN: 2047-7538. DOI: 10.1038/s41377-022-00717-8. URL: <https://www.nature.com/articles/s41377-022-00717-8> (retrieval: 03/18/2024).

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)