# Background and Discussion on Use and Implementation of Deep Learning with FPGA

1st Dimitri Häring
*School of Engineering (Electrical Engineering)*
*Grand Valley State University*
Grand Rapids, MI, United States of America
haringd@mail.gvsu.edu

***Abstract*—The propose of the paper is to introduce the audience to the background of deep learning with an focus on computational challenges and approaches.**

**Using advanced searches in the IEEE explorer allowed to gain access to published papers about the researched topics to outline the boundaries of the general concepts that have to be explained, elaborated and discussed. Furthermore, Furthermore, actual newspaper articles as well as free lecturers from Standford university where used to build a knowledge base that allowed the author to boldly go where a lot of engineers have been gone before. In addition, the comparison of multiple sources allowed to gain new knowledge.**

**Results indicate that the implementation of deep learning with artificial neural network can be used as accelerator and improve energy efficiency significantly.**

***Index Terms*—Deep Learning, Artificial Neural Network, Planetary Rover**

## I. Introduction

This paper is an introduction into deep learning. First, an overview is presented with a comparison between machine learning and deep learning. This allows readers with less background to follow the reasons why deep learning often is associated with neural networks in particular artificial neural networks.

Second, computational challenges are discussed in terms of performance and power efficiency are discussed and figures given to give an scene of magnitude of different sizes of neural networks.

Third, computational approaches are discussed by using two examples that differs in complexity and performance. Therefor, Section II introduces the audience to the necessary background required to understand what machine and deep learning means.

## II. Background

### A. General Concepts

Learning is a commonly used word and means according to Merriam Webster the following [1]:

1 : the act or experience of one that learns
\\ a computer program that makes learning fun
2 : knowledge or skill acquired by instruction or study
\\ people of good education and considerable learning
3 : modification of a behavioral tendency by experience (such as exposure to conditioning)

This means that a machine that learns receives either instructions or experiences something. This leads to the not unreasonable assumption that if a machine learns, somebody teaches the machine in form of instructions or exploits the machine to experience. To give a machine instructions is most likely the essence of the todays computing. While a machine is capable of receiving instructions in many different forms the most common at the time is still the users instruction, according to the authors opinion. By applying this concept to a machine that learns a instruction in terms of machine learning not a single instruction is given a algorithm is implemented that similar to adaptive filtering in signal processing wights an input $x$ and will provide an output $z$. The a function $W$ is used to define the weight according to different methods as example linear regression or a second order polynomial. The issue for such an algorithm is the weighting function $W$ needs to be defined by a human that invests a lot of time studying cases and tries to figure out what might be the best function to weight the input for a certain situation. By introducing this thought of weight which is implemented as a simple multiplication of an input value $x$. This allows to use a vector to describe inputs hence the weights would be a vector of the same length. Now if as example an image of 128x128 pixels would be wighted a vector of length 16384 would have to be used, assuming it is a chromatic picture with a single integer that defines each pixel. This is a real simple example as you start thinking about it but nobody in the world would try to do that with no strategy on hand. To make a point, a machine would have to learn how to interpret a picture depending on the state of the wight of each pixel or most likely groups of pixels. A machine could improve the resolution of his weighting function by learning known pictures where the result is known which is known as **supervised learning**. Due to the fact that a machine can learn such sets and define a wight for each situation very fast a machine is capable of learning more situations in days than a human is capable of in years.

At this point lets create a hypothetical scenario to understand learning furthermore. Assume that a human with all cognitive abilities that we have just start to exist on a plane with a river and a sun. As a human needs a lot of water most likely this human would experience a sense of thirstiness at some point. The human has no teacher and no previous

knowledge so how can the human know that he can drink from the river if not learned from a teacher.

Assume there is an animal too on this plane that drinks from the water. If there is no teacher and a similar situation and obviously no other option than to explore the river the human might be try to do the same even not knowing that it will help to still the sense of thirstiness. The human **learned unsupervised** by experience of a situation as a witness.

This rises the question would make a human brain the decision that it could try to drink from the river with having a single reference. The authors best guess is that most likely a decision would be either drinking, not drinking or something else not expected like trying to communicate with the animal and ask it why it drinks. This shows that a solution can have more possibilities then just yes or no which can be **classified** as a vector.

In summary, there are three important terms to know, highlighted in the previous Section II-A which are input, output, supervised learning, unsupervised learning, classification. Notice, this is not meant to be a full list of concepts there is more to explore and to know. Due to the fact that the paper shall not exceed a certain length not every concept is discussed rather than a subset is chosen that seems in the authors opinion the most important to understand the background of the matter.

The next Section II-B is dedicated to machine learning.

### *B. Machine Learning*

Where as general concepts has been discussed in the previous section II-A in this section machine learning will be discussed. In machine learning all the previous discussed concepts can be applied. As the previous discussed example of an 128x 128 picture introduced the goal is to find functions that wights the input there for in machine learning which involves usually no more then one non linear logic layer that is applied to each input vector. The goal is to optimize the weight function $W$ in case of using the concept of supervised learning it would be a predefined set of pictures with known result that is instructed to machine. A commonly known and widely used set is Caltech 101 [2].

Know most decisions are not based on a simple algorithm or one set of pixels it is usually a combination of those patterns. Due to the fact that humans are not capable of performing complex tasks only simple ones in series or parallel, it might not be the best way to analyze a picture as such.

Therefor, a artificial neural network is discussed in Section II-C.

### *C. Artificial Neural Network*

As the previous Section II-B discussed briefly machine learning and introduced the process of decisions this section is discussing artificial neural networks (ANN).

A brain is build with neurons, a single neuron is shown in Figure 1. According to Merriam Webster a neuron is defined as the following [3]:

> : a grayish or reddish granular cell that is the fundamental functional unit of nervous tissue transmitting and receiving nerve impulses and having cytoplasmic processes which are highly differentiated frequently as multiple dendrites or usually as solitary axons which conduct impulses to and away from the cell body : NERVE CELL sense 1

In the authors interpretation this means that a neuron receives a input signal $x$ from another neuron. A neuron outputs a signal $z$ to the next neuron. The neuron does weight the signal with function $W$ by modifying it as shown in Equation 1.

$$\vec{W}(\vec{x}) = \vec{z} \tag{1}$$

At this point it seams to be appropriate that it is commonly known that the eight function $W$ is a non linear function. The reason is kind of simple if it would be a linear function then you could use Gauss to solve for an equation if the matrices are invertible. However, the point is that the brain does network with a single neuron it works with many neurons each connected to each other in an three dimensional space, which makes it kind of complex at the end. Therefor, do not try to copy a machine that you do not understand it is not an engineer that will solve the brains complexity most likely it will be an neuron scientist and the engineer will use his research and apply the research to common technology. To

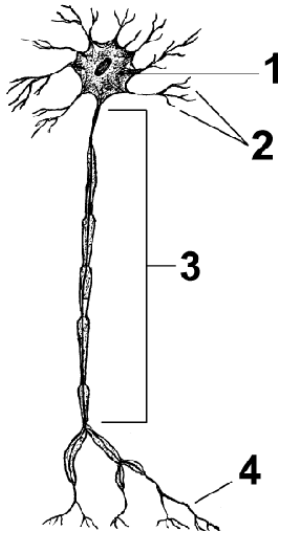


Fig. 1. Illustration of Neuron: 1 cell body, 2 dendrite, 3 axon, 4 nerve ending [3].

emphasis the functions of multiple neurons an ANN is build out of multiple non linear logic layers. Figure 2 shows how one logic layer would look like based on formula presented in Equation 1. Notice, that in this example not a specific algorithm commonly known is used because the algorithm would be chosen or have been designed to specific problem. By accumulating multiple non linear logic layers to a vast network where as usually each layer is dedicated to a specific task an artificial neural network is build.

Section II-D will discuss the use of neural networks in a field called deep learning.

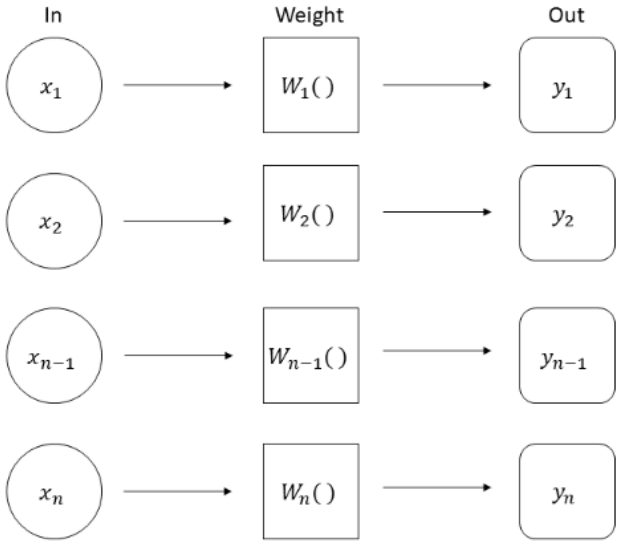


Fig. 2. Visualized algorithm of Equation 1.

### *D. Deep Learning*

As ANN has been discussed in Section II-C this section will provide inside why deep learning is a subset of machine learning.

As the knowledge has been gathered that an ANN is a cumulation of non liner logic layers and an Machine learning algorithm was defined as a single layer of a non linear logic which determines the main difference between machine learning and deep learning. Deep refers to the number of non linear logic levels that are implied in the ANN. Therefor, Schmidhuber introduced a concept of credit assignment for each stage and depending on the number of credits assigned a distinction between shallow learning and deep learning can be made [4]. Due to the scope of the paper the shallow learning will not be further discussed.

Section III will provide a deeper inside to specific applications where deep learning is applied in combination with an FPGA to accelerate process speed and energy efficiency.

## III. Computational challenges

As discussed in Section II-D deep learning uses in general a form of an ANN. In this section discuss why it is useful to use an FPGA for the ANN implementation.

As discussed each layer of an ANN represents an weighting function so that in general one layer could be reduced to the thought of an flip-flop of an shift register. Each time the shift register shifts the information trough the layers more information about the input will be gathered by applying different methods. As we can identify that each layer has similar patterns to general logic it makes sense to implement it on an FPGA platform. Hedge et. al [5] performed performance measurements for different supervised learning sets, to name one Caltech101 [2]. To implement the neural network (NN) the team used an MXP-enhanced FPGA platform because it delivers 1.4-5x higher energy efficiency than all other platforms. Furthermore, the supervised learning sets where run on a GPU, CPU, and DSP platform.

Gankidi et. al [6] discusses in his work FPGA Architecture for deep learning and its application to planetary robotics where the problem of constrained embedded system and the power efficiency is considered critical. The comparison between an i5 2.3 GHz CPU and an Xilinx-Space-garde Virtex FPGA is made in terms of energy performance. As introduction Gankidi et. al provides also interesting figures to clarify the need for FPGA architecture on constrained devises by stating that Google brain uses around 16,000 CPUs and consumes around 5 MW of power. While in comparison an a planetary rower with an radiation hardened processor RAD750 uses 5 W of power which differs in the magnitude of $10^6$ in power consumption. This clearly states that there are boundaries to computing and paramount challenges to implement a big enough ANN onto a continent embedded systems. Therefore an FPGA could significantly enhance performance depending on the algorithm and task it is assigned for.

Section IV will discuss an approach of an implemented convolution neural network on FPGA.

## IV. Computational approaches

As Section III introduced performance advantages of the FPGA platforms other other architectures this section will focused on two applied examples on how to implement deep learning ANN on FPGA architecture.

Bacis et. al [7] used a pipelined and scalable dataflow implementation of convolutional neural networks on FPGA with ANN for image classification and image recognition. The convolutional neural network structure is shown in Figure 3. To the left the image is shown that is used as input. The image is first processed by the convolutional layer. The next layer is a sub-sampling layer followed by an additional convolutional layer. Then a linear layer is applied that pipes into the classification layer and presents afterwards an output. As shown on this example of an applied algorithm the further discussed model apply in reality and can be used with real algorithm to implement deep learning with ANN on an FPGA platform.

Furthermore, Gankidi et. al [6] used an Vertex 7 FPGA platform to implement an Q-learning preceptor and used the FPGA platform as accelerator. The term accelerator is used because compared to the CPU architecture the FPGA accelerates the task in time domain by a multiple factor depending on what task is performed. The most impressive result could be seen in his results section in Table 4, where FPGA -Virtex 7, Fixed point outperforms an CPU - Intel i5 2.3 GHz processor by 95x.

## V. Conclusion

After introducing the audition to the topic and sate common definitions with simple scenarios two papers where chosen

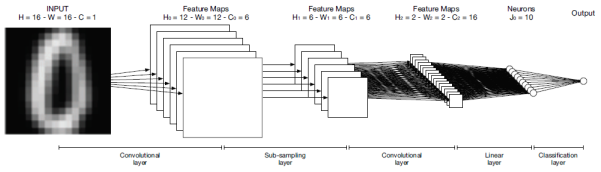


Fig. 3. Convolutional Neural Network structure [7].

to present as example. The pipelined convolutional FPGA approach as well as the planetary rover application shows that FPGA show significant advantages in power efficiency and processing speed by 95x. In addition, the reader could gain a sense of what is it about when scientist talk about machine and deep learning and it allows to draw a line that for sure a small network one million connections is not an artificial intelligence (AI).

## REFERENCES


[1] [1] "Learning — Definition of Learning by Merriam-Webster." [Online]. Available: https://www.merriam-webster.com/dictionary/learning. [Accessed: 26-Sep-2018].

[2] [2] "Caltech101." [Online]. Available: http://www.vision.caltech.edu/Image_Datasets/Caltech101/. [Accessed: 26-Sep-2018].

[3] [3] "Neuron — Definition of Neuron by Merriam-Webster." [Online]. Available: https://www.merriam-webster.com/dictionary/neuron. [Accessed: 27-Sep-2018].

[4] [4] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, Jan. 2015.

[5] [5] G. Hegde, Siddhartha, N. Ramasamy, V. Buddha, and N. Kapre, "Evaluating Embedded FPGA Accelerators for Deep Learning Applications," in 2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), 2016, pp. 25–25.

[6] [6] P. R. Gankidi and J. Thangavelautham, "FPGA architecture for deep learning and its application to planetary robotics," in 2017 IEEE Aerospace Conference, 2017, pp. 1–9.

[7] [7] M. Bacis, G. Natale, E. Del Sozzo, and M. D. Santambrogio, "A pipelined and scalable dataflow implementation of convolutional neural networks on FPGA," in 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017, pp. 90–97.