Technical University of Crete

Diploma Thesis

Embedded System Architecture for the Acceleration of Collaborative Learning in Neural Networks

Author:
Emmanouil Petrakos

Thesis Committee:
Prof. Apostolos Dollas
Associate Prof. Michail G. Lagoudakis
Dr. Vassilis Papaefstathiou (FORTH-ICS)

A thesis submitted in fulfillment of the requirements for the diploma of Electrical and Computer Engineer

in the

School of Electrical and Computer Engineering
Microprocessor and Hardware Laboratory

May 25, 2022

TECHNICAL UNIVERSITY OF CRETE

Abstract

School of Electrical and Computer Engineering

Electrical and Computer Engineer

Embedded System Architecture for the Acceleration of Collaborative Learning in Neural Networks

by Emmanouil PETRAKOS

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

TECHNICAL UNIVERSITY OF CRETE

Abstract

School of Electrical and Computer Engineering

Electrical and Computer Engineer

Embedded System Architecture for the Acceleration of Collaborative Learning in Neural Networks

by Emmanouil PETRAKOS

Η περίληψη της διπλωματικής γράφεται εδώ (και συνήθως αποτελεί αυτή την μία μόνο σελίδα). Η σελίδα αυτή κρατάται στοιχισμένη στην μέση οριζόντια και κάθετα, ώστε να μπορεί να επεκτίνεται στον κενό χώρο και πάνω από τον τίτλο...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor…

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
CCPA	**C**alifornia **C**onsumer **P**rivacy **A**ct
CNN	**C**onvolutional **N**eural **N**etwork
CPU	**C**entral **P**rocessor **U**nit
DNN	**D**eep **N**eural **N**etwork
FL	**F**ederated **L**earning
FPGA	**F**ield **P**rogrammable **G**ate **A**rray
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
MEC	**M**ulti-access **E**dge **C**omputing
ML	**M**achine **L**earning
RNN	**R**ecurrent **N**eural **N**etwork

Dedicated to my family and friends…

Chapter 1

Introduction

In recent years, edge devices with advanced computing and data collection capabilities are becoming commonplace. As a result, massive volumes of new and useful data are generated, which can be exploited in Machine Learning (ML). When combined with recent advances and techniques in ML, new opportunities emerge in a variety of fields, including self-driving automobiles and medical applications.

Traditional ML approaches demand the data to be consolidated in a single entity where learning takes place. However, due to unacceptable latency and storage requirements of centralizing huge amounts of raw data, this may be undesirable. To address the inefficiency of data silos, cloud computing architectures such as Multi-access edge computing (MEC) [1] have been proposed in order to transfer the learning closer to where the data is produced. Unfortunately, these techniques still require raw data to be shared between the edge devices and intermediate servers.

Due to growing privacy concerns, recent legislation like General Data Protection Regulation (GDPR) [2] and California Consumer Privacy Act (CCPA) [3] have severely limited the usage of technologies that transfer private data. To continue leveraging the increasing real-world data while adhering to such regulations, the concept of Federated Learning (FL) [4] has been introduced. FL is a collaboratively decentralized privacy-preserving technology, in which learning takes place at the data collection point, i.e. the edge device. The edge devices train a ML model provided by the server and share model updates instead of raw data. As a result, collaborative and distributed ML is possible while maintaining the privacy of the participating devices.

1.1 Motivation

Most FL research, to our knowledge, focuses on simulations and treats edge devices as black boxes; generally ignoring their nature and constrains. Taking in consideration the complexities from implementing ML on hardware, recent advancements in FL might be diminished or invalidated. The main motivation of this thesis is to identify, explore and possibly overcome the intrinsic conflicts that exist between FL and Artificial Neural Network (ANN) training in Field Programmable Gate Arrays (FPGA)s.

Beside being incompatible, these two technologies may complement each other, which is something worth investigating. Frequently in FL, transformations are applied on the generated ANN variables to reduce network utilization and enhance privacy. These transformations, which include quantization [5], adding Gaussian noise [6] and others, tend to be spatially independent and could be implemented highly efficiently in hardware accelerators like FPGAs.

Finally, FL literature is almost devoid of wall-clock time examples. This thesis aims to provide a real world FL implementation that may be considered as a benchmark for future research. Furthermore, in order to be extendable and utilized in future works, the implementation is modular and platform independent.

1.2 Scientific Contributions

The main focus of this thesis is combining FL training with FPGA implementations of ANN, while exploring and overcoming their inherent conflicts.
Furthermore, it focus on the mostly unexplored FL setting of small client pools and its inherent difficulties.
Finally, it gives a real world implementation of FL that can be used as a benchmark for future works. It provides an FL implementation that is agnostic to the ANN training implementation and can be used as a starting point for future works.

1.3 Thesis Outline

- **Chapter 2 - Theoretical Background:** Description of the theoretical background of ML and FL.

- **Chapter 3 - Related Work:** Related works on FL, optimization techniques and hardware implementations of it.
- **Chapter 4 - Robustness Analysis:** Chapter 4 description
- **Chapter 5 - FPGA Implementation:** Chapter 5 description
- **Chapter 6 - Results:** Chapter 6 description
- **Chapter 7 - Conclusions and Related Work:** Chapter 7 description

Chapter 2

Theoretical Background

2.1 Artificial Intelligence & Machine Learning

Various researchers and textbooks may provide different definitions of Artificial Intelligence (AI). Depending the school of though, AI is an artificial actor that thinks or acts, rationally or human-like, depending on what it knows. Generally, AI can be described as the study of intelligence agents. It is a modern science that encompasses a large variety of sub-fields, ranging from general-purpose areas, such as learning, to specific tasks like playing chess and giving medical diagnoses. AI can be relevant to any intellectual field, as it systematizes and automates intellectual tasks. [7]

Machine learning (ML) is an AI field in which agents, in addition to the performance element, include a learning element that utilises their past experiences to enhance their behaviour. The core idea behind ML is that perception should be used to improve the ability to act in the future, not simply react in the present. Designing a learning element is a multi-facet problem that is affected by three major issues. [8]

Information management

The first issue is determining what information what information is useful and how it should be utilized. Different components of the input and output data should be learnt depending on the context in which the learning actor operates. One method is to directly link the current state of the actor or the world to their actions. Sometimes it can be more appropriate to infer relevant patterns from the data while ignoring unnecessary information. Another way is to collect action-value information indicating the desirability of actions based on their effect in the world state. These and other options may need to be combined in order to extract the most meaningful knowledge from the available data.

Another key factor when designing learning systems is the availability of prior knowledge. Researchers have extensively looked into the issue where the agent uses only information that they encounter, but ways for transferring prior knowledge have been devised to speed up learning and improve decision-making.[9]

Feedback mechanism

The type of feedback available has a significant impact on the design and is perhaps the most crucial aspect of the learning problem. Usually three major types are distinguished: supervised, unsupervised, and reinforcement learning.

Supervised learning problems involve learning functions between sets of inputs and outputs. This is the case of a fully observable environments where the effects of the actors actions are immediately visible or the existence of a third party providing the correct solutions.

Unsupervised learning problems, on the other hand, do not supply output values and learning patterns are solely based on the input. As it has no knowledge of what constitutes a correct action or a desired state, an unsupervised learning agent cannot learn what to do. This is a common scenario for probabilistic reasoning systems or when generating output data is prohibitively expensive. For the last case, a semi-supervised learning setting, in which only a subset of the outputs is generated, might be useful.

In the reinforcement learning setting there is no correct output provided, instead a reward is given to actor appropriate to the desirability of their actions. This is common when the world which the actor take part in continuously change according to their actions, or a desirable or undesirable state may be reached after a series of actions.

Representation of the learned information

The representation of the learned information is another important factor in establishing how the learning algorithm should operate. Common schemes include linear weighted polynomials for utility functions, propositional or first order logic, probabilistic representations like Bayesian Networks[10] and ANNs[11], and other methods have all been created.

2.2 Deep learning

Deep learning is a sub-field of ML, partially overlapping with big data science. It consists of algorithms that use the perceptron as their basic building block, which is a mathematical function based on the McCulloch-Pitts model of biological neurons. They typically have hundreds of thousands to millions of perceptors with a variety of designs and topologies. Deep learning architectures include Deep Neural Networks (DNN)s, Convolutional Neural Networks (CNN)s, Recurrent neural networks (RNN)s and others, each one offering different capabilities and options.

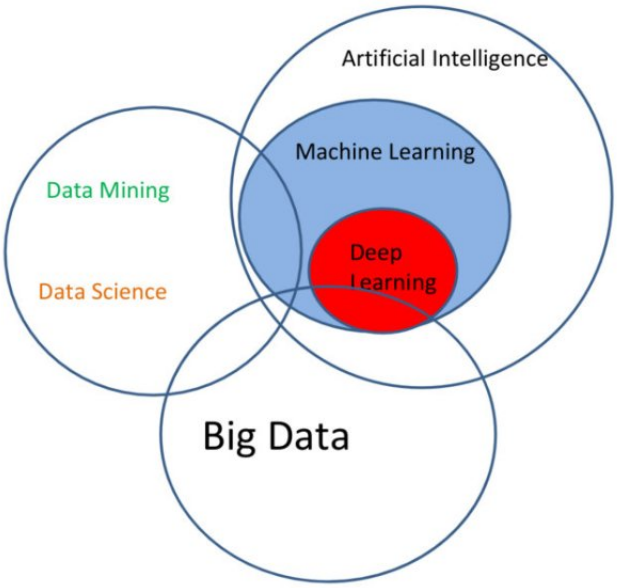

FIGURE 2.1: AI, ML, DL, Data Mining, Data Science, and Big Data: URL.

Deep learning applications have demonstrated human-like or superior capabilities in several scientific and commercial fields such as image[12] and speech[13] recognition, natural language processing[14], climatology[15] and biotechnology[16]. Due to these exceptional capabilities and wide range of applications, deep learning has attracted a large number of researchers from various scientific domains, resulting in its tremendous expansion. However, the science is still young and there are a number of challenges to be overcome. Expecting deep learning combined with improved data processing being a solution to computers gaining generic human-like intelligence (human equivalent AI) is still a distant dream.[17]

Historically, the field of deep learning emerged in 1943 with the inception of the aforementioned McCulloch-Pitts perceptron. In 1949, Donald Hebb noted out in his book "The Organization of Behavior" that neural pathways are strengthened each time they are utilized, a principle that is crucial to how humans learn. He claimed that when two nerves fire at the same moment, the link between them is strengthened. This progress resulted in the creation of the first real-world application of ANNs, "MADALINE" an adaptive filter that eliminates echoes on phone lines. In 1962, Widrow & Hoff developed a learning procedure that distributed the error across the network, resulting in its eventual elimination. Despite these advances, deep learning research plummeted due to a variety of internal and external factor, including the widespread use of fundamentally faulty learning function and the adoption of von Neumann architecture across computer science.

Deep learning research stagnated until 1975, when developments such as Werbos' backpropagation and the building of the first multilayered network reignited interest in the field. Since then, the field continues to expand with innovations like hybrid models and ANN pooling layers. The current focus is on developing deep learning-specific hardware, as fast and efficient ANNs rely on it being defined for their use. Generally, architectures based on accelerators such as GPUs and FPGAs, or VLSI hardware-based designs, outperform CPU-based architectures. [18]

2.2.1 Artificial Neuron & activation function

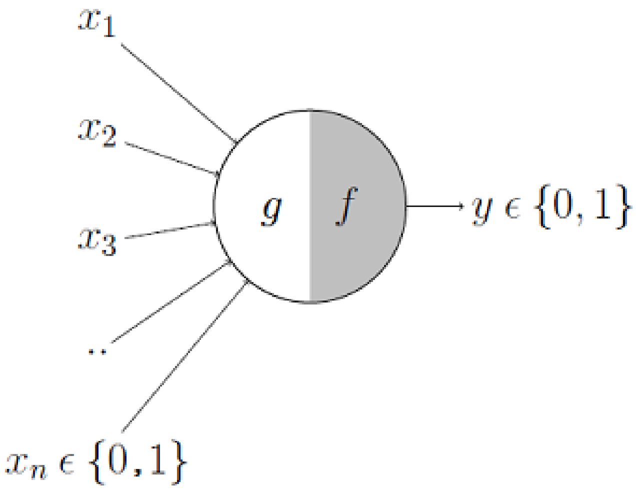

FIGURE 2.2: The McCulloch-Pitts Neuron: URL.

2.2.2 ANN architectures

dnn cnn rnn

2.2.3 Training Neural Networks

2.2.4 Model Overfitting

FL smooths overfiting a bit, better talk about it. Maybe subsection of previous section?

2.3 Federated Learning

Lifecycle of a Model in Federated Learning / Typical Federated Training Process

2.3.1 Data distribution

iid - non iid

2.3.2 Categorization

horizontal - vertical - transfer learning?

2.3.3 Architectures for a federated learning system

Chapter 3

Related Work

3.1 Training Datasets

Common datasets and FL datasets from TF.

3.2 FedAvg

3.3 CE-FedAvg

3.4 Evolution of FL

Fig 5 from A review of applications in federated learning

3.5 quantization?

3.6 The FPGA Perspective

3.7 Thesis Approach

Chapter 4

Robustness Analysis

Maybe this need its own chapter
Developed FL architecture. Server - Client architecture, communication protocol, tf embeddment. Interface - code agnostic of NN design and training.

The experiments from the reviews. Add a prologue to show why they exist. Remake the supplementary ones with the latest settings.

4.1 Experiment A

4.2 Experiment B

Chapter 5

FPGA Implementation

5.1 Tools Used

5.1.1 Vivado IDE

5.1.2 Vivado High Level Synthesis (HLS)

5.1.3 Vivado SDK

5.2 FPGA Platforms

Chapter 6

Results

6.1 Specification of Compared Platforms

6.2 Power Consumption

6.3 Energy Consumption

6.4 Throughput and Latency Speedup

6.5 Final Performance

Chapter 7

Conclusions and Future Work

7.1 Conclusions

7.2 Future Work

Bibliography

[1] Yun Chao Hu et al. *Mobile Edge Computing A key technology towards 5G*. Tech. rep. 11. 06921 Sophia Antipolis CEDEX, France: European Telecommunications Standards Institute, Sept. 2015. URL: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp11_mec_a_key_technology_towards_5g.pdf.

[2] European Parliament and Council of the European Union. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. May 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj.

[3] Chau A., Hertzberg S., and Dodd S. *The California Consumer Privacy Act of 2018*. June 2018. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[4] H. Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: (2016). DOI: 10.48550/ARXIV.1602.05629. URL: https://arxiv.org/abs/1602.05629.

[5] Jed Mills, Jia Hu, and Geyong Min. "Communication-Efficient Federated Learning for Wireless Edge Intelligence in IoT". In: *IEEE Internet of Things Journal* 7.7 (Aug. 2020), pp. 5986–5994. DOI: 10.1109/jiot.2019.2956615. URL: https://doi.org/10.1109/jiot.2019.2956615.

[6] Kang Wei et al. "Federated Learning With Differential Privacy: Algorithms and Performance Analysis". In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 3454–3469. DOI: 10.1109/tifs.2020.2988575. URL: https://doi.org/10.1109/tifs.2020.2988575.

[7] Stuart J. Russell and Peter Norvig. "Introduction". In: *Artificial Intelligence: A modern approach*. 2nd ed. Pearson Education, Inc., 2003, pp. 31–32. ISBN: 0-13-790395-2.

[8] Stuart J. Russell and Peter Norvig. "Learning from Observations". In: *Artificial Intelligence: A modern approach*. 2nd ed. Pearson Education, Inc., 2003, pp. 649–651. ISBN: 0-13-790395-2.

[9] Fuzhen Zhuang et al. *A Comprehensive Survey on Transfer Learning*. 2019. DOI: 10.48550/ARXIV.1911.02685. URL: https://arxiv.org/abs/1911.02685.

[10] Dan Geiger and David Heckerman. *Advances in Probabilistic Reasoning*. 2013. DOI: 10.48550/ARXIV.1303.5718. URL: https://arxiv.org/abs/1303.5718.

[11] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The Bulletin of Mathematical Biophysics* 5.4 (Dec. 1943), pp. 115–133. DOI: 10.1007/bf02478259. URL: https://doi.org/10.1007/bf02478259.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[13] Yu Zhang et al. *Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition*. 2020. DOI: 10.48550/ARXIV.2010.10504. URL: https://arxiv.org/abs/2010.10504.

[14] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. *A Survey of the Usages of Deep Learning in Natural Language Processing*. 2018. DOI: 10.48550/ARXIV.1807.10854. URL: https://arxiv.org/abs/1807.10854.

[15] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. "Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data". In: *Journal of Advances in Modeling Earth Systems* 11.8 (Aug. 2019), pp. 2680–2693. DOI: 10.1029/2019ms001705. URL: https://doi.org/10.1029/2019ms001705.

[16] Jason Riordon et al. "Deep Learning with Microfluidics for Biotechnology". In: *Trends in Biotechnology* 37.3 (Mar. 2019), pp. 310–324. DOI: 10.1016/j.tibtech.2018.08.005. URL: https://doi.org/10.1016/j.tibtech.2018.08.005.

[17] Ritika Wason. "Deep learning: Evolution and expansion". In: *Cognitive Systems Research* 52 (Dec. 2018), pp. 701–708. DOI: 10.1016/j.cogsys.2018.08.023. URL: https://doi.org/10.1016/j.cogsys.2018.08.023.

[18] *History of Neural Networks*. URL: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/index.html.