Deep Learning on SpiNNaker

MASTER THESIS

Jonas Fassbender jonas@fassbender.dev

In the course of studies
HIGH PERFORMANCE COMPUTING WITH DATA SCIENCE

For the degree of
MASTER OF SCIENCE

The University of Edinburgh

First supervisor: Caoimhín Laoide-Kemp
EPCC, University of Edinburgh

Second supervisor: Dr Kevin Stratford
EPCC, University of Edinburgh

Third supervisor: Dr Alan Stokes
APT, University of Manchester

Edinburgh, August 2020

Declaration

I declare that this dissertation was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Jonas Fassbender
August 2020

Abstract

Contents

1. Introduction

Deep learning is revolutionizing the world. It has become part of our daily lives as consumers, powering major software products—from recommendation systems over translation tools to web search (LeCun et al., 2015). Major breakthroughs in fields like computer vision or natural language processing were achieved through the use of deep learning (Krizhevsky et al., 2012; Hinton et al., 2012). It has emerged as a driving force behind discoveries in numerous domains like particle physics, drug discovery, genomics or gaming (Ciodaro et al., 2012; Ma et al., 2015; Leung et al., 2014; Silver et al., 2016).

Deep learning has become so ubiquitous that we are changing the way we build modern hardware to account for its computational demands. From the way edge devices like mobile phones or embedded systems are build over modern CPUs to specialized hardware designed only for deep learning models (Deng, 2019; Boitano, 2020; Perez, 2017; Jouppi et al., 2017). Whole state-of-the-art supercomputers are build solely for deep learning (Langston, 2020). Hardware manifacturers are faced with a major challenge in meeting the computational demands arising from inference and more importantly training deep learning models. OpenAI researchers have estimated that the computational costs of training increases exponentially; approximately every 3.4 months the cost doubles (Amodei et al., 2019). With the end of Moore's Law chip makers have to get creative in scaling up computing the same way machine learning researchers are scaling up their models (Simonite, 2016). Production and research into new hardware designs for deep learning are well on the way.

Another field which has high computational demands for very specific tasks and algorithms is neuroscience. Neuroscience has long been linked to deep learning, which has its origin in reasearch done by neuroscientists (McCulloch and Pitts, 1943). While in the recent past deep learning research has been more focused on mathematical topics like statistcs and probability theory, optimization or linear algebra, researchers are again looking over to neuroscience in order to further improve the capabilities of deep learning models (Marblestone et al., 2016).

But not only the algorithms used by neuroscientists are drawing interest from the deep learning community. Computational neuroscience has long been trying to develop hardware for the efficient modeling of the human brain. Neuromorphic computer architectures are being developed to meet the demands for efficient computing for running large-scale spiking neural networks used for modeling brain functions (Furber, 2016). While being developed mainly for the task of modeling the human brain, deep learning has been linked to neuromorphic computing for some time, especially in the context of commercial usability. Both the low energy demands of neuromorphic computer chips and their general design concerning scalability and massive-parallelism are intriguing for two very important use cases of deep learning: (i) edge computing, for example robotics and mobile devices and (ii) supercomputers and the cloud-era (Gomes, 2017).

One of this neuromorphic computer architectures is the Spiking Neural Network Architecture (SpiNNaker). This thesis investigates the performance of SpiNNaker machines for deep learning by training the state-of-the-art computer vision model ResNet-50 under the closed division rules of the MLPerf benchmark (He et al., 2015; Mattson et al., 2019). In order to benchmark ResNet-50 on SpiNNaker a prototypical implementation was developed as part of this thesis.

- here a paragraph about the results

Section 2 presents the background of this thesis. An introduction to deep learning is given, as well as an overview over the benchmark presented in Section 4. The SpiNNaker architecture is also described and compared to current deep learning hardware. Related work can be found in Section 2.5. Section 3 presents the architecture of the prototype developed for benchmarking. In Section 5 the results of the benchmark are discussed as well as the development process. Section 6 contains the conclusion while Section 7 outlines the next steps for further increasing the performance of SpiNNaker by enhancing the prototype.

2. Background

2.1 An Introduction to Deep Learning

1. history of DL
2. clearify that DNNs are statistical methods (glorified non-linear classifiers) not biological like SNNs
3. concepts of the MLP:
   - layers
   - activations
   - forward- and backward-pass
   - SGD
   - ...
4. CNNs

2.2 Computer Vision: ImageNet and the ILSVRC

1. short section about imagenet and ilsvrc and their importance for computer vision
2. ResNet50 and residual stuff

2.3 Benchmarking Deep Learning Systems: The MLPerf Benchmark

1. short section about MLPerf (so short that I maybe add it to previous section. Could maybe be only a single paragraph.

2.4 SpiNNaker as a Neuromorphic Computer Architecture

1. describe spinnaker and the spinnaker architecture
2. compare to other DL accelerators (GPGPUs and TPUs)

2.5 Related Work

1. SNNToolbox for translating DNNs to SNNs (only inference)
2. TrueNorth has a paper about its DL implementation

3. Deep Learning on SpiNNaker

4. Benchmark

5. Discussion

6. Conclusion

7. Next Steps

References

Dario Amodei, Danny Hernandez, Girish Sastry, Jack Clark, Greg Brockman, and Ilya Sutskever. AI and Compute. `https://openai.com/blog/ai-and-compute/`, 2019.

Justin Boitano. How NVIDIA EGX Is Forming Central Nervous System of Global Industries, 05 2020. URL `https://blogs.nvidia.com/blog/2020/05/15/egx-security-resiliency/`.

T Ciodaro, D Deva, Joao Seixas, and Denis Oliveira Damazio. Online particle detection with neural networks based on topological calorimetry information. *Journal of Physics: Conference Series*, 368, 06 2012. doi: 10.1088/1742-6596/368/1/012030.

Yunbin Deng. Deep learning on mobile devices - A review. *CoRR*, abs/1904.09274, 2019. URL `http://arxiv.org/abs/1904.09274`.

Steve Furber. Large-scale neuromorphic computing systems. *Journal of Neural Engineering*, 13, 08 2016. doi: 10.1088/1741-2560/13/5/051001.

Lee Gomes. Neuromorphic Chips Are Destined for Deep Learning—or Obscurity, 05 2017. URL `https://spectrum.ieee.org/semiconductors/design/neuromorphic-chips-are-destined-for-deep-learningor-obscurity`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Norman P. Jouppi, Cliff Young, Nishant Patil, David A. Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, Richard C. Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. In-datacenter performance analysis of a tensor processing unit. *CoRR*, abs/1704.04760, 2017. URL `http://arxiv.org/abs/1704.04760`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Jennifer Langston. Microsoft announces new supercomputer, lays out vision for future AI work, 05 2020. URL `https://blogs.microsoft.com/ai/openai-azure-supercomputer/`.

Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

Michael Leung, Hui Xiong, Leo Lee, and Brendan Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, 30:i121–i129, 06 2014. doi: 10.1093/bioinformatics/btu277.

Junshui Ma, Robert Sheridan, Andy Liaw, George Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55, 01 2015. doi: 10.1021/ci500747n.

Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Towards an integration of deep learning and neuroscience. *bioRxiv*, 2016. doi: 10.1101/058545. URL `https://www.biorxiv.org/content/early/2016/06/13/058545`.

Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim Hazelwood, Andrew Hock, Xinyuan Huang, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. Mlperf training benchmark, 2019.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, pages 115–133, 12 1943. doi: 10.1007/BF02478259. URL `http://link.springer.com/10.1007/BF02478259`.

Andres Felipe Rodriguez Perez. Intel Processors for Deep Learning Training, 11 2017. URL `https://software.intel.com/content/www/us/en/develop/articles/intel-processors-for-deep-learning-training.html`.

David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

Tom Simonite. Moore's Law Is Dead. Now What?, 05 2016. URL `https://www.technologyreview.com/2016/05/13/245938/moores-law-is-dead-now-what/`.