Approximation accuracy of different FPNN types

Martin Krcma, Zdenek Kotasek
Brno University of Technology, Faculty of Information Technology, Centre of Excellence IT4Innovations
Bozetechova 2, 612 66 Brno, Czech Republic

{ikrcma, kotasek}@fit.vutbr.cz

Abstract

The artificial neural networks [1] are one of the important models of softcomputing and artificial intelligence. They are structures composed of *neurons* interconnected by weighted *synapses*. Basically, the goal of the networks is to learn the relation between two sets of data vectors, to generalize the relation, to determine its features and to use it for the determining the relation of the unknown vectors belonging to the same problem. This capability can be used for classification tasks, for time series and functional prediction, to control tasks, to image recognition, clustering and other tasks.

The implementation of neural networks is challenged with two great neural networks complexities - space complexity and time complexity. The usual solution of both is to use a powerful hardware, such as graphical processor units or processor clusters, which suffer from a high power consumption. For some networks, FPGAs can be one of the possible solutions if a lower power consumption is desired. In this case, the time complexity is solvable by parallelism which is easy to achieve in both FPGAs and neural networks since both are parallel by their nature. The space complexity is bigger problem since an FPGA has limited resources. Thus, there is a need for such designs that exploit the neural networks parallel character for fast computations and save the FPGA resources as well. A Field Programmable Neural Networks (FPNN) concept can be seen as one of the possible solutions. The goal of this paper is to describe the types of FPNNs and compare their capabilities.

The concept of FPNNs [5] is meant to simplify the implementation of artificial neural networks in FPGAs by adjusting their properties to be more suitable for implementation into them. The simplification originates from its main feature - a highly customizable structure which makes it possible to establish resource sharing between the original synaptic connections of the neural network. The FPNNs are composed of dedicated interconnected units called neural resources which approximate the original neurons and synaptic interconnections. The units of the first type are called *activators* and represent the original neural network neurons. The other units are called *links* and serve as an approximation of the original synaptic interconnection. Every link disposes of a set of affine operators serving as an approximation of the original synaptic weights.

The FPNNs are not the same structures as neural networks, although they can be constructed in that way. The FPNNs represent a different model which can structurally differ from the implemented neural network. They can also have different capabilities which means that they are not only an implementation of the neural networks, they are an approximation of neural networks as well - with different structure and properties, they can provide similar results as the networks. The accuracy is the main problem here.

The approximation capabilities depend on the number of affine operators belonging to links. This number depends on the FPNN structure directly. However, the model can be altered to dispose of different number of affine operators. Two different models with different approximation capabilities exist. The original model disposes of as many affine operators as the number of directly connected preceding units. These operators are shared between groups of synapses approximated by the particular preceding units. This type of an FPNN is called *Standard FPNN*. We derived a stronger model that has the number of affine operators that allows it to reach the precise approximation accuracy. This type of an FPNN

is called *Full FPNN*. In case of a full FPNN, every link disposes of dedicated affine operator for every synapse it approximates. There is no sharing of affine operators between synapses, therefore the accurate approximation is ensured. Although, this type of FPNN demands more FPGA resources.

We experimented with the presented models and algorithms, the experiments and results will be described and summarized. The experiments were focused on the approximation capabilities of the standard and full FPNNs model. The goal of the experiments was to show and compare the capabilities of both models and their space complexity.

Paper origin

The original paper has been accepted and presented at the Work in Progress session of the 19th Euromicro Conference on Digital Systems Design in Cyprus [2].

Acknowledgement

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II); project IT4Innovations excellence in science - LQ1602, ARTEMIS JU under grant agreement no 621439 (ALMARVI) and BUT project FIT-S-14-2297.

References

[1] Munakata, T.: Neural Networks: Fundamentals and the Backpropagation Model. In *Fundamentals of the New Artificial Intelligence*, editace T. Munakata, Texts in Computer Science, Springer London, 2007, ISBN 978-1-84628-839-5, s. 7–36, 10.1007/978-1-84628-839-5–2.

[2] Krcma, M.; Kotasek, Z.; Lojda, J.; Kastil, J.: Comparison of FPNNs Approximation Capabilities. In *Proceedings of the Work in progress Session held in connection with DSD 2016*, Aug 2016, ISBN 978-3-902457-46-2, pp. 1–2.

[3] KRCMA Martin, KASTIL Jan a KOTASEK Zdenek: *Mapping trained neural networks to FPNNs.* In: IEEE 18th International Symposium on Design and Diagnostics of Electronic Circuits and Systems. Belgrade: IEEE Computer Society, 2015, pp. 157–160. ISBN 978-1-4799-6779-7.

[4] Krcma, M.; Kotasek, Z.; Kastil, J.: Fault tolerant Field Programmable Neural Networks. In *Nordic Circuits and Systems Conference (NORCAS): NORCHIP International Symposium on System-on-Chip (SoC), 2015*, Oct 2015, ISBN 978-1-4673-6576-5, pp. 1–4.

[5] Girau, B.: FPNA: Concepts and Properties. In *FPGA Implementations of Neural Networks*, editted by A. R. Omondi; J. C. Rajapakse, Springer US, 2006, ISBN 978-0-387-28487-3, p. 71–123, 10.1007/0-387-28487-7-3.