ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Fast Deep Neural Networks Convergence using a Weightless Neural Model

Alan T. L. Bacellar, Brunno F. Goldstein, Victor C. Ferreira,
Leandro Santiago, Priscila M.V. Lima and Felipe M. G. França *

Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro - Brazil

Abstract. Deep Neural Networks (DNNs) have surged as a promising technique for AI applications combining a huge parametric space with efficient learning algorithms. The efficiency of the training procedure relies on some optimization algorithms which adjust the initial weights to minimize the loss of the model. Such strategies are essential to speed up the convergence of the optimization steps. Nonetheless, a general initialization procedure is still an open problem since the proposed techniques either require a long processing time or take a considerable number of iterations to figure out an acceptable model. This work presents a weight initialization strategy using transfer learning via Weightless Neural Network (WNN). This WNN initialization strategy reaches up to 5.5× accuracy and 15× loss reduction at the first iterations when compared against well-known techniques such as Xavier and He.

1 Introduction

Over the past years, Deep Neural Networks (DNNs) had a massive impact in several fields like image classification, segmentation, speech recognition and natural language processing. It has seized the attention of the industry and research community with a series of learning breakthroughs, opening the horizon for an unprecedented problem solving through neural networks. However, training such DNNs workloads in a reasonable time is a challenging process that relies on a close-to-optimal choice of a set of hyperparameters.

Weight initialization is critical for fast and stable convergence of a neural network model. Without proper initialization [1], weights could have little or even no updates, hindering the learning process, or even stalling it. The model architecture should also be taken into account by the chosen technique. Hidden layers with more neurons should be initialized with smaller weight and vice-versa [2, 3].

WiSARD [4] is the most adopted Weightless Neural Network (WNN) model based on Random Access Memory (RAM). In contrast to DNN, training procedure of WiSARD relies on pseudo-random mapping from the binary input to a set of locations into RAM nodes, that is performed as one-shot training. It has been explored in various applications resulting in simple implementations and real-time performance [5, 6, 7, 8, 9].

*The authors thank CAPES, CNPq and FAPERJ for the financial support for this work.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

The main contribution of the paper is a proposal for a new weight initialization technique using transfer learning from a WNN to a DNN. Our approach leverages the output feature vectors from a pre-trained DNN as a WiSARD training input data. Then, we translate the RAM units content to the DNN weights, using it as initialization values for the DNN classifier. We compared the proposed WiSARD based initialization with widely known methods like Xavier[2], and He[3], showing that it can deliver up to 5.5x better accuracy and 15x in loss decrease during the first set of iterations. This approach speeds up DNN convergence.

2 Background

2.1 Neural Network Initialization

Training neural networks rely upon a set of hyperparameters to achieve state-of-the-art accuracy. Learning rate, activation functions, number of epochs, and batch size are some of these parameters that determine if the network converges or not, and how fast it happens. However, even selecting the best options, the number of layers in current deep neural networks makes it infeasible to train without proper initialization.

Training DNN is a non-convex optimization problem, so selecting a good starting point is crucial. The idea of initializing the weights using small random numbers from Gaussian distribution with zero mean and $1e-2$ standard deviation became popular at the beginning, where networks had a small size with a few numbers of layers (less than five). Yet, training large ones was still challenging due to the *vanishing gradient* problem that appears in the last layers. In 2010, Glorot et al. [2] proposed the Xavier initialization by adding a scaling factor to standard deviation based on the number of input neurons on each layer, mitigating the *vanishing* effects. Furthermore, He et al. [3] extended this approach to deal with current nonlinear activation functions, e.g., $ReLU$, which was not possible with the Xavier method.

2.2 WiSARD

Weightless Neural Network (WNN) [10] represents each neuron as a Random Access Memory (RAM) node. It has been applied as attractive solution for pattern-recognition and artificial consciousness applications. WiSARD (Wilkie, Stoneham and Aleksander's Recognition Device) is the pioneering WNN model [4] proposed as multi-discriminator classifier that is able to determine similar patterns from binary input. Several works have shown the potential performance of WiSARD for different applications such as facial emotion classification [11], robot global localization [6], lip animation processing [7] and online tracking of multiple objects [8]. Each discriminator is associated to a class that is comprised of a set of RAM units. An initial transformation, called mapping, converts the incoming data to its binary representation in order to allow the correct training and classification. The binary data has $N \times M$ bits where N is the number of

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

tuples and M is the tuple length so that each tuple n, $n = 1, \ldots, N$, addresses one entry of the n-th RAM. Each RAM has 2^M locations.

During training and classification phases, a *pseudo-random mapping* function maps the input binary matrix to N tuples with M bits each. Different functions may be associated to the discriminators, however the same pseudo-random mapping of the same discriminator must be used in both phases. In the training phase, all RAMs' locations are initialized with zero (0). Since WiSARD is a supervised learning model, each sample is learned by the corresponding discriminator where all selected entries (determined by the tuples) are updated to one (1). In the classification phase, the input data is submitted to all discriminators that generate responses by summing all selected RAM entries. By analysing all the discriminators' responses, the highest one indicates the discriminator with appropriate class of the input.

3 Methodology

The main goal of the proposed initialization method is to transfer knowledge from a weightless to a weighted neuron model. But first, we need to understand in a simple way how both models work. In a weighted one, layer inputs are associated with learnable weights that measure how important that input is to each class. Bigger weights mean more relevant and smaller ones, less critical. In a weightless model (e.g., WiSARD), the learning process does not take into account single input points, but the whole pattern that is mapped to its RAM structures. The relevance of single points is intrinsically stored within the pattern, making it possible to extract within some loss.

In this work, we deploy a WiSARD model as an initialization method. First, we train the model using a small porting of the training set. During this process, the model access some RAM addresses, incrementing the memory content with the hit frequency[12], meaning that addresses are relevant to recognize that class. Finally, we transfer the WiSARD knowledge to a fully connected perceptron by calculating a mean concerning all activated bit inside a RAM into a neuron weight that corresponds to that input point. Then, we scale down all weights values so that they stick below a specific threshold. A set of metrics were tested, such as least squares, geometric mean, weights normalization with variance, but simple arithmetic mean yielded the best results.

For non-binary inputs, a thermometer technique was applied to the training data, bucketing each value into N bits on base 1. In this case, during the knowledge transfer process, weights are calculated as an average of its corresponding binary inputs in the thermometer representation.

Figure 1 exemplify how a generic WiSARD RAM unit with $M = 3$ is converted into neuron weights. The `X` values on the left side represent the available address, while their contents are represented by letters (a dash represent no changes in the address content during training). Each weight W_i is correlated to an address bit X_i, where i = 0, ... $M - 1$. For example, W_2 will receive the mean of the stored values of addresses that have $X_2 = 1$ as shown in green.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

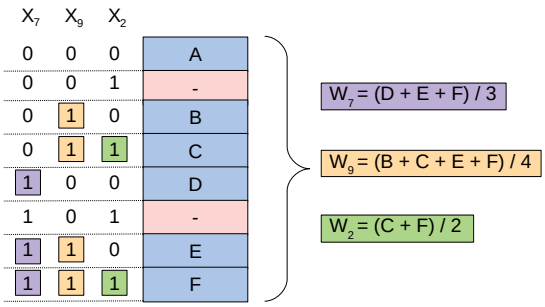

Fig. 1: A trained WiSARD RAM with $M = 3$ and how the weights are extracted. X values represent address bits, the letters inside the RAM represent trained content stored, and W are the weights before normalization.

4 Experiments & Results

To evaluate our proposed initialization method, we selected the 42 layers deep Convolutional Neural Network (CNN), InceptionV3[13] which has been previously pre-trained on the ImageNet [14] dataset. We remove the last layer, responsible for classifying the ImageNet classes, and replace it with a new one with the intuit of training the network on the CIFAR-10 dataset. Before training the new classifier, we extract some feature vectors by inputting some CIFAR-10 train samples on the pre-trained network. These feature vectors are then used as the WiSARD input training set from which the initialization weights of CIFAR-10 classifier are going to be translated. To map these vectors into WiSARD inputs we used thermometers.

Figure 2 depicts the InceptionV3 learning curves using three types of initializers. The proposed WiSARD based initializer trained with only 100 samples, outperforms the other methods by reaching almost 60% of Accuracy and reducing the Loss in 15x with fewer interations as seen on the graphs. After around 50 iterations, all methods reach the same stable convergence curve until the end.

We also evaluate how the WiSARD transfer learning approach works using different types of learning optimizers. Each row of the Figure 2 compares the learning curves of the InceptionV3 network initialized with the WiSARD method using Adam, Stochastic Gradient Descent (SGD), and Adadelta optimizers. For SGD we can see that the WiSARD becomes more stable faster, while the Adam provides the best overall accuracy and loss values. The Adadelta seems to be the middle term, where it appears to be more stable then the Adam but takes longer to achieve higher accuracy.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

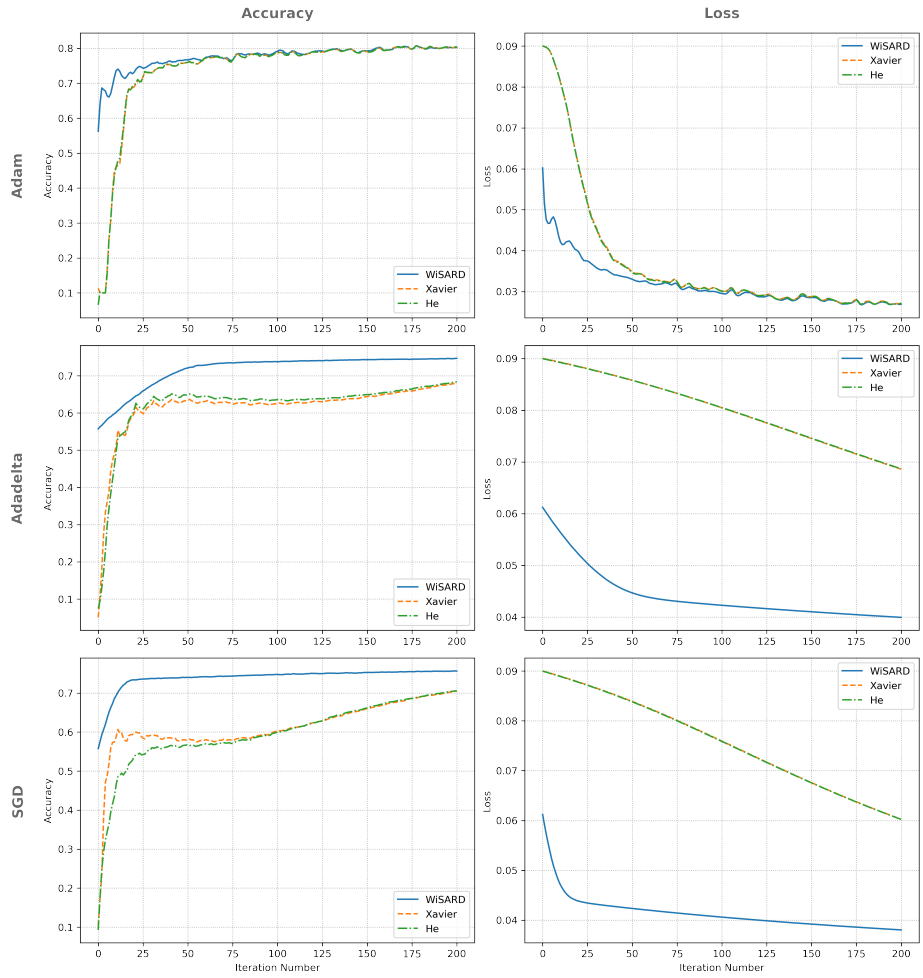

Fig. 2: Accuracy and Loss results for WiSARD (blue), Xavier (orange), and He (green) initializers over three different types of optimizers, Adam (first row), Adadelta (second row) and SGD (third row).

5 Conclusion and Future Work

Weight initialization is an important step to boost the convergence of accurate deep learning models. However, current initialization techniques suffer from performance constraints and are model dependent. In this work, we have proposed a new neural network initialization method based on transfer learning with WiSARD, a RAM-based weightless neural network. We show that our method outperforms state-of-the-art techniques when initializing a CIFAR-10 classification layer using a pre-trained InceptionV3 network, quickly reaching 60% of

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

accuracy with a 15× lower loss.

As future work, we are currently updating the presented method with a hybrid neural network that incorporates the WiSARD itself as a DNN classifier, enabling a more natural training and weight initialization.

References

[1] Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015.

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, 2010.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[4] I. Aleksander, W.V. Thomas, and P.A. Bowden. Wisard·a radical step forward in image recognition. *Sensor Review*, 4(3):120–124, 1984.

[5] Massimo De Gregorio. An intelligent active video surveillance system based on the integration of virtual neural sensors and bdi agents. *IEICE TRANSACTIONS on Information and Systems*, 91(7):1914–1921, 2008.

[6] Paolo Coraggio and Massimo De Gregorio. Wisard and nsp for robot global localization. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 449–458. Springer, 2007.

[7] Charles B Do Prado, Felipe MG Franca, Eduardo Costa, and Luiz Vasconcelos. A new intelligent systems approach to 3d animation in television. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 117–119. ACM, 2007.

[8] Rafael Lima De Carvalho, Danilo SC Carvalho, Felix Antonio Claudio Mora-Camino, Priscila VM Lima, and Felipe MG França. Online tracking of multiple objects using wisard. In *ESANN 2014, 22st European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning*, pages pp–541, 2014.

[9] Victor C Ferreira, Alexandre S Nery, Leandro AJ Marzulo, Leandro Santiago, Diego Souza, Brunno F Goldstein, Felipe MG França, and Vladimir Alves. A feasible fpga weightless neural accelerator. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019.

[10] I. Aleksander, M. De Gregorio, F. Maia Galvão França, P. Machado Vieira Lima, and H. Morton. A brief introduction to weightless neural systems. In *ESANN 2009, 17th European Symposium on Artificial Neural Networks*, 2009.

[11] Leopoldo Lusquino Filho, Felipe M. G. França, and Priscila M. V. Lima. Near-optimal facial emotion classification using a wisard-based weightless system. In *26th European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium, April 25-27, 2018*, 2018.

[12] Bruno P. A. Grieco, Priscila M. V. Lima, Massimo De Gregorio, and Felipe M. G. França. Producing pattern examples from "mental" images. *Neurocomput.*, 73(7-9):1057–1064, March 2010.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.