ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Distributive Thermometer: A New Unary Encoding for Weightless Neural Networks

Alan T. L. Bacellar[1], Zachary Susskind[2], Luis A. Q. Villon[1],
Igor D. S. Miranda[3], Leandro S. de Araújo[4], Diego L. C. Dutra[1],
Mauricio Breternitz Jr.[5], Lizy K. John[2],
Priscila M. V. Lima[1] and Felipe M. G. França[1,6] *

1- UFRJ, Rio de Janeiro, Brazil, 2- UT Austin, Austin, USA,
3- UFRB, Cruz das Almas, Brazil, 4- UFF, Niterói, Brazil,
5- ISTAR/ISCTE-IUL, Lisbon, Portugal, 6- Instituto de Telecomunicações, Portugal

Abstract. The binary encoding of real valued inputs is a crucial part of Weightless Neural Networks. The Linear Thermometer and its variations are the most prominent methods to determine binary encoding for input data but, as they make assumptions about the input distribution, the resulting encoding is sub-optimal and possibly wasteful when the assumption is incorrect. We propose a new thermometer approach that doesn't require such assumptions. Our results show that it achieves similar or better accuracy when compared to a thermometer that correctly assumes the distribution, and accuracy gains up to 26.3% when other thermometer representations assume an unsound distribution.

1 Introduction

Weightless neural networks (WNNs) are a type of neural model that utilizes a random access memory (RAM) to determine neuron activation, as opposed to weights and dot products commonly used in modern deep learning approaches. Because it only uses lookup tables, instead of multiply and accumulate operations which are comparably expensive, they can offer much lower latencies and energy costs [1], making them an attractive solution, especially for usage on edge, and it has been explored in various applications resulting in simple implementations and real-time performance [2, 3, 4, 5].

As using RAMs implicitly requires inputs to be binary, the encoding of real valued inputs is a crucial part of a WNN model and a naive approach can be detrimental to learning [6]. The literature presents many binary encoding techniques [6], with the linear thermometer [7] being the most prominent one. The linear thermometer works by encoding the real value inputs in unary code, under a uniform distribution assumption. A variation of the linear thermometer technique was proposed in [8], where different distributions were used as priors, such as a normal distribution, allowing for an increased resolution of information near the mean of the distribution, showing a significant increase in accuracy for the problem at hand. The major problem with this approach is the need to know

*[*] **Acknowledgement**: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This article was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020; Project FLOYD: POCI-01-0247-FEDER-045912.

the input distribution in advance, or it may not fit the data well, giving increased resolution to regions where it is not needed, and lower resolution where a higher resolution would be beneficial. With this in mind, we proposed the distributive thermometer, a thermometer variation wherein it is not necessary to know the input distribution in advance, but instead use the training data to determine regions of high resolution. The results show this approach achieves greater and similar accuracies to thermometers using prior knowledge about the distribution and a significant gain in accuracy compared to when those assumed priors are incorrect.

2 Background

2.1 WiSARD

WiSARD (Wilkie, Stoneham and Aleksander's Recognition Device) [9] is the most adopted WNN, proposed as a multi-discriminator classifier that is able to recognize similar patterns in binary input. WiSARD represents each class with a discriminator, each of which is composed of multiple RAMs. Each discriminator's RAM is addressed by a unique subset of the binary input. This subset, called the mapping, is chosen pseudo-randomly. A subset may be the same between discriminators but must be unique within each one. During training, a sample is presented to its corresponding discriminator, and the address designated by each binary subset is applied to the corresponding RAM, setting its value to 1. On inference, the binary pattern is presented to all discriminators, and the addressed content of each discriminator's RAMs is summed. The discriminator that yields the highest response is the output class of the model.

Over-fitting happens within the WiSARD model when too much training data is given, or when the size of the binary subset is too small, causing most of the RAMs contents to be set to 1. To solve this, the bleaching tiebreaker technique [10] was proposed. It works by changing the binary value of the RAMs contents to a counter that increments as the same subset is repeatedly presented, and adding a new variable to the model, denominated bleaching-threshold, to be used during inference. On inference, the output of a RAM is 1 if the addressed counter is greater than the bleaching-threshold and 0 otherwise. For each sample during inference, the value of the bleaching-threshold is set to 0, and if the model outputs a draw between discriminators, the bleaching-threshold is incremented and the classification happens again. This process continues until no draw happens, or until all discriminators output is zero.

2.2 Encodings

Proper binary encoding of real valued inputs is a crucial part of a WNN model, and a naive approach is detrimental to learning. As WiSARD learns from similar patterns in data, making a certain number of bit flips in the encoded input must directly correspond to a similar change in their actual values as well [6].

With this in mind, next, we will examine some competing approaches, namely Threshold, Binary Representation, and Hamming Code, and then the leading ones in the state of the art, namely Thermometer, Gaussian Thermometer, and Thermometer with different distribution assumptions.

Threshold: This binarization is the most simplistic way of binarizing an input. It assigns a single bit to a feature by checking if it's above or below a predetermined value (i.e., threshold). It was the binarization used in the early days of WiSARD with great success in datasets like MNIST, but it fails when applied to real valued data, as a significant loss of information is implied.

Binary representation: It is simply using the input binary representation (i.e., integer or floating-point). Although this seems like a good representation to use, it is not a binarization scheme that is appropriate for the WiSARD model, as a single bit flip can yield a large change in value, instead of a close one, because not all bits in the representation convey similar weights [6].

Hamming code: It encodes the input using Hamming code. It has the property that adjacent values are 1-bit flip apart, being a better candidate for the WiSARD model than the Binary Representation, but it still does not satisfy the condition of local smoothness, as Hamming Code does not satisfy that N distance values are N bit flips apart, rendering it unsuitable for the WiSARD model as well.

Linear thermometer: It encodes an input using a unary code, satisfying the N flip condition. Given the number of thermometer bits B, it splits the input space uniformly into B+1 buckets, just like a histogram plot, and assigns the unary code to the input based on which bucket it belongs [7].

Gaussian thermometer: It encodes data in a similar fashion to the thermometer, but instead of assuming a uniform distribution in the input space, it assumes a normal distribution. It calculates the training data mean and standard deviation, and then divides the Gaussian curve into B+1 regions of the same probability, assigning a unary code to each input based on the region it belongs to. This technique provides increased resolution for values near the center of the curve [8].

Other distributions thermometer: Different distributions can also be used alongside the thermometer as an assumption or a prior about the input data. It works by approximating the desired distribution using the training data, and then splitting it into regions of equal probability, just like the gaussian thermometer, and then assigning a unary code to each input based on the region it belongs to. In the experiments section, we utilize an exponential thermometer, as some of the datasets used are known to follow an exponential distribution.

3 The Distributive Thermometer

All thermometer variations yield significant improvements in accuracy compared to the other binarization schemes, but all of them assume a distribution over the input space. The linear thermometer, a uniform distribution, the gaussian thermometer, a normal distribution, etc. This can be highly beneficial where these assumptions are true, as it gives higher resolution to inputs near the mean of the distribution.

Nevertheless, such strategies can be detrimental to accuracy when the input distribution does not match these assumptions, as high density areas in the value scale may not get as high resolution from the thermometer as needed and less dense areas may end up getting excess resolution instead. In order to address this, we propose a simple yet effective solution, the distributive thermometer, a thermometer encoding that splits the input space purely according to the training data, without making an assumption about the distribution. It splits the training data into B+1 percentiles of same probability, then assigns the unary code based on the percentile the input data belongs to. This has the desirable property that high density areas will have higher resolution (bit assignments), and low density areas, a lesser one.

4 Experiment and Results

To verify our thermometer technique indeed approximates the input distribution from the training data, and that the other thermometers are detrimental to training when they assume an incorrect distribution over the input, we compare the accuracy of a WiSARD model using the bleaching tiebreaker while using the Linear Thermometer, the Gaussian Thermometer, an Exponential Thermometer and the Distributive Thermometer on a wide range of datasets [11, 12, 13, 14]. We choose to include an Exponential thermometer as some of the datasets are known to follow an exponential distribution. For each dataset and encoding, we run the WiSARD model on 1000 different mappings, calculate the mean and standard deviation of the results, and perform a t-test to verify the significance in the change in accuracy between thermometers.

Table 1 shows the means and standard deviations of the results. All t-tests between thermometers with accuracies with different means was found to have a p-value < 0.0001 (i.e., they indeed represent a statistically significant change in accuracy). We see that the Distributive Thermometer wins in most datasets, only being in 2nd by a tiny amount (at most 0.004). We also verify that the Gaussian thermometer performs best at different datasets than the Exponential, as both assume different distributions, and that when the assumption is wrong it is highly detrimental to accuracy, as expected. The Distributive wins by a big margin when the assumptions of the Gaussian and Exponential are wrong in relation to one another (up to 0.03), and keeps up with them or even wins when the assumptions are true. In the EEG Eye State dataset, where most attributes neither follow a uniform, gaussian or exponential distribution, the Distributive

	Accuracy			
Dataset	Linear	Gaussian	Exponential	Distributive
Glass	0.71 ± 0.04	0.75 ± 0.04	0.72 ± 0.04	0.75 ± 0.04
Ecoli	0.85 ± 0.02	0.85 ± 0.02	0.83 ± 0.03	0.86 ± 0.02
Vehicle	0.73 ± 0.02	0.74 ± 0.02	0.74 ±0.02	0.75 ± 0.02
Fetal Health	0.909 ± 0.009	0.910 ± 0.009	0.912 ± 0.007	0.914 ± 0.007
SatImage	0.877 ± 0.005	0.887 ± 0.004	0.871 ±0.006	0.887 ± 0.004
EEG Eye State	0.589 ±0.002	0.65 ± 0.01	0.603 ± 0.008	0.852 ± 0.007
Maggic Gamma Telescope	0.823 ± 0.007	0.835 ± 0.006	0.841 ± 0.006	0.837 ± 0.006
Fashion MNIST	0.835 ± 0.002	0.837 ± 0.002	0.852 ± 0.002	0.848 ± 0.002
MNIST	0.958 ± 0.001	0.948 ± 0.002	0.9628 ± 0.0008	0.9617 ± 0.0008

Table 1: Mean accuracy and standard deviation of the Linear Thermometer (Linear), Gaussian Thermometer (Gaussian), Exponential Thermometer (Exponential) and Distributive Thermometer (Distributive) on the different datasets.

Thermometer wins by 0.263, 0.202, and 0.249 respectively.

In Fig. 1 we are able to visualize how each of the thermometers is approximating the distribution and where they are giving higher resolution to the input. The Gaussian and Exponential Thermometers are able to properly approximate the distribution only when their priors are true, while the Distributive is able to do so in all of them, splitting the input space in the same manner as Gaussian in the first column, Exponential in the second column, and the only one to give higher resolution to the two peaks in the third column.

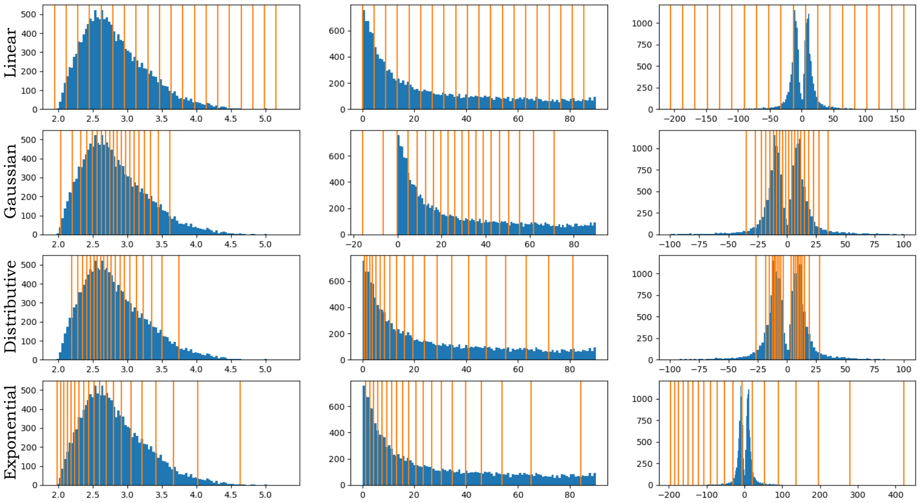

Fig. 1: Division of the input space by the thermometers. On first column, how they divide a normal distribution, and on the second column, an exponential distribution, and on third column, a distribution with two peaks.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

5 Conclusion and Future Work

In this work, we propose a thermometer variation that doesn't make assumptions about the inputs distribution, but instead follows the training data, providing high resolution where it is needed. The results show it achieves similar or better accuracy when compared to a thermometer that has prior knowledge about the distribution, and up to 26.3% more accurate when other thermometers assume an unsound distribution. As immediate further work, one can take labels of the training data into account as well, as a region of high resolution may not be needed if there is only a single label there, and extra bits may be useful in regions where there is a higher density of different labels close to each other.

References

[1] Zachary Susskind, Aman Arora, Igor D. S. Miranda, Luis A. Q. Villon, Rafael F. Katopodis, Leandro S. Araujo, Diego L. C. Dutra, Priscila M. V. Lima, Felipe M. G. França, Mauricio Breternitz Jr, and Lizy K. John. Weightless neural networks for efficient edge inference. *arXiv:2203.01479*, 2022.

[2] 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.

[3] Charles B. Do Prado, Felipe M. G. França, 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.

[4] Rafael L. Carvalho, Danilo S. C. Carvalho, Felix A. C. Mora-Camino, Priscila V. M. Lima, and Felipe M. G. França. Online tracking of multiple objects using WiSARD. In *ESANN 2014*, pages pp–541, 2014.

[5] 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.

[6] Andressa Kappaun, Karine Camargo, Fabio Rangel, Fabricio Faria, Priscila Lima, and Jonice Oliveira. Evaluating binary encoding techniques for WiSARD. In *BRACIS*, pages 103–108, 10 2016.

[7] Hugo Carneiro, Felipe França, and Priscila Lima. Multilingual part-of-speech tagging with weightless neural networks. *Neural Networks*, 66, 03 2015.

[8] Pedro Xavier, Massimo De Gregorio, Felipe M. G. França, and Priscila M. V. Lima. Detection of elementary particles with the WiSARD n-tuple classifier. In *ESANN 2020, Bruges, Belgium, October 2-4, 2020*, pages 643–648, 2020.

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

[10] Danilo Carvalho, Hugo Carneiro, Felipe França, and Priscila Lima. B-bleaching : Agile overtraining avoidance in the WiSARD weightless neural classifier. In *ESANN*, 04 2013.

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[12] Diogo A. Campos, João Bernardes, Antonio Garrido, Joaquim M. Sá, and Luis P. Leite. Sisporto 2.0: A program for automated analysis of cardiotocograms. *The Journal of Maternal-Fetal Medicine*, 9(5):311–318, 2000.

[13] Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[14] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.