ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Performance Evaluation of Activation Functions in Extreme Learning Machine

Karol Struniawski[1], Aleksandra Konopka[1] and Ryszard Kozera[1,2]

1 - Warsaw University of Life Sciences - SGGW
Institute of Information Technology
ul. Nowoursynowska 159, Warsaw, Poland

2 - The University of Western Australia
School of Physics, Mathematics and Computing
35 Stirling Highway, Crawley, Perth, Australia

Abstract. This study investigates the performance of 36 different activation functions applied in Extreme Learning Machine on 10 distinct datasets. Results show that Mish and Sexp activation functions exhibit outstanding generalization abilities and consistently perform well across most datasets, while other functions are more dependent on the characteristics of the task at hand. The selection of an activation function is intricately linked to the applied dataset and novel activation functions may possess superior generalization capabilities comparing to commonly employed alternatives. This study provides valuable insight for researchers and practitioners seeking to optimize Extreme Learning Machine performance for solving classification tasks.

1 Introduction

Extreme Learning Machine (ELM) is a type of neural network that was introduced by Huang et al. in 2004 [1]. The ELM architecture comprises of an input layer, a single hidden layer and an output layer of neurons. The number of neurons in the input and output layer are adapted to the specific task at hand. Due to the scarcity of theoretical methods, it is difficult to determine upfront the optimal number of hidden units for the ELM. Consequently, the latter is usually established through empirical evaluations. ELM has been widely applied in various fields, including image classification [2], medical diagnosis [3] and soil microorganism identification [4]. It is shown to be highly computationally efficient in both classification and regression tasks [5]. The ELM has become an increasingly popular Machine Learning (ML) technique in recent years as its versatility and effectiveness make it a valuable tool for a wide range of applications.

Extreme Learning Machine utilizes the McCulloch-Pitts neurons [6] for which an activation function needs to be determined. Huang et al. proved that in contrast to conventional gradient-based learning algorithms that are exclusively applicable to differentiable activation functions, ELM may also employ non-differentiable or piecewise differentiable activation functions [7].

In recent years, new activation functions are proposed that yield promising results in ML. The Rectified Linear Unit (ReLU) [8] is one of the activation functions that is adopted in Convolution Neural Networks (CNN). Another activation function that has gained popularity in recent years is the Exponential

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Linear Unit (ELU) function that can achieve better performance than the ReLU in certain scenarios e.g. in networks with more than 5 layers [9].

The choice of activation function depends on the specific input task and researchers are continuously exploring new activation functions that can improve the performance of ELM applied to a wide range of problems [10]. In practical applications of ELM, literature overview shows that despite novel activation functions being developed the sigmoid and hyperbolic tangent functions remain the most widely used in ELM [11]. The insufficient treatment of complex comparison of activation functions applied for different datasets has been detected in the field in question. In one of the works that provide comparison between 11 different activation functions only one dataset is examined [12]. The observed activation function's performance on a single dataset raises concerns regarding its generalizability. To address this issue, we present a comprehensive performance investigation of the 36 different activation functions on 10 distinct datasets. The aim of this study is to determine whether certain activation functions outperform others and to assess whether the optimal selection varies depending on the dataset. Our hypothesis is that the activation function selection is intricately linked to the characteristics of the dataset used for which subset of functions can be identified that consistently exhibit superior or inferior performance. Noteworthy, this paper has utilized a diverse set of activation functions for the ELM, many of which have not been previously investigated in the literature like Mish, originally introduced in 2019 [13]. Novel activation functions may possess superior generalization capabilities in comparison to the commonly employed alternatives, rendering remarkable candidates for enhancing the performance of classification tasks using ELMs.

2 Extreme Learning Machine Classifier

In a supervised classification task N observations are represented as pairs of values denoted by $\{(x_i, t_i)\}_{i=1}^N$. The i-th vector x_i is composed of d features, while the corresponding i-th label t_i identifies the class to which the vector belongs. For a classification problem with M distinctive classes, t_i ranges from 0 to $M-1$. The input data is used to construct a matrix $X = (x_1, x_2, \ldots, x_N) \in \mathbb{M}_{d\times N}(\mathbb{R})$ with each $x_i \in \mathbb{R}^d$, along with a vector $T = (t_1, \ldots, t_N)$.

The input layer of an ELM neural network is composed of d neurons, while its output layer has a number of units equal to M. The network calculates N values $\{y_i\}_{i=1}^N$ as its output, which are then used to form the matrix $Y = (y_1, y_2, \ldots, y_N) \in \mathbb{M}_{N\times M}(\mathbb{R})$. To recognize a given input x_i, the maximal value of y_i observed on the p-th index is extracted. This assigns x_i to the p-th class. Suppose that a fixed number of neurons, denoted by L is selected for the hidden layer in advance. The weights connecting the input and hidden layers define the matrix $W \in \mathbb{M}_{d\times L}(\mathbb{R})$, where w_{ij} corresponds to the weight associated with the connection between the i-th input layer neuron and the j-th neuron in the hidden layer. The bias connections are represented by a vector $b = (b_1, \ldots, b_N)$. During the learning process of the ELM, the coefficients of W and b are determined

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

using a uniform distribution function $U(-1,1)$. The outputs of the hidden layer neurons are stored in the matrix H. In ELM, the activation function $f:\mathbb{R}\rightarrow\mathbb{R}$ introduces non-linearity to the hidden layer output, which is crucial for the network's performance on tasks that involve complex relationships between input and output variables [1]. The weights β between the hidden and output layer in ELM can be calculated by solving the equation $Y = H\beta$. This system cannot be directly solved since H is non-invertible and $\|H\beta - Y\| = 0$ (see Huang et al. [1]). Instead, we estimate β as the minimizer of the mean residual square error: $\hat{\beta} = \arg\min_{\beta}\|H\beta - T\|^2 = H^{\dagger}T$, where $H^{\dagger}$ is the Moore-Penrose generalized inverse of H [14]. The pseudo-inverse of matrix $H^{\dagger}$ is uniquely determined and in the case of a non-singular matrix H, it coincides with an ordinary inverse, i.e. $H^{\dagger} = H^{-1}$. The matrix $H^{\dagger}$ gives the solution $\hat{\beta}$ so that $H\hat{\beta}$ is close to Y in terms of mean square error.

3 Methodology and results

The use of ELM's techniques necessitates the selection of an appropriate activation function and requires determination of the number of hidden layer units denoted by L. To enable clear comparison of results, we propose a lucidity experiment involving running ELM on a specific dataset using a fixed activation function and conducting 50 repetitions of 10% cross-validation. A search over the range of L values, from 100 to 5000 in increments of 100, is then performed to identify the optimal classification accuracy. The reported results reflect the highest accuracy obtained using a particular dataset and activation function at various values of L. In light of the extensive use of various activation functions in this study, it is recommended that each applied activation function should be referenced to the relevant scientific literature. Activation functions taken here into consideration are: identity $f(x) = x$, Binary Step Function (***BSF***) $f(x) = \{1 : x \geq 0; 0 : x < 0\}$, ***TanhRe*** $f(x) = tanh(x) + x$, ***HTan1*** $f(x) = \min(\max(x, -1), 1)$, ***Sine*** $f(x) = sin(x)$, ***ASin*** (Inverse Sine) $f(x) = \arcsin(\min(\max(x, -1), 1))$, ***Cosine*** $f(x) = cos(x)$, Soft Exponential (***Sexp***) $f(x) = \max(x, 0) + \ln(1 + exp\{-|x|\})$, Inverse Square Root Linear Unit (***ISRLU***) $f(x) = \frac{x}{1+e^{-1.5x}}$, Inverse Square Root Linear Units (***ISRLUs***) $f_\alpha(x) = \frac{x}{\sqrt{1+\alpha x^2}}$, Asymmetric Rectified Linear Unit (***AReLU***) $f(x) = \{x : x \geq 0; 0.1x : x < 0\}$, Bent's Exponential Linear Units (***BELUs***) $f(x) = \frac{\sqrt{x^2+1}-1}{2} + x$, Exponential Linear Units with Maxout (***Max-ELUs***) $f_\alpha(x) = \max(x, \alpha e^x - 1)$, Tilted Exponential Linear Units (***TELUs***) $f_\alpha(x) = \{x : x \geq 0; \alpha e^x - 1 : x < 0\}$, Soft Clip Exponential Linear Units (***SCELU***) $f_\alpha(x) = \{x : x \geq 0; \alpha e^x : x < 0\}$, Scaled Exponential Sine Linear Units (***SESLUs***) $f_{\alpha,\beta}(x) = \{x : x \geq 0; \alpha sin(\beta x) : x < 0\}$, Square Non-Linearity (***SQNL***) $f(x) = \{-1 : x < -2; \frac{x+x^2}{4} : x < 0; \frac{x-x^2}{4} : x \geq 2\}$, ***Soft Clipping*** $f(x) = \{-1 : x \leq -1; x : x > -1 \text{ and } x > 1; 1 : x \geq 1\}$, ***SineReLU*** $f(x) = \max(0, \sin(x))$, Rectified Square Root (***ReSQRT***) $f(x) = \sqrt{\max(0, x)}$,

# Samples	690	569	1k	6k	208	14k	13k	351	3k	846
# Features	14	30	11	166	60	14	16	34	180	6
Activation	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
identity	**77**	**95**	*77*	**96**	74	64	90	82	**95**	76
BSF	*60*	*57*	*2*	*56*	*55*	*55*	*16*	*74*	*39*	*31*
Sigmoid	76	94	84	**95**	74	88	91	84	**94**	77
Swish	77	94	84	93	**77**	88	91	83	93	**79**
ELiSH	76	93	81	94	74	77	90	85	94	75
TanH	75	94	82	94	*72*	86	90	*80*	**93**	75
HardTanH	70	88	*76*	86	72	*61*	*64*	81	90	*62*
ReLU	72	90	79	87	74	62	89	**88**	88	67
TanhRe	75	94	82	**94**	71	86	90	81	94	75
ELUs	76	93	81	94	74	77	90	85	94	75
***Soft*-Plus**	**77**	**95**	**84**	94	77	**89**	**91**	86	94	**79**
LReLU	71	90	79	88	74	63	90	**88**	88	68
SeLU	76	94	83	93	72	86	90	81	93	74
ReLU6	71	90	79	87	74	62	89	**88**	88	66
HTan1	73	90	79	92	74	63	74	82	93	70
Sinusoidal	75	93	**84**	94	**77**	**90**	**91**	*80*	93	77
Asin	*69*	*88*	78	89	73	*61*	*67*	83	93	63
Cosine	*47*	*48*	*1*	*40*	*37*	*41*	*1*	*34*	*12*	*14*
Sexp	**78**	**95**	**84**	**95**	**77**	**89**	**91**	85	94	**79**
Mish	**78**	**95**	**84**	94	**77**	**89**	**91**	83	**94**	**80**
ISRLU	77	94	83	92	76	87	91	82	92	77
RReLU	73	92	80	87	75	73	90	**86**	*86*	70
GELU	76	94	83	92	76	87	91	83	92	78
SELU	74	91	79	94	*71*	72	90	82	94	72
ISRLU	75	94	82	94	72	85	90	*80*	93	74
AReLU	72	91	79	89	74	62	90	**87**	90	69
BELU	77	94	83	**95**	**77**	82	91	84	**95**	78
Max-ELU	**79**	91	80	94	74	**91**	91	81	90	**80**
TELUs	76	93	81	94	74	74	90	85	94	74
SCELU	*61*	*86*	*24*	*71*	*66*	67	*21*	84	*56*	*37*
SESLU	76	94	82	94	72	78	90	82	94	76
SQNL	76	93	81	94	73	70	87	81	94	74
Soft Clip	73	90	79	92	73	63	74	82	93	70
SineReLU	71	89	79	*86*	72	65	88	84	*86*	64
ReSQRT	*68*	*88*	*77*	84	73	*61*	84	86	89	*58*
SiLU	77	**95**	**84**	93	77	86	91	83	93	79

Table 1: The accuracy (ACC) [%] of the Extreme Learning Machine (ELM) for a particular activation function and dataset with given samples and features number. The bold values indicate the top five ACC for a selected dataset among activation functions, while the italicized values represent the five functions with the lowest ACC.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Sigmoid Linear Unit (***SiLU***) $f(x) = \frac{x}{1+e^{-x}}$,. For the following activation functions details can be obtained in [15]: ***Sigmoid***, ***Swish***, Exponential Linear Squashing (***ELiSH***), ***TanH***, ***HTanH***, Rectified Linear Unit (***ReLU***), Exponential Linear Units (***ELUs***), ***SoftPlus***, Leaky ReLU (***LReLU***), Scaled Exponential Linear Unit (***SeLU***), ***ReLU6***, ***Mish***, Gaussian Error Linear Units (***GELUs***), Scaled Exponential Linear Units (***SELU***), Randomized Leaky ReLU (***RReLU***). In this paper, α and β are set to 1 as the selection of activation function parameters is beyond the scope of this research.

To provide the meaningful comparison between various activation functions 10 different datasets were used. The datasets are made publicly available by the UCI Machine Learning Repository for the purpose of classification tasks and are commonly used in ML [16]. For simplicity in further considerations datasets are marked as $A, \ldots, J$, where A - Australian credit card applications, B - Breast Cancer, C - Wine-Red, D - Musk, E - Sonar, F - EyeState, G - Dry Bean, H - Ionosphore, I - DNA and J - Vehicle. The experiment's results are presented in Tab. 1 and are analyzed in the conclusions section below.

4 Conclusions

The results (see Tab.1) indicate that some activation functions performed poorly, including BSF, Cosine and SCELU, which exhibited the worst results for all datasets (being 10 times in bottom 5 ACC). With high confidentiality we can exclude these functions from practical usage for ELM. On the other hand, the Mish (9 times), Sexp (8), SoftPlus (6), Maxout-ELUs (4) exhibited superior performance across datasets being in top 5. Noteworthy, each of these functions has never been chosen as the worst 5 for a given dataset. Identity and sinusoidal functions were in the top 5 for the 4 times, but simultaneously each of these was also once noted as the bottom 5. Their usage may be beneficial only for some of the assignments. Our experimental results demonstrate the exceptional generalization capabilities of Mish and Sexp activation functions for ELM. These functions have consistently performed well across 10 diverse datasets, encompassing various classification tasks. They have exhibited strong performance even when applied to datasets with varying numbers of features and samples. In contrast, the effectiveness of other activation functions has been more reliant on the specific characteristics of the tasks. Notably, the accuracy of the ELM classifier can differ significantly, up to 80 percentage points, depending on the chosen activation function. Therefore, it is crucial to evaluate the efficacy of activation functions for each task to ensure optimal classifier performance. Our experimental results highlight the promising potential of Mish activation function for ELM. Mish, a relatively new activation function in Deep Learning, has not been extensively utilized in ELM until now. Based on the obtained results we can straightforward conclude that the ELM's performance with selected activation cannot be measured on a single dataset as there is no guarantee that the activation function's generalization abilities will be sufficient for a given task. The analysis did not reveal any significant correlations between the number of features or samples and the performance of activation functions. Instead, the results suggest that the performance is closely tied to the characteristics of the specific classification task. This observation opens up avenues for future research, particularly in exploring the implications of Universal Approximation theorems for ELM. Further research on the Mish and Sexp activation functions applied in ELM should be conducted to examine their performance on more datasets. It would also be beneficial to introduce optimization strategies to the values of α and β for specific activation functions.

References

[1] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme Learning Machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[2] Wei Bao, Yuan Lin, and Ming Cheng. Deep Convolutional Extreme Learning Machine and its application in image classification. *IEEE Trans. Cybern.*, 48(6):1886–1898, 2018.

[3] Guodong Li, Yanyu Zhao, Jie Liu, and Huaiqing Wang. A new Extreme Learning Machine with application to medical diagnosis. *J. Med. Syst.*, 41(3):1–8, 2017.

[4] Aleksandra Konopka, Karol Struniawski, Ryszard Kozera, Paweł Trzciński, Lidia Sas-Paszt, Anna Lisek, Krzysztof Górnik, Edyta Derkowska, Sławomir Głuszek, Beata Sumorok, and Magdalena Frac. Classification of soil bacteria based on machine learning and image processing. In *ICCS 2022*, pages 263–277, 2022.

[5] Gang Chen, Xuerong Mao, and Yingkai Zhang. Extreme Learning Machine: a review. In *NeurIPS*, pages 477–490, 2019.

[6] Mohamad H. Hassoun. *Fundamentals of Artificial Neural Networks*. MIT Press, 1st edition, 1995.

[7] Guang-Bin Huang, Qin-Yu Zhu, Kudo Mao, Chee Siew, P. Saratchandran, and Narasimhan Sundararajan. Can threshold networks be trained directly? *IEEE Trans. Circuits Syst. II Express Briefs*, 53:187 – 191, 04 2006.

[8] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. *Proc. Int. Conf. Electron. Comput. Artif. Intell.*, 15:315–323, 2011.

[9] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by Exponential Linear Units (ELUs). In *ICLR*, 2016.

[10] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme Learning Machines: A survey. *IEEE Trans. Neural Netw. Learn Syst.*, 30(10):2822–2840, 2019.

[11] Guang-Bin Huang, Hong Zhou, Xiao-Tong Ding, and Rui Zhang. Trends in Extreme Learning Machines: a review. *Neural Netw.*, 61:32–48, 2015.

[12] Dian Eka Ratnawati, Marjono, Widodo, and Syaiful Anam. Comparison of activation function on Extreme Learning Machine (ELM) performance for classifying the active compound. *AIP Conf. Proc.*, 2264(1), 2020.

[13] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *CoRR*, abs/1908.08681, 2019.

[14] C. Radhakrishna Rao and Sujit Mitra. *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, 1971.

[15] An overview of activation functions. paperswithcode.com/methods/category/activation-functions.

[16] Dheeru Dua and Casey Graff. UCI machine learning repository. archive.ics.uci.edu/ml, 2017.