Sum of Step Approximation of a Novel Non Linear Activation Function

Venkatappareddy.P, M. Deepthi

Abstract—In this manuscript, we propose sum of steps approximation of a novel nonlinear activation function tunable ReLU for VLSI architecture implementation of neural networks. The characteristics of the proposed activation function depend on a tunable parameter and input data set values. Also, we propose a linear-in-the-parameter model for the proposed activation function using an even mirror Fourier nonlinear filter. Finally, simulation results are presented to show performance of the proposed activation function on various data sets and observe its superiority against to other activation functions.

Keywords— Activation function, perceptron, tunable ReLU, deep neural network, EMFN filter.

I. INTRODUCTION

Single and multi-layer neural networks are computational training architectures that propagate data through linear and nonlinear operators. Nonlinear operators are also called activation functions that have been found to be very effective when the network has more than one layer. In [1], LeCun introduces deep learning networks that consist of more than five layers. As the number of layers in the deep neural networks (DNN) increase, its performance gets better at the cost of computational complexity. In [2], authors discuss interpretation of DNN model and its applications on various datasets for industrial applications. In [3], authors discuss advantages of deep leaning for classification and regression problems in high dimension. DNNs are used in various applications such as monaural speech enhancement in various back ground noise conditions [4], image quality assessment [5], natural language understanding [6], speech recognition [7], and others.

In the literature, various nonlinear activation functions have been studied viz., sigmoid, *tanh* and rectified linear unit (ReLU). It is observed that an activation function [8] which gives better classification accuracy for one dataset may not give better classification accuracy for other in industrial applications. Sigmoid and *tanh* are two smoothly varying nonlinear activation functions however they cannot be effectively used in DNN architecture due to gradient vanishing problem. Currently, ReLU is a popular activation function since it eases the gradient vanishing problem. It was proposed for Boltzmann machines and was used in the neural networks by Glorot et al., [9]. Sigmoid and *tanh* are continuous differential functions with point-wise contractive nature. Since Lipschitz constants for these activation functions are 0.25 and 0.5 respectively [10]. Whereas ReLU is not differentiable at zero and its Lipschitz constant is 1. Most recently other nonlinear activation functions exponential linear unit (ELU) and scaled exponential linear units (SELU) were introduced in [11, 12] that speeds up training in neural network and performs better than ReLU. A closed form approximation for sigmoid and *tanh* functions by sum of steps for hardware implementation are discussed in [13]. An even mirror Fourier nonlinear (EMFN) filter [14] is used to obtain a linear-in-the-parameter (LIP) model for any strong nonlinear characteristic activation function in DNN for better mathematical tractability.

A small percentage increment in the training accuracy requires an additional number of layers in the network which increases computational complexity. In this manuscript, we propose a novel nonlinear activation function tunable ReLU for characterization of neural networks mathematically and training networks efficiently. First, we study the behavior of the proposed activation function against its tunable parameter. We also approximate tunable ReLU by sum of steps representation for VLSI implementation. Further, we discuss a LIP model for the proposed activation function using an EMFN filter. This LIP model is used to completely represent neural networks with linear and LIP operators.

This manuscript is organized as follows. In Section II, we discuss various existing activation functions and an EMFN filter basis to approximate strong nonlinear systems as preliminaries. In Section III, we discuss the definition and a polynomial approximation of tunable ReLU. Also, we discuss an approximation of tunable ReLU using sum of steps and its LIP model using an EMFN filter. In Section IV, we present the performance results of the proposed activation function. Conclusions follow in Section V.

II. PRELIMINARIES

In this section, we present a brief summary of two concepts that are used in this manuscript. We first talk of definition of an activation function and representations of various activation functions.

Definition: A function is called an activation function $\rho(x):R\rightarrow R$ that is differentiable everywhere mostly [15].

Some of the activation functions used in neural networks to train the datasets are:

- Sigmoid ($\sigma(x)$): $\dfrac{1}{1+e^{-x}}$

Revised Manuscript Received on February 11, 2019.

VENKATAPPAREDDY.P, DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING, VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH, Vadlamudi, Guntur, AP, INDIA-522213

M. DEEPTHI, Department of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur Guntur, AP, INDIA- 522508 ({siri.deepu9, venkatappareddyp}@gmail.com)

- Hyperbolic Tangent: $tanh(x)$
- ReLU: $max(0, x)$
- Exponential linear unit (ELU):

$$g(x)= x \ \ if\ x\geq 0$$
$$= \alpha x\, e^{x} \ \ if\ x\leq 0$$

where α is a parameter that controls the value to which an ELU saturates for negative net inputs. Scaled exponential linear units (SELU) is introduced in [12] as an activation function and it is defined as

$$g(x)= \lambda x \ \ if\ x\geq 0$$
$$= \lambda\alpha\,(e^{x} -1) \ \ if\ x\leq 0$$

Where, λ is a parameter for controlling mean of the activation function.

(a) Even Mirror Fourier nonlinear filter

In this manuscript, an EMFN filter is used to obtain a LIP model for the proposed activation function. A LIP model is used to characterize nonlinear activation functions in NN, DNN and LSTM architectures mathematically. The properties of Fourier series such as an even mirror repetition and periodicity are used to obtain an EMFN filter basis [14]. Let $h(x)$ is a periodic signal with period 4 and an even mirror repetition property such that

$$h(x + 4) = h(x) \tag{1}$$
$$h(1 + x) = h(1 - x)$$

Then the Fourier basis of $h(x)$ that satisfy properties of an Algebra is of the form [16, 17],

$$\left\{1,\ \sin\left(\frac{\pi x(n)}{2}\right),\ \cos\left(\frac{2\pi x(n)}{2}\right),\ \sin\left(\frac{3\pi x(n)}{2}\right),\ ..\right\} \tag{2}$$

The above basis functions are useful only when the input sample values are in the range of [-1 1]. The EMFN filter basis is used to approximate any continuous function based on Stone-Weierstrass theorem which states that:

"Given $h : [a, b]\rightarrow R$ continuous and an arbitrary $\varepsilon > 0$ there exits an algebraic polynomial k such that $|h(x)-k(x)| \leq \varepsilon,\ \forall\ x \in [a, b]$" [18].

III. A NOVEL TUNABLE RELU AS ACTIVATION FUNCTION

In this section, we introduce the tunable ReLU as a novel nonlinear activation function to classify datasets for various industrial applications and mathematical understanding of NN, DNN and LSTM architectures. The proposed activation function is differentiable at every point and eases the gradient vanishing problem. The proposed tunable ReLU representation is obtained by considering the definition of a ReLU activation function. It is denoted as $R(x)$ and represented as

$$R(x)= max(0, x)\ \forall\ x \in R \tag{3}$$

ReLU activation function in (3) can be written in terms of signum function as

$$R(x)=\frac{x}{2}\left(1 + sign(x)\right) \tag{4}$$

where, sign(.) represents the signum function defined as

$$sign(x)=\begin{cases}1 & for\ \ x > 0\\ 0 & for\ \ x = 0\\ -1 & for\ x < 0\end{cases}$$

The signum function is approximated using tanh function as

$$sign(x) \approx\ \tanh(kx)\ \text{for}\ \ |kx| \gg 0 \tag{5}$$

where, $k \in R^{+}$. By substituting (5) in (4), we obtain the proposed tunable ReLU $\rho(x)$ as

$$\rho(x)= \frac{x}{2}\,(1 + \tanh(kx))\ \text{for}\ \ |kx| \gg 0 \tag{6}$$
$$= \frac{xe^{kx}\,\mathrm{sech}(kx)}{2} \quad \text{for}\ \ |kx| \gg 0 \tag{7}$$

We use the equality $\mathrm{sech}(x) = \frac{2}{e^{x} + e^{-x}}$ in (7) to obtain another form of the tunable ReLU based on a parameter value k which is differentiable at every point as

$$\rho(x) = \frac{x}{1+e^{-2kx}}\ \text{for}\ \ |kx| \gg 0 \tag{8}$$

The tunable ReLU $\rho(x)$ and its derivative $\rho'(x)$ are also expressed using sigmoid functions as

$$\rho(x) = x\sigma(2kx)\ \text{for}\ \ |kx| \gg 0 \tag{9}$$

Where, $\sigma(2kx)$ is defined as

$$\sigma(2kx) = \begin{cases}1 & if\ kx \gg 0\\ 0 & if\ kx \ll 0\\ \frac{1}{2} & if\ kx\ = 0\end{cases} \tag{10}$$

using this $\rho'(x)$ is found as

$$\rho'(x) = 2kx\sigma'(2kx) + \sigma(2kx)\ for\ \ |kx| \gg 0 \tag{11}$$

where, $\sigma'(x)$ is the derivate of σ(x).

We present another form of activation function as

$$\rho(x) = \begin{cases}x & if\ kx \gg 0\\ xe^{2kx} & if\ kx \ll 0\\ 0 & if\ x = 0\end{cases} \tag{12}$$

The proposed activation function is also represented as

$$\rho(x) = xe^{k(x-|x|)} \qquad \text{for}\ \ |kx| \gg 0 \tag{13}$$

where | . | denotes modulus operator. By using the $e^{x} = 2^{x ln_2 e}$ and the value of $ln_2 e$=1.4428, we obtain another form of the proposed activation function as

$$\rho(x) = x2^{1.4428k(x-|x|)}\ \text{for}\ \ |kx| \gg 0 \tag{14}$$

The proposed activation function is represented using a modulus operator and powers of '2' as shown in Fig. 1. Where, the multiplication operator is denoted as 'X'. The linearized functions l_s and l_t of sigmoid and tanh functions are obtained using Taylor series expansion about zero. The linear approximation of sigmoid function is given as

$$l_s \approx 0.25x+0.5$$

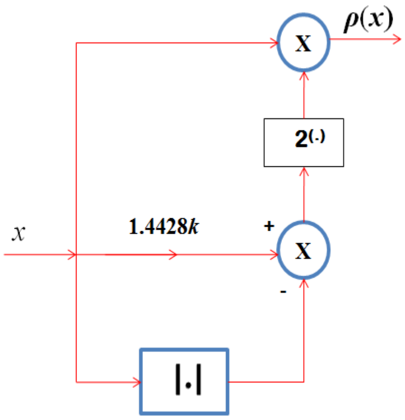

Fig. 1: Representation of tunable ReLU.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

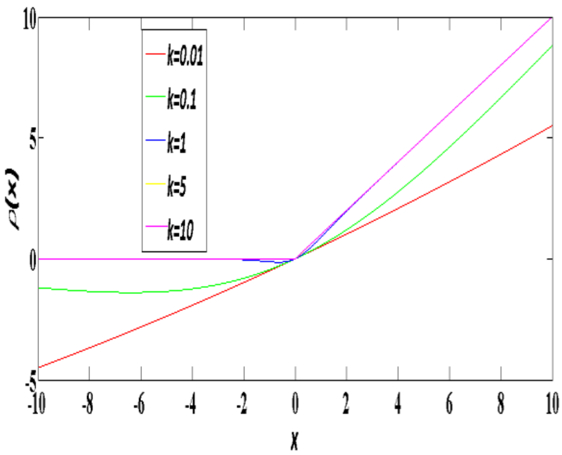

Fig.2 Input-Output relationship of tunable ReLU for various parameter values (k) with input range [-10 10].

The linear approximation of tanh function is given as

$$l_t \approx x$$

Similarly, we obtain a polynomial approximation for the proposed activation function in (8) using Taylor series approximation around zero. The polynomial expression for the tunable ReLU is given as

$$\rho(x) \approx 0.5x + 0.5kx^2 - k^3\, \frac{x^4}{3!} + \qquad for\ \ |kx|>>0 \qquad (15)$$

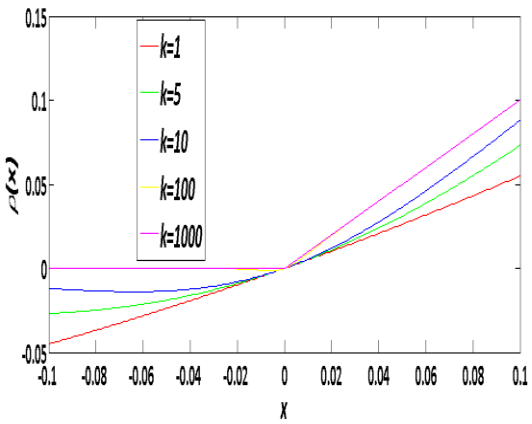

Fig.2 Input-Output relationship of tunable ReLU for various parameter values (k) with input range [-0.1 0.1].

In this polynomial approximation except first order, all other higher odd order variables vanish at zero. From (15) a linear approximation for $\rho(x)$ is obtained as

$$l_p = 0.5x \qquad (16)$$

The input-output characteristics of tunable ReLU for different k and x values are shown in Figs. 2 and 3. In Fig. 2, we consider values of x are in the range [-10 10] with parameter value 0.1 and 0.01 in (8). For these values the characteristics of tunable ReLU are similar to ELU as shown in Fig. 2. If the value of k= 5 or larger, the proposed activation function characteristics are similar to ReLU as shown in Fig. 2. From this figure, we also observe that for smaller values of k mean value of the proposed nonlinear activation function moves closer to zero resulting in fast learning. Similarly, if the values of x are in the range of [-0.1 0.1] and with parameter values of k (>100), then the characteristics of the proposed tunable ReLU are similar to a ReLU as shown in Fig. 3.

A. A close approximation of the tunable ReLU by sum of steps

In this subsection, we find an efficient method to approximate the proposed tunable ReLU function using sum of steps to aid hardware implementation. We consider following representation for the proposed activation function.

$$\rho(x) = \frac{x}{1+e^{-2kx}} \quad \forall\ x \in R \qquad (17)$$

From this function to obtain sum of steps representation we use the following two steps:

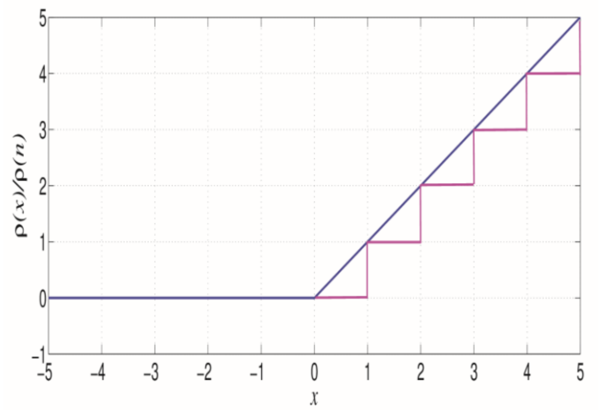

Fig. 4 Continuous tunable ReLU $\rho(x)$ with $k = 10$.

(i) We replace e by 2 since computers use binary number system.

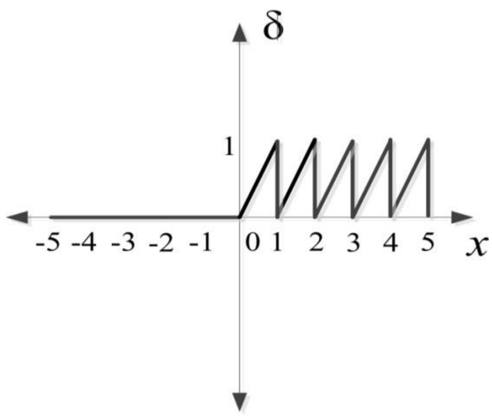

Fig. 5: The quantization error $\delta = \rho(x) - \rho(n)$.

(ii) We replace x by n^1 to obtain fixed point representation of tunable ReLU as

$$\rho(x) = \frac{x}{1+2^{-2kxln_2(e)}} \qquad [\because\ e^{-x} = 2^{-2xln_2 e}]$$

$$\rho(n) = \frac{n}{1+2^{-kn}} \qquad \forall\ n\ \in\ Z$$

$$= \frac{n2^{nk}}{1+2^{kn}}$$

$$= \frac{n2^{nk}(2^{nk}-1)}{2^{2nk}-1}$$

$$= \frac{n(2^{nk}-1)}{2^{nk}}\,\frac{1}{1-2^{-2nk}}$$

$$= n(2^{nk}-1)\sum_{l=0}^{\infty}\frac{1}{2^{2kln+nk}}$$

$$\rho(n) = \sum_{l=0}^{\infty}\sum_{i=1}^{nk}\frac{n}{2^{2kln+i}} \qquad (18)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

[∵ kx is a large value, we have approximated $(2ln_2e)kx$ by kn]. For larger values of kn the above approximation becomes sum of steps approximation for ReLU function as shown in Fig. 4. The quantization error δ = $\rho(x)$ - $\rho(n)$ lies between [0 1] as shown in Fig. 5. By using two's complement form of n and rounding, we reduce this error.

B. A linear-in-the-parameter model for the proposed tunable ReLU

In this subsection, a LIP model for the tunable ReLU function in (8) is obtained using an even mirror Fourier nonlinear filter. A time invariant, nonlinear,

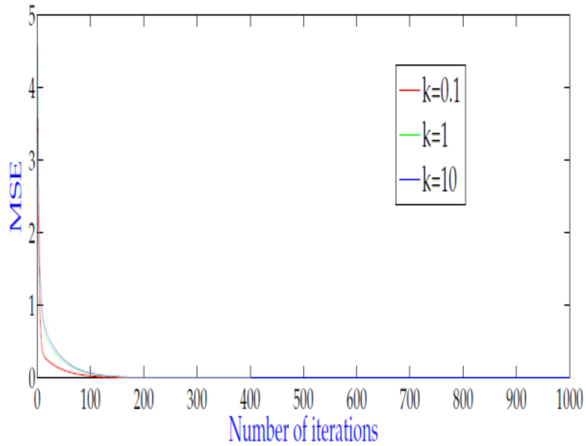

Fig. 6 Mean square error (MSE) Vs Number of iterations for obtaining Fourier nonlinear filter coefficients of tunable ReLU with different *k*.

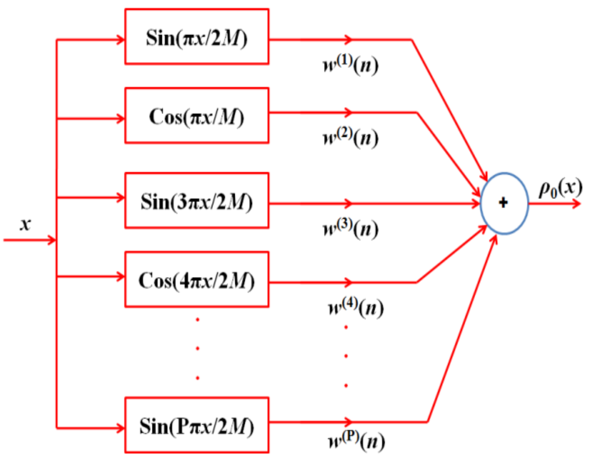

Fig. 7: Representation of tunable ReLU using an even mirror Fourier nonlinear filter basis.

causal, memoryless system is approximated using an EMFN filter basis functions that satisfy the conditions stated in Stone-Weierstrass theorem. If $\rho_0(x)$ is a continuous function that maps from [-M M] to R, then by using its periodicity and even mirror repetition properties, such that

$$\rho_0(x+4M) = \rho_0(x)$$
$$\rho_0(x+M) = \rho_0(M-x)$$

We obtain the following EMFN basis,

$$\sin\left(\frac{P\pi x}{2M}\right) \quad for \quad P \quad odd$$
$$\cos\left(\frac{P\pi x}{2M}\right) \quad for \quad P \quad even \tag{19}$$

here P is the order of an EMFN filter. Let $w^{(1)}, w^{(2)}, \dots \dots \dots, w^{(P)}$ be the $1^{st}, 2^{nd}, \dots \dots ., P^{th}$ order EMFN filter coefficients. The input-output relationship of an EMFN filter is given by

$$\rho_0(x) = \; w^{(1)}\sin\left(\frac{\pi x}{2M}\right) + w^{(2)}\cos\left(\frac{2\pi x}{2M}\right) + \; ... \tag{20}$$

The coefficients of an EMFN filter are obtained by minimizing the mean square error (MSE) by solving an overdetermined system of linear equations. We denote x as the input, $\rho_0(x)$ as the output and $\tilde{\rho}_0(x)$ as the desired response of the nonlinear activation function. Let us consider the input x=[-1 4 1 1 -1 -2-1 0 1 0 -1 0 -1 0 0 1 -2 0 -1 1 -3] and desired response $\tilde{\rho}_0(x)$ is obtained using (8) for a specific value of k. For these values, an EMFN filter coefficients are obtained by minimizing the mean square error between $\rho_0(x)$ and $\tilde{\rho}_0(x)$ with step size of 0.004. The MSE plot to obtain an EMFN filter coefficients for different k values in (8) is shown in Fig. 6. From this plot, we observe that for smaller values of k MSE converges faster than higher values of k. The representation of the proposed nonlinear activation function using coefficients of an EMFN filter and its basis is shown in Fig. 7. The proposed nonlinear activation function response is proportional to the coefficients of an EMFN filter.

IV. PERFORMANCE ANALYSIS & RESULTS

The simulation results show that the proposed activation function tunable ReLU gives better training accuracy in classifying data sets than other activation functions such as sigmoid, *tanh*, ReLU and ELU. In the proposed activation function, we consider k =0.01 for which the learning rate of network is faster than other activation functions as shown in Fig. 8. The training accuracies using various activation functions for MNIST data set is listed in Table I. From Fig. 9, we observe that the training accuracy of the network using the proposed activation function increases as the number epochs
increase.

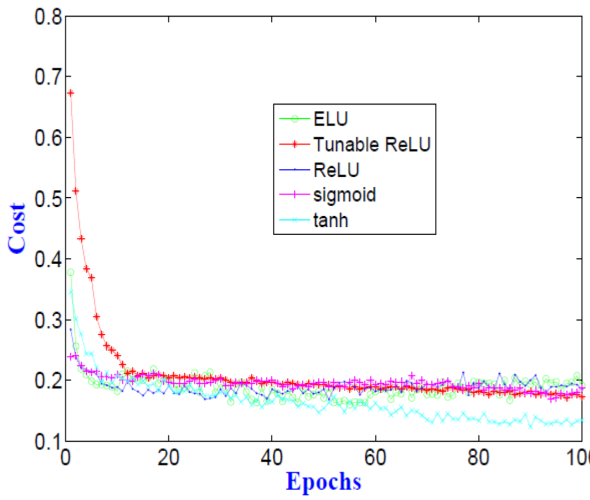

Fig. 8 Epochs vs Cost for Acoustic data in [19] using various activation functions.

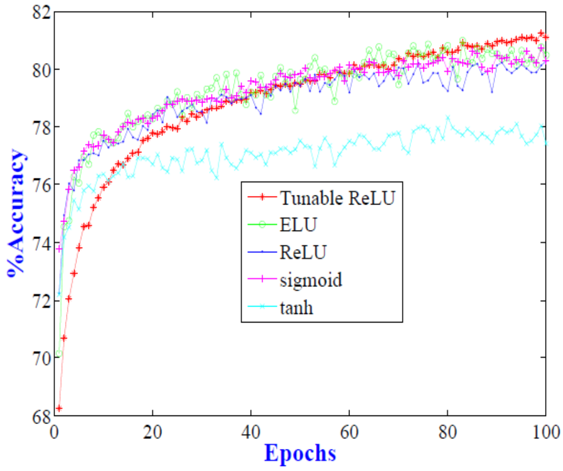

Fig. 8 Epochs vs Accuracy for Acoustic data in [19] using various activation functions.

TABLE I: Performance comparison of various activatio n functions on Acoustic data set in [19].

Activation function	%Accuracy
Sigmoid	80.1548
tanh	78.0837
ReLU	80.2261
ELU	80.8177
Tunable ReLU	80.8224

V. CONCLUSIONS

In this paper, we have proposed tunable ReLU as a novel nonlinear activation function. We have obtained a polynomial approximation of this activation function. An approximation of the tunable ReLU has obtained using sum of steps for hardware implementations. We have studied characteristics of the proposed activation function, it acts as a ReLU for larger tunable parameters and input data values. We have obtained a linear-in-the-parameter model for this function using an even mirror Fourier nonlinear filter for better mathematical tractability. Finally, the simulation results have shown that the proposed activation function gives better classification accuracy.

REFERENCES

1. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton,"Deep learning". nature 521, no. 7553 (2015): 436.
2. Montavon, Grgoire, Wojciech Samek, and Klaus-Robert Mller, "Methods for interpreting and understanding deep neural networks". Digital Signal Processing (2017).
3. Hou, Weilong, Xinbo Gao, Dacheng Tao, and Xuelong Li, "Blind image quality assessment via deep learning". IEEE transactions on neural networks and learning systems 26, no. 6 (2015): 1275-1286.
4. Liu, Qingju, Wenwu Wang, Philip JB Jackson, and Yan Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions". In Signal Processing Conference (EUSIPCO), 2017 25th European, pp. 1270-1274. IEEE, 2017.
5. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks". In Advances in neural information processing systems, pp. 1097-1105. 2012.
6. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le, "Sequence to sequence learning with neural networks". In Advances in neural information processing systems, pp. 3104-3112. 2014.
7. Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". IEEE Signal Processing Magazine 29, no. 6 (2012): 82-97.
8. Duch, W. and Jankowski, N., 1997. New neural transfer functions. Applied Mathematics and Computer Science, 7, pp.639-658.
9. Glorot, Xavier, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks". In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315-323. 2011.
10. Wiatowski T, Blcskei H. A mathematical theory of deep convolutional neural networks for feature extraction. IEEE Transactions on Information Theory. 2018 Mar;64(3):1845-66.
11. Clevert, Djork-Arn, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)". 4th international conference on ICLR, San Juan, US, (2016).
12. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. InAdvances in Neural Information Processing Systems 2017 (pp. 971-980).
13. Beiu, Valeriu, J. A. Peperstraete, Joos Vandewalle, and Rudy Lauwereins, "Closse approximations of sigmoid functions by sum of step for VLSI implementation of neural networks". Sci. Ann. Cuza Univ. 3 (1994): 5-34.
14. A. Carini, G. L. Sicuranza, "Recursive even mirror fourier nonlinear filters and simplified structures". IEEE Trans. Signal Process., 62(24), 6534-6544 (2014).
15. Gulcehre, Caglar, Marcin Moczulski, Misha Denil, and Yoshua Bengio,"Noisy activation functions". In International Conference on Machine Learning, pp. 3059-3068. 2016.
16. Atkinson, Kendall E, "An introduction to numerical analysis". John Wiley & Sons, 2008.
17. A. Friedman, "Foundations of modern analysis". Courier Corporation,(1970).
18. W. Rudin, "Principles of mathematical analysis". New York: McGraw- Hill, Vol. 3, (1964).
19. https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/./.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*