

Article

An n-Sigmoid Activation Function to Improve the Squeeze-and-Excitation for 2D and 3D Deep Networks

Desire Burume Mulindwa  and Shengzhi Du *

Department of Electrical Engineering, Tshwane University of Technology, Pretoria 0183, South Africa

* Correspondence: dushengzhi@gmail.com

Abstract: The Squeeze-and-Excitation (SE) structure has been designed to enhance the neural network performance by allowing it to execute positive channel-wise feature recalibration and suppress less useful features. SE structures are generally adopted in a plethora of tasks directly in existing models and have shown actual performance enhancements. However, the various sigmoid functions used in artificial neural networks are intrinsically restricted by vanishing gradients. The purpose of this paper is to further improve the network by introducing a new SE block with a custom activation function resulting from the integration of a piecewise shifted sigmoid function. The proposed activation function aims to improve the learning and generalization capacity of 2D and 3D neural networks for classification and segmentation, by reducing the vanishing gradient problem. Comparisons were made between the networks with the original design, the addition of the SE block, and the proposed n-sigmoid SE block. To evaluate the performance of this new method, commonly used datasets, CIFAR-10 and Carvana for 2D data and Sandstone Dataset for 3D data, were considered. Experiments conducted using SE showed that the new n-sigmoid function results in performance improvements in the training accuracy score for UNet (up 0.25% to 99.67%), ResNet (up 0.9% to 95.1%), and DenseNet (up 1.1% to 98.87%) for the 2D cases, and the 3D UNet (up 0.2% to 99.67%) for the 3D cases. The n-sigmoid SE block not only reduces the vanishing gradient problem but also develops valuable features by combining channel-wise and spatial information.

Keywords: squeeze and excitation; channel and spatial correlation; sigmoid activation function; neural network models; datasets



Citation: Mulindwa, D.B.; Du, S. An n-Sigmoid Activation Function to Improve the Squeeze-and-Excitation for 2D and 3D Deep Networks. *Electronics* **2023**, *12*, 911. <https://doi.org/10.3390/electronics12040911>

Received: 5 December 2022

Revised: 7 February 2023

Accepted: 9 February 2023

Published: 10 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Activation functions (AFs) refer to the idea that neurons in the brain form input–output dynamic systems [1]. The choice of such computational transfer function mappings is important for artificial neural networks (ANN). Sigmoid functions are commonly used as activation functions, due to their derivatives being adaptable to gradient descent algorithms [2]. Additionally, they possess intrinsic output boundedness and are continuously differentiable $C^\infty(\mathbb{R})$, monotonic with nonlinearity properties [3,4], which fulfills important Lyapunov stability conditions [5].

Remarkably, for the best function approximation and faster convergence, sigmoid functions ought to be symmetric about the origin of the function’s input–output map [6]. In the literature, the function approximation power of the sigmoid has also been explored exhaustively. For instance, Ohn I. and Kim Y. [7] investigated the smooth function approximation capacity of deep neural networks and grouped the activation functions into two groups: piecewise linear [8] (rectified linear unit (ReLU) variants) and locally quadratic (sigmoid variants).

Sigmoid functions remain the favorite choice in architectures such as recurrent neural networks (RNNs) [9,10], mostly because parallel computing techniques (such as the GPU power) have made these computations faster. Moreover, the use of the computing techniques of function approximation has drastically reduced the computational complexity to adequate levels of accuracy [11].

Nevertheless, two important problems affect the sigmoid functions in general. First, representations of sigmoid functions hold a fundamental learning or optimization downside whereby their derivatives exponentially move towards zero, particularly as the networks grows deeper. This conundrum is termed the vanishing (exponentially decreasing) gradient problem.

The second problem is the numerical indeterminacy of the log-sum-exp [10]. In naïve computation, one encounters underflows or even overflows depending on the scale of the input. The proposed n-sigmoid in this paper aims to resolve both issues.

To decrease both the computational and model complexity, several neural network architectures [12,13] centered on the network channels in many recent works have demonstrated that the efficacy of very deep networks is very low [14,15]. In contrast, the Squeeze-and-Excitation Network (SENet) [16], winner of the ILSVRC 2017 classification challenge, centers on channel-wise feature recalibration to refine the network's accuracy.

The SE block has two major steps, squeeze followed by excitation. The squeeze part relies on global average pooling and brings about the channel descriptor. As for the excitation segment, it has a straightforward gating structure using a sigmoid activating function. A scalar magnified by the feature map in a channel-wise manner forms the output.

In summary, the key contributions of this proposed method are as follows:

- (1) This paper proposes a novel SE block with the n-sigmoid activating function. The proposed n-sigmoid function satisfies the requirements of an activating function, such as boundedness, smoothness, and sensitivity to input values.
- (2) The proposed SE block boosts the system representation capacity by clearly modeling the interconnections between the channels.
- (3) The proposed n-sigmoid activating function is in the range of $[0, 1]$, which allows it to preserve meaningful features and subdue less beneficial features.
- (4) The proposed SE structure can be simply merged with existing models and boost their performance.

This paper is arranged as follows. Section 1 is the introduction. In Section 2, the related works are discussed to first introduce the squeeze-and-excitation networks as seen in the vision fraternity, secondly to explore the various activation functions and their enhancements, and finally to briefly investigate the channel and spatial correlations in the convolutional neural network (CNN). In Section 3, the proposed methodology is discussed, where the traditional sigmoid activation function and the vanishing gradient problem are explained before the new proposed n-sigmoid is introduced together with a solution to the numerical stability of the log-sum-exp and to the vanishing gradient. In Section 4, the experiments with their results are explained in detail, and, finally, in Section 5, we discuss the conclusions and future works.

2. Related Works

2.1. Squeeze-and-Excitation (SE)

The Squeeze-and-Excitation Network has been accepted in the vision fraternity. The SE structure recalibrates channel-wise features and reduces the less valuable ones. At each layer of the network, a set of filters displays neighborhood spatial connectivity designed through input channels—combining channel-wise and spatial information around neighboring receptive fields.

The convolution part (the main building block of convolutional neural networks) allows networks to build valuable features by combining channel-wise and spatial information in the internal receptive fields. A large category of previous research has explored the spatial relationship by trying to reinforce the representational strength of a convolutional neural network (CNN) by improving the nature of spatial encodings throughout its feature ladder. Conversely, the “Squeeze-and-Excitation” (SE) block (Figure 1) is based on the channel relationship and introduces a new building unit to recalibrate adaptively channel-wise features. SE blocks significantly improve current state-of-the-art convolutional neural networks' performance, adding a small additional computational cost.

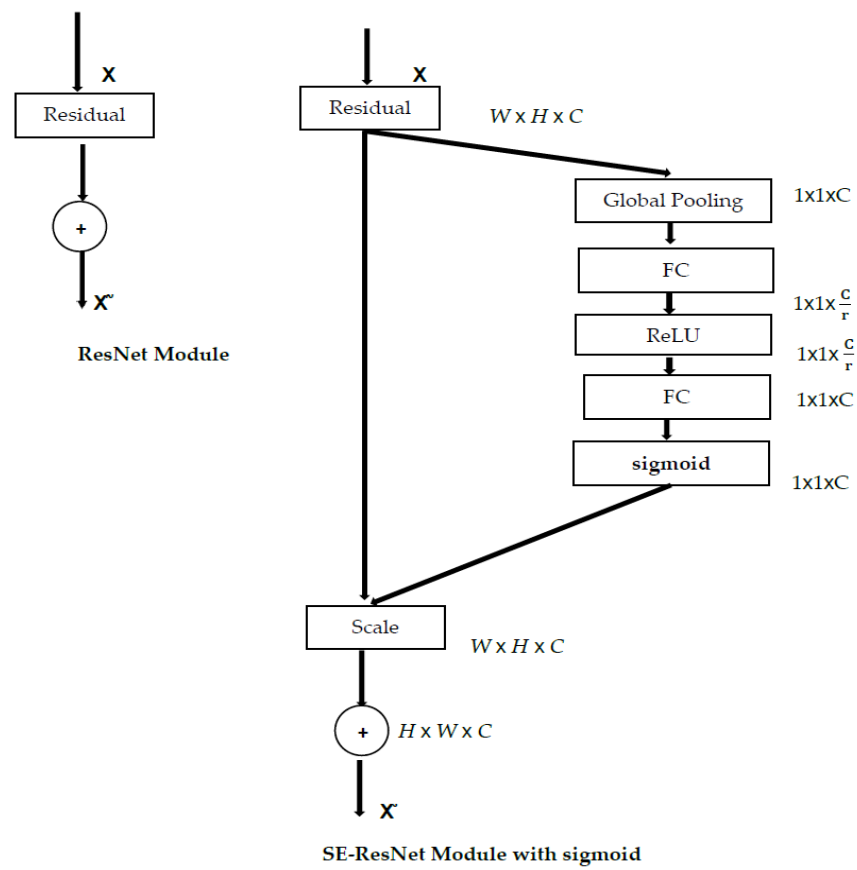


Figure 1. The schematic of the ResNet module with its SE.

The Squeeze-and-Excitation formula is as follows:

$$SE = \sigma(W_2 \gamma(W_1 z + b_1) + b_2) \quad (1)$$

where σ is the sigmoid, γ is the ReLU function, and $\{W_1 \in \mathbb{R}^{C/r \times C}, b_1\}$ and $\{W_2 \in \mathbb{R}^{C \times C/r}, b_2\}$ are the weights connecting two fully connected layers.

Inside the SE, the computational cost and capacity are controlled by a reduction ratio r . The input X and the scalar SE are multiplied channel-by-channel to obtain the final output:

$$X_{SE} = X \times SE \quad (2)$$

where “ $\times$ ” symbolizes channel-wise multiplication. The SE structure will retain useful features and reduce unnecessary features to boost the network’s performance.

Figure 1 shows that the sigmoid function plays an important role inside the squeeze segment of an SE structure, where it acts as a gating mechanism with a sigmoid activation function to increase the representational ability of the network’s dynamic channel-wise feature recalibration.

The different uses of the sigmoid activation function in artificial neural networks are intrinsically limited by gradients that are vanishing. The word “vanishing” means that, in a feedforward network (FFN), the backpropagated error signal naturally declines (or rises) exponentially in relation to the distance from the last layer [17]. The vanishing gradient arises when the artificial neural network is trained with backpropagation and gradient-based learning methods. In the worst case, this would stop the neural network from training more and better.

2.2. Activation Functions

Activation functions have been interminably the cause of nonlinearity in neural networks [18]. Conventional activation functions (such as tanh and sigmoid) are differentiable and continuous, but the sigmoid has only a positive value, while the tanh has a negative one. Most enhancements of the sigmoid function normally focus on changing the slope of the sigmoid or shifting the original sigmoid, as opposed to the new proposed sigmoid that is a piecewise log-shifted function in a finite input–output space.

The sigmoid activation function is often used in feed-forward neural networks (FFNN) [19] to introduce nonlinearity. To accelerate network convergence, CNNs mostly use the hyperbolic tangent as an activation function [20]. One of the latest advances in activation functions is the non-negative rectified linear unit (ReLU) [21], where the identity map in the positive portion solves the gradient vanishing challenge.

To conquer the challenge of the inability of ReLU to multiply negative values because of the constant zero in its negative parts, the parametric rectified linear unit (PReLU) [22] includes a learnable variable in the negative portion. New exponential linear unit (ELU) enhancements [23] similarly insert additional learnable variables to expand the significance of the ELU. These variables/parameters are either channel-wise or channel-shared. Lately, the addition of the learnable parameters has been the new direction in activation function research. The p-sigmoid [24] showcases parameters related exclusively to function boundaries (which allows positive, zero, or negative contributions from each hidden unit), horizontal displacement, and the steepness of the curve.

2.3. Channel or Spatial Correlations

With the revival of the convolutional neural network (CNN), the vision fraternity prioritized two hyperparameters, which are the depth [8,12] of the network and the filter sizes [25]. The depth has continued to increase, to the extent that it is now possible to train very deep networks.

Aside from the depth of the network, Wide ResNet [26] has shown that the channel width is also a very important consideration when it comes to performance. Old convolutions have integrated channel and spatial correlations, while more recent work has attempted to dissociate them to improve the efficiency. The Inception module [27] first discovered cross-channel interaction and later led to a set of spatial correlations.

ResNext [28] investigated the trade-offs between width (the number of channels in a layer), depth, and group. Xception [29] introduced the depth-wise separable convolution, where the number of group convolutions equals the number of channels. ShuffleNet [30] introduced a channel shuffle task to improve the depth-wise separable convolution. Shi et al. [31] introduced a spatial and channel attention unit. The input features were improved using spatial attention and channel attention separately, where the features were acquired with hybrid attention.

2.4. Adaptation and Improvements of the SE Block

Since its inception, due to its large success, various attempts have been made to improve the SE block. Several avenues have been exploited, such as the reduction of its architecture [32], optimization of the reduction ratio [33], alteration of the global pooling operation [34], combination of the squeeze and excitation stages [35], an improved gating mechanism by introducing a more powerful activation function [24], etc., as can be seen in Table 1. The proposed n-sigmoid aims to challenge the latter by presenting a very competitive activation function.

Table 1. Summary of various SE methods.

SE Methods	Dataset Used	Result %	Limitations
PSigmoid [24]	CIFAR10	93.03	Over-reliance on the learned attention scores, computational complexity, and overfitting if not properly generalized
SE SqueezeNext [32]	CIFAR10	92.6	Tradeoff between model's size, accuracy, and speed, and no change in the accuracy even when the model depth is reduced.
LRSE-Net [33]	Deep Stenosis Detection	96.05	Need for further improvements, such as evaluating for an object-based recognition system.
CaSE Efficient-NetBO [34]	VTAB + MD	98.02	Requires iterative hardware-dependent gradient updates and falls short on structured datasets.
FuSENet [35]	IP, UP, and SA HSI	98.64%	Needs addition to a residual learning network to improve learning efficiency.

p-sigmoid [24] (PSE) aims to enhance the SE and introduces the notion of combined channel-wise and channel-shared parameters. In [32], the SE SqueezeNext presents a more sophisticated block with a condensed architecture for higher accuracy, but it is found that there is a trade-off between the model's accuracy, size, and speed for various resolution and width multipliers, even though there is no improvement in the model accuracy when reducing the depth of the model. Ovalle-Magallanes et al. [33] presented a Lightweight Residual Squeeze-and-Excitation Network (LRSE-Net) for stenosis classification in X-ray coronary angiography images, where the reduction ratios of individual Squeeze-and-Excitation modules are optimized to enhance the feature recalibration. In [34], the authors present a new adaptive block called Contextual Squeeze-and-Excitation (CaSE) that regulates a pretrained neural network to drastically enhance the performance with a user data single forward pass as context. The execution of SENet relies on the global pooling squeeze operation, which could lead to reduced performance. In [35], the authors propose a bilinear fusion mechanism over various types of squeeze operation (max pooling and global pooling). The excitation operation is executed with the fused output of the squeeze operation.

3. Proposed Method

3.1. Squeeze-and-Excitation (SE) Block with n -Sigmoid

The proposed methodology seeks to improve the SE block by using the n -sigmoid (explained in Section 3.2) as a gating mechanism instead of the classic sigmoid. Figure 2 shows the schematic of the proposed solution, which will allow the SE block to keep meaningful features and suppress less valuable ones, as well as increase its capacity to learn and generalize 2D and 3D networks for overall enhanced performance.

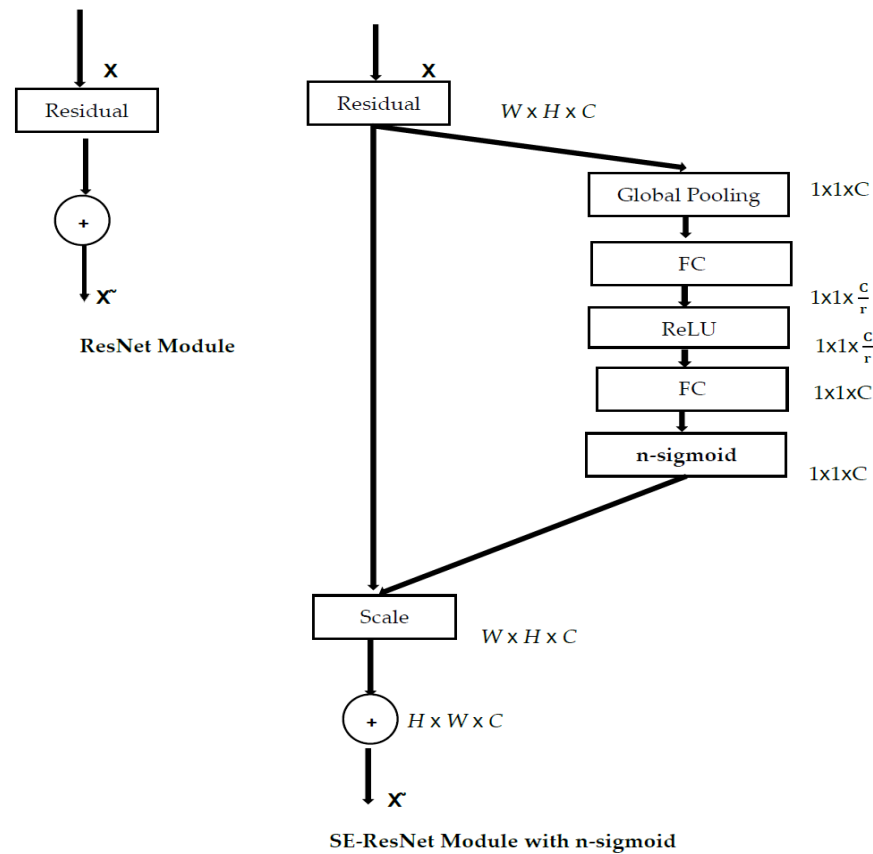


Figure 2. The schematic of the ResNet module using an SE block with an n-sigmoid.

The idea of replacing the sigmoid activation function with the n-sigmoid presents several advantages:

- (1) The network performance is improved. The training and validation loss and accuracy are slightly upgraded.
- (2) The famous gradient vanishing/exploding challenge is resolved.
- (3) The spatial and channel relationships in the network are enhanced since the n-sigmoid allows the SE block to further model channel interdependencies.

3.2. The Proposed n-Sigmoid Activation Function

The proposed generalized form of the n-sigmoid is

$$\begin{aligned}
 f(x) &= -6pt \int_0^\infty \sum_{k=1}^n \frac{\alpha}{(e^{(ax \pm \beta)} + 1)} \\
 &= \pm \alpha * \int_0^\infty \sum_{k=1}^n \frac{1}{(e^{(ax \pm \beta)} + 1)} dx \\
 &= \pm a * \frac{\sum_1^n \log_n (e^{(aax \pm \beta)} + 1)}{b}
 \end{aligned} \tag{3}$$

where a is a constant that depends on α and results from the computation of the integration of the derivative, and b is a constant depending on β and resulting from the computation of the integration of the derivative.

β is the offset parameter that centers the input domain x , while $\alpha (\in \Re)$ is the logistic growth rate parameter. The finite input–output space that is contained within $x_{\min} - x_{\max}$, and $y_{\min} - y_{\max}$ defines the boundaries of the function.

An n-sigmoid (Figure 3) is presented in this finite input–output surface, which yields the resulting integration of an n -equal sigmoid (Equation (3)). However, the naïve computation of the logarithm sum may lead to numerical challenges that will need a solution, as explained in Section 3.2.1.

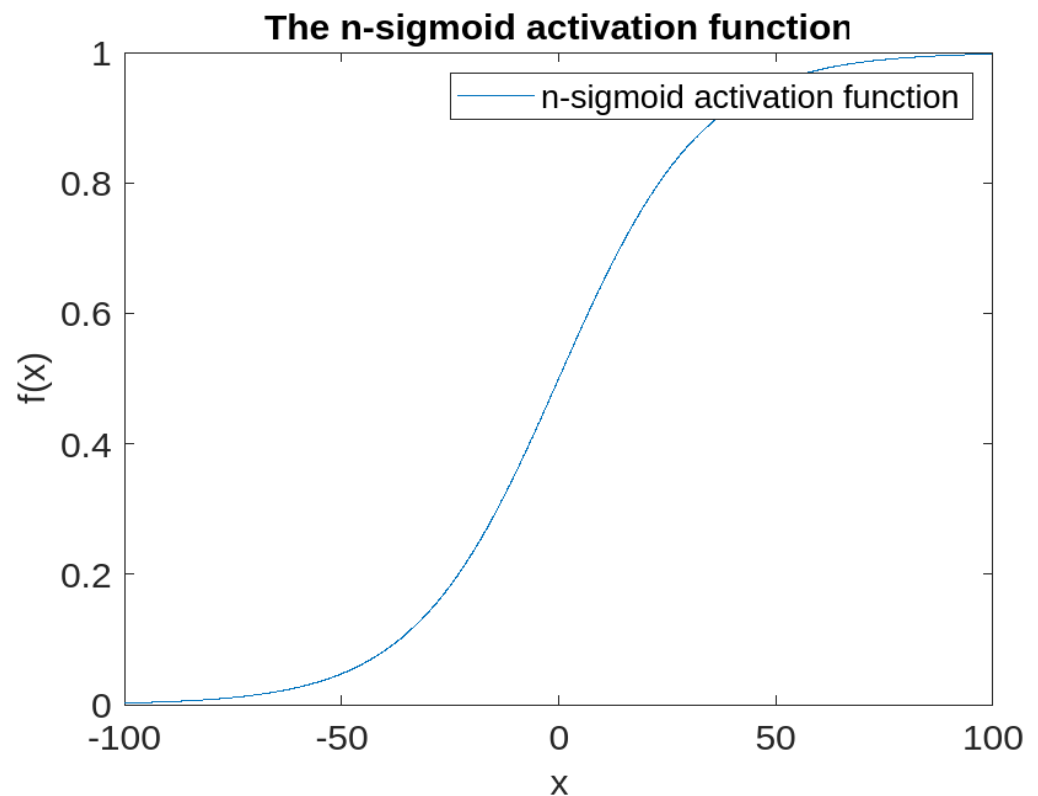


Figure 3. The n-sigmoid activation function.

The novel n-sigmoid function has a derivative a series of n shifted sigmoids in a defined and finite input–output space. This results in restricting the vanishing gradient problem, as explained in Section 3.2.2.

3.2.1. Numerical Stability of the Log-Sum-Exp

A naïve implementation of the log-sum-exp can result in numerical indeterminacy even for moderate values. In deep learning and statistical modeling, it is possible to work in a logarithmic scale. There are many possible situations for this numerical explosion. For instance, if x and y are small numbers, their multiplication may result in underflow.

Moreover, if x and y are small numbers, their multiplication may result in identity

$$\log \sum_{n=1}^N \exp\{x_n\} = a + \log \sum_{n=1}^N \exp\{x_n - a\} \quad (4)$$

for any a . It is now possible to shift the center of the exponentiated variables up and down as needed. A typical action to take is to perform

$$a = \max_n x_n \quad (5)$$

which ensures that the largest value that we exponentiate will be zero. This means that the result will not overflow, and even if the rest underflow, one will still obtain a reasonable value.

3.2.2. Solution to the Vanishing/Exploding Gradient

Presenting the novel activation function in a finite input–output could reduce drastically the vanishing gradient limitation and enhance the accuracy of a fully convolutional neural network [36].

As an activation function, the maximum gradient is set to or falls within a value that ensures that the gradient neither explodes nor vanishes. This is achieved since, in this case, the maximum value of the derivative is 0.5. Hence, the maximum gradient is solely dependent on the input/output limits of the derivative.

Therefore, with the n-sigmoid activation function definition, gradients do not vanish, except at the input value x , which charts the maximum or minimum boundary of the output $f(x)$ space, as proven in Figure 4. Consequently, due to the way in which the gradient is restricted and presented piecewise over the input–output map, the common occurrence of vanishing gradients will be far diminished compared to the traditional sigmoid use case.

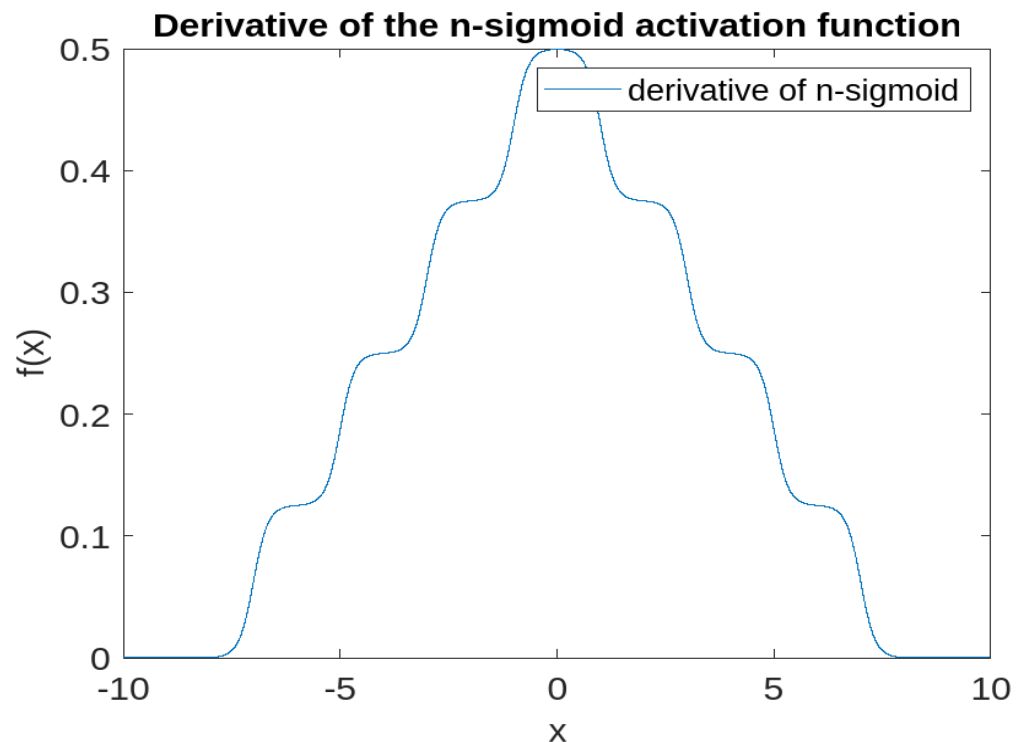


Figure 4. n-sigmoid derivative, which is piecewise of sigmoid.

Figure 4 shows the derivative of the proposed n-sigmoid, which is a piecewise shifted sigmoid, and its integration results in the new proposed activation function (Figure 3).

The derivative g of the n-sigmoid is a piecewise n shifted sigmoid function defined as

$$g = \pm \sum_{k=1}^n \frac{\alpha}{(e^{(\pm \alpha x \pm \beta)})} \quad (6)$$

where n is the number of the subdivisions of the sigmoid function.

3.3. The Advantages of Using the n-Sigmoid over Other Logistic Activation Functions

The n-sigmoid present many advantages compared to existing sigmoid activation functions such as the n-logistic sigmoid [36]:

1. The peak of the derivative is present at the most sensitive point (middle at zero).
2. This novel activation function presents a lower risk of gradient vanishing problems since the derivative glides only in the defined space, which results from the definition of the derivative piecewise sigmoid function.
3. It is also opposite to the n-logistics sigmoid [36], which has multiple peaks that cannot be physically motivated or explained, since the value of the activation function is not constantly at the mid-length of the n th sigmoid, as can be seen in the graph presented in the paper. The question is what will cause the presence of so many

peaks to be relevant, especially because there are no peaks at zero. There is a lack of physical motivation.

4. Experiments and Results

In this section, the datasets used are described and the n-sigmoid is compared to other traditional activation functions, and finally a performance comparison of the n-sigmoid is established when used in 2D and 3D CNN models and when adding the Squeeze-and-Excitation Network.

The vanishing gradient problem is resolved by the restriction of the input–output bounded n-sigmoid function in a finite map.

4.1. Datasets and Evaluation Metrics

The datasets used for the 2D implementation are the MNIST [37], CIFAR10 [38], and Carvana [39] datasets. As for the 3D implementation, the 3D Sandstone dataset have been used.

The MNIST dataset (Figure 5) is a vast handwritten digit dataset that is frequently utilized to train different image processing structures. The database is extensively utilized for training and testing in several machine learning tasks.



Figure 5. An illustration of the MNIST dataset.

The CIFAR10 dataset is composed of 60,000 (32×32) color images divided into 10 classes. It has 50,000 training images and 10,000 test images. The dataset has one test batch and five training batches, each with 10,000 images in random order.

The Carvana dataset has been designed for a successful online used car start-up to develop long-term trust with clients and streamline the online buying operation. The dataset consists of a custom revolving car photo studio of 16 standard pictures of each vehicle in their inventory.

Sandstone 2015 is a 3D model series having both 3D and 2D geoscientific information that supplements the GSWA Record, while the lung CT scan dataset comprises 20 CT scans of COVID-19 patients gathered from Radiopaedia and coronavirus cases.

To evaluate the proposed approach, the metrics considered were the accuracy, loss, and dice similarity coefficient (DSC) or dice score, also known as the F1 score. In the case of image segmentation, DSC is one of the most used metrics and is proposed to evaluate the models. In general, the DSC is meant to indicate the results' similarity regarding the segmentation area and the predicted result and the ground truth. In addition to using DSC, this study also used a classification metric, accuracy, which measures the ratio of correctly identified predicted pixels to all predicted pixels.

4.2. *n*-Sigmoid Compared to Traditional Activation Functions

A common convolutional model (details of the model) was implemented and used with the CIFAR dataset to compare the activation functions and show that the *n*-sigmoid activation function performs better.

A simple convolutional neural network (CNN) consists of a pile of Conv2D and MaxPooling2D layers. As input, a CNN has tensors of shape image_height, image_width, and color_channels.

To analyze and access the various training results, different activation functions, such as the sigmoid function, the hyperbolic tangent function, the ReLU function, the leaky-ReLU function, and the *n*-sigmoid, were selected. In this case, the *n*-sigmoid function was configured as a hidden layer activated with the values of $n = 4$. The main metric used to compare their performance on the convolutional neural network was the loss (validation and training) for the different activation functions (Figures 6 and 7) using the same CIFAR dataset on a simple deep learning framework. It can be observed from Figure 6 that the *n*-sigmoid (purple line) ensured a low validation loss throughout the experiment and only ReLU (green line) performed better. The same can be observed in Figure 7, where the training loss is lower than for the other activation functions for the simple convolutional network throughout the training.

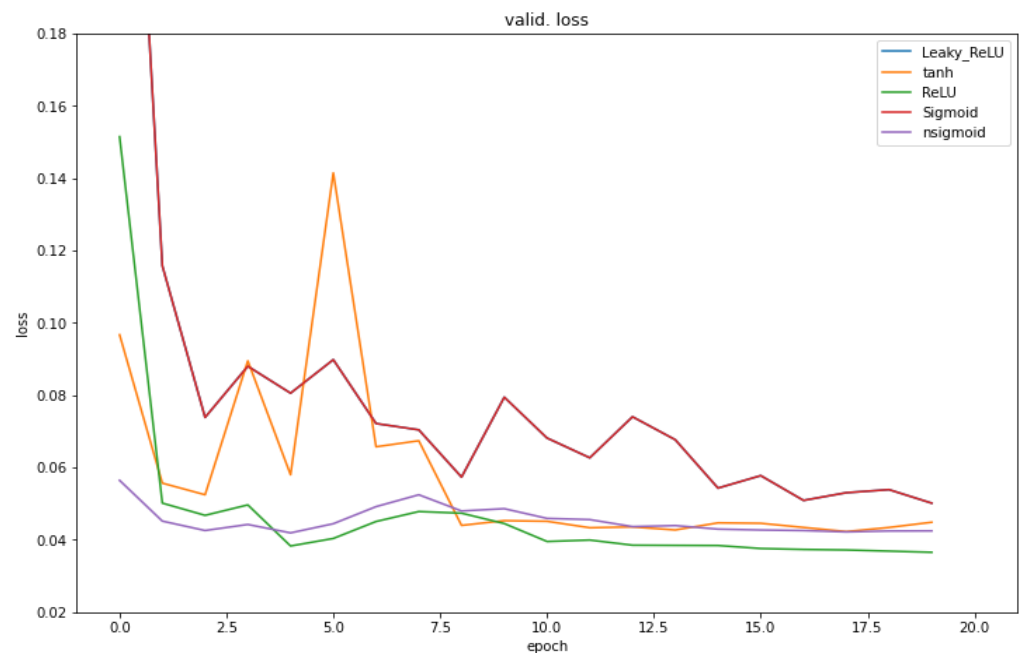


Figure 6. Comparison of the validation loss of the convolutional neural network (CNN) training for all the activation functions.

Another evaluation aimed to compare the effects of changing the sigmoid activation function in the SE block with other activation functions such as the rectified linear unit (ReLU), exponential linear unit (ELU), sigmoid, and p-sigmoid (Table 2) when using the Resnet network on the CIFAR10 dataset.

It is observable that the *n*-sigmoid greatly improves the performance of the SE structure, because it enhances the network attributes by specifically modeling the interdependencies between channels.

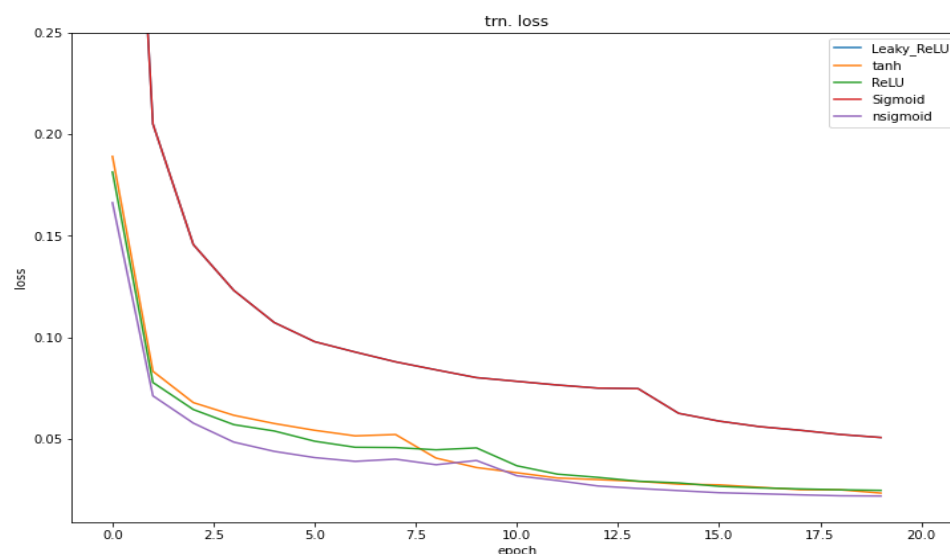


Figure 7. Comparison of the training loss of the convolutional neural network (CNN) training for all the activation functions.

Table 2. Effects of other activation functions on the SE structure using CIFAR10.

Model	Activation Function	CIFAR10 Results (%)
Resnet	Original	92.41
	ReLU	*
	ELU	85.46
	Sigmoid	92.97
	p-sigmoid	93.03
	n-sigmoid	95.1

* indicates failure to converge.

4.3. Implementation Details

All training and validations were performed using either the TensorFlow backend with the Keras library or the Pytorch library. The idea was to experiment and assess the performance of the existing leading architectures in both 2D and 3D when using the Squeeze-and-Excitation networks with the n-sigmoid as the activation function.

In the 2D area, the architectures selected were ResNet (classification), DenseNet (classification and segmentation), and UNet (segmentation). The purpose was to compare the accuracy training results to confirm that the Squeeze-and-Excitation network with the new proposed sigmoid activation function (n-sigmoid SE) would outclass the high-performing Squeeze-and-Excitation network (SE).

The training replicated the parameters of the existing leading networks but added a few necessary changes related to the specifications and hardware used. The comparison process consisted of first training the model as originally designed, and then later with the addition of an SE block or layer, and, finally, the n-sigmoid_SE block was added (Table 2), and all results were compared mostly in terms of the training accuracy score.

For the training process, the 2D UNet [40] employed the Adam optimization method with an initial learning rate of $1e-4$, while ResNet [12] and DenseNet [41] used Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight_decay of $5e-4$. The 3D UNet was trained using Keras, where the kernel_initializer was set to 'he_uniform' and the model was compiled with the Adam optimizer with a learning rate of $1e-3$. In these experiments, all models used simple data augmentation and conventional parameter settings.

As for the loss function, the 2D UNet was trained with binary cross-entropy (BCE) with logits loss, while ResNet and DenseNet used the cross-entropy loss function. Binary cross-entropy was used as a loss function for the 3D UNet.

4.4. Squeeze-and-Excitation Block Using n-Sigmoid Analysis

The Squeeze-and-Excitation block using n-sigmoid is the essential improvement in our proposed methodology. To demonstrate its effect on different networks in 2D and 3D, several ablation experiments were carried out.

4.4.1. Results Analysis of the 2D Squeeze-and-Excitation Block with n-Sigmoid Selected Models' Definition

UNet is a convolutional neural network developed initially for biomedical image segmentation but has gained a great deal of popularity. The network architecture was improved and expanded to work with even small training images and to give more accurate segmentations.

ResNet (Residual Network) [12] is a neural network that allows the training of very deep neural networks with over 150 layers with success. Before ResNet (Figure 8), it was very difficult to train very deep neural networks because of vanishing gradients. ResNet is the first to introduce the idea of a skip connection, where the first input is added to the output of the convolution block.

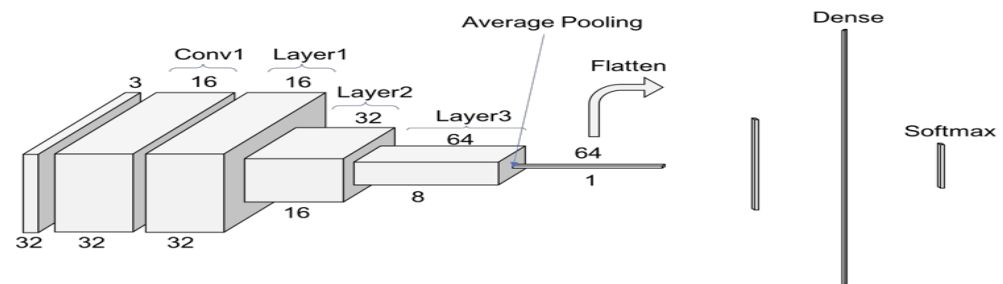


Figure 8. ResNet structure on CIFAR10.

A DenseNet [41] is a convolutional neural network with dense connections between layers (dense blocks), where all layers (with matching feature-map sizes) are connected and used for segmentation (Figure 9).

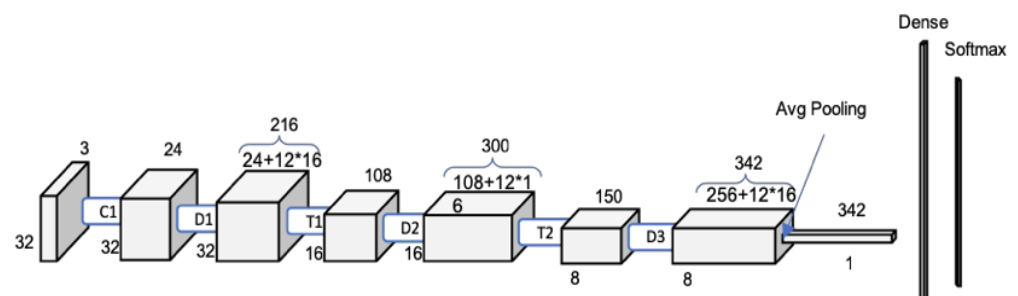


Figure 9. DenseNet on CIFAR10 [with a first single convolutional layer, followed by two pairs of dense blocks, and a third dense block followed by the global average pooling].

Comparative Analysis

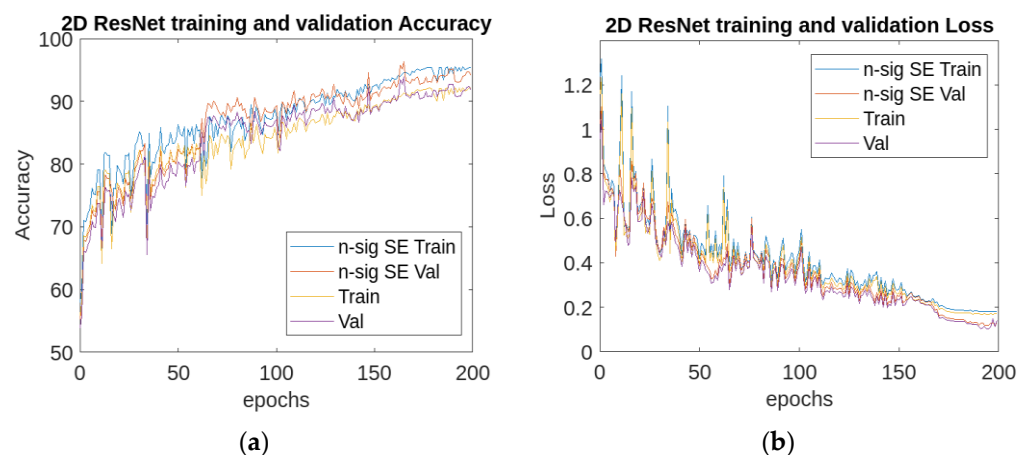
Various parameters (learning rate, epochs, the number of repetitions in the input–output space of the sigmoid, etc.) were used, but the comparison is confined within the network, keeping the design characteristics unchanged.

This comparison work was performed on an NVIDIA Tesla K80 GPU. Table 3 shows that there was a performance improvement when the models added a Squeeze-and-Excitation (SE) block using the n-sigmoid as the activation function.

Table 3. Performance comparison of the state-of-the-art models for training accuracy on classification (a) and segmentation (b).

Model	Category	Accuracy Results (%)
Resnet20 (Cifar10)	Original	92.7
	SE	94.2
	Parametric SE	93.03
	n-sigmoid SE ($n = 4$)	95.5
ResNeXt (Cifar10)	Original	96.4
	SE	96.47
	Parametric SE	96.54
	n-sigmoid SE ($n = 4$)	97.41
(a)		
Model	Category	Accuracy Results (%)
Unet (Carvana)	Original	99.3
	SE	99.48
	Parametric SE	–
	n-sigmoid SE ($n = 4$)	99.67
DenseNet (Cifar10)	Original	99.69
	SE	99.77
	Parametric SE	–
	n-sigmoid SE ($n = 4$)	98.87
(b)		

The 2D UNet training accuracy improved by 0.25% from 99.67%, while ResNet20 with the n-sigmoid SE increased by 0.9% from 94.2% to 95.1%, and DenseNet with the n-sigmoid SE was enhanced by 1.1% from 96.7% to 98.87%. The training and validation curves of DenseNet and ResNet for both the SE models and the n-sigmoid SE models are shown in Figures 10 and 11, and the proposed model obtained the highest accuracy curves and the lowest loss.

**Figure 10.** (a) Training and (b) validation accuracy and loss curves of the SE ResNet and the n-sigmoid SE ResNet.

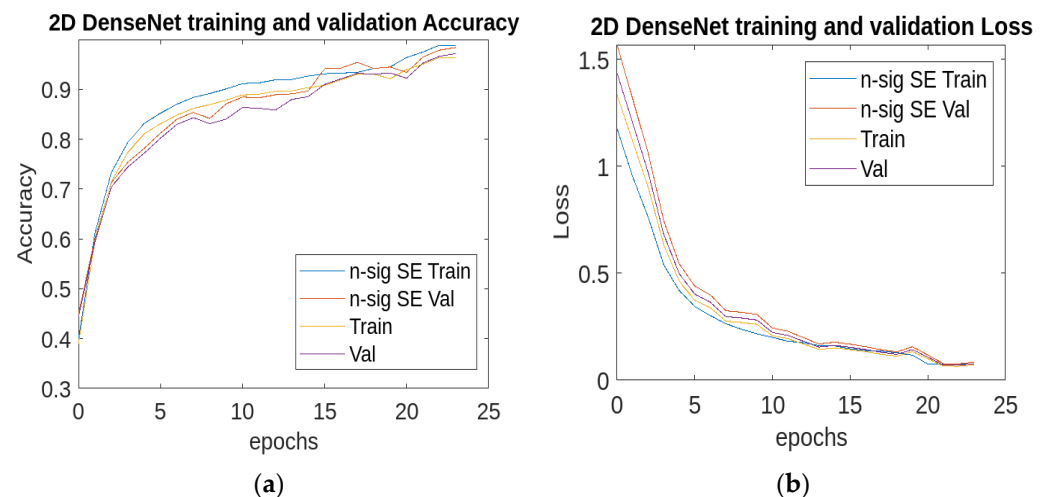


Figure 11. (a) Training and (b) validation accuracy and loss curves of the SE DenseNet and the n-sigmoid SE DenseNet.

Although the difference in the results was not of great importance, the experiments showed that a variation in the value of α slightly affected the results obtained. For instance, at α equal to 0.06, the training accuracy result was 95.360% for ResNet with n-sigmoid, which is a decrease from 95.500% when $\alpha = 0.03$. This shows that the lower the value chosen for α , the lower the accuracy becomes.

The compared testing results showcase that the n-sigmoid SE improves the testing accuracy results when compared to the parametric SE from 93.03% to 95.1%.

n = number of repetitions (sigmoid) in the input–output map of the n-sigmoid activation function.

4.4.2. Results Analysis of the 3D Squeeze-and-Excitation Block with n-Sigmoid Selected Model Definition

Firstly, 3D-UNet [42] has a contractive and expanding path to build a bottleneck in its pivot by using a combination of convolution and pooling operations. After this bottleneck, the image is reconstructed through an amalgamation of convolutions and upsampling. Skip connections are brought in to help the backward flow of gradients and improve the training (Figure 12).

Comparative Analysis

This work also was performed on a Tesla K80 GPU. As discussed in the previous sub-section, even in 3D data, the 3D n-sigmoid affects positively the performance of the neural network model. Here, 3D state-of-the-art deep learning models were tested, and the results are shown in Table 4. The results shown in Table 4 illustrate that the proposed n-sigmoid SE UNet achieved accuracy of 99.67% and DSC of 88.91. Essentially, when using the n-sigmoid SE, the DenseNet achieved 99.71% (from 99.12% on original 3D DenseUNet) accuracy and 95.26 DSC (from 92.54 on original 3D DenseUNet), while ResUNet achieved 99.01% (from 97.93% on original 3D ResUNet) accuracy and 89.52 DSC (from 89.09 on original 3D ResUNet).

Table 4 also shows the Dice coefficients of all models used. It is worth noting that there are limited improvements in the DSC results over the 3D UNet baseline using Carvana dataset [43], which improves only by 0.8 from 88.08 to 88.93. The same is true for the 3D ResUNet DSC, which only improves by 0.43 from the 3D ResUNet baseline. It is only for the 3D Dense UNet that the improvement is sizeable, possibly due to its structure of dense connections between layers.

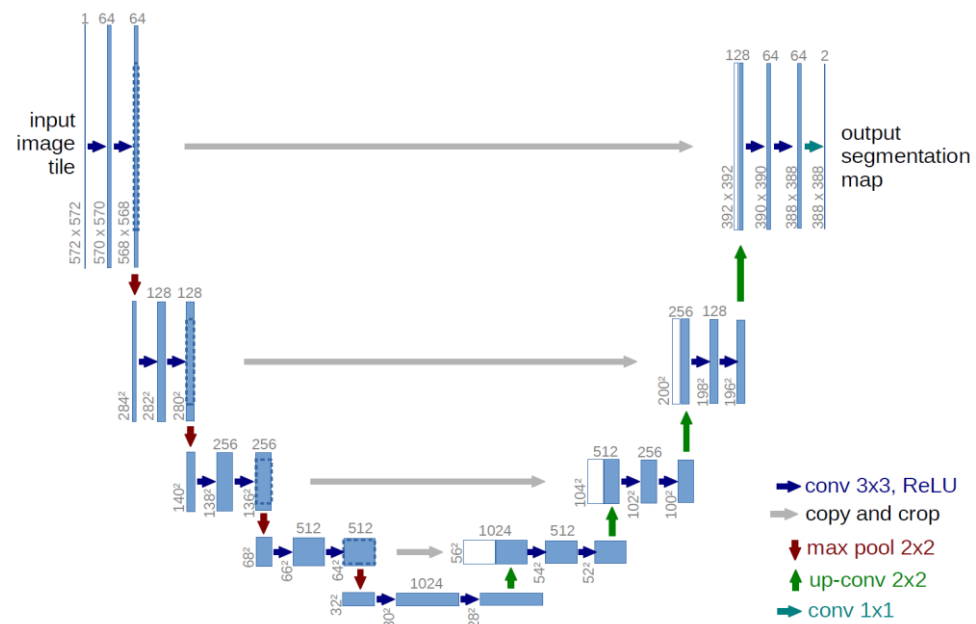


Figure 12. Scheme of 3D-UNet.

Table 4. Segmentation accuracy and Dice similarity coefficient (DSC) results comparison.

Model	Dataset	Dice Similarity Coefficient	Accuracy (%)
3D UNet	Sandstone	88.08 [44]	99.43 [45]
3D SE UNet		88.62	99.48
3D n-sigmoid SE UNet		88.91	99.67
3D DenseUNet	LiTS	92.54 [44]	99.12 [46]
3D SE DenseUNet		94.56 [47]	99.35
3D n-sigmoid SE DenseUNet		95.26	99.71
3D ResUNet	Lung CT scan dataset	89.09 [44]	97.93 [46]
3D SE ResUNet		89.31 [44]	98.29
3D n-sigmoid SE ResUNet		89.52	99.01

One can see that the n-sigmoid SE model performs well in all models studied, with an acceptable margin for both accuracy and DSC. Note that all the results shown here are from three different datasets, namely the Sandstone, LiTS, and lung CT scan datasets. Fairness will dictate that all the models be trained with the same dataset, but the large difference in performance shows that n-sigmoid SE will still perform better, even if using similar datasets.

The training and validation curves of the 3D UNet together with the 3D n-sigmoid SE UNet's accuracy and loss are shown in Figure 13. In essence, our proposed n-sigmoid SE yields the best performance throughout the training, while it presents the lowest curve for the loss too.

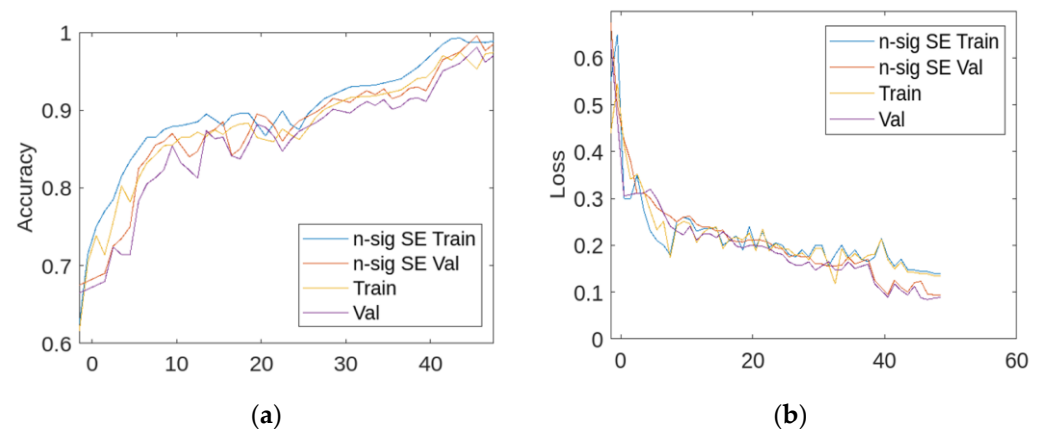


Figure 13. Evolution of the training and validation accuracy (a) and loss (b) curves of the 3D SE UNet and the n-sigmoid 3D SE UNet.

4.5. Discussion

The performance results validate the capability of the proposed method in the different datasets used, such as the CIFAR10, Carvana, LiTS, and Lung CT scan datasets. Several performance metrics were used to evaluate the performance, such as the accuracy and loss, as well as the Dice similarity coefficient (DSC), where the n-sigmoid has consistently shown better results compared to the other models throughout the experiments. By visualizing the training and validation curves of the 2D and 3D experiments (Figures 10, 11, and 13), it is worth noting that the network performance is directly enhanced when using the n-sigmoid SE. The data augmentation approach and hyperparameter settings were set as described in Section 4.3.

4.5.1. Impact of Using Other Activation Functions

The authors have evaluated the effect of replacing the sigmoid activation function in the SE block with other activation functions. They considered four other activation functions—the original sigmoid, ReLU, ELU, and p-sigmoid—and compared them to the n-sigmoid. Replacing the sigmoid in the SE block with other well-known activation functions will greatly reduce the performance. This results from the triumph of the SE block, which has significantly improved the quality of network representation by clearly modeling the interdependencies between channels.

4.5.2. The Vanishing Gradient Solution

The most important challenge of the sigmoid activation function is the vanishing/exploding gradients. It essentially creates a situation wherein a deep neural network does not have the ability to propagate useful gradient information from the output end of the model back to the layers near the input end of the model. This paper has demonstrated that by constraining the n-sigmoid in a finite input–output space, the problem of vanishing gradients could be drastically limited. The proposed SE with n-sigmoid has drastically lowered the risk of gradient vanishing, as explained in Section 3.2.2.

4.5.3. Future Research Directions

In the future, although the n-sigmoid activation function has shown good results in the SE, there is still a need for further improvements, such as overcoming robustly numerical instability for extreme values of the input, evaluating the proposed activation function in a Transformer as the backbone for a self-attention mechanism, and detecting scenes or partial scenes in 3D indoor segmentation. Additionally, a future direction of this work regarding model reduction could be to investigate more methodologies, such as their decomposition, quantization, etc. Finally, another research direction could be to consider

using n-sigmoid in more complex tasks such as 2D–3D object detection, feature extraction for video identification, etc.

5. Conclusions

In this paper, the authors introduced and defined an n-sigmoid Squeeze-and-Excitation block for use as an activation function. It introduces the log of shifted sigmoid functions to be n-times summed up over a finite input–output map. This function largely overcomes the famous vanishing gradient problem of sigmoid functions since it is constrained in a defined space.

The n-sigmoid used in the SE network has shown that it is possible to resolve the vanishing or exploding gradient dilemma to allow deep neural networks to train adequately. The vanishing gradient problem was essentially constrained with the use of the n-sigmoid function in a finite input–output bounded space.

Furthermore, the authors integrated the novel n-sigmoid into the SE structure to improve the SE's performance in its feature recalibration task. Experiments conducted on four different leading networks using a set of benchmark datasets (CIFAR10, Carvana, 3D Sandstone) show that this new module is an advancement from the SE block as well as the parametric SE.

Author Contributions: Conceptualization, D.B.M. and S.D.; methodology, D.B.M. and S.D.; software, D.B.M.; validation, D.B.M. and S.D.; formal analysis, D.B.M.; investigation, D.B.M.; resources, D.B.M.; data curation, D.B.M.; writing—original draft preparation, D.B.M.; writing—review and editing, D.B.M. and S.D.; visualization, D.B.M.; supervision, S.D.; project administration, D.B.M. and S.D.; funding acquisition, S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Sharma, S.; Athaiya, A. Activation functions in neural networks, *International Journal of Engineering Applied Sciences and Technology*. *IJEAST* **2020**, *4*, 310–316. Available online: <http://www.ijeast.com> (accessed on 12 November 2022). [[CrossRef](#)]
- Yuen, B.; Hoang, M.T.; Dong, X.; Lu, T. Universal activation function for machine learning. *Sci. Rep.* **2021**, *11*, 18757. [[CrossRef](#)]
- Runje, D.; Sharath, M.S. Constrained Monotonic Neural Networks. *arXiv* **2023**, arXiv:2205.11775.
- Chibole, J.P. Performance Analysis of the Sigmoid and Fibonacci Activation Functions in NGA Architecture for a Generalized Independent Component Analysis. *IOSR J. VLSI Signal Process.* **2017**, *7*, 26–33. [[CrossRef](#)]
- Wang, Y.; Gao, O.; Pajic, M. Learning Monotone Dynamics by Neural Networks. *arXiv* **2022**, arXiv:2006.06417.
- Chai, E.; Yu, W.; Cui, T.; Ren, J.; Ding, S. An Efficient Asymmetric Nonlinear Activation Function for Deep Neural Networks. *Symmetry* **2022**, *14*, 1027. [[CrossRef](#)]
- Ohn, I.; Kim, Y. Smooth Function Approximation by Deep Neural Networks with General Activation Functions. *Entropy* **2019**, *21*, 627. [[CrossRef](#)]
- Blanchard, P.; Higham, D.J.; Higham, N.J. *Accurate Computation of the Log-Sum-Exp and Softmax Functions*; The University of Manchester: Manchester, UK, 2019; MIMS EPrint:2019.16.
- Timmons, N.G.; Rice, A. Approximating Activation Functions. *arXiv* **2020**, arXiv:2001.06370.
- Apaydin, H.; Feizi, H.; Sattari, M.T.; Colak, M.S.; Shamshirband, S.; Chau, K.-W. Comparative Analysis of Recurrent Neural Network Architectures for Reservoir Inflow Forecasting. *Water* **2020**, *12*, 1500. [[CrossRef](#)]
- Chiluveru, S.R.; Tripathy, M.; Mohapatra, B. Accuracy controlled iterative method for efficient sigmoid function approximation. *Electron. Lett.* **2020**, *56*, 914–916. [[CrossRef](#)]
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
- Gottapu, R.D.; Dagli, C.H. System Architecting Approach for Designing Deep Learning Models. *Procedia Comput. Sci.* **2019**, *153*, 37–44. [[CrossRef](#)]
- Alaeddine, H.; Jihene, M. Deep network in network. *Neural Comput. Appl.* **2021**, *33*, 1453–1465. [[CrossRef](#)]
- Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 420. [[CrossRef](#)]
- Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [[CrossRef](#)]

17. Li, Y.; Fan, C.; Li, Y.; Wu, Q.; Ming, Y. Improving deep neural network with multiple parametric exponential linear units. *Neurocomputing* **2018**, *301*, 11–24. [CrossRef]
18. Zhao, Z.; Feng, F.; Tingting, H. FNNS: An Effective Feedforward Neural Network Scheme with Random Weights for Processing Large-Scale Datasets. *Appl. Sci.* **2022**, *12*, 12478.
19. Liu, X.; Di, X. TanhExp: A smooth activation function with high convergence speed for lightweight neural networks. *IET Comput.* **2021**, *15*, 136–150. [CrossRef]
20. Agarap, A.F. Deep Learning using Rectified Linear Units (RELU). *arXiv* **2018**, arXiv:1803.08375.
21. Trottier, L.; Giguere, P.; Chaib-draa, B. Parametric Exponential Linear Unit for Deep Convolutional Neural Networks. In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18 December 2017.
22. Ven, L.; Lederer, J. Regularization and Reparameterization Avoid Vanishing Gradients in Sigmoid-Type Networks. *arXiv* **2021**, arXiv:2106.02260.
23. Ying, Y.; Zhang, N.; Shan, P.; Miao, L.; Sun, P.; Peng, S. PSigmoid: Improving squeeze-and-excitation block with parametric sigmoid. *Appl. Intell.* **2021**, *51*, 7427–7439. [CrossRef]
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, Diego, CA, USA, 7–9 May 2015.
25. Zagoruyko, S.; Komodakis, N. Wide residual networks. In Proceedings of the British Machine Vision Conference, York, UK, 19–22 September 2016. [CrossRef]
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
27. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500. [CrossRef]
28. Chollet, F. Xception: Deep learning with depth-wise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 550–558. [CrossRef]
29. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856. [CrossRef]
30. Shi, C.; Zhang, X.; Sun, J.; Wang, L. A Lightweight Convolutional Neural Network Based on Group-Wise Hybrid Attention for Remote Sensing Scene Classification. *Remote Sens.* **2022**, *14*, 161. [CrossRef]
31. Chappa, R.T.N.V.S.; El-Sharkawy, M. Squeeze-and-Excitation SqueezeNext: An Efficient DNN for Hardware Deployment. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 0691–0697. [CrossRef]
32. Ovalle-Magallanes, E.; Avina-Cervantes, J.G.; Cruz-Aceves, I.; Ruiz-Pinales, J. LRSE-Net: Lightweight Residual Squeeze-and-Excitation Network for Stenosis Detection in X-ray Coronary Angiography. *Electronics* **2022**, *11*, 3570. [CrossRef]
33. Patacchiola, M.; Bronskill, J.; Shyshey, A.A.; Hofmann, K.; Nowozin, S.; Turner, R.E. Contextual Squeeze-and-Excitation for Efficient Few-Shot Image Classification, Advances in Neural Information Processing Systems (NeurIPS 2022). *arXiv* **2022**, arXiv:2206.09843.
34. Roy, S.K.; Dubey, S.R.; Chatterje, S.E.; Chaudhuri, B.B. *FuSENet: Fused Squeeze-and-Excitation Network for Spectral-Spatial Hyperspectral Image Classification*; The Institution of Engineering and Technology: London, UK, 2020. [CrossRef]
35. Somefun, O.A.; Dahunsi, F. The nlogistic-sigmoid function. *Eur. PMC.* 2020. *preprint*. Available online: www.researchgate.net/publication/343568534 (accessed on 4 December 2022).
36. Yann, L.; Cortes, C. The Mnist Database of Handwritten Digits. 2005. Available online: <http://yann.lecun.com/exdb/mnist/> (accessed on 4 December 2022).
37. Giuste, F.O.; Vizcarra, J.C. CIFAR-10 Image Classification Using Feature Ensembles. *arXiv* **2020**, arXiv:2002.03846.
38. Pandolfi, G.M.; Saliaj, L. Exploratory and Predictive Analysis for Carvana Auction Dataset. 2020. Available online: academia.edu (accessed on 4 December 2022).
39. Zhao, R.; Chen, W.; Cao, G. Edge-Boosted U-Net for Medical Image Segmentation. *IEEE Access* **2019**, *99*, 1. [CrossRef]
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
41. Huang, G.; Liu, Z.; van der Maaten, L. Densely Connected Convolutional Networks. *arXiv* **2018**, arXiv:1608.06993.
42. Cicek, O.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv* **2016**, arXiv:1606.06650.
43. Kaggle. Available online: <https://www.kaggle.com/code/alanyu223/unet-segmentation-on-carvana-dataset> (accessed on 21 September 2022).
44. Github. Available online: https://github.com/zhouyuangan/SE_DenseNet (accessed on 22 September 2022).
45. Cao, Z.; Yu, B.; Lei, B.; Ying, H.; Zhang, X.; Chen, D.; Wu, J. Cascaded SE-ResUnet for segmentation of thoracic organs at risk. In *Neurocomputing*; Elsevier: Amsterdam, The Netherlands, 2020.

46. Asnawi, M.H.; Pravitasari, A.A.; Darmawan, G.; Hendrawati, T.; Yulita, I.N.; Suprijadi, J.; Nugraha, F.A.L. Lung and Infection CT-Scan-Based Segmentation with 3D UNet Architecture and Its Modification. *Healthcare* **2023**, *11*, 213. [[CrossRef](#)]
47. Wang, J.; Zhang, X.; Wang, H. MAD-Unet: Multi-scale attention and deep supervision based on 3D Unet for automatic liver segmentation from CT. *Math. Biosci. Eng. MBE* **2023**, *20*, 1297–1316. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.