PASS: A NOVEL ELASTIC PERIODIC ACTIVATION FOR DEEP NEURAL NETWORKS

Shouyi Wang, Jie Han, Linh Ho Manh, Abhijit Anand Deshpande
Industrial Manufacturing and Systems Engineering
The University of Texas at Arlington
Arlington, TX 76019
shouyw@uta.edu

ABSTRACT

The choice of activation functions is critical to train neural network models for machine learning tasks. Currently, the most widely-used activation function is the Rectified Linear Unit (ReLU). Although there are various alternative activation functions proposed in recent years, most of them are still close approximations or variants of the ReLU function with inconsistent performance in different studies. It is noted that a majority of activation functions used today are non-oscillatory and monotonically increasing. Some recent studies showed that the ReLU-based neural network architectures are incapable of modeling data with fine details, and are not efficient to learn complex periodic patterns, such as multivariate time series data modeling and forecasting problems with many periodic factors and their complex interactions. On the other hand, periodic functions such as Sine waves can be more expressive activation functions than ReLU-like functions to learn complex periodic relationships with fine details using a smaller network size. This study proposes a novel parameterized periodic activation function, which combines the shape flexibility strength of the Swish activation and the notable monotonic-periodic property of the SNAKE activation using two learnable parameters. We call it Parameterized Snake Swish (PASS) in our study. The adaptability and learnability of the proposed PASS activation make it promising to solve many complex real world data modeling problems with both periodic and non-periodic components. We evaluated the proposed activation function on both synthesized data and real-world data. Compared to popular alternative activation functions, experimental results showed that the proposed new PASS activation is indeed effective to largely improve the extrapolation property of a learned neural network, and also demonstrated superior performance for the challenging long-sequence multivariate time series forecasting problem.

***K*eywords** Periodic Activation Functions · Multivariate Time Series · Neural Networks

1 Introduction

Neural networks have potentials to learn arbitrarily complex nonlinear mappings between high-dimensional input to a target output. The universal approximation feature is critically dependent on the nature of the activation function non-linearity used by each layer in the neural network. Training a neural network can be viewed as adjusting a set of parameters to scale, compress, dilate, combine and compose many nonlinear activation functions to model complex relationships between input and target. Despite the critical importance of the nature of the activation function in determining the performance of neural networks, simple monotonic non-decreasing nonlinear activation functions are dominantly used, such as the Rectified Linear Unit (ReLU) function. In this study, we investigate the effects of introducing periodic components to activation functions in deep neural networks. We hypothesize that activation functions with an appropriate periodic component are more efficient to learn complex periodic pattern mappings using less training samples and a smaller network size.

Traditionally, the sigmoid functions were widely used due to their property of approximating the step function while being differentiable. The S-shaped saturating activation functions have the important property of being interpretable as a binary yes/no decision. However, deep neural networks with sigmoid activation functions are hard to train, due to the vanishing gradient phenomenon which arises when saturating activation functions are used. It is considered a milestone to adopt the non-saturating ReLU activation function to alleviate the vanishing gradient issues in deep learning problems. However, a ReLU activation function is only active when inputs are positive during back-propagation. This could leads to two issues:

- Dead Neurons: If the units are not activated initially, then they are always in the off-state as zero gradients flow through them (Dead Neurons) [1].
- Bias Shift: From ReLU, there is a positive bias in the network for subsequent layers, as the mean activation is larger than zero. Since the outputs of all ReLU units are non-negative the outputs may combine to produce very large positive inputs to subsequent layers. The positive mean shift in the next layers may slow down the learning process and lead to numerical accuracy issues [2].

In recent years, several variants of ReLU, such as Leaky ReLU, PReLU[3], SELU [4], ELU [5], have been successful to an extent in mitigating these shortcomings. In addition, a variety of non-monotonic activation functions have also been explored in recent years. Swish [6] and Mish [7] are emerging non-monotonic functions that showed promising results in different benchmark tasks. While promising in various machine learning tasks, the current non-monotonic functions are still close approximations of the ReLU function. The current neural network architectures typically lack the capacity to represent fine details for periodic patterns. This is partly due to the fact that ReLU-like networks are piecewise linear, their second derivative is zero everywhere, and they are thus incapable of modeling information contained in higher-order derivatives of natural signals. In particular, periodical patterns exhibit the intrinsic nature of many complex systems in physics, science, and engineering. For example, many biological and neural systems can be seen as networks of interacting periodic processes, the weather follows daily, seasonal, and yearly modulations, the energy price and demand have complex periodic fluctuations, the stock and economy have complicated and superimposed cycles. The capability to model a periodic system is important to predict future events and evolution based on current and past observations. Although deep neural networks using non-periodic and monotonic activation are powerful to model complex nonlinear patterns from existing data, they are generally inefficient to learn periodic systems and cannot extrapolate well beyond the training range.

On the other hand, periodic functions (such as sin/cos) have been widely used in science and engineering to model complex periodic systems. However, their role in deep neural networks remains largely ignored. This is largely due to the general consideration that periodic neuron activation function may introduce many extra local minima, and the resulting neural networks can be difficult to train and optimize. While there are still repeated investigations of periodic activation functions over the past decades. Some studies, in particular several most recent ones, have shown the great potential of periodic activation functions in deep neural networks. In 1999, Sopena et al. [8] showed that a multi-layer perception with one hidden layer using sinusoids improves accuracy and shortens training times compared to sigmoid activation. The studies of McCaughan [9] and Wong et al. [10] also showed that periodic activation could improve the performance of multi-layer feedforward neural networks on some pattern recognition tasks. Periodic activation on recurrent neural networks (RNNs) was also studied. Sopena Alquezar [11], and Alquezar Mancho et al. [12] showed that replacement of sigmoid to sine activation in the last fully connected layer of an RNN led to higher accuracy on a sequential prediction task. Koplon Sontag [13] used a sinusoidal activation in an RNN to fit sequential input and output data, and Choueiki et al. [14] constructed a sinusoidal activation based RNN for short-term load forecast. Another line of research is called Fourier neural networks, which use sin, cos, and their linear combinations as activation functions in single-hidden-layer networks to mimic the Fourier transform [15, 16, 17]. It has been shown that such models are universal function approximators [18, 19].

Some most recent works have shown the potential of periodic activation functions in neural networks. Giambattista et al. [20] reviewed sine as activation in deep neural networks and showed that sinusoidal activation functions can potentially learn faster and better than those using established monotonic functions on the popular benchmark handwritten recognition task using the MNIST dataset. Sitzmann et al. proposed a Sinusoidal Representation Network (SIREN) using sine as an activation function, which showed superior capability and performance to model fine details of data patterns (e.g., images, wavefields, video, sound) and their spatial and temporal derivatives [21]. Maennel demonstrated that replacing ReLU by sine as activation could greatly enhance a neural network's ability to detect model uncertainty outside of the training distribution. Noel et al. [22] explored a cos-based periodic activation function and showed that replacing ReLU in the convolutional layers of a deep convolutional neural network improved performance on benchmark pattern recognition datasets CIFAR-10, CIFAR-100, and Imagenette. Liu et al. [23] introduced a semi-periodic activation function called Snake, which combines a linear function with a sine component to learn periodic patterns. It showed that the periodic activation could largely improve the extrapolation capability of a learned neural network to make more accurate predictions beyond the range of observed training data points on both simulated and real datasets.

Inspired by these seminal works, we propose a new periodic activation function, which is designed to increase the capability of neural networks to learn both complex periodic and non-periodic data patterns efficiently. In brief, the main contributions of this work can be summarized as follows:

- A new activation function, PArameterized Sinusoidal Swish (PASS), defined as $f(x) = \left(x + \frac{1}{a}\sin^2(ax)\right) / \left(1 + e^{-bx}\right)$, has been proposed. The proposed PASS activation incorporates a periodic component to enable the capability of each neuron to capture both periodic and non-periodic pattern details from the data.
- One learnable parameter a has been introduced to control the periodic component to increase the adaptability of each neuron to capture periodic patterns at different frequencies.
- Another learnable parameter b has been introduced to control the main curve of the activation function with high flexibility, which could adaptively adjust the activation from ReLu-like one-side bounded curves to both-side unbounded curves.
- We investigated the extrapolation properties of neural networks beyond training data, which has been little studied in current deep learning literature. This study shows that neural networks with standard non-periodic activation functions are inefficient to learn periodic patterns, and are incapable of extrapolating periodic patterns outside the training data range. The proposed PASS activation provides an effective solution for this problem.
- The proposed PASS activation and a set of most popular activation functions have been compared and validated on a variety of synthetic and real data. The experimental results clearly indicate that the new PASS activation effectively improves performance compared to popular alternative functions for the challenging long-sequence multivariate time series modeling and forecasting problems.

2 Proposed Activation Function

The use of periodic and and non-monotonic functions in neural networks is not new, while most studies just explored sin and cos or their linear combinations as activation functions. Although the sinusoidal activation functions have been shown successfully for some specific applications and tasks with periodic data, their performance still cannot compete against ReLU-based monotonic activation functions on general practical deep learning tasks. This is because few real-world data problems can be considered as pure periodic. In most of the cases, we do not know exactly if the problem is periodic or contains one or many periodic components for complex big data problems. Thus, basic sinusoidal activation functions are limited to handle different data patterns adaptively in practical applications. For these cases, it is highly desirable to design a flexible activation function that can achieve adaptive learning of both periodic and non-periodic components from the data. Inspired by this motivation, we propose a new activation function to fill the gap in this field.

The proposed activation function is defined as follows:

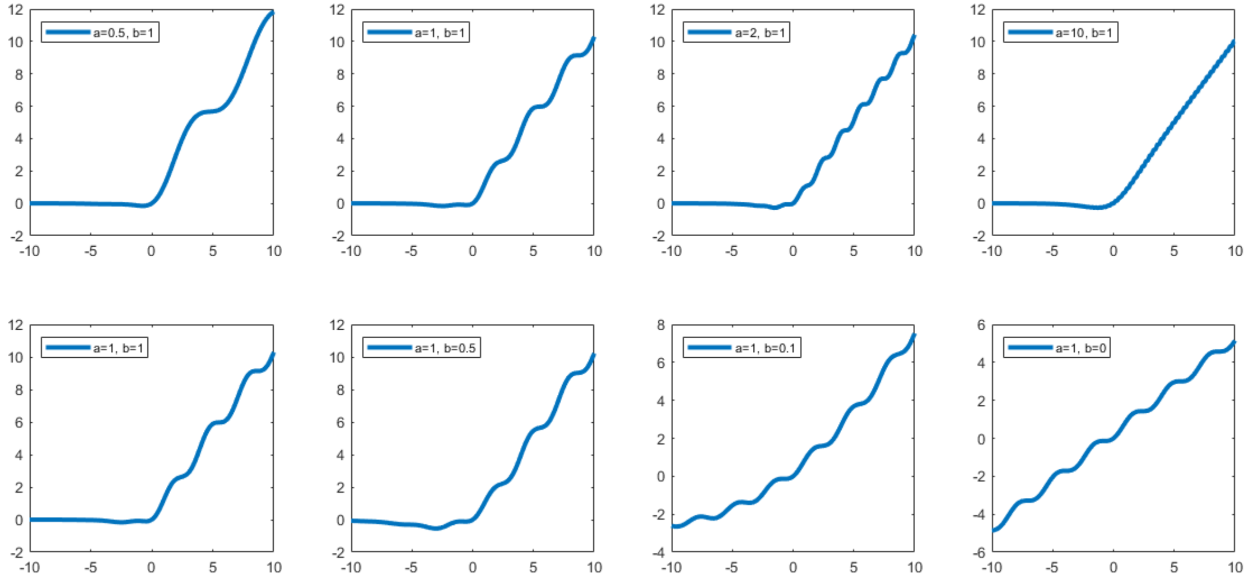
$$f(x) = \left(x + \frac{1}{a}\sin^2(ax)\right) / \left(1 + e^{-bx}\right) \tag{1}$$

The design of the proposed activation function was inspired by two seminal works Swish [6] and SNAKE [23]. The Swish activation, defined as $f(x) = \frac{x}{1+e^{-bx}}$, was discovered by the Google Brain Team with an extensive investigation of various activation functions using a reinforcement learning-based search method. Swish was ranked as the top function and tends to work better than ReLU across a number of challenging tasks. The most striking difference between Swish and ReLU is the non-monotonic "bump" curve at $x < 0$. Their experiments indicate that the monotonic property for activation functions may not be as highly important as people long believed. The non-monotonic properties may be an important aspect of Swish to increase representation power and win the activation competitions in challenging deep learning problems. For periodic activation functions, SNAKE is an impressive work to show that the activation function $f(x) = x + \frac{1}{a}\sin^2(ax)$ is suited to enable a learned neural network to extrapolate well beyond the training range, especially for periodic signals. Compared with other sinusoidal functions, one notable property of the SNAKE function is that it is a monotonic function. Thus, it achieves monotonicity and periodicity at the same time. This design enables neural networks to learn a periodic patterns efficiently while reserving favorable optimization properties of monotonic activation functions.

We observe that Swish could represent a flexible family of non-periodic activation functions using a learnable parameter, while it is limited to learn and extrapolate periodic patterns efficiently. On the other hand, the SNAKE function demonstrated its effectiveness for neural networks to learn and extrapolate periodic patterns efficiently, while it is an unbounded function on both positive and negative sides, and it is not flexible to learn complex non-periodic patterns for general purposes. In this study, we propose to combine the strengths of the Swish and SNAKE functions and create a new activation function defined in Equation 1, which has a learnable parameter a to control the periodic component at different frequencies for complex periodic patterns learning, and another learnable parameter b to control the main shape of the function curve (similar to Swish) for nonlinear non-periodic pattern learning. We call the proposed new activation function as PArameterized Snake-Swish (PASS). Given the flexibility of the two learnable parameters, the proposed PASS function is highly adaptive to achieve efficient learning for both periodic and non-periodic components from the data. Given different machine learning tasks and datasets, the activation function has the potential to evolve to the studied task and data at suited curve shape and periodic frequency adaptively without tedious manual tuning. In particular, as shown in Figure 1, when a becomes larger, the frequency of the periodic component will be higher while the wave amplitude is smaller. When $a \to \infty$ the periodic component effect will vanish, the PASS function will become a Swish function; on the other hand, when $b \to 0$ the denominator will approach a constant, the PASS function will become a SNAKE function. Since the PASS function is continuous and differentiable for all parameter values, the learnable parameters can be optimized using gradient descent algorithms. The first-order derivative of the PASS function is as follows:

$$f'(x) = \frac{be^{-bx} \cdot \left(x + \frac{\sin^2(ax)}{a}\right)}{\left(1+e^{-bx}\right)^2} + \frac{2\cos(ax)\sin(ax)+1}{1+e^{-bx}} \tag{2}$$

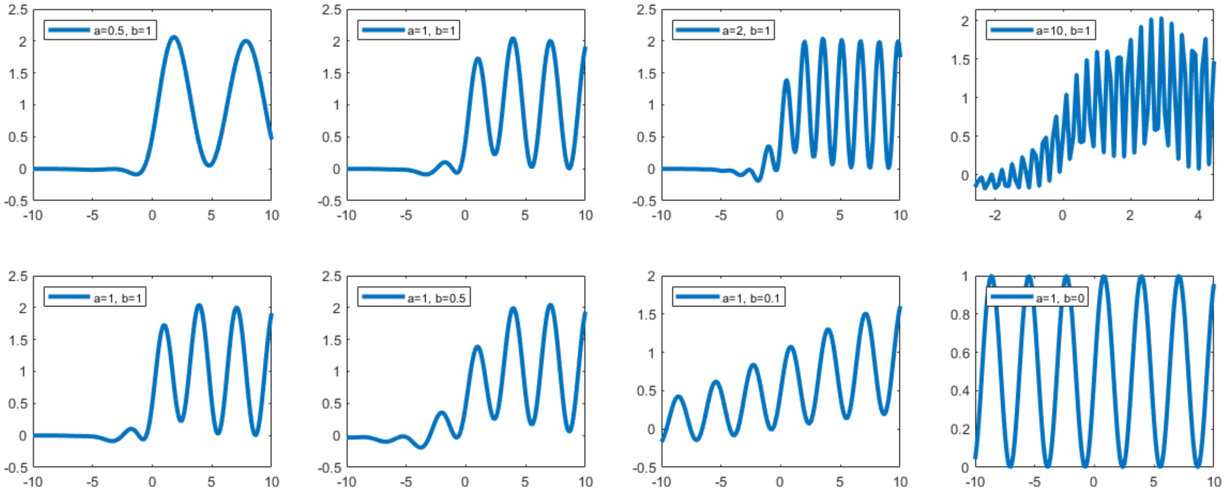
Figure 1: The curve shape of the PASS activation function at different values of a and b. When a becomes larger, the frequency of the periodic component will be higher while the wave amplitude is smaller. When $a \to \infty$, the PASS function will become a Swish function (top-right plot). When $b \to 0$, the PASS function will approach a SNAKE function (bottom-right plot).



The first order derivatives of the PASS function at different values of a and b are also shown in Figure 2. Past deep learning research indicates that activation functions with larger derivative values for a wider set of input values tend to perform better. For example, the use of the ReLU-like activation functions result in faster training compared to saturating sigmoidal type activation functions as they do not saturate for a wider range of inputs. While ReLu function has zero gradient on the negative side which leads to inefficient dead neurons in many cases. The derivatives of the PASS function are non-zero and do not saturate for a wider range of inputs. Noel et al. [22] showed that periodic derivative of periodic activation functions could improve gradient flow and alleviate the vanishing gradient problem in backpropagation. While more detailed theoretical analysis is necessary to validate the potential benefits of having oscillatory activation derivatives for deep neural network training.

In general, given a big data problem, it is hard to know beforehand that if the data contains periodic components. With many alternative activation functions, selecting the best activation function that suits a specific task and data is hard and time-consuming by experimenting the activation functions one by one. Every time the activation function changes, the neural networks need to be retrained from scratch. The proposed PASS activation function provides a novel solution for this problem to achieve adaptive learning of both periodic and non-periodic components from the data using a flexible function structure with two learnable parameters. The co-training of adaptive activation functions and the neural network weights showed faster convergence and superior performance in some recent studies [24, 25].

Figure 2: The first order derivatives of the PASS function at different values of a and b. The derivatives of the PASS function are non-zero and do not saturate for a wider range of inputs. According to Noel et al. [22], periodic derivatives could improve gradient flow and alleviate the vanishing gradient problem in backpropagation.



3 Experimental Validation

This paper investigates the properties of the PASS activation function and compares with most popular alternative functions as listed in Figure 3. Since the PASS function is equivalent to the SNAKE function when $b = 0$, we did not list SNAKE as a separate comparing function. Instead, we consider SNAKE is a special case of the PASS function with $b = 0$ in our experiments. Three groups of preliminary investigations have been conducted to answer the following research questions:

- If the PASS activation is effective to improve extrapolation of a learned neural network?
- If the PASS activation is effective to learn multivariate periodic data patterns (interpolation)?
- If the PASS activation is effective and practical for real data problems?

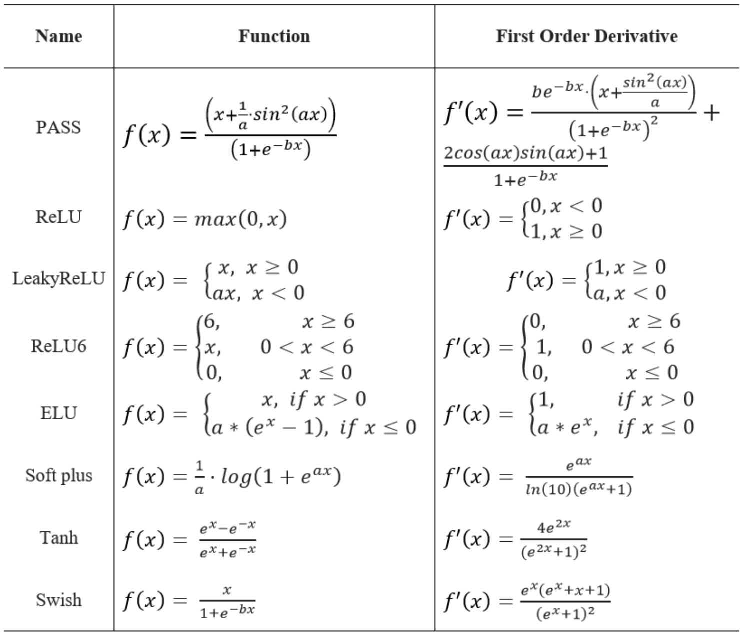
3.1 Extrapolation Experiments on Synthetic Data

A key difference between periodic and other non-periodic functions is the extrapolation property. If intrinsic periodic components of a complex system are learned, it can be highly useful to predict future evolution based on past observations. To evaluate the extrapolation property of neural networks with different activation functions, We set up a numerical experiment to generate a set of periodic signals to train a fully connected feedforward neural network with two hidden layers with 256 neurons in each layer. The periodic signals were generated using three analytical functions as follows:

$$f_1(x) = x + \sin(c_1 x)/c_1 + \varepsilon_1 \tag{3}$$

$$f_2(x) = x + \sin(c_1 x)/c_1 + \sin(c_2 x)/c_2 + \varepsilon_2 \tag{4}$$

Figure 3: A list of activation functions considered in this paper and their definitions.



Name	Function	First Order Derivative
PASS	$f(x) = \frac{\left(x+\frac{1}{a}\cdot sin^2(ax)\right)}{(1+e^{-bx})}$	$f'(x) = \frac{be^{-bx}.\left(x+\frac{sin^2(ax)}{a}\right)}{(1+e^{-bx})^2} + \frac{2cos(ax)sin(ax)+1}{1+e^{-bx}}$
ReLU	$f(x) = max(0,x)$	$f'(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}$
LeakyReLU	$f(x) = \begin{cases} x, \ x \geq 0 \\ ax, \ x < 0 \end{cases}$	$f'(x) = \begin{cases} 1, x \geq 0 \\ a, x < 0 \end{cases}$
ReLU6	$f(x) = \begin{cases} 6, & x \geq 6 \\ x, & 0 < x < 6 \\ 0, & x \leq 0 \end{cases}$	$f'(x) = \begin{cases} 0, & x \geq 6 \\ 1, & 0 < x < 6 \\ 0, & x \leq 0 \end{cases}$
ELU	$f(x) = \begin{cases} x, \ if \ x > 0 \\ a * (e^x - 1), \ if \ x \leq 0 \end{cases}$	$f'(x) = \begin{cases} 1, & if \ x > 0 \\ a * e^x, & if \ x \leq 0 \end{cases}$
Soft plus	$f(x) = \frac{1}{a} \cdot log(1 + e^{ax})$	$f'(x) = \frac{e^{ax}}{ln(10)(e^{ax}+1)}$
Tanh	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$f'(x) = \frac{4e^{2x}}{(e^{2x}+1)^2}$
Swish	$f(x) = \frac{x}{1+e^{-bx}}$	$f'(x) = \frac{e^x(e^x+x+1)}{(e^x+1)^2}$

$$f_3(x) = x + \sin(c_1 x)/c_1 + \sin(c_2 x)/c_2 + \sin(c_3 x)/c_3 + \varepsilon_3 \tag{5}$$

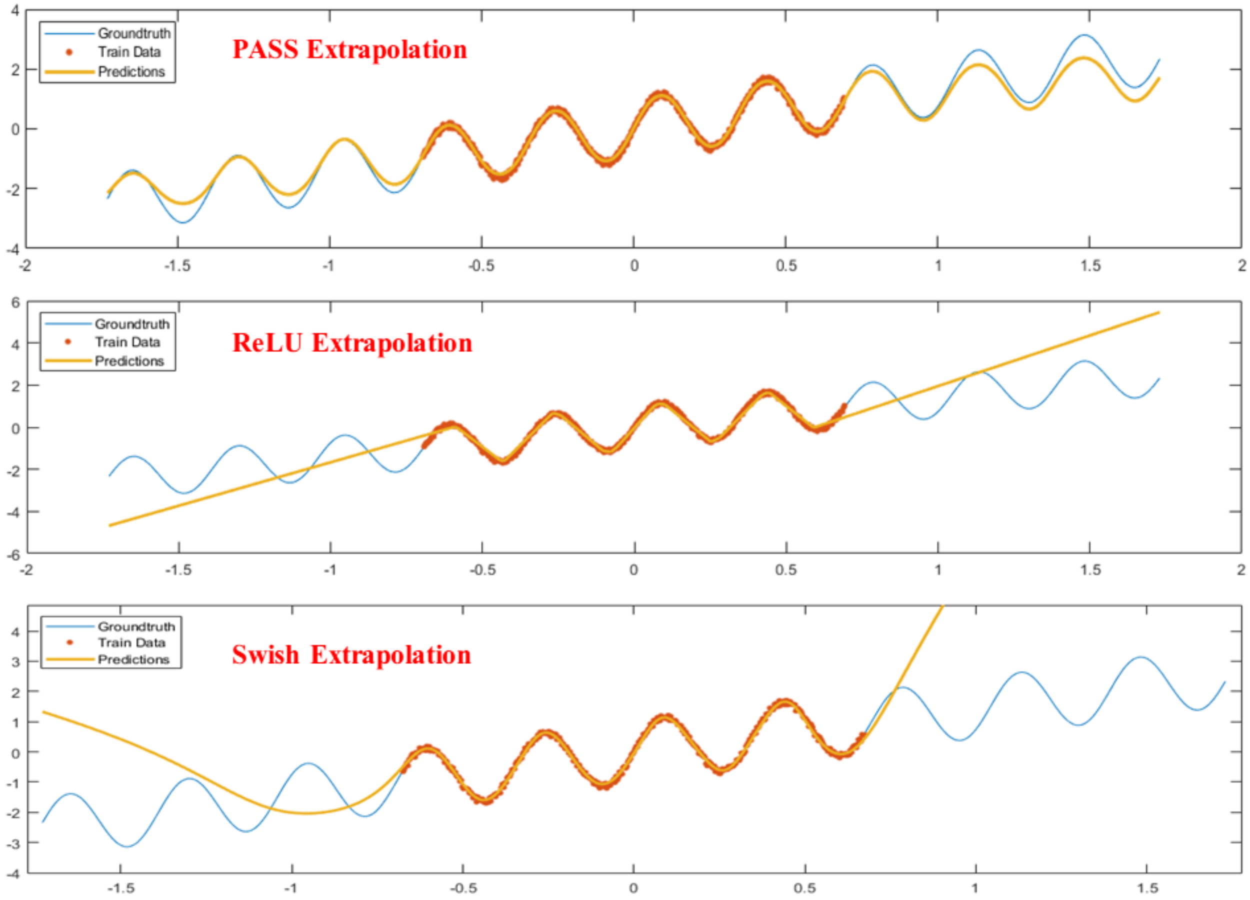
Each generated signal has a nonlinear component and one or multiple periodic components, which allow us to study the interpolation and extrapolation behaviour of neural networks with various activation functions. The generated data has 1000 pairs of independent and response variables, 40% of the data were used to train and validate the neural network model and the remaining 60% were used to test the extrapolation performance. Given random initialization, we run experiments five times for each activation function and calculated the mean RMSE and the standard deviation over the five runs.

The experimental results are summarized in Table 1. The PASS activation shows clearly superior extrapolation performance on all the three generated data with one, two, and three periodic components, respectively. The Figure 4 demonstrates the extrapolation performances of the PASS, ReLU, and Swish activation. One can observe that all three activation functions successfully learned the periodic pattern within the training data range. While the ReLu and Swish (as well as other comparing functions) did not capture the intrinsic periodic component of the data. With those functions, the learned neural networks is incapable to predict the response variable at all When data go beyond the training data range. On the other hand, the PASS activation function achieved the best extrapolation performance in all three synthetic data with vary number of different periodic components. The proposed PASS function successfully learned the intrinsic patterns from the data and extrapolated well to predict response variable values beyond the training data.

Table 1: Test Performance Summary on the Extrapolation Task using Synthetic Data with combined Periodic and linear trend patterns.

	Dataset 1		**Dataset 2**		**Dataset 3**	
Activation	**RMSE**	**STD**	**RMSE**	**STD**	**RMSE**	**STD**
PASS	**0.2695**	0.0484	**0.1958**	0.0344	**0.6268**	0.1527
ReLU	5.5431	0.8311	9.5053	3.2288	3.5542	0.6578
Leaky ReLU	6.0634	0.0823	4.3045	0.3369	2.8262	0.4586
ReLU6	2.4312	0.4529	2.5002	0.0803	3.1554	0.1485
ELU	2.2359	1.0535	5.7373	0.3265	7.1086	2.8620
Softplus	2.5795	0.0853	4.6454	1.1269	4.1948	0.4026
Tanh	2.3987	0.0917	2.5947	0.0426	3.8797	0.1343
Swish	1.4098	0.3171	6.6020	0.3342	10.5063	3.1849

Figure 4: Extrapolation performance comparison of PASS, ReLU, and Swish on the dataset 1 generated by the analytical function 3



In addition, we notice that the best performances of PASS in this set of experiments were achieved with $b = 0$, which means the SNAKE function best suited the data. This is not surprising that the SNAKE function has a linear and periodic component, which coincidentally matches the generated data patterns very well. To further evaluate the adaptability of the PASS activation, we introduce a nonlinear pattern x^2 to the analytical function of data generation. The modified analytical functions are:

$$f_4(x) = x^2 + \sin\left(c_1 x\right)/c_1 + \varepsilon_1 \tag{6}$$

$$f_5(x) = x^2 + \sin\left(c_1 x\right)/c_1 + \sin\left(c_2 x\right)/c_2 + \varepsilon_2 \tag{7}$$

$$f_6(x) = x^2 + \sin\left(c_1 x\right)/c_1 + \sin\left(c_2 x\right)/c_2 + \sin\left(c_3 x\right)/c_3 + \varepsilon_3 \tag{8}$$

Three new synthetic data have been generated named as Dataset 4, Dataset 5, Dataset 6, respectively. There are 1000 pairs of independent and response variables that were generated. As shown in Figure 4, 40% of the data were used to train and validate the neural network model and the remaining 60% were used to test the extrapolation performance.

The experimental results are summarized in Table 2. Again, the PASS activation achieved the best extrapolation RMSE in all three datasets. In this round of experiments, the best parameter setting is $a = 0.5, b = 0.1$ for dataset 4, $a = 0.5, b = 0.3$ for dataset 5, and $a = 0.5, b = 0.3$ for dataset 6. Since the main data pattern x^2 is bounded on one side, and learned PASS activation tends to approach the shape of Swish, which is bounded on one side. These experimental results confirmed the capability and adaptability of the PASS activation to learn both periodic and non-periodic patterns from the data, which is a highly desirable but lacked property in existing activation functions for deep neural networks.

Table 2: Test Performance Summary on the Extrapolation Task using Synthetic Data with both Periodic and Non-linear Non-Periodic patterns.

	Dataset 4		**Dataset 5**		**Dataset 6**	
Activation	**RMSE**	**STD**	**RMSE**	**STD**	**RMSE**	**STD**
PASS	**0.1322**	0.0452	**0.0945**	0.0147	**0.1360**	0.0306
ReLU	2.5807	0.1265	0.9269	0.0919	0.4734	0.0067
Leaky ReLU	1.9301	0.2848	0.7914	0.1747	0.2661	0.1154
ReLU6	0.5498	0.5343	0.2755	0.1571	0.4726	0.2001
ELU	0.5665	0.2449	1.3268	0.5350	0.5908	0.9654
Softplus	0.8441	0.0810	0.4483	0.0608	0.3737	0.0454
Tanh	0.3396	0.0466	0.3236	0.0876	0.4174	0.0185
Swish	0.9170	0.2241	1.2463	0.2455	0.3798	0.1339

3.2 Multivariate Data Modeling Experiments on Synthetic Data

We have demonstrated the superior extrapolation property of the PASS activation on a univariate forecasting problem. The second experiment is designed to evaluate the extrapolation capability on multivariate data. In particular, we formulated a synthetic long-sequence multivariate time series forecasting task, which is helpful for many real world problems but challenging for most machine learning models. The data consists of four time series variables with periodic components and one response variable sequence. The response variable was generated by a nonlinear mapping of the four input variables plus Gaussian random noises with a noise to signal ratio of 0.2. The governing analytical functions for the four input variables are summarized as follows:

$$X_1(x) = d_{11}\sin(e_{11}x) \times d_{12}\sin(e_{12}x) + \varepsilon \tag{9}$$

$$X_2(x) = d_{21}\sin(e_{21}x) \times d_{22}\sin(e_{22}x) + \varepsilon \tag{10}$$

$$X_3(x) = d_{31}\sin(e_{31}x) \times d_{32}\sin(e_{32}x) + \varepsilon \tag{11}$$

$$X_4(x) = d_{41}\cos(e_{41}x) + \varepsilon \tag{12}$$

Given the above four periodic variables, three response variables were created using the following non-linear mapping functions:

$$Y_1(x) = \sin(X_1 + X_1 + X_1 + X_1) + \varepsilon \tag{13}$$

$$Y_2(x) = \sin(X_1^3) + \varepsilon_1 \tag{14}$$

$$Y_3(x) = \sin(X_1 + 2X_2 + 3X_3 + 4X_4) + \varepsilon \tag{15}$$

The machine learning task is to predict long-sequence future values of the three response variables given a window of multivariate time series sequential data at different time stamps. In particular, we set the monitoring window length is 60 and the prediction horizon is 120. In other words, the model input is a multivariate time series epoch with a dimension of 4×60, and the model output is a sequential forecasted future values of a response variable in the next 120 steps. We name the datasets corresponding to the three response variables as MTS Dataset 1, 2, 3, respectively. There are 340 multivariate time series data samples generated for each dataset. The train/validation/test ratio were 0.4, 0.3, 0.3, respectively. Since the input are sequential data, we adopted the long short-term memory (LSTM) neural network structure for this machine learning task. The experimented neural network model has one LSTM layer with 128 hidden nodes to process the sequential input data, and a fully connected layer with 128 hidden nodes is used to connect the LSTM layer and the output layer. In the experiment, different activation functions were only placed to the fully connected layer, and the LSTM layer keeps the original design. We ran experiments five times for each activation function and recorded the mean RMSE and the standard deviation over five runs.

The experimental results are summarized in Table 3. We notice that most activation functions performed very well in this simulated multivariate time series forecasting problem. Even though the PASS activation still achieved the best performance on the three datasets, the performance advantages are very small compared to the extrapolation task. Since the train and test data are within the same value range, this machine learning task is equivalent to a regular interpolation task to learn a nonlinear mapping between multivariate input and output. Thus, all non-periodic activation functions have strong non-linear mapping approximation capabilities for this type of task, which is the most common machine learning task in deep learning literature. The slight performance improvements introduced by the PASS activation may be due to the superior capability of periodic functions to learn and represent fine details of the underlying signal patterns. As shown in Sitzman et al. [21], the sinusoidal function based SIREN activation was able to represent and reconstruct smooth fine details of a set of 3D objects, while ReLu activation cannot reconstruct such smooth surfaces due to its inherent piecewise linear design. The generated response variable coupled a sinusoidal mapping function of multivariate time series data with seven intrinsic periodic components. The slight performance improvements by PASS may confirm the findings in [21] that the periodic activation has improved capability to learn and reconstruct fine details of underlying data patterns.

Table 3: Test Performance Summary for the Long-Sequence Multivariate Time Series Forecasting Task Using Synthetic Data with known analytical functions that governing the data patterns.

	MTS Dataset 1		**MTS Dataset 2**		**MTS Dataset 3**	
Activation	**RMSE**	**STD**	**RMSE**	**STD**	**RMSE**	**STD**
PASS	**0.4022**	0.0044	**0.5734**	0.0164	**0.7009**	0.0163
ReLU	0.4215	0.0189	0.6092	0.0132	0.7090	0.0175
Leaky ReLU	0.4182	0.0107	0.5939	0.0357	0.6993	0.0162
ReLU6	0.4126	0.0101	0.6075	0.0200	0.7183	0.0172
ELU	0.4296	0.0096	0.5941	0.0132	0.7337	0.0196
Softplus	0.4369	0.0116	0.6136	0.0202	0.7155	0.0169
Tanh	0.4102	0.0166	0.6038	0.0193	0.7102	0.0120
Swish	0.4120	0.0117	0.5986	0.0371	0.7233	0.0040

3.3 Real Data Analysis

In the last experiment, we evaluated the activation functions for long sequence multivariate time series forecasting using three real benchmark datasets, including Beijing PM2.5 [26], air temperature data of Dallas Fort Worth (DFW) [27], and the Minamitorishima Atmospheric Temperature data [23, 28].

- Beijing PM2.5 data: it contains hourly weather data with 12 time series variables from Beijing Capital International Airport, and the response variable is the hourly PM2.5 concentration of US Embassy in Beijing, from 1:00 on Jan 2nd, 2010, to 24:00 on Dec 31st, 2014 (43800 hours in total). The machine learning task is to predict the PM 2.5 concentrations for the next 24 hours given the past 144 hours of 12 time series predictors. The train/validation/test ratios are 0.4, 0.3, 0.3, respectively.
- DFW Temperature Data: air temperature data of the Redbird airport station in Dallas Fort Worth was retrieved from National Oceanic and Atmospheric Administration between Jan 1st, 2017 and Dec 31st, 2020. The dataset contains one attribute (the temperature variable) and 35059 hours in total. The machine learning task is to predict future temperatures in next 24 hours using the temperature observed in the past 144 hours. We took 1083 random samples from the data and split the data using train/validation/test ratios of 0.4, 0.3, 0.3, respectively.
- Minamitorishima Atmospheric Temperature Data: daily average temperature data was collected at longitude 153.98 and latitude 24.28, in Minamitorishima, an island south of Tokyo for 3287 days starting from 2009. The data only has one time series attribute (atmospheric temperature). The machine learning task is to predict future temperatures in next 24 hours using the temperature observed in past 144 hours. We took 134 samples from the data, and the train/validation/test ratios are 0.4, 0.3, 0.3, respectively.

The real data tasks are all long-sequence time series forecasting problems. Since the input are sequential data, we still adopted the long short-term memory (LSTM) recurrent neural network structure in the real data experiment. The employed neural network model has one LSTM layer with 256 hidden nodes to process sequential input data, and a fully connected layer with 256 hidden nodes is used to connect the LSTM layer and the output layer. Different activation functions were only replaced in the fully connected layer. For each dataset and activation function, we repeated experiments five times with random initialization and recorded the mean RMSE and the standard deviation over five runs.

The experimental results are summarized in Table 4. One can observe that the PASS activation also achieved the best prediction performance on all the three real datasets. These results further confirm our hypothesis that the proposed periodic activation function may have improved capability to learn and reconstruct fine pattern details of the intrinsic dynamics of a complex underlying system.

Table 4: Test Performance Summary for Long-Sequence Time Series Forecasting Using Three Real Datasets

	BEIJING PM2.5		**DFW Temperature**		**Atmospheric Temperature**	
Activation	**RMSE**	**STD**	**RMSE**	**STD**	**RMSE**	**STD**
PASS	**83.8374**	0.4663	**3.8637**	0.0493	**1.3706**	0.1202
ReLU	86.3220	2.0064	4.0702	0.1256	1.9499	0.4472
Leaky ReLU	87.3705	2.0491	4.3345	0.2721	2.2422	0.8838
ReLU6	85.5464	1.9554	4.0716	0.0822	2.2154	0.6695
ELU	87.1368	1.1690	4.1768	0.2144	1.8018	0.0304
Softplus	84.7140	0.4760	4.0819	0.0952	2.2785	0.8557
Tanh	87.7655	2.0798	3.9419	0.0566	1.4192	0.3571
Swish	86.1602	1.5441	4.2966	0.4492	1.7301	0.0948

4 Conclusion

Periodic patterns exhibit the intrinsic nature of many complex systems in physics, science, and engineering. However, it has been shown that the current neural networks using ReLu or other popular alternatives are inefficient to learn intrinsic periodic patterns from data. On the other hand, periodic functions have been successfully used to model complex periodic systems. However, their role in deep neural networks remains largely ignored. Traditional periodic neural networks are long believed to be difficult to train and optimize. In this work, we conducted extensive research on periodic activation function and propose a novel elastic periodic activation function PASS, which combines the shape flexibility strength of the Swish activation and the notable monotonic-periodic property of the SNAKE activation using two learnable parameters. The co-training of the activation parameters and the neural network weights make it highly adaptive to adjust its morphology to fit the studied machine learning task and data. This adaptability is highly desirable to solve enormous real world data problems with both periodic and non-periodic components in the data. The experimental results confirmed our research hypothesis that periodic activation functions are indeed effective to learn periodic patterns from data. The three sets of experiments demonstrated that 1) the PASS activation is effective to improve extrapolation of a learned neural network; 2) the PASS activation is effective to learn and present fine details of multivariate periodic data patterns (interpolation); 3) the PASS activation is practical to solve real data problems with a highly flexible and learnable structure. It noted that the datasets used in this study all have strong periodic properties and are relatively small. More research investigation and detailed theoretical analysis are necessary to study the properties of periodic activation for general pattern recognition problems. This study shed a light to attract more attention to periodic functions for deep neural networks.

Acknowledgments

This work was supported in part by NSF Grant ECCS 1938895.

References

[1] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[2] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

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

[4] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). arxiv 2015. *arXiv preprint arXiv:1511.07289*, 2, 2016.

[6] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[7] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[8] Josep M Sopena, Enrique Romero, and Rene Alquezar. Neural networks with periodic and monotonic activation functions: a comparative study in classification problems. 1999.

[9] David B McCaughan. On the properties of periodic perceptrons. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 1, pages 188–193. IEEE, 1997.

[10] Kwok-wo Wong, Chi-sing Leung, and Sheng-jiang Chang. Handwritten digit recognition using multilayer feedforward neural networks with periodic and monotonic activation functions. In *Object recognition supported by user interaction for service robots*, volume 3, pages 106–109. IEEE, 2002.

[11] JM Sopena and R Alquezar. Improvement of learning in recurrent networks by substituting the sigmoid activation function. In *International Conference on Artificial Neural Networks*, pages 417–420. Springer, 1994.

[12] René Alquézar Mancho. *Symbolic and connectionist learning techniques for grammatical inference*. Universitat Politècnica de Catalunya, 1997.

[13] Renée Koplon and Eduardo D Sontag. Using fourier-neural recurrent networks to fit sequential input/output data. *Neurocomputing*, 15(3-4):225–248, 1997.

[14] M Hisham Choueiki, Clark A Mount-Campbell, and Stanley C Ahalt. Implementing a weighted least squares procedure in training a neural network to solve the short-term load forecasting problem. *IEEE Transactions on Power systems*, 12(4):1689–1694, 1997.

[15] E Rafajlowicz and M Pawlak. On function recovery by neural networks based on orthogonal expansions. 1997.

[16] Krzysztof Halawa. Fast and robust way of learning the fourier series neural networks on the basis of multidimensional discrete fourier transform. In *International Conference on Artificial Intelligence and Soft Computing*, pages 62–70. Springer, 2008.

[17] Abylay Zhumekenov, Malika Uteuliyeva, Olzhas Kabdolov, Rustem Takhanov, Zhenisbek Assylbekov, and Alejandro J Castro. Fourier neural networks: A comparative study. *arXiv preprint arXiv:1902.03011*, 2019.

[18] Emmanuel J Candès. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis*, 6(2):197–218, 1999.

[19] Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

[20] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Taming the waves: sine as activation function in deep neural networks. 2016.

[21] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[22] Mathew Mithra Noel, Advait Trivedi, Praneet Dutta, et al. Growing cosine unit: A novel oscillatory activation function that can speedup training and reduce parameters in convolutional neural networks. *arXiv preprint arXiv:2108.12943*, 2021.

[23] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594, 2020.

[24] Ameya D Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020.

[25] Ameya D Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks. *Proceedings of the Royal Society A*, 476(2239):20200334, 2020.

[26] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015.

[27] National Oceanic and Atmospheric Administration. Integrated surface dataset(global). `https://www.ncei.noaa.gov/access/search/data-search/global-hourly`. Accessed: 2022-01-10.

[28] Jülich open web interface (join). `https://join.fz-juelich.de/access/`.