Abstract

A very complex task in deep learning such as image classification must be solved with the help of neural networks and activation functions. The backpropagation algorithm advances backward from the output layer towards the input layer; the gradients often get smaller and smaller and approach zero which eventually leaves the weights of the initial or lower layers nearly unchanged, as a result, the gradient descent never converges to the optimum. We propose a two-factor non-saturating activation functions known as Bea-Mish for machine learning applications in deep neural networks. Our method uses two factors, beta (β) and alpha (α), to normalize the area below the boundary in the Mish activation function and we regard these elements as Bea. Bea-Mish provide a clear understanding of the behaviors and conditions governing this regularization term can lead to a more principled approach for constructing better performing activation functions. We evaluate Bea-Mish results against Mish and Swish activation functions in various models and data sets. Empirical results show that our approach (Bea-Mish) outperforms native Mish using SqueezeNet backbone with an average precision (AP50val) of 2.51% in CIFAR-10 and top-1 accuracy in ResNet-50 on ImageNet-1k. shows an improvement of 1.20%.

Keywords: Neural Network, Machine Learning Applications, α and β Regularizers, Activation Function.
computation values of sigmoid to -1 up to +1, however at the end it did not work perfect for deep neural networks to resolve vanishing gradients. Rectified Linear Unit (ReLU) showed better generalization and improved convergence speed compared to Sigmoid and tanh [2,3]. Various alternatives to ReLU have been proposed, these include LReLU, ReLU, ELU and SELU [4-7]. However, none of these variants has managed to replace ReLU due to inconsistent gains.

Figure 1 shows the process of applying an activation function in a layer of a neural network.

![Figure 1](image)

**Figure 1. The process of applying an activation function in a layer of a neural network can be mathematically realized as**

\[ y = a(z) = a\left(\sum_i w_i x_i + b\right) \]  

**where y is the output of the activation function a(z)**

Recently, a set of non-monotonic activation functions, such as the Gaussian Error Linear Unit (GELU)[8] use \( \Phi(x) \), known as the standard Gaussian cumulative distribution function. The Google brain team proposed an activation function called Swish that was discovered in a nonlinear function space by Neural Architecture Search (NAS) which use a controlled search agent [9,10]. HardSwish is a kind of activation function based on Swish, but it replaces the computationally expensive sigmoid with a piecewise linear analog [11]. Rectified Exponential Unit (REU) and Mish have been proposed with similar mathematical formulas and show robust and improved results in various models and data sets [12,13]. The main difference between monotonic and non-monotonic activation functions is their non-monotonic properties in the negative quadrant. ReLU is described by equation (1), which is not constrained on the positive axis of the input and sparse on the negative part [2].

\[ f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \]  

(1)

From equation (1) \( x \) is the input and \( f(x) \) is the output (the activation value), ReLU uses identification mapping in the positive quadrant and output 0 in the negative part, identification mapping can alleviate the problem of disappearing gradients. GELU nonlinearity weights inputs by their value, rather than gates inputs by their sign as in ReLUs (\( x > 0 \)) [8].

Mish, a novel self-regularized non-monotonic activation function inspired by the self-gating property of Swish. Mish is mathematically defined as:

\[ f(x) = x \cdot \tanh(\text{softplus}(x)) \]  

[13]. However, self-regulation did
not solve gradient convergence as an optimal problem in deep neural networks.

2. Theory

In this work, we propose to Bea-Mish a two-factor regularized non-monotonic activation function for machine learning applications in deep neural networks to solve gradient descent converges to the optimum challenge. Our method uses two factors, beta (β) and alpha (α), to regularize the region below the boundary in the Mish activation function. In this method, together these factors are called Bea regularizer. Bea-Mish uses a common Mish mathematical expression, but we introduce additional definitions:

\[ \text{softplus}(r) = \ln(1 + e^r) \text{ where } r = e^{\frac{ax}{\sqrt{\beta+x^2}}} \]
\[ f(x) = x \cdot \tanh \left( \ln \left( 1 + e^{\frac{ax}{\sqrt{\beta+x^2}}} \right) \right) \]

First derivative of the equation (3) is described by the equation (6) below.

\[ f'(x) = \frac{\alpha x \beta x - \beta \sim^2 e^{\frac{ax}{\sqrt{\beta+x^2}}}}{\cosh^2(\text{softplus}(x))} + (x^2 + \beta)^2 (1 + e^{\frac{ax}{\sqrt{\beta+x^2}}}) \tanh(\text{softplus}(x)) \]

\[ (x^2 + \beta)^2 (1 + e^{\frac{ax}{\sqrt{\beta+x^2}}}) \]

The value of α is determined by the value of β using the ratio α:β=1:5, and the value of β ranges from 1 to 200. Bea-Mish avoids saturation, which typically slows down training quickly due to near-zero gradients [14].

3. Experiment Settings

We compared Bea-Mish results against Mish and Swish activation functions in various models and data sets. Our results show especially in computer vision tasks such as image classification, Bea-Mish consistently exceeded the highest performing networks. Multiple runs were used to observe the statistical significance of the results. In this section we analyze the statistical results and the ablation study of β and α.

4. Results and Discussion

4.1 CIFAR-10 Dataset on Various Baseline Activation Functions

Bea-Mish activation function compared to baseline activation functions, we compared the mean test accuracy, mean test loss, and standard deviation of test accuracy for CIFAR-10 image classification task using SqueezeNet following the parameter settings in [13, 15, 16]. Table 1 shows that Bea-Mish outperforms the other activation functions with the highest mean accuracy (μ acc), the second lowest mean loss(μ loss), and the third lowest accuracy standard deviation (σ acc).

<table>
<thead>
<tr>
<th>Activation Function</th>
<th>μ acc</th>
<th>μ loss</th>
<th>σ acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bea-Mish (Ours)</td>
<td>90.97%</td>
<td>4.05%</td>
<td>0.33</td>
</tr>
<tr>
<td>Mish[13]</td>
<td>87.48%</td>
<td>4.13%</td>
<td>0.40</td>
</tr>
<tr>
<td>Swish[9]</td>
<td>82.32%</td>
<td>4.22%</td>
<td>0.41</td>
</tr>
</tbody>
</table>
4.2 CIFAR-10 Dataset on Various Standard Neural Network Architectures

We compare the performance of various criterion activation functions for image classification tasks in the CIFAR-10 dataset using different standard neural network architectures [17]. We alternated the activation function and keep all other network parameters and training parameter constant. We evaluate the performance of Bea-Mish compared to Mish and Swish using empirical results such as in native Mish for various standard network architectures [9,13]. Table 2 shows that Bea-Mish activation consistently outperforms Mish, and Swish activation in all standard architectures used in the experiments, providing performance gains up to 1% compared to native Mish network architectures.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Bea-Mish (ours)</th>
<th>Mish</th>
<th>Swish</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-20</td>
<td>92.69%</td>
<td>92.02%</td>
<td>91.61%</td>
</tr>
<tr>
<td>WRN-10-2</td>
<td>86.97%</td>
<td>86.83%</td>
<td>86.83%</td>
</tr>
<tr>
<td>SimpleNet</td>
<td>92.31%</td>
<td>91.70%</td>
<td>91.44%</td>
</tr>
<tr>
<td>Xception Net</td>
<td>88.94%</td>
<td>88.73%</td>
<td>88.56%</td>
</tr>
</tbody>
</table>

4.3 ImageNet-1k Dataset on Various Standard Neural Network Architectures

We evaluate the performance of Bea-Mish against native Mish and Swish for ImageNet-2012 dataset classification task. ImageNet is considered one of the most challenging and important classification tasks in the domain of computer vision [18]. ImageNet consists of 1.28 million training images spread across 1,000 classes. To evaluate the performance of the trained network, we use a validation set of 50,000 images. We set the batch size, mini-batch size, initial learning rate, momentum, and weight decay to 128, 32, 0.01, 0.9, and 5e-4 respectively with total number of 6 million training steps. Table 3 results are more consistent with Bea-Mish and generally guaranteed performance improvement in all neural network architecture for ImageNet classification.

<table>
<thead>
<tr>
<th>Model</th>
<th>Augmentation</th>
<th>Bea-Mish (ours)</th>
<th>Mish</th>
<th>Swish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>ResNet-18[15]</td>
<td>No</td>
<td>72.40%</td>
<td>90.47%</td>
<td>71.20%</td>
</tr>
<tr>
<td>ResNet-50[15]</td>
<td>No</td>
<td>76.84%</td>
<td>93.01%</td>
<td>76.10%</td>
</tr>
<tr>
<td>SpineNet-49[19]</td>
<td>Yes</td>
<td>73.62%</td>
<td>91.02%</td>
<td>71.40%</td>
</tr>
<tr>
<td>CSP-ResNet-50[20]</td>
<td>Yes</td>
<td>78.45%</td>
<td>95.08%</td>
<td>78.10%</td>
</tr>
</tbody>
</table>

4.4 Ablation Study of $\alpha$ and $\beta$ on CIFAR-10

We observe the effect of increasing the size of $\alpha$ and $\beta$ by considering the ratio $\alpha: \beta=1:5$ on CIFAR-10 dataset using ResNet framework. The network was optimized using SGD with a batch size of 128 [21]. For a fair comparison, the same learning rate was maintained for the different size of $\alpha$ and $\beta$. Table 4 show that the value of $\beta$, from 1 to 200, provides high test accuracy with increasing of values $\alpha$ and $\beta$ in ResNet framework.
5. Conclusion

In this work, we propose to Bea-Mish a two-factor regularized non-monotonic activation function for machine learning applications in deep neural networks to solve gradient descent converges to the optimum challenge. We propose the Bea-Mish activation function with two-element regularizer of the Mish activation function. Bea-Mish demonstrates a clear understanding of the behaviors and conditions governing this regularization term may motivate a more principled approach to constructing better performing activation functions. Bea-Mish shares many properties with Mish, such as infinite positive domain, limited negative domain, non-monotonic form, and smooth out-growth. Empirical results show that Bea-Mish outperforms native Mish in all aspects of different datasets and neural architecture models. The future study includes evaluating the performance of the Bea-Mish activation function another state-of-the-art model on various tasks in the domain of computer vision such as object detection [22].

Acknowledgement

This work was supported by Dongseo University, “Dongseo Cluster Project” Research Fund of 2021 (DSU-20210001).

References


DOI: http://dx.doi.org/10.7236/IJIBC.2021.13.2.156