A REPLICATION STUDY OF SWISH ACTIVATION FUNCTION

Hussein Hasan *hussein.hasan@bahcesehir.edu.tr*
Artificial Intelligence Engineering
Bahcesehir University

Majd Owda *majd.owda@bahcesehir.edu.tr*
Artificial Intelligence Engineering
Bahcesehir University

1 Reproducibility Summary

1.1 Scope of Reproducibility

In this paper, we examine the reproducibility of the paper "Searching for Activation Functions"[1]. The paper in the study proposes that novel activation functions acquired by search methods could perform with a higher accuracy than conventional and popular ones. We, therefore, aim to validate whether such a claim is correct.

1.2 Methodology

We fine-tuned a community-written code [2] that resembles the original code of the paper. And performed some of the paper's experiments on the same datasets they were performed on.

1.3 Results

The main claim of the paper is verified; the activation function SWISH (which was acquired by searched method) did, in fact, perform with higher accuracy than traditional activation functions (ReLU and LeakyReLU) when were used on the same models and same datasets.

2 Introduction

In this work, we reproduce the paper "Searching for Activation Functions". Specifically, we compare the performance of the novel activation function SWISH to two of the most used conventional activation functions; ReLU and Leaky ReLU. By doing so, we can determine whether novel activation functions that are acquired by search method can outperform traditional and commonly used ones, as the paper claims.

3 Methodology

3.1 Datasets

3.1.1 CIFAR-10

The CIFAR-10 [3] is a dataset that consists of 60000 tiny images (32 x 32) colored images that are divided into 10 classes. The data set is commonly used in machine learning and computer vision algorithms. The data set is open source and provided by the Canadian Institute For Advanced Research (Thus, the name CIFAR).

3.1.2 CIFAR-100

3.1.2 – CIFAR-100 The CIFAR-100 [4] is a dataset similar to CIFAR-10, except that the images in it are divided into 100 class instead of 10.

3.2 Models

Two models were used in our experiment: ResNet and Wide ResNet.

3.2.1 ResNet

ResNet (short for Residual Network) is a deep learning model that is commonly used, especially in tasks of image recognition (which is within the scope of our study).

3.2.2 Wide ResNet

Wide ResNet is a later variation of the ResNet model. It was introduced to improve on the original model.

3.3 Experiments

Our experiments included running the two models (ResNet and Wide ResNet) on the two data sets (CIFAR-10 and CIFAR-100). Each model was implemented on each dataset three times; each time using one of the three activation functions: SWISH, ReLU, and Leaky ReLU.

By doing so, we acquired the accuracy of the implementation of every model when run using a different activation function. Thus, providing us with enough information to compare the performance of different activation functions, which ultimately allowed us to determine which ones perform better.

4 Results

The results of our experiment include the accuracy of each model when performed on a certain dataset with a certain activation function, which are shown in table 1.

Table 1: Comparison of Performance with Different Activation Functions on ResNet model

Activation Function	**CIFAR-10**	**CIFAR-100**
ReLU	94.61	76.05
LeakyReLU	94.23	76.01
Swish	94.74	75.98

Table 2: Comparison of Performance with Different Activation Functions on Wide ResNet model

Activation Function	**CIFAR-10**	**CIFAR-100**
ReLU	94.41	76.34
LeakyReLU	93.49	74.28
Swish	94.75	75.39

Furthermore, the results are illustrated through the figures below.

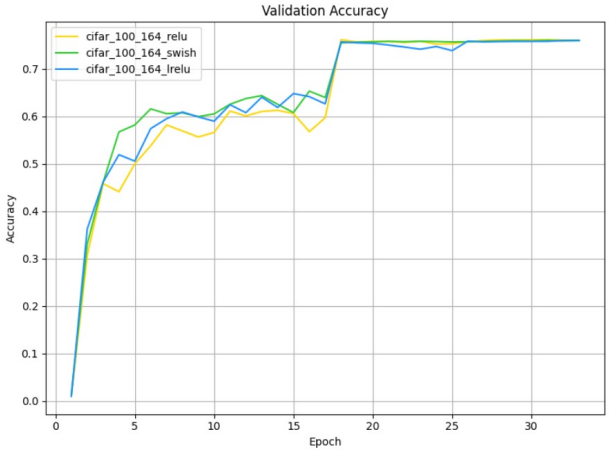

Figure 1: Validation Accuracy of ResNet model

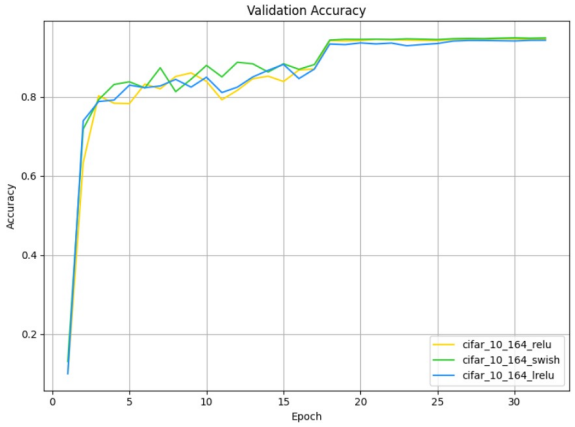

Figure 2: Validation Accuracy of ResNet model

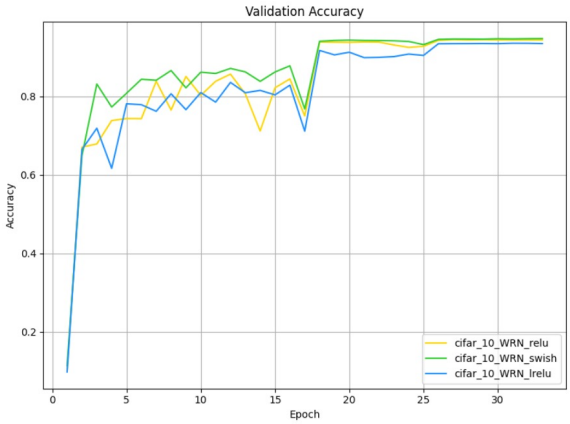

Figure 3: Validation Accuracy of Wide ResNet model

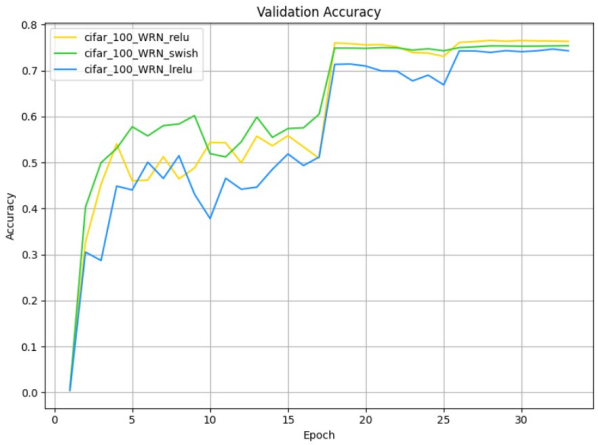

Figure 4: Validation Accuracy of Wide ResNet model

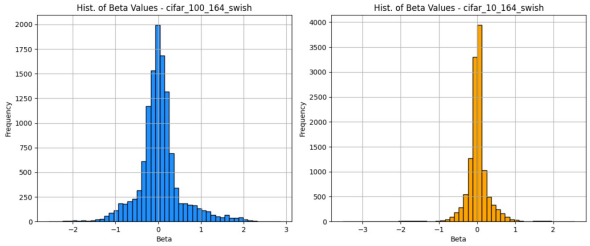

Figure 5: Hirtogram of SWISH Beta Values on ResNet model

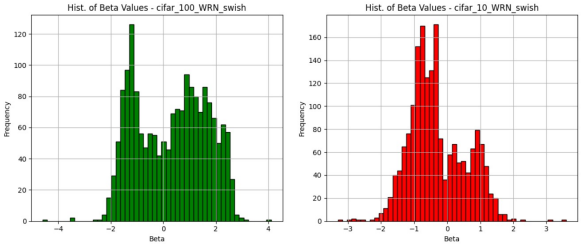

Figure 6: Hirtogram of SWISH Beta Values on Wide ResNet model

5 Conclusion

The results show that the models performed with a higher accuracy when implemented with the activation function SWISH. The comparison between SWISH and the other activation functions is illustrated in table 2.

Table 3: Comparison of performance of different activation functions with relevence to SWISH*

Results	**ReLU**	**LeakyReLU**
Under-performed	1	0
outperformed	2	3
app. equal	1	1

This confirms the claim that was proposed in the original paper; a novel activation function (in the context of this analysis, SWISH) performed better than the conventional and wide-spread ReLU and LeakyReLU. The SWISH activation functions tends to work better with deeper neural networks.

6 References

1 - https://arxiv.org/abs/1710.05941
2 - https://github.com/sjmikler/resnets/tree/main
3 - https://paperswithcode.com/dataset/cifar-10
4 - https://paperswithcode.com/dataset/cifar-100

*The threshold for performance evaluation is 0.1 .