Smooth Maximum Unit: Smooth Activation Function for Deep Networks using Smoothing Maximum Technique

Koushik Biswas[1], Sandeep Kumar[1,2], Shilpak Banerjee[3], Ashish Kumar Pandey[1]

[1]IIIT Delhi   [2] Shaheed Bhagat Singh College, University of Delhi   [3]IIT Tirupati

{koushikb,ashish.pandey}@iiitd.ac.in, sandeep_kumar@sbs.du.ac.in, shilpak@iittp.ac.in

Supplementary Material

Figure 1, 2 represents approximation of ReLU ($\alpha = 0$), Leaky ReLU ($\alpha = 0.25$) by SMU-1 for different values of μ and as $\mu \to 0$, SMU-1 overlap ReLU or Leaky ReLU depending on the value of α. Figure 3 represents the derivatives of SMU-1 for $\alpha = 0.25$ and different valus of μ.

There are many known smooth approximation to the $|x|$ function like $x \operatorname{erf}(\mu x)$, $\sqrt{x^2 + \mu^2}$, $\frac{x^2}{\sqrt{x^2+\mu^2}}$ etc. As $\mu \to 0$, $\frac{x^2}{\sqrt{x^2+\mu^2}}$ gives smooth approximation of $|x|$ from below. We give a plot of well known approximation to $|x|$ in Figure 4.
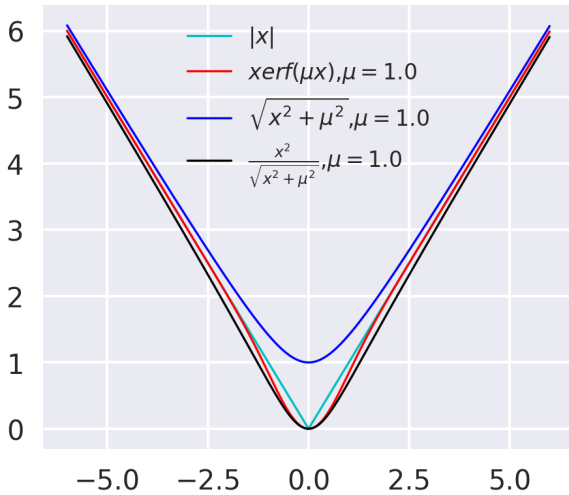

Figure 4. Smooth approximation of $|x|$

Replace $x_1 = x$ and $x_2 = \alpha x$ in equation (3), we have a smooth approximation of Leaky ReLU or Parametric ReLU depending on whether α is a hyperparameter or a learnable parameter. We call it SMU-1 and is defined as

$$f_2(x, \alpha x; \mu) = \frac{(1+\alpha)x + \sqrt{(1-\alpha)^2 x^2 + \mu^2}}{2}$$

and the corresponding derivative with respect to input variable x is

$$\frac{d}{dx} f_2(x, \alpha x; \mu) = \frac{(1+\alpha) + \frac{(1-\alpha)^2 x}{\sqrt{(1-\alpha)^2 x^2 + \mu^2}}}{2}$$

.

A. More results on MNIST, Fashion MNIST, and SVHN datasets

We report more results on MNIST, Fashion MNIST, and SVHN datasets with SMU, SMU-1 and baseline activations with LeNet, AlexNet, and a custom-designed model in Table 4, Table 6, and Table 8 respectively. Our custom homogeneous convolutional neural network has maxpooling layers(thrice), channel depths of size 128 (twice), 64 (thrice), 32 (twice), and a dense layer of size 128. Batchnormalization is applied before the activation function layer. We use 3×3 kernels in CNN layers and 2×2 kernels in max-pooling layers.

B. More results on CIFAR10 and CIFAR100

We report more results on CIFAR100 and CIFAR10 datasets with AlexNet, ShuffleNet V1, GoogleNEt, Inception V3, DenseNet-121, WideResNet 28-10, squeezeNet, VGG 16, and LeNet architectures in Table 1 and Table 7 respectively. An extension to the Table (2) on CIFAR100 dataset with baseline activations are reported in Table 2 & Table 3 and an extension to Table (3) on CIFAR10 dataset with baseline activations are reported in Table 5 & Table 9.

B.1. Mixup

We run experiments with Mixup augmentation method on CIFAR100 dataset with ShuffleNet V2 (2.0x), MobileNet V2, AlexNet, Xception, ResNet-50 models and results are reported in Table 10. The results are reported with the same experimental setup is reported in the CIFAR section. From table 10, it is clear that the proposed activations perform better than the baseline activations in all the models.
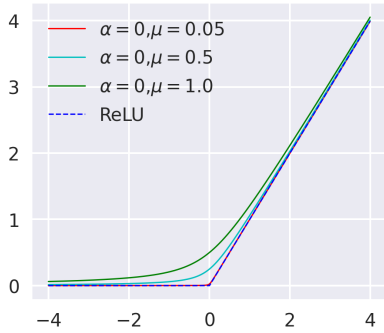
Figure 1. Approximation of ReLU using SMU-1 ($\alpha = 0$) for different values of μ. As $\mu \to 0$, SMU-1 smoothly approximate ReLU

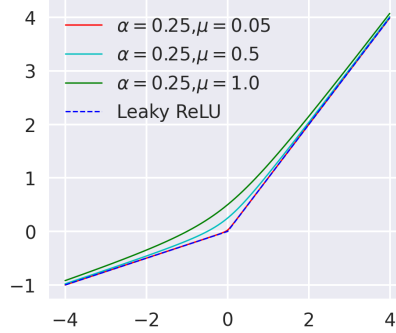Figure 2. Approximation of Leaky ReLU ($\alpha = 0.25$) using SMU-1 for different values of μ. As $\mu \to 0$, SMU-1 smoothly approximate Leaky ReLU
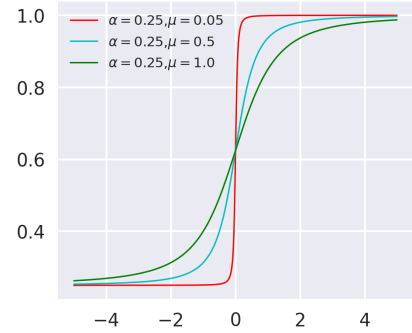

Figure 3. First order derivatives of SMU-1 for $\alpha = 0.25$ and different values of μ.

Activation Function	Alex Net	Shuffle Net V1	Google Net	Inception V3	Dense Net 121	WideRes Net 28-10	Squeeze Net	VGG 16	LeNet
SMU	**61.27** ±0.21	**69.15** ±0.22	**74.61** ±0.25	**77.52** ±0.24	**78.57** ±0.23	**78.89** ±0.24	**68.51** ±0.24	**73.26** ±0.22	47.20 ±0.25
SMU-1	**60.98** ±0.23	**68.71** ±0.22	**74.29** ±0.24	**76.88** ±0.23	**78.01** ±0.24	**78.30** ±0.23	**68.07** ±0.25	**72.79** ±0.23	47.03 ±0.24
ReLU	54.89 ±0.28	65.79 ±0.29	72.52 ±0.30	74.12 ±0.27	75.81 ±0.28	76.45 ±0.26	66.22 ±0.29	71.87 ±0.30	45.54 ±0.28
Leaky ReLU	55.26 ±0.27	65.99 ±0.30	72.42 ±0.31	74.49 ±0.28	75.93 ±0.27	76.61 ±0.27	66.15 ±0.27	71.92 ±0.29	45.77 ±0.29
ReLU6	55.89 ±0.26	66.19 ±0.28	72.47 ±0.28	74.51 ±0.25	75.98 ±0.28	76.71 ±0.27	66.39 ±0.26	71.95 ±0.28	45.79 ±0.27
PReLU	55.47 ±0.29	65.87 ±0.32	72.69 ±0.29	74.39 ±0.30	76.06 ±0.29	76.71 ±0.27	66.35 ±0.28	71.96 ±0.32	45.59 ±0.31
ELU	55.91 ±0.26	65.72 ±0.28	72.92 ±0.28	74.65 ±0.26	75.72 ±0.25	76.25 ±0.26	66.39 ±0.28	71.79 ±0.30	46.02 ±0.28
Softplus	54.99 ±0.39	65.11 ±0.38	71.81 ±0.38	74.25 ±0.35	75.19 ±0.35	75.42 ±0.37	65.73 ±0.35	70.92 ±0.32	44.12 ±0.39
GELU	57.32 ±0.26	67.22 ±0.25	73.16 ±0.26	75.66 ±0.26	76.68 ±0.26	77.07 ±0.25	66.99 ±0.29	71.88 ±0.27	47.27 ±0.25
Swish	57.55 ±0.27	67.01 ±0.26	73.32 ±0.26	75.47 ±0.28	76.51 ±0.29	77.35 ±0.24	66.56 ±0.27	71.94 ±0.28	47.34 ±0.23
PAU	57.35 ±0.29	67.45 ±0.28	73.68 ±0.27	75.85 ±0.31	76.72 ±0.28	77.02 ±0.26	66.89 ±0.24	71.79 ±0.25	47.30 ±0.29
Mish	58.22 ±0.23	67.85 ±0.24	73.97 ±0.24	76.29 ±0.25	77.25 ±0.24	77.45 ±0.23	67.35 ±0.23	72.45 ±0.22	**47.42** ±0.27

Table 1. Comparison between SMU, SMU-1 activations and other baseline activations on CIFAR100 dataset for image classification problem. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

C. ImageNet-1k

We also evaluate the performance of proposed and baseline activation functions on bookmarking the ImageNet-1k dataset. The dataset consists of 1,281,167 training images and 50,000 validation images with 1000 classes. The images have a resolution of 224×224. Results are reported on Table 11 with ShuffleNet V2 (1.0x) and ResNet-50 models. We use four NVIDIA V100 GPUs with 32GB RAM each to run these experiments. We trained the models up to 600k iterations with a batch size of 256 and SGD optimizer, 0.9 momentum, $5e^{-4}$ weight decay rate.

Activation Function	SF V2 0.5x	SF V2 1.0x	SF V2 1.5x	SF V2 2.0x	SeNet 18	SeNet 34	SeNet 50	Res-Next	Xcep-tion	EffitientNet B0
Leaky ReLU	62.25 ±0.33	65.39 ±0.34	67.39 ±0.29	67.79 ±0.29	74.51 ±0.23	75.14 ±0.24	76.23 ±0.22	74.58 ±0.23	71.01 ±0.26	76.81 ±0.28
ReLU6	62.39 ±0.30	65.71 ±0.29	67.65 ±0.29	68.10 ±0.26	74.69 ±0.20	75.34 ±0.22	76.61 ±0.21	74.65 ±0.24	71.39 ±0.23	76.67 ±0.24
PReLU	62.01 ±0.34	65.10 ±0.33	67.37 ±0.34	67.99 ±0.29	74.39 ±0.27	75.19 ±0.23	76.39 ±0.24	74.48 ±0.26	71.19 ±0.28	76.68 ±0.30
ELU	62.61 ±0.30	65.60 ±0.28	67.71 ±0.27	67.91 ±0.31	74.79 ±0.22	75.10 ±0.21	76.49 ±0.23	74.69 ±0.22	71.45 ±0.24	76.71 ±0.30
Softplus	61.87 ±0.35	64.45 ±0.37	67.19 ±0.32	68.79 ±0.30	74.36 ±0.34	74.78 ±0.34	75.22 ±0.37	74.31 ±0.35	71.30 ±0.39	76.56 ±0.36
GELU	64.40 ±0.26	66.79 ±0.23	69.79 ±0.29	70.10 ±0.28	74.82 ±0.19	76.20 ±0.21	77.20 ±0.21	75.17 ±0.24	72.07 ±0.22	77.31 ±0.22
Swish	63.79 ±0.25	66.99 ±0.25	69.59 ±0.27	70.29 ±0.24	74.62 ±0.19	75.77 ±0.22	76.89 ±0.24	75.17 ±0.25	72.19 ±0.21	77.17 ±0.20
PAU	64.10 ±0.26	66.77 ±0.27	69.52 ±0.25	70.54 ±0.26	74.89 ±0.20	75.92 ±0.24	77.10 ±0.23	75.66 ±0.26	72.62 ±0.27	77.41 ±0.24
Mish	64.91 ±0.24	67.78 ±0.24	70.44 ±0.25	71.49 ±0.22	75.32 ±0.19	76.52 ±0.23	77.69 ±0.23	76.20 ±0.24	73.49 ±0.22	78.15 ±0.22

Table 2. This is an extension to the Table-2. We report Top-1 test accuracy (in %) on CIFAR100 dataset for baseline functions for the mean of 15 different runs. mean±std is reported in the table. SF V2 stands for ShuffleNet v2.

Activation Function	ResNet 18	ResNet 34	ResNet 50	PreAct ResNet 18	PreAct ResNet 34	PreAct ResNet 50	MobileNet V1	MobileNet V2
Leaky ReLU	73.12 ±0.25	73.41 ±0.28	74.19 ±0.25	73.29 ±0.23	73.33 ±0.24	74.02 ±0.24	71.22 ±0.26	74.03 ±0.25
ReLU6	73.35 ±0.24	73.59 ±0.26	74.23 ±0.23	73.47 ±0.23	73.56 ±0.22	74.46 ±0.23	71.56 ±0.24	74.51 ±0.23
PReLU	73.02 ±0.27	73.52 ±0.29	74.32 ±0.28	73.21 ±0.25	73.45 ±0.26	74.29 ±0.25	71.41 ±0.29	74.45 ±0.30
ELU	73.42 ±0.24	73.68 ±0.26	74.48 ±0.24	73.32 ±0.20	73.49 ±0.25	74.44 ±0.25	71.32 ±0.24	74.22 ±0.23
Softplus	72.86 ±0.39	73.20 ±0.38	74.10 ±0.40	72.99 ±0.41	73.10 ±0.35	73.96 ±0.38	71.04 ±0.38	74.27 ±0.36
GELU	73.89 ±0.22	74.10 ±0.25	75.59 ±0.22	74.98 ±0.23	74.41 ±0.21	74.92 ±0.22	71.74 ±0.22	75.01 ±0.23
Swish	73.68 ±0.23	74.17 ±0.24	75.35 ±0.24	75.12 ±0.25	74.81 ±0.22	75.10 ±0.21	71.92 ±0.21	75.15 ±0.22
PAU	74.10 ±0.20	74.44 ±0.22	75.87 ±0.21	74.92 ±0.21	74.72 ±0.19	75.68 ±0.17	71.83 ±0.22	75.19 ±0.19
Mish	74.59 ±0.20	74.70 ±0.21	76.22 ±0.22	75.11 ±0.23	75.34 ±0.21	76.98 ±0.19	72.24 ±0.20	75.45 ±0.20

Table 3. This is an extension to the Table-2. We report Top-1 test accuracy (in %) on CIFAR100 dataset for baseline functions for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	MNIST	Fashion MNIST	SVHN
ReLU	99.16 ± 0.11	91.42 ± 0.18	92.10 ± 0.21
Leaky ReLU	99.12 ± 0.12	91.43 ± 0.22	92.27 ± 0.20
ReLU6	99.21 ± 0.10	91.47 ± 0.19	92.28 ± 0.16
PReLU	99.23 ± 0.10	91.40 ± 0.20	92.09 ± 0.20
ELU	99.30 ± 0.10	91.41 ± 0.21	92.28 ± 0.19
Softplus	99.01 ± 0.19	91.11 ± 0.25	91.92 ± 0.26
GELU	99.33 ± 0.08	91.60 ± 0.13	92.47 ± 0.17
Swish	99.29 ± 0.09	91.66 ± 0.15	92.35 ± 0.20
PAU	99.37 ± 0.10	91.56 ± 0.14	92.37 ± 0.21
Mish	99.36 ± 0.06	$\mathbf{91.68} \pm 0.13$	92.41 ± 0.17
SMU	$\mathbf{99.47} \pm 0.04$	91.58 ± 0.16	$\mathbf{92.79} \pm 0.16$
SMU-1	$\mathbf{99.41} \pm 0.05$	91.51 ± 0.14	$\mathbf{92.66} \pm 0.17$

Table 4. Comparison between SMU, SMU-1 activations and other baseline activations on MNIST, Fashion MNIST, and SVHN datasets for image classification problem on LeNet architecture. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	ResNet 18	ResNet 34	ResNet 50	PreAct ResNet 18	PreAct ResNet 34	PreAct ResNet 50	MobileNet V1	MobileNet V2
Leaky ReLU	94.00 ±0.25	94.18 ±0.24	94.29 ±0.24	93.51 ±0.20	94.29 ±0.22	94.32 ±0.22	92.54 ±0.21	94.10 ±0.19
ReLU6	94.19 ±0.26	94.20 ±0.25	94.26 ±0.27	93.69 ±0.21	94.19 ±0.25	94.52 ±0.23	92.69 ±0.20	94.21 ±0.20
PReLU	94.22 ±0.28	94.29 ±0.29	94.17 ±0.27	93.58 ±0.23	94.31 ±0.25	94.48 ±0.28	92.50 ±0.20	94.29 ±0.23
ELU	94.15 ±0.23	94.24 ±0.22	94.20 ±0.24	93.59 ±0.22	94.42 ±0.24	94.45 ±0.20	92.69 ±0.21	94.04 ±0.19
Softplus	93.82 ±0.29	93.99 ±0.31	93.77 ±0.31	93.09 ±0.28	94.01 ±0.35	94.08 ±0.32	92.01 ±0.32	93.91 ±0.27
GELU	94.38 ±0.22	94.41 ±0.23	94.59 ±0.23	93.70 ±0.21	94.24 ±0.25	94.69 ±0.23	92.81 ±0.20	94.20 ±0.16
Swish	94.31 ±0.21	94.32 ±0.20	94.64 ±0.22	93.80 ±0.21	94.14 ±0.24	94.61 ±0.23	92.69 ±0.22	94.22 ±0.17
PAU	94.40 ±0.20	94.46 ±0.22	94.59 ±0.22	93.84 ±0.20	94.29 ±0.22	94.73 ±0.24	93.01 ±0.15	94.54 ±0.13
Mish	94.52 ±0.23	94.39 ±0.22	94.79 ±0.22	93.78 ±0.22	94.51 ±0.22	94.81 ±0.24	92.78 ±0.20	94.77 ±0.18

Table 5. This is an extension to the Table-3. We report Top-1 test accuracy (in %) on CIFAR10 dataset for baseline functions for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	MNIST	Fashion MNIST	SVHN
ReLU	99.48 ± 0.07	92.70 ± 0.20	95.03 ± 0.16
Leaky ReLU	99.47 ± 0.07	92.81 ± 0.19	95.10 ± 0.18
ReLU6	99.52 ± 0.05	92.94 ± 0.14	95.16 ± 0.14
PReLU	99.45 ± 0.09	92.79 ± 0.20	95.12 ± 0.16
ELU	99.51 ± 0.06	92.96 ± 0.15	95.19 ± 0.16
Softplus	99.27 ± 0.11	92.30 ± 0.27	94.71 ± 0.20
GELU	99.57 ± 0.07	93.09 ± 0.12	95.20 ± 0.13
Swish	99.59 ± 0.06	92.90 ± 0.17	95.35 ± 0.16
PAU	99.51 ± 0.10	93.06 ± 0.18	95.29 ± 0.15
Mish	99.61 ± 0.06	93.12 ± 0.15	95.31 ± 0.12
SMU	$\mathbf{99.68} \pm 0.04$	$\mathbf{93.31} \pm 0.15$	$\mathbf{95.59} \pm 0.12$
SMU-1	$\mathbf{99.65} \pm 0.05$	$\mathbf{93.20} \pm 0.11$	$\mathbf{95.46} \pm 0.13$

Table 6. Comparison between SMU, SMU-1 activations and other baseline activations on MNIST, Fashion MNIST, and SVHN datasets for image classification problem on AlexNet architecture. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	Alex Net	Shuffle Net V1	Google Net	Inception V3	Dense Net 121	WideRes Net 28-10	Squeeze Net	VGG 16	LeNet
SMU	$\mathbf{87.25}$ ±0.15	$\mathbf{92.42}$ ±0.14	$\mathbf{94.10}$ ±0.17	$\mathbf{95.59}$ ±0.14	$\mathbf{96.07}$ ±0.12	$\mathbf{96.23}$ ±0.14	$\mathbf{91.77}$ ±0.16	$\mathbf{94.54}$ ±0.14	77.66 ±0.16
SMU-1	$\mathbf{86.77}$ ±0.16	$\mathbf{92.01}$ ±0.15	$\mathbf{93.69}$ ±0.16	$\mathbf{95.11}$ ±0.15	$\mathbf{95.65}$ ±0.12	$\mathbf{95.71}$ ±0.13	$\mathbf{91.38}$ ±0.15	$\mathbf{94.32}$ ±0.15	77.39 ±0.16
ReLU	84.10 ±0.20	91.34 ±0.19	92.91 ±0.18	94.04 ±0.18	94.77 ±0.19	95.08 ±0.21	90.59 ±0.20	93.59 ±0.18	75.80 ±0.21
Leaky ReLU	84.22 ±0.22	91.56 ±0.20	92.79 ±0.17	94.29 ±0.22	94.68 ±0.22	95.01 ±0.20	90.71 ±0.20	93.71 ±0.19	75.99 ±0.20
ReLU6	84.79 ±0.19	91.68 ±0.18	92.97 ±0.16	94.21 ±0.19	94.59 ±0.20	95.39 ±0.20	90.87 ±0.19	93.70 ±0.17	75.88 ±0.18
PReLU	84.30 ±0.24	91.74 ±0.23	92.91 ±0.24	94.45 ±0.20	94.59 ±0.23	95.10 ±0.20	90.79 ±0.23	93.58 ±0.22	75.90 ±0.21
ELU	84.89 ±0.19	91.89 ±0.18	92.99 ±0.16	94.45 ±0.17	94.72 ±0.18	95.23 ±0.17	90.87 ±0.15	93.78 ±0.16	75.88 ±0.18
Softplus	84.01 ±0.30	91.10 ±0.29	92.56 ±0.32	94.17 ±0.31	94.54 ±0.29	94.89 ±0.28	90.55 ±0.33	93.39 ±0.29	75.45 ±0.35
GELU	85.02 ±0.19	91.77 ±0.18	93.36 ±0.18	94.32 ±0.17	94.71 ±0.20	95.19 ±0.18	90.89 ±0.16	93.64 ±0.16	77.71 ±0.19
Swish	85.19 ±0.18	91.49 ±0.20	93.26 ±0.19	94.40 ±0.19	94.69 ±0.17	95.47 ±0.17	91.12 ±0.19	93.68 ±0.17	77.70 ±0.18
PAU	84.91 ±0.20	91.95 ±0.21	93.20 ±0.19	94.32 ±0.23	94.50 ±0.22	95.07 ±0.20	90.51 ±0.19	93.50 ±0.21	77.68 ±0.20
Mish	85.78 ±0.17	91.96 ±0.15	93.29 ±0.17	94.49 ±0.16	95.03 ±0.13	95.39 ±0.16	91.14 ±0.16	93.77 ±0.17	$\mathbf{77.79}$ ±0.15

Table 7. Comparison between SMU, SMU-1 activations and other baseline activations on CIFAR10 dataset for image classification problem. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	MNIST	Fashion MNIST	SVHN
ReLU	99.41 ± 0.09	92.97 ± 0.16	94.48 ± 0.14
Leaky ReLU	99.40 ± 0.07	93.17 ± 0.14	94.59 ± 0.18
ReLU6	99.46 ± 0.07	93.35 ± 0.16	94.61 ± 0.12
PReLU	99.37 ± 0.11	93.18 ± 0.13	94.58 ± 0.15
ELU	99.49 ± 0.07	93.24 ± 0.14	94.57 ± 0.15
Softplus	99.21 ± 0.14	92.99 ± 0.24	94.34 ± 0.22
GELU	99.52± 0.05	93.39 ± 0.14	94.87 ± 0.10
Swish	99.54±0.07	93.34 ± 0.15	94.84 ± 0.14
PAU	99.55±0.12	93.37 ± 0.17	94.79 ± 0.14
Mish	**99.64**±0.06	93.43 ± 0.12	94.87 ± 0.10
SMU	99.61 ± 0.06	**93.61** ± 0.09	**95.06** ± 0.10
SMU-1	99.57 ± 0.07	**93.49** ± 0.11	**95.18** ± 0.12

Table 8. Comparison between SMU, SMU-1 activations and other baseline activations on MNIST, Fashion MNIST, and SVHN datasets for image classification problem on custom designed architecture. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	SF V2 0.5x	SF V2 1.0x	SF V2 1.5x	SF V2 2.0x	SeNet 18	SeNet 34	SeNet 50	Res- Next	Xcep- tion	EffitientNet B0
Leaky ReLU	88.32 ±0.24	91.20 ±0.26	91.24 ±0.24	91.70 ±0.24	94.18 ±0.24	94.52 ±0.23	94.51 ±0.21	93.25 ±0.20	90.81 ±0.25	95.35 ±0.15
ReLU6	88.52 ±0.22	91.15 ±0.23	91.32 ±0.20	91.64 ±0.22	94.39 ±0.22	94.50 ±0.24	94.61 ±0.22	93.49 ±0.21	91.20 ±0.22	95.40 ±0.16
PReLU	88.28 ±0.24	91.02 ±0.22	91.29 ±0.24	91.77 ±0.24	94.35 ±0.24	94.57 ±0.25	94.62 ±0.24	93.35 ±0.24	91.07 ±0.23	95.37 ±0.19
ELU	88.20 ±0.22	91.17 ±0.26	91.40 ±0.22	91.81 ±0.24	94.22 ±0.22	94.42 ±0.24	94.71 ±0.22	93.52 ±0.23	91.45 ±0.22	95.19 ±0.19
Softplus	87.95 ±0.30	90.42 ±0.30	91.01 ±0.28	91.00 ±0.30	93.82 ±0.29	94.05 ±0.30	94.22 ±0.27	93.10 ±0.29	90.56 ±0.27	95.07 ±0.25
GELU	88.92 ±0.20	91.62 ±0.24	91.77 ±0.20	92.29 ±0.19	94.49 ±0.20	94.77 ±0.20	94.79 ±0.16	93.61 ±0.20	91.99 ±0.22	95.45 ±0.15
Swish	89.04 ±0.20	91.71 ±0.22	91.81 ±0.20	92.20 ±0.19	94.30 ±0.18	94.69 ±0.18	94.55 ±0.17	93.61 ±0.19	91.69 ±0.19	95.56 ±0.16
PAU	89.18 ±0.21	91.70 ±0.24	92.20 ±0.20	92.31 ±0.19	94.32 ±0.21	94.77 ±0.22	94.70 ±0.20	93.50 ±0.19	91.91 ±0.22	95.49 ±0.15
Mish	89.42 ±0.20	91.98 ±0.18	92.18 ±0.17	92.47 ±0.18	94.49 ±0.19	94.81 ±0.18	94.97 ±0.15	93.89 ±0.17	92.07 ±0.20	95.70 ±0.12

Table 9. This is an extension to the Table-3. We report Top-1 test accuracy (in %) on CIFAR10 dataset for baseline functions for the mean of 15 different runs. mean±std is reported in the table. SF V2 stands for ShuffleNet v2.

Activation Function	ShuffleNet V2 (2.0x)	ResNet-50	Xception	AlexNet	MobileNet V2
ReLU	70.21 ± 0.23	75.61 ± 0.26	72.10 ± 0.20	55.80 ± 0.27	75.72 ± 0.23
Leaky ReLU	70.09 ± 0.25	75.74 ± 0.27	72.22 ± 0.22	56.10 ± 0.28	75.81 ± 0.25
PReLU	70.17 ± 0.24	75.82 ± 0.28	72.18 ± 0.24	56.52 ± 0.26	75.98 ± 0.27
ReLU6	70.21 ± 0.23	76.14 ± 0.25	72.35 ± 0.19	56.69 ± 0.26	75.87 ± 0.22
ELU	70.34 ± 0.24	76.15 ± 0.26	72.41 ± 0.23	56.97 ± 0.25	75.79 ± 0.25
Softplus	69.91 ± 0.26	75.51 ± 0.30	71.94 ± 0.26	55.65 ± 0.35	75.60 ± 0.27
Swish	73.64 ± 0.21	76.80 ± 0.24	73.45 ± 0.20	58.77 ± 0.24	76.67 ± 0.21
Mish	74.25 ± 0.22	77.30 ± 0.24	74.34 ± 0.21	59.87 ± 0.25	77.02 ± 0.22
GELU	73.51 ± 0.21	76.85 ± 0.25	73.71 ± 0.18	58.50 ± 0.26	76.61 ± 0.22
PAU	73.85 ± 0.20	77.07 ± 0.24	73.87 ± 0.20	58.80 ± 0.25	76.81 ± 0.21
SMU	**75.78** ± 0.20	**78.71** ± 0.24	**75.30** ± 0.18	**62.42** ± 0.23	**77.83** ± 0.20
SMU-1	**75.01** ± 0.21	**77.81** ± 0.24	**74.84** ± 0.20	**61.93** ± 0.25	**77.49** ± 0.22

Table 10. Comparison between SMU, SMU-1 activations and other baseline activations on CIFAR100 dataset for image classification problem with Mixup augmentation method. We report Top-1 test accuracy (in %) for the mean of 15 different runs. mean±std is reported in the table.

Activation Function	ShuffleNet V2 (1.0x)	ResNet-50
ReLU	69.21	75.52
Leaky ReLU	69.28	75.67
PReLU	69.01	75.40
ReLU6	69.45	75.70
ELU	69.49	75.62
Softplus	69.01	75.30
Swish	70.35	76.06
Mish	70.53	76.45
GELU	70.12	76.01
PAU	70.28	76.14
SMU	**71.93**	**77.48**
SMU-1	**71.17**	**76.89**

Table 11. Top-1 accuracy reported on ImageNet-1k dataset.