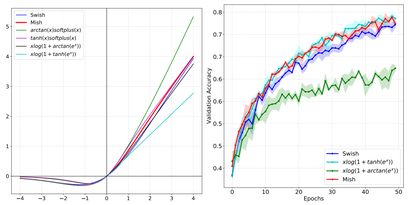
Mish: A Self Regularized Non-Monotonic Neural Activation Function

Diganta Misra

mishradiganta91@gmail.com



Contribution: Improved Generalization *and* Smoother Profile

ReLU activation function is fast but not reliable often leading to information loss due to thresholding of negative weights. We investigate new smoother candidate non-linearities which leads us to an optimal non-monotonic smooth alternative to ReLU which we call Mish. Mish results in much better generalization and often faster convergence and avoids critical information loss caused by ReLU.

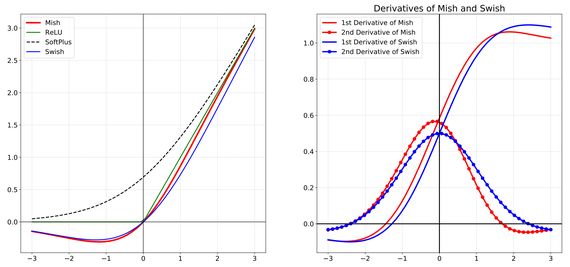
General definition

Mish can be defined as

$$f(x) = x \tanh(softplus(x)), \quad \text{where } softplus(x) = \ln(1 + e^x) \tag{1}$$

and its derivative can be defined as

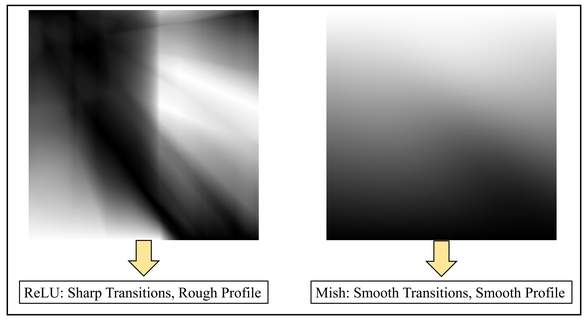
$$f'(x) = \Delta(x) swish(x) + \frac{f(x)}{x}, \quad \text{where } \Delta(x) = sech^2(softplus(x)) \tag{2}$$



(a)

(b)

Output Landscape



Get the weights, paper, and source code at:
`https://github.com/digantamisra98/Mish`

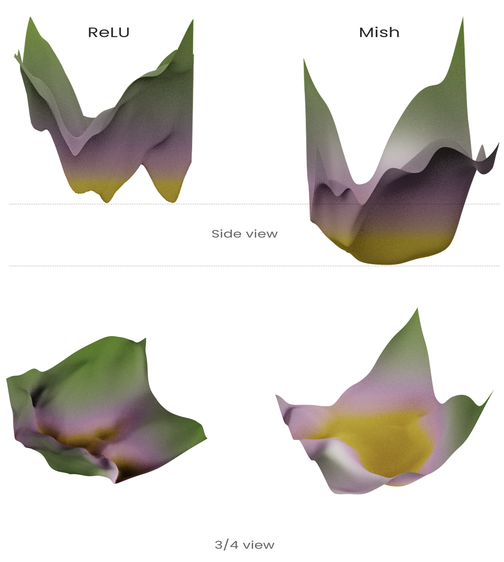
References

[1] Ramachandran, Prajit and Zoph, Barret and Le, Quoc V Searching for activation functions, 2017

Motivation

Recent theoretical advancements into non-linear dynamics in deep neural networks have prompted the discoveries of new smooth activation functions which improve information propagation in the network. Swish was discovered by a Neural Architecture Search (NAS) over the space of the non-linear functions by a controlled search agent. The design of Mish, while influenced by the work performed by Swish, was found by systematic analysis and experimentation over the characteristics that made Swish so effective. We found five candidate functions which met the requirements and post rigorous evaluation finalized on Mish.

Loss Landscapes



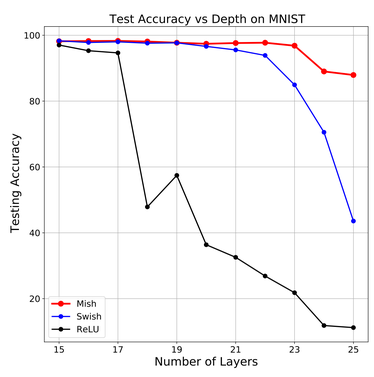
Mish has a smooth, convex, easy to optimize loss landscape when compared with ReLU as visualized above for a ResNet-20 on CIFAR-10 dataset. Further details at: `https://losslandscape.com/`

ImageNet-1k results

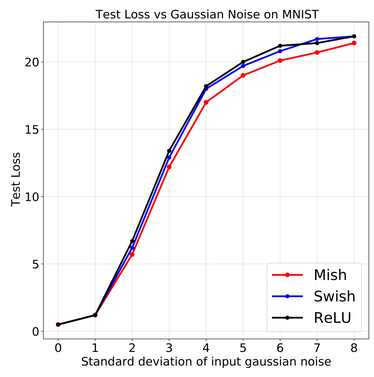
Model	Data Augmentation	LReLU		Swish		Mish	
		Top-1	Top- 5	Top-1	Top- 5	Top-1	Top- 5
SpineNet-49	Yes	-	-	78.1%	94%	**78.3%**	**94.6%**
CSPResNext-50	No	77.9%	94%	64.5%	86%	**78.9%**	**94.5%**
CSPResNext-50	Yes	78.5%	94.8%	-	-	**79.8%**	**95.2%**
CSPResNet-50	Yes	77.1%	94.1%	-	-	**78.1%**	**94.2%**
CSPDarkNet-53	Yes	77.8%	94.4%	-	-	**78.7%**	**94.8%**
Pelee Net	No	70.7%	90%	**71.5%**	**90.7%**	71.4%	90.4%
CSPPelee Net	No	70.9%	90.2%	-	-	**71.2%**	**90.3%**

Image classification of ImageNet-1k dataset. Models equipped with Mish demonstrated stable learning curves with high generalization capability. This characteristic was consistent in larger models when evaluated for the ImageNet-1k classification task. Further, we evaluated the performance of Mish, Swish and Leaky ReLU under different settings of data augmentations like that of CutMix and Label Smoothing. Although Swish marginally beats Mish in Pelee Net, we observed it's performance to be inconsistent especially in larger networks with increasing depth which is evident in the drop in performance seen in CSP-ResNext-50 model.
All the models were evaluated under fair settings used in standard benchmarking practices on a cluster of 4 Nvidia-V100 Tesla GPUs.

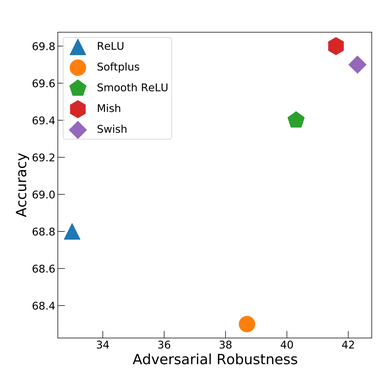
Robustness



(a)



(b)



(c)

Mish is more consistent across increasing depth and provides adversarial robustness. We compare Mish with Swish and ReLU in network with linearly increasing depth and observe a general trend of Mish maintaining stronger generalization capability as compared to Swish and ReLU, both of which suffer from a large drop in accuracies. Additionally, Mish also provides better performance (lower loss) under the scenario of increasing Gaussian noise in the input data to the network.
Further, we also observe Mish provide strong adversarial robustness with strong accuracy as compared to non-smooth non-linear counterparts. This further emphasizes the importance of smooth functions like Mish and Swish as compared to piece-wise linear alternatives like ReLU or Leaky ReLU.