Tanh Works Better with Asymmetry

Dongjin Kim[1,3], Woojeong Kim[2], Suhyun Kim[3]

[1]Korea University, [2]Cornell University, [3]Korea Institute of Science and Technology

NEURAL INFORMATION PROCESSING SYSTEMS

KOREA UNIVERSITY

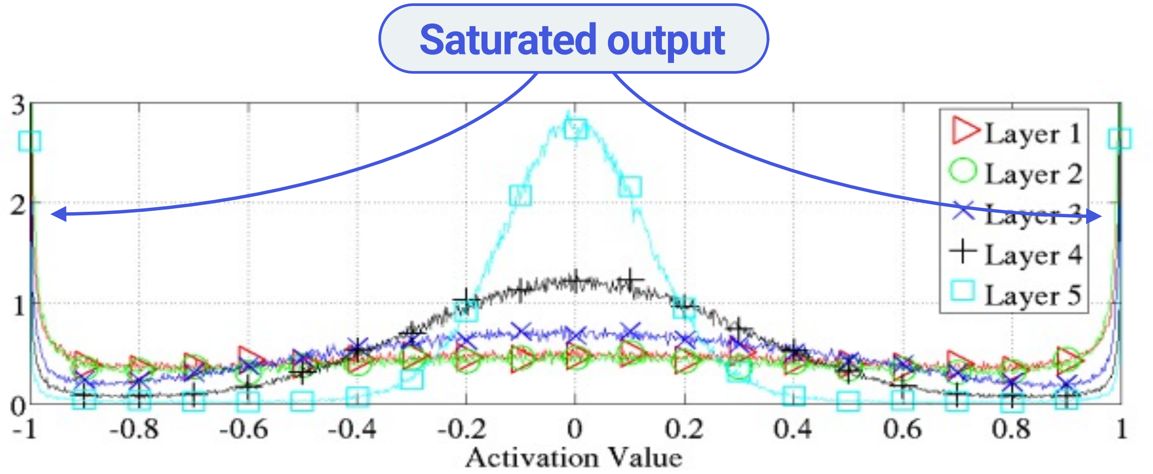
CORNELL UNIVERSITY

KIST Korea Institute of Science and Technology

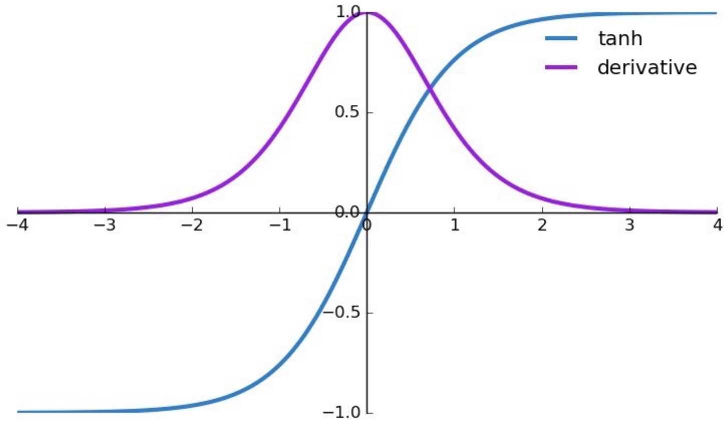
An activation function with two boundaries

- It has a **saturation state**.

  The activation is close to the asymptotic value.



- The saturated output suffered from the vanishing gradients problem

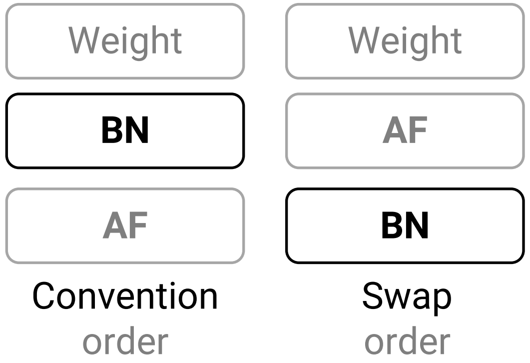


- Due to poor performance, Tanh becomes forgotten.

Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010.

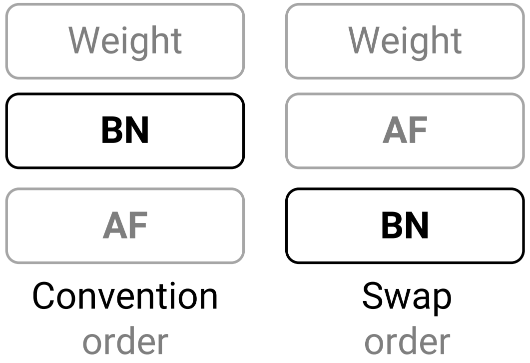
Order of Batch Normalization

Batch Normalization(BN) places between the weight and activation function(AF) .



Order of Batch Normalization

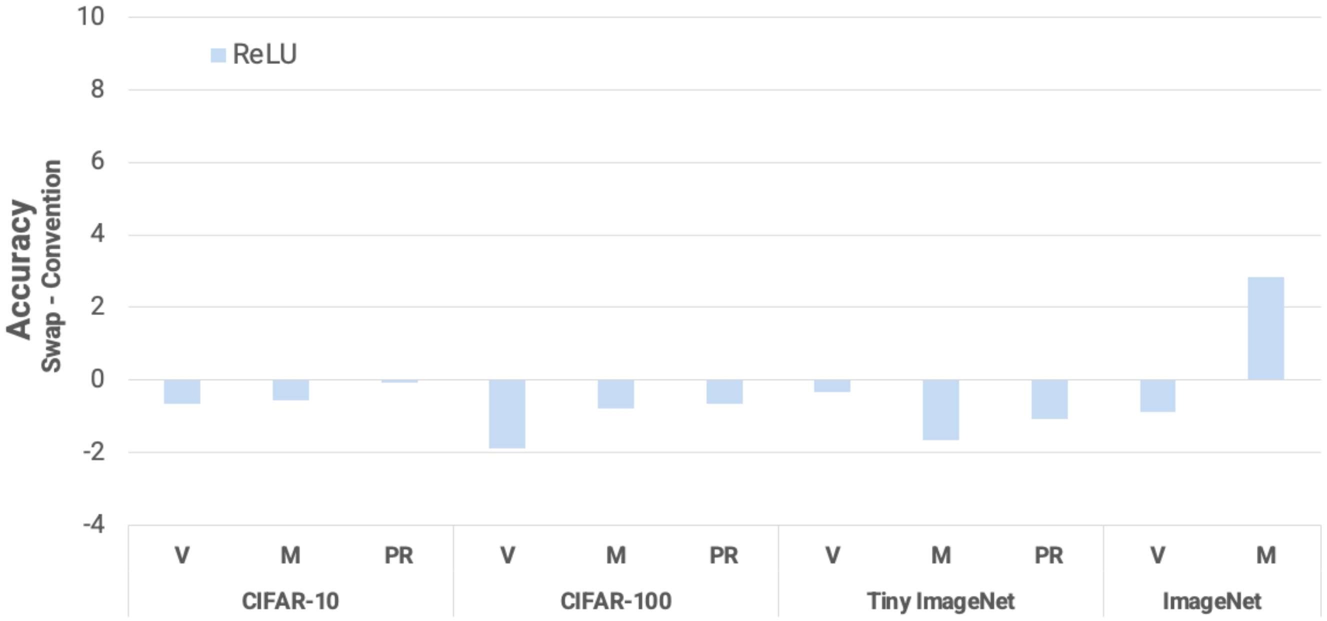
Batch Normalization(BN) places between the weight and activation function(AF) .



Datasets	Models	ReLU	
		Convention	Swap
CIFAR-10	VGG16	93.69	93.04
	MobileNet	92.48	91.93
	PreAct-ResNet18	94.94	94.86
CIFAR-100	VGG16	73.68	71.79
	MobileNet	70.27	69.49
	PreAct-ResNet18	78.06	77.39
Tiny ImageNet	VGG16	59.37	59.05
	MobileNet	51.90	50.25
	PreAct-ResNet34	67.28	66.21
ImageNet	VGG16	73.83	72.95
	MobileNet	68.27	71.1

ReLU

The layer order does not dramatically change accuracy.



V: VGG16, M: MobileNet, PR: PreAct-ResNet

Order of Batch Normalization

Batch Normalization(BN) places between the weight and activation function(AF) .

Convention order	Swap order
Weight	Weight
BN	AF
AF	BN

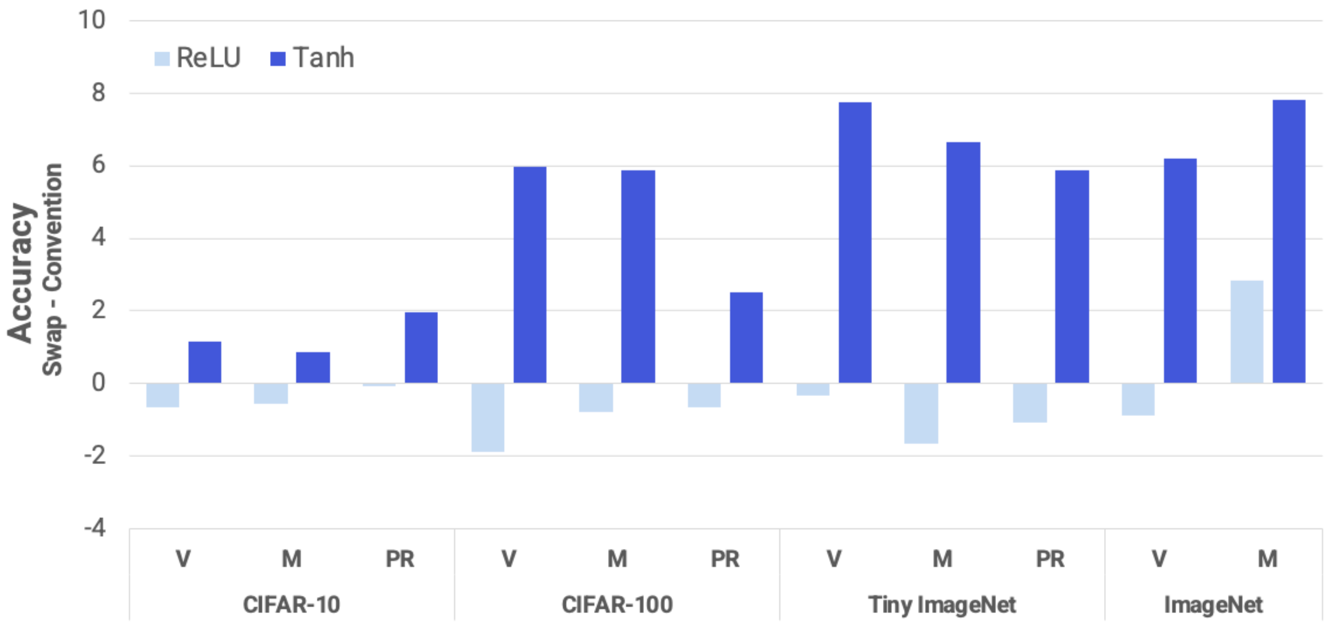
Datasets	Models	ReLU		Tanh	
		Convention	Swap	Convention	Swap
CIFAR-10	VGG16	93.69	93.04	91.75	92.90
	MobileNet	92.48	91.93	91.66	92.53
	PreAct-ResNet18	94.94	94.86	92.46	94.41
CIFAR-100	VGG16	73.68	71.79	64.95	70.93
	MobileNet	70.27	69.49	64.50	70.39
	PreAct-ResNet18	78.06	77.39	73.26	75.76
Tiny ImageNet	VGG16	59.37	59.05	49.29	57.05
	MobileNet	51.90	50.25	45.38	52.05
	PreAct-ResNet34	67.28	66.21	59.06	64.94
ImageNet	VGG16	73.83	72.95	60.85	67.04
	MobileNet	68.27	71.1	64.26	72.07

ReLU

The layer order does not dramatically change accuracy.

Tanh

The Swap order significantly outperforms the Convention order.



V: VGG16, M: MobileNet, PR: PreAct-ResNet

Order of Batch Normalization

Batch Normalization(BN) places between the weight and activation function(AF) .

Weight	Weight
BN	**AF**

Datasets	Models	ReLU		Tanh	
		Convention	Swap	Convention	Swap
CIFAR-10	VGG16	93.69	93.04	91.75	92.90
	MobileNet	92.48	91.93	91.66	92.53
	PreAct-ResNet18	94.94	94.86	92.46	94.41
CIFAR-100	VGG16	73.68	71.79	64.95	70.93
	MobileNet	70.27	69.49	64.50	70.39

Goal

1. Reveal a hidden property

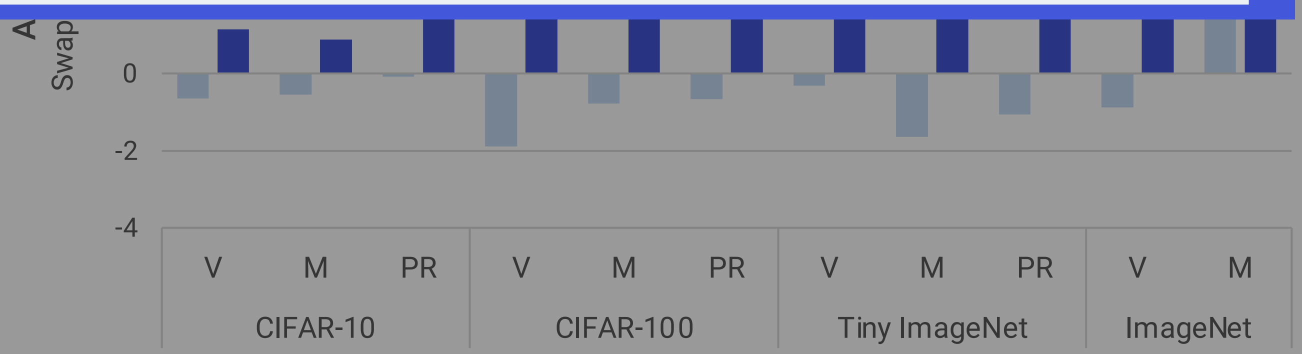
Why is the Swap order effective on Tanh?

2. Modified Activation Function

How can we redesign the order-agnostic Tanh with improved accuracy?

Tanh

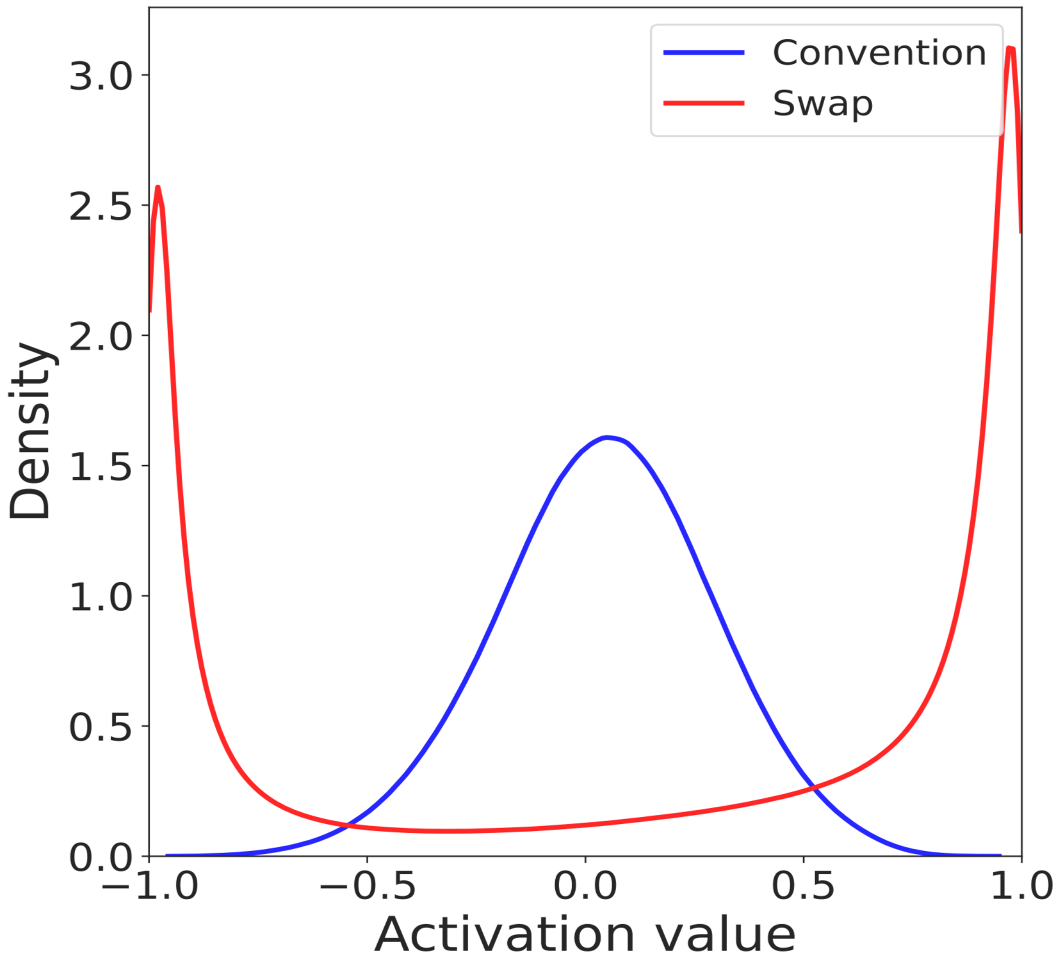
The Swap order significantly outperforms the Convention order.



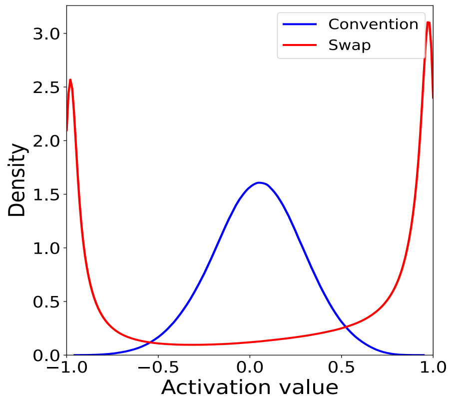
Layer-wise Activations of Tanh

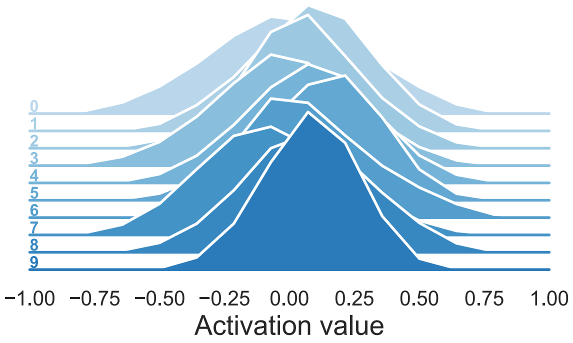
↓

Symmetrically Distributed



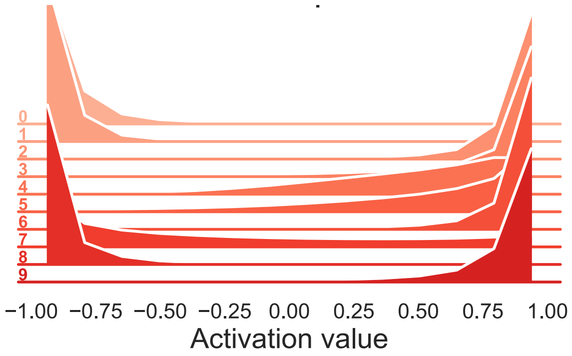
Channel-wise Activation of Tanh





Convention model

Symmetric activation



Swap model

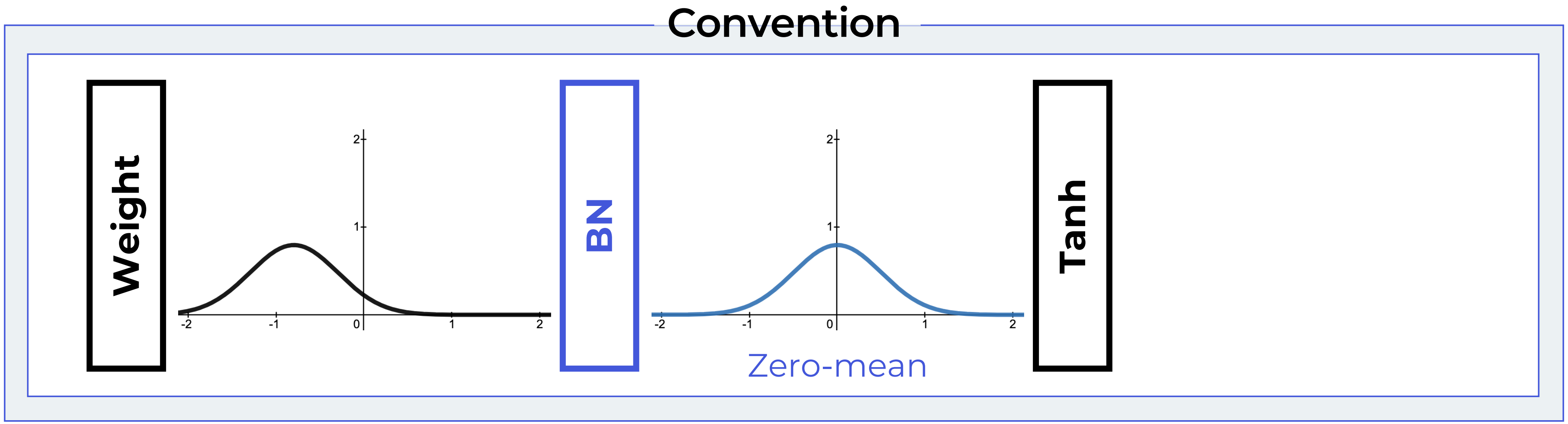
Asymmetric activation

Channel-wise activation of the Swap order is asymmetrically distributed.

The Reason for Low Asymmetry in the Convention Order

- Batch Normalization shifts
  the biased weighted sum outputs to **zero**.

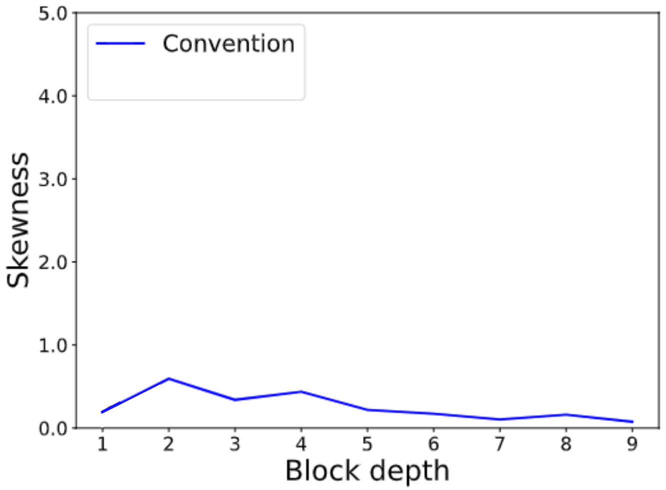
- In the Convention order,
  the zero mean distribution generates
  symmetric activation on Tanh.

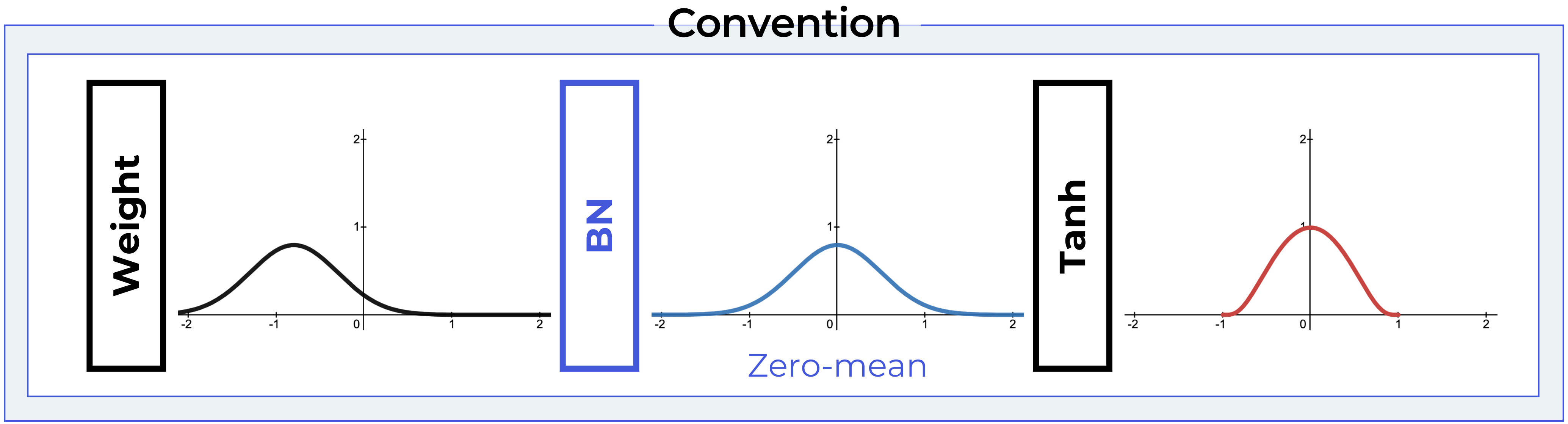


The Reason for Low Asymmetry in the Convention Order

- Batch Normalization shifts the biased weighted sum outputs to **zero**.

- In the Convention order, the zero mean distribution generates **symmetric activation on Tanh**.

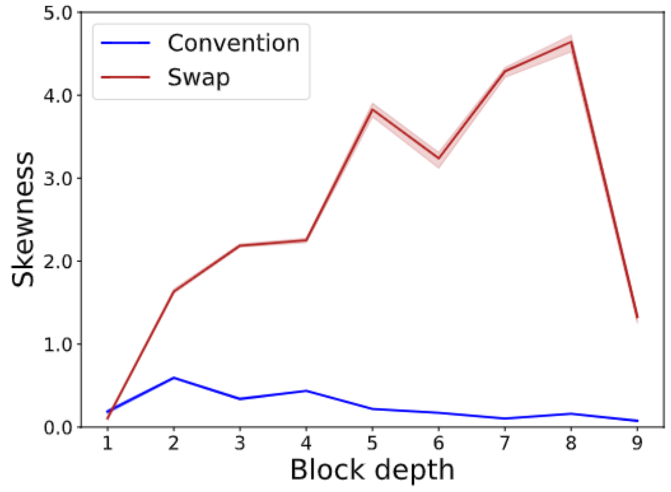




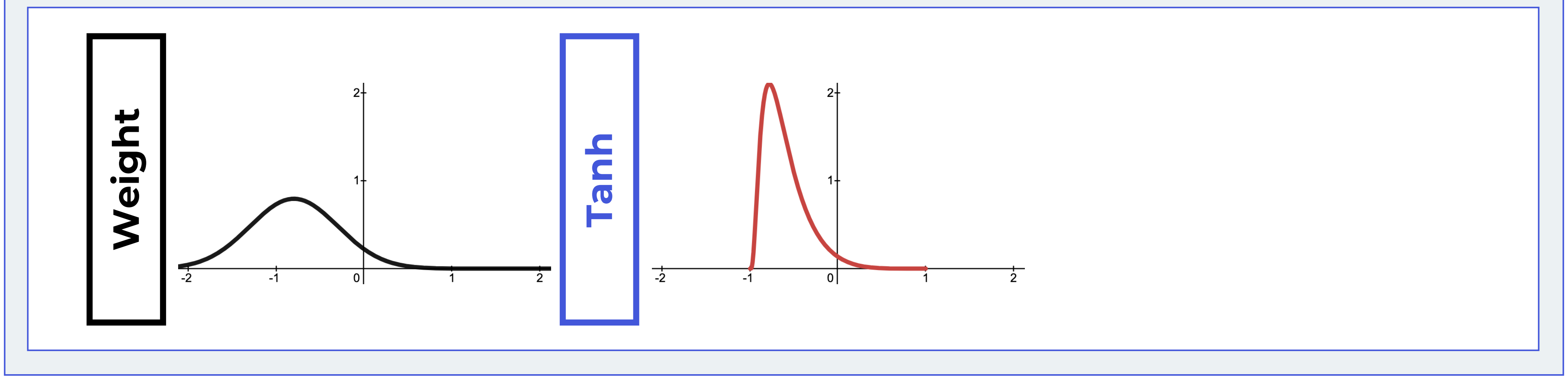
What Brings Asymmetry in the Swap Order?

The elimination of the preceding Batch Normalization

The biased distribution to Tanh encourages asymmetric saturation in the Swap order.



Swap



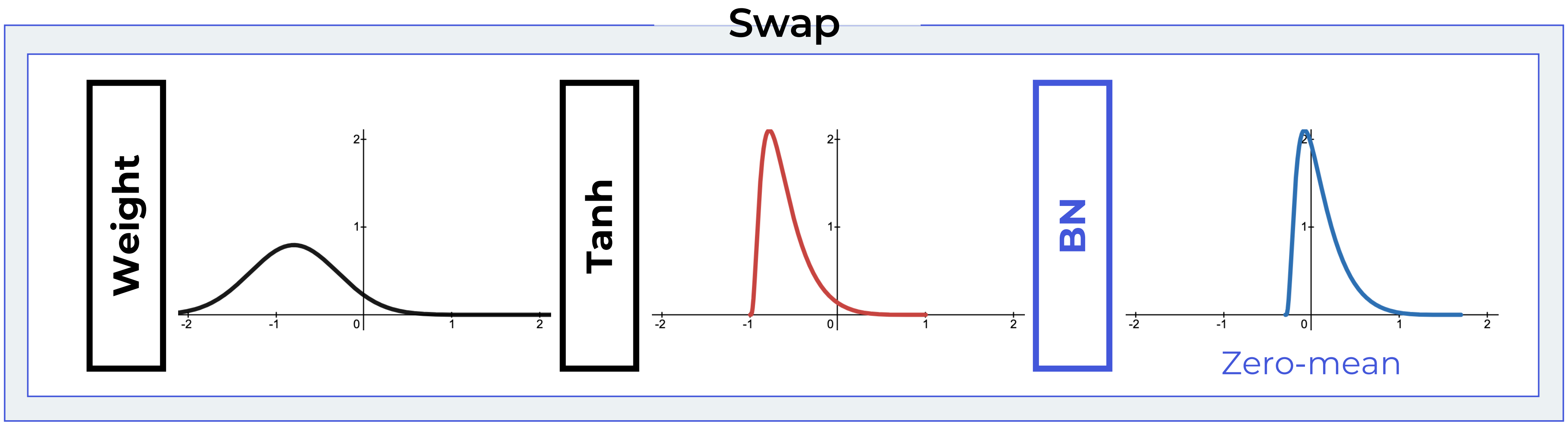
Additional Improvement incurred by Asymmetry

The elimination of Batch Normalization before the Tanh

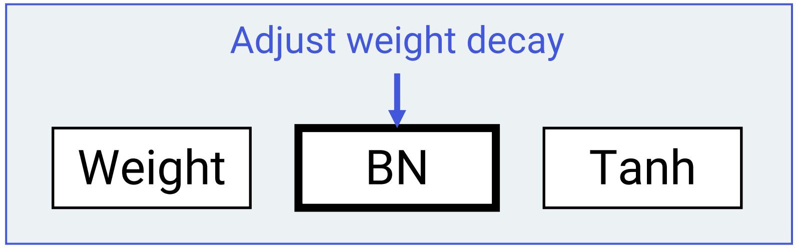
The biased distribution to Tanh encourages asymmetric saturation in the Swap order.

Asymmetric saturation in the Swap order incurs sparsity

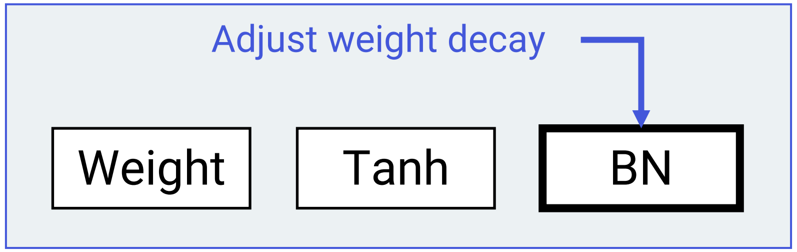
Asymmetric saturation incurs sparsity by a zero mean shifting in normalization.



The Effect of Asymmetry and Sparsity on Accuracy

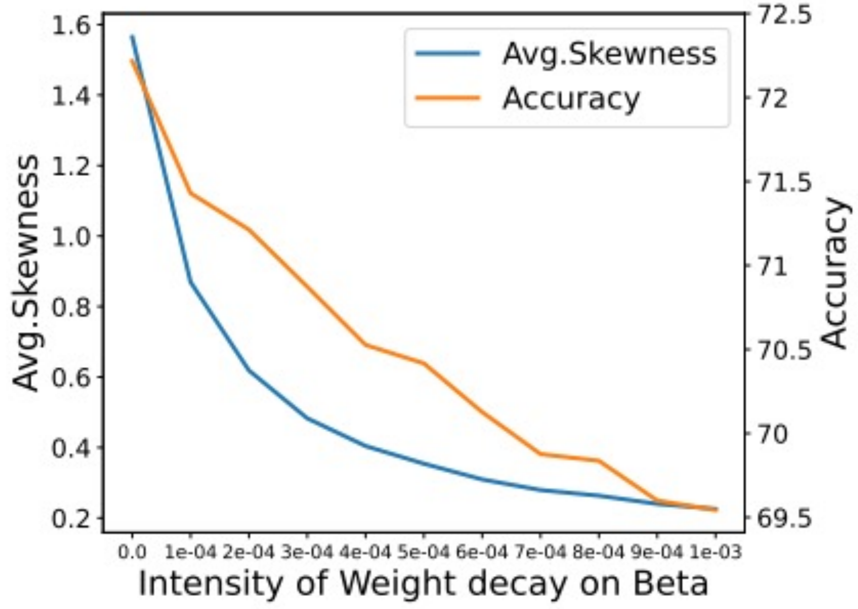


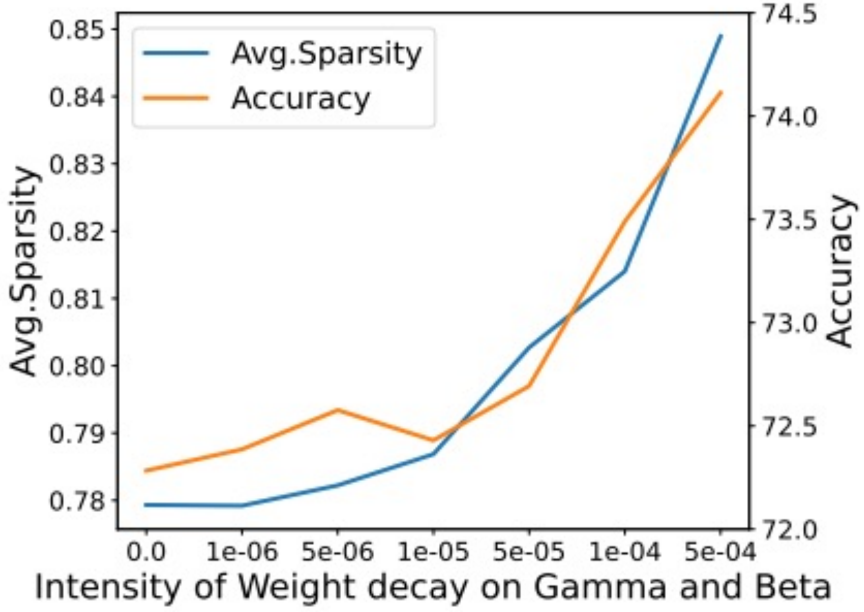
Convention order

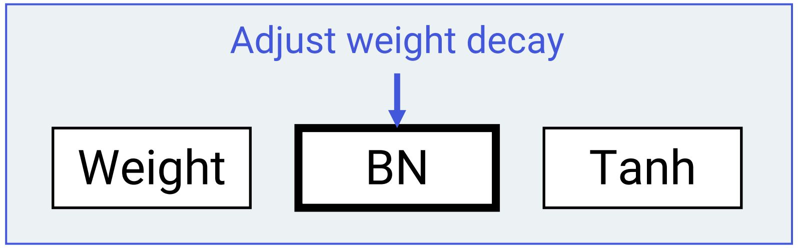


Swap order

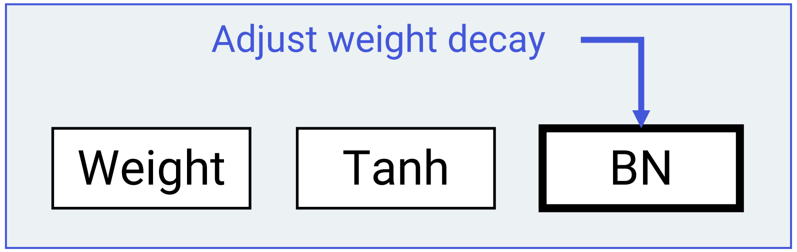
The Effect of Asymmetry and Sparsity on Accuracy







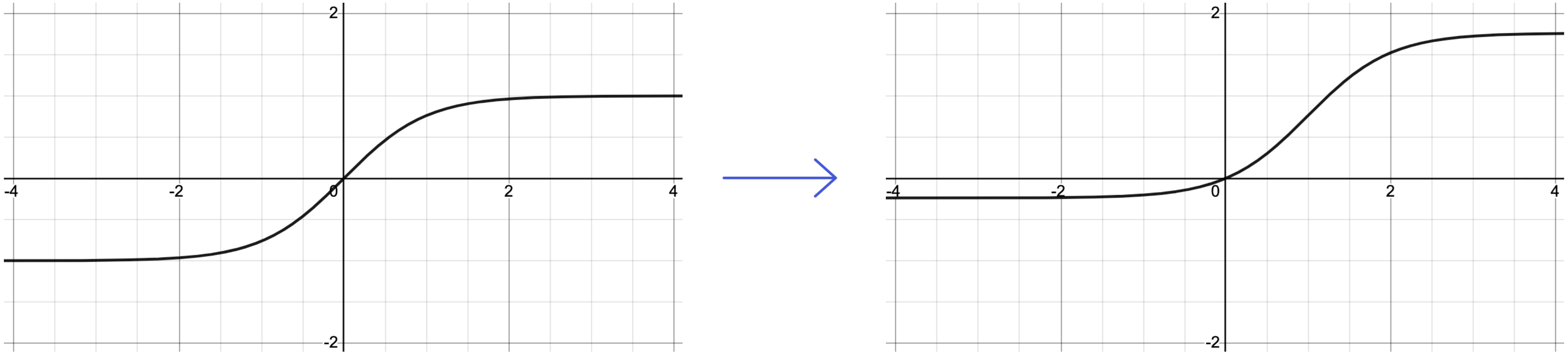
Convention order



Swap order

The shifted Tanh introduce asymmetric and sparse activation easily.

$$tanh(x+\boldsymbol{\tau}) - \boldsymbol{tanh}(\boldsymbol{\tau})$$

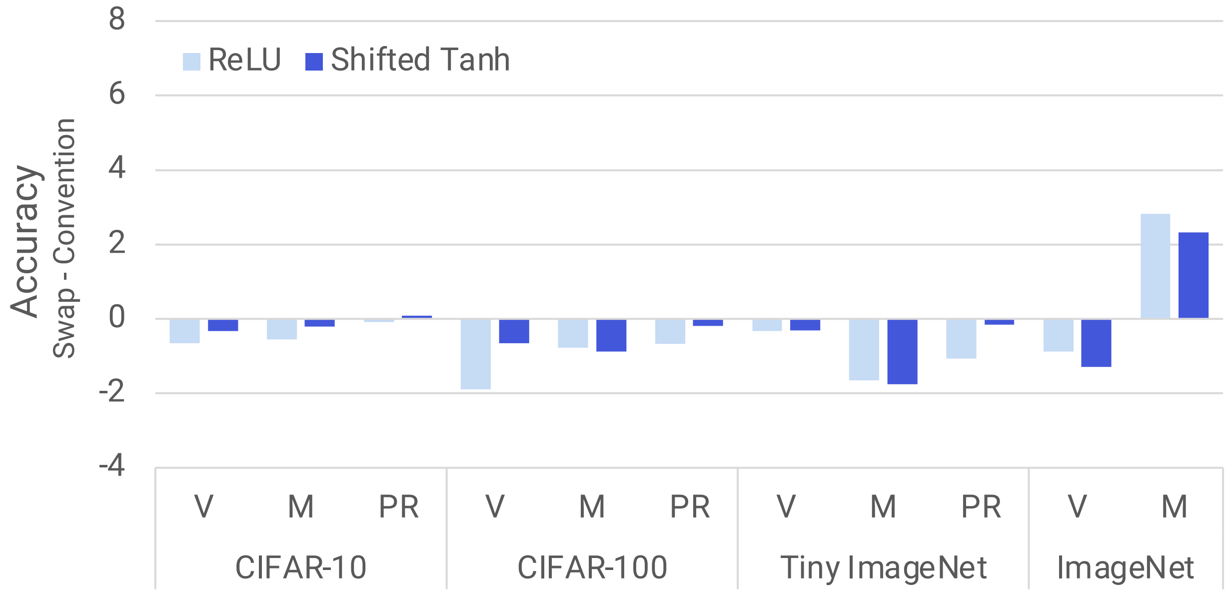


$\tau = 0$ | Original Tanh

$\tau = -1$ | Shifted Tanh

The properties of the shifted Tanh

- It shows improved accuracy comparable with the ReLU model.
- The accuracy discrepancy between orders decreased.



Dataset	Model	Shifted Tanh		ReLU	
		Convention	Swap	Convention	Swap
CIFAR-10	VGG16	93.92	93.60	93.69	93.04
	MobileNet	92.85	92.64	92.2	91.93
	PreAct-ResNet18	94.48	94.57	94.94	94.86
CIFAR-100	VGG16	73.87	73.21	73.68	71.79
	MobileNet	72.75	71.88	70.06	69.49
	PreAct-ResNet18	76.70	76.51	78.06	77.39
Tiny ImageNet	VGG16	60.68	60.37	61.54	59.05
	MobileNet	59.95	58.20	59.79	59.1
	PreAct-ResNet34	65.82	65.66	67.28	66.21
ImageNet	VGG16	73.4	72.11	73.83	72.95
	MobileNet	70.4	72.72	70.48	71.1

Other Bounded Activation Functions

- The Swap model with other bounded functions outperforms the Convention model.

- Softsign, which is a slower approach to its asymptotes than Tanh, underperforms Tanh on the Swap order, even if it performs better on the Convention order.

Activation functions	**Order**		**Δavg. Skewness**
	Convention	Swap	(Swap - Convention)
Tanh	69.5	74.11	2.38
Softsign	70.01	73.65	1.28
LeCun Tanh	67.82	74.46	1.90

Other Activation Functions

Other Bounded Activation Functions

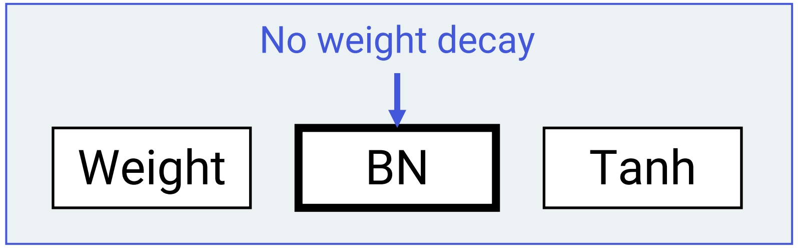
- The Swap model with other bounded functions outperforms the Convention model.

- Softsign, which is a slower approach to its asymptotes than Tanh, underperforms Tanh on the Swap order, even if it performs better on the Convention order.

Activation functions	Order		Δavg. Skewness
	Convention	Swap	(Swap - Convention)
Tanh	69.5	74.11	2.38
Softsign	70.01	73.65	1.28
LeCun Tanh	67.82	74.46	1.90

Dominance Between Asymmetry and Sparsity

- The NWDBN model with encouraged asymmetry outperforms the Convention model even if the sparsity is decreased.



NWDBN

Configurations	Accuracy	Avg. Skewness	Avg. Sparsity
Convention	69.5	0.254	**0.718**
NWDBN	72.22	**0.718**	0.288

Thank You!

npclinic3@gmail.com