

# ShakeDrop Regularization

11/29/2018

Paper By: Yoshihiro Yamada, Masakazu Iwamura, and Koichi Kise

Presented by: Travis Bender



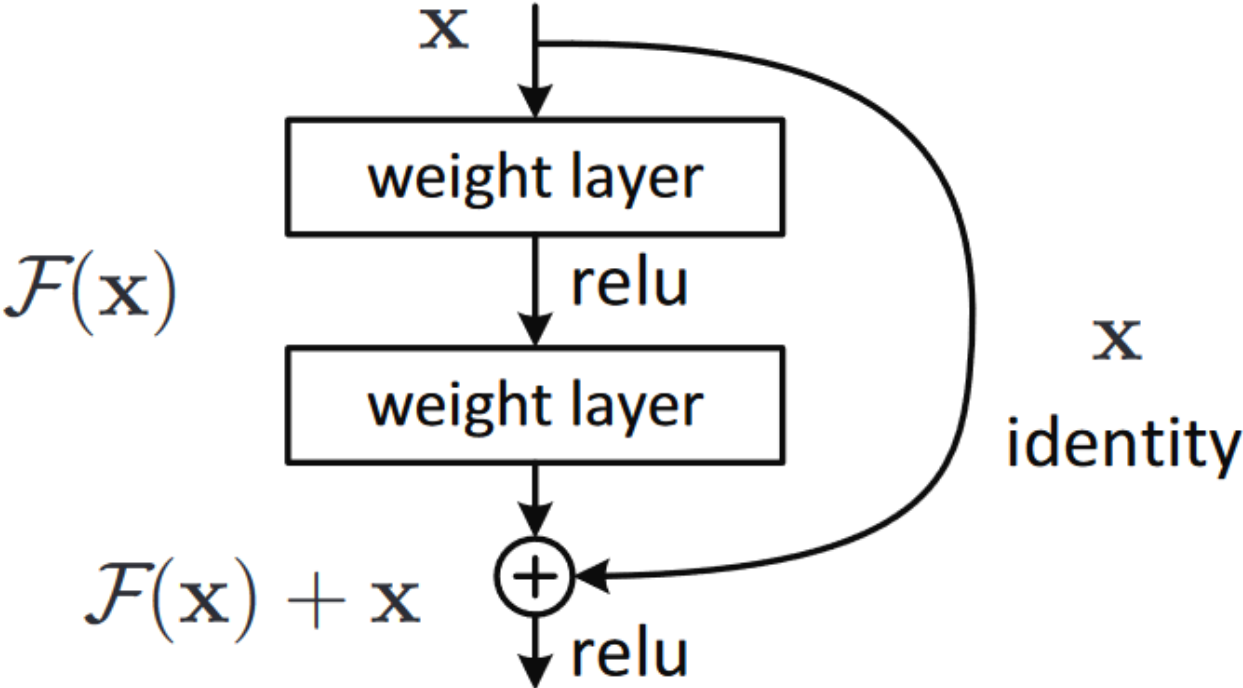
**UNIVERSITY OF WATERLOO**  
FACULTY OF MATHEMATICS

# Motivation

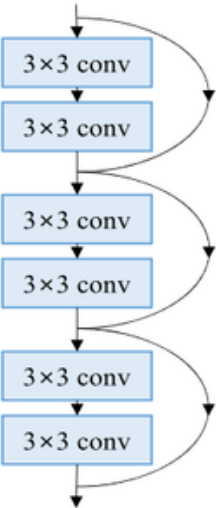
- Current state-of-the-art architectures for image classification utilize residual blocks to perform well
- Very deep strategies suffer from overfitting, and have been shown improvement when subjected to regularization
- The previous best regularization approach is limited to a small subset of network architectures and could be generalized.



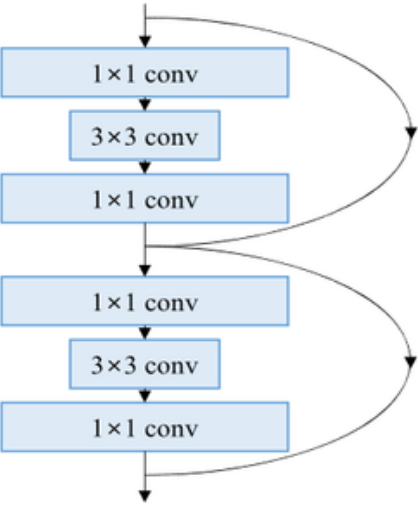
# Basic Residual Block



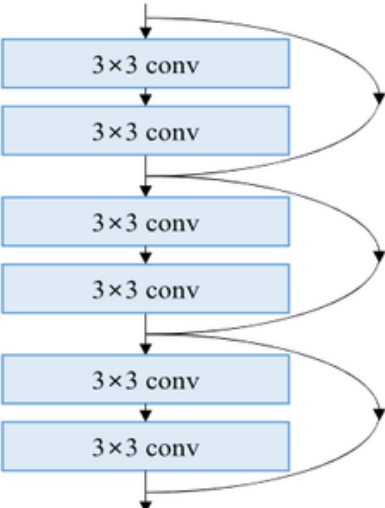
# Residual Network Architectures



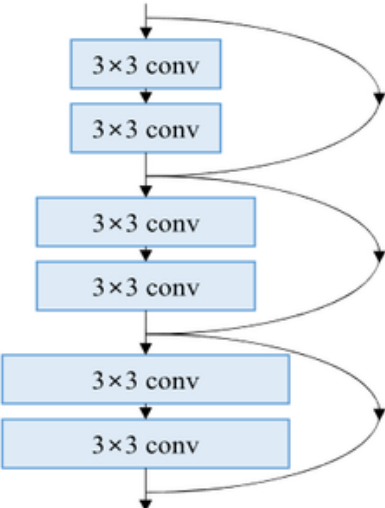
(a) basic



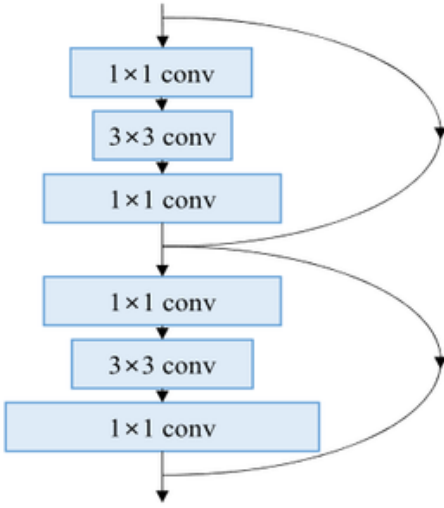
(b) bottleneck



(c) wide



(d) pyramidal



(e) pyramidal bottleneck



# Regularization Approaches

## Stochastic Depth

- Developed to address vanishing gradients in ResNet

$$G(x) = x + b_l F(x)$$

- $b_l \in \{0,1\}$  is a Bernoulli random variable with probability  $p_l$ . Uses a linear decay rule that defines  $p_l$  as

$$p_l = 1 - \frac{l}{L}(1 - p_L)$$

- $L$  is the number of layers and  $p_L$  is an initial parameter
- Originally developed for ResNet, and later adapted to PyramidNet

## ShakeShake Regularization

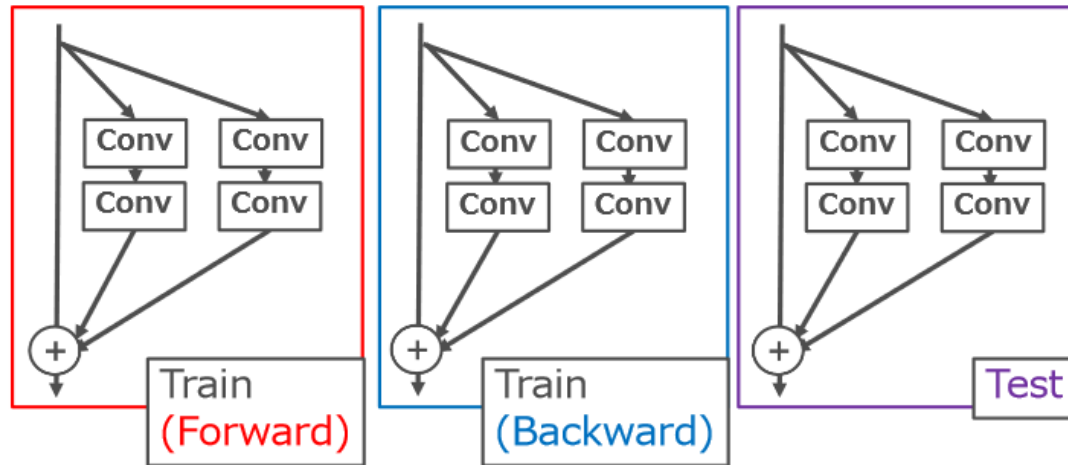
- Developed for use with the ResNeXt architecture

$$G(x) = x + \alpha F_1(x) + (1 - \alpha)F_2(x)$$

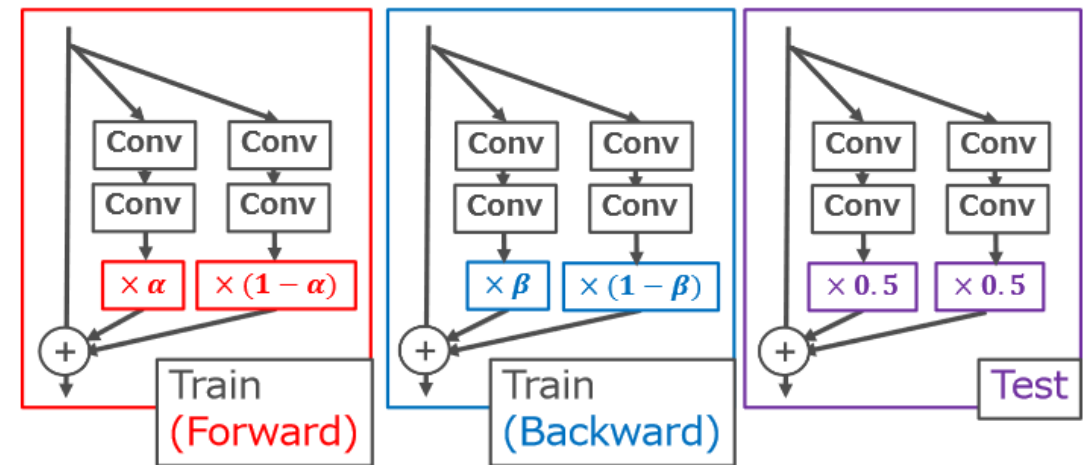
- $\alpha \in [0,1]$
- Backwards pass is formulated identically but with another random parameter  $\beta$
- Interpolates results between each branch, almost like the network is using augmented data



# ShakeShake Regularization



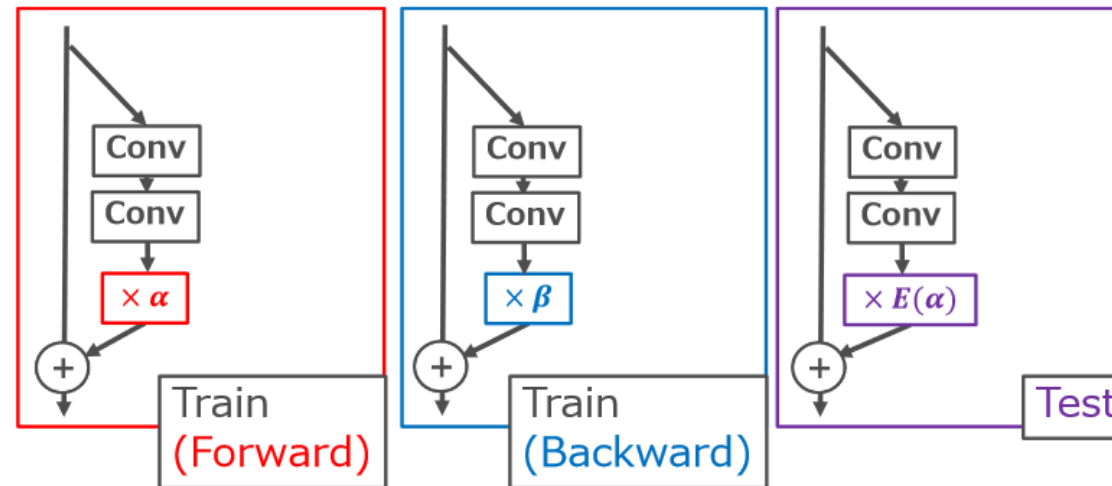
ResNeXt (Xie et al., 2017)



ResNeXt + ShakeShake (Gastaldi, 2017)

# 1-Branch Shake

- An Adaptation of ShakeShake for use in single branch architectures
- $G(x) = x + \alpha F(x)$  for forward pass
- $G(x) = x + \beta F(x)$  for backwards pass



# 1-Branch Shake Cont.

- Unfortunately, 1-Branch Shake performs horribly when applied in this basic form, achieving an error rate of 77.99% on CIFAR-100
- Failure is caused by perturbation that is too strong.
- Could be improved by combining 1-Branch Shake with Stochastic Depth/ResDrop





# ShakeDrop Regularization

- Based on a combination of 1-branch shake and stochastic depth
- Given as

$$G(x) = x + (b_l + \alpha - b_l\alpha)F(x)$$

- Or alternatively, (replace  $\alpha$  with  $\beta$  for backwards pass)

$$G(x) = \begin{cases} x + F(x), & \text{if } b_l = 1 \\ x + \alpha F(x), & \text{otherwise (i.e., if } b_l = 0) \end{cases}$$

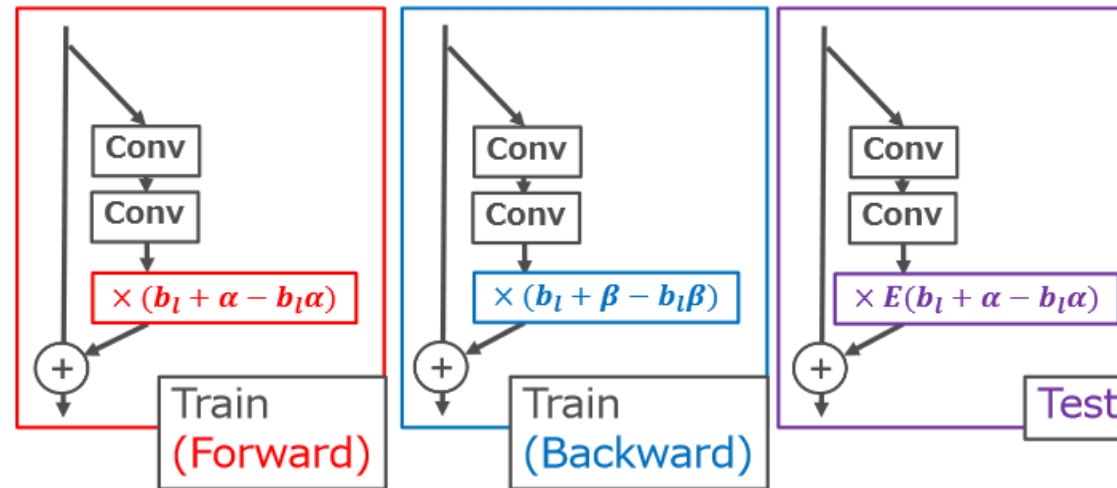
- $b_l$  is a Bernoulli random variable utilizing linear decay rule

$$p_l = 1 - \frac{l}{L}(1 - p_L)$$



# ShakeDrop Regularization Cont.

- Causes a single branch to behave as if there are two networks, the original and the one with dropped residual blocks.
- Uses parameter  $\beta$  on backwards pass
- Layers with  $b_l = 0$  are not update during a step of training,



# ShakeDrop Parameter Search

Average Top-1 errors(%) of PyramidNet + ShakeDrop for different parameters

	$\alpha$	$\beta$	Error (%)	Note
A	1	1	18.01	Equivalent to PyramidNet
B	0	0	17.74	Equivalent to PyramidDrop
C	[0, 1]	[-1, 1]	20.61	
D	[0, 1]	[0, 1]	18.27	
E	[-1, 1]	1	18.68	
F	[-1, 1]	0	17.28	
G	[-1, 1]	[-1, 1]	18.26	
H	[-1, 1]	[0, 1]	<b>16.22</b>	

# ShakeDrop Level Setting

Batch – Same scaling coefficients for all images in mini-batch

Image – Same scaling coefficients for each image for each residual block

Channel – Same scaling coefficients for each channel for each residual block

Pixel – Same scaling coefficients for each element in each residual block

$\alpha$	$\beta$	Level	Error (%)
[-1, 1]	[0, 1]	Batch	16.22
		Image	16.04
		Channel	16.12
		Pixel	<b>15.78</b>

# Experiments Setup

- Attempted to match setup as closely as possible to make results comparable
- Learning rate is either determined with a schedule and 300 epochs, or using cosine annealing with an 1800 epochs where specified.
- CIFAR-10/100 was color normalized, horizontally flipped with probability 0.5, and is zero padded to be 40x40, then randomly cropped back to 32x32
- Where specified, data additionally is augmented with either Cutout or Random Erasing
- Wide ResNet added batch normalization to residual blocks
- Type A means the regularization term is inserted before the add term in residual branches, Type B adds the regularization term afterwards.



# CIFAR-100 Top-1 Errors

Methods	Regularization	Original (%)	EraseReLU (%)
<b>ResNet-110</b> <Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	28.51	24.93
	ResDrop	24.09	22.88
	1-branch Shake	24.18	23.80
	ShakeDrop	×	<b>22.68</b>
<b>ResNet-164 Bottleneck</b> <Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	22.00	21.96
	ResDrop	21.96	20.35
	1-branch Shake	22.20	21.60
	ShakeDrop	×	<b>19.89</b>
<b>ResNeXt-29 8-64d Bottleneck</b> <Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	20.90	20.25
	ResDrop	20.66	20.28
	1-branch Shake	22.70	24.00
	ShakeDrop	×	<b>19.90</b>
<b>PyramidNet-272 <math>\alpha</math>200 Bottleneck</b> <BN-Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add>	Vanilla	*16.35	N/A
	ResDrop	15.94	
	1-branch Shake	71.51	
	ShakeDrop	<b>14.90</b>	



# CIFAR-100 Top-1 Errors Continued

Methods	Regularization	Original (%)	w/ BN (%)
<b>Wide-ResNet-28-10k</b> <BN-ReLU-Conv-BN-ReLU-Conv-(BN)-add>	Vanilla	26.49	24.24
	ResDrop	34.19	26.64
	1-branch Shake	90.73	58.89
	ShakeDrop	76.87	<b>19.12</b>

Methods	Regularization	Original (%)	EraseReLU (%)
<b>ResNeXt-164 2-1-40d Bottleneck</b> <Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	23.82	21.75
	ResDrop Type-A	21.38	20.44
	ResDrop Type-B	21.34	20.21
	Shake-Shake	22.35	22.51
	ShakeDrop Type-A	×	<b>19.98</b>
	ShakeDrop Type-B	×	<b>19.83</b>
<b>ResNeXt-29 2-4-64d Bottleneck</b> <Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	21.19	×
	ResDrop Type-A	21.12	20.13
	ResDrop Type-B	19.27	19.01
	Shake-Shake	19.16	18.82
	ShakeDrop Type-A	×	20.07
	ShakeDrop Type-B	×	<b>18.17</b>



# Tiny ImageNet

Methods	Regularization	Original (%)	EraseReLU (%)
<b>ResNet-110</b> <Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	42.07	<b>41.24</b>
	ResDrop	43.74	42.50
	1-branch Shake	45.56	45.16
	ShakeDrop	×	48.92
<b>ResNet-164 Bottleneck</b> <Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add-(ReLU)>	Vanilla	38.20	<b>36.52</b>
	ResDrop	37.17	38.09
	1-branch Shake	39.29	42.10
	ShakeDrop	×	42.80
<b>PyramidNet-110 <math>\alpha 270</math></b> <BN-Conv-BN-ReLU-Conv-BN-add>	Vanilla	36.52	N/A
	ResDrop	33.97	
	1-branch Shake	85.84	
	ShakeDrop	<b>32.44</b>	
<b>PyramidNet-200 <math>\alpha 300</math> Bottleneck</b> <BN-Conv-BN-ReLU-Conv-BN-ReLU-Conv-BN-add>	Vanilla	32.92	N/A
	ResDrop	32.17	
	1-branch Shake	78.12	
	ShakeDrop	<b>31.15</b>	
Methods	Regularization	Original (%)	w/ BN (%)
<b>Wide-ResNet-28-10k</b> <BN-ReLU-Conv-BN-ReLU-Conv-(BN)-add>	Vanilla	99.50	37.88
	ResDrop	99.50	45.80
	1-branch Shake	98.68	93.62
	ShakeDrop	91.11	<b>36.39</b>





# State-of-the-Art Comparisons

Method	Reg	Cos	Fil	Depth	#Param	CIFAR -10 (%)	CIFAR -100 (%)
Coupled Ensemble (Dutt et al., 2017)				118	25.7M	*2.99	*16.18
				106	25.1M	*2.99	*15.68
				76	24.6M	*2.92	*15.76
				64	24.9M	*3.13	*15.95
				-	50M	*2.72	*15.13
				-	75M	*2.68	*15.04
				-	100M	*2.73	*15.05
ResNeXt (Xie et al., 2017)		✓		26	26.2M	+3.58	-
				29	34.4M	-	+16.34
ResNeXt + Shake-Shake (Gastaldi, 2017)	SS	✓		26	26.2M	*2.86	-
				29	34.4M	-	*15.85
ResNeXt + Shake-Shake + Cutout (DeVries & Taylor, 2017b)	SS	✓	CO	26	26.2M	*2.56	-
				29	34.4M	-	*15.20
PyramidNet (Han et al., 2017b)				272	26.0M	*3.31	*16.35
		✓	RE	272	26.0M	3.42	16.66
PyramidDrop (Yamada et al., 2016)	RD			272	26.0M	3.83	15.94
	RD	✓	RE	272	26.0M	2.91	15.48
PyramidNet + ShakeDrop (Proposed)	SD			272	26.0M	3.41	<b>14.90</b>
	SD		RE	272	26.0M	2.89	<b>13.85</b>
	SD	✓		272	26.0M	2.67	<b>13.99</b>
	SD	✓	RE	272	26.0M	<b>2.31</b>	<b>12.19</b>

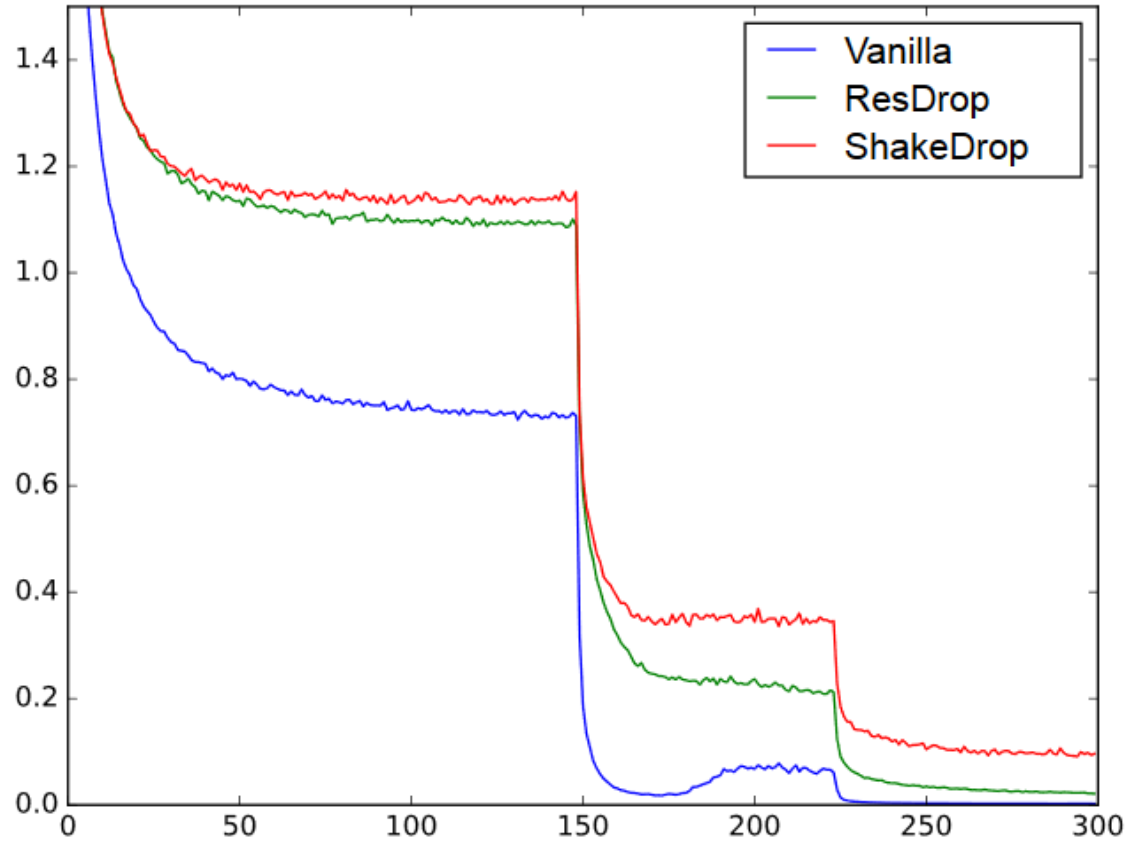


# State-of-the-Art Comparisons

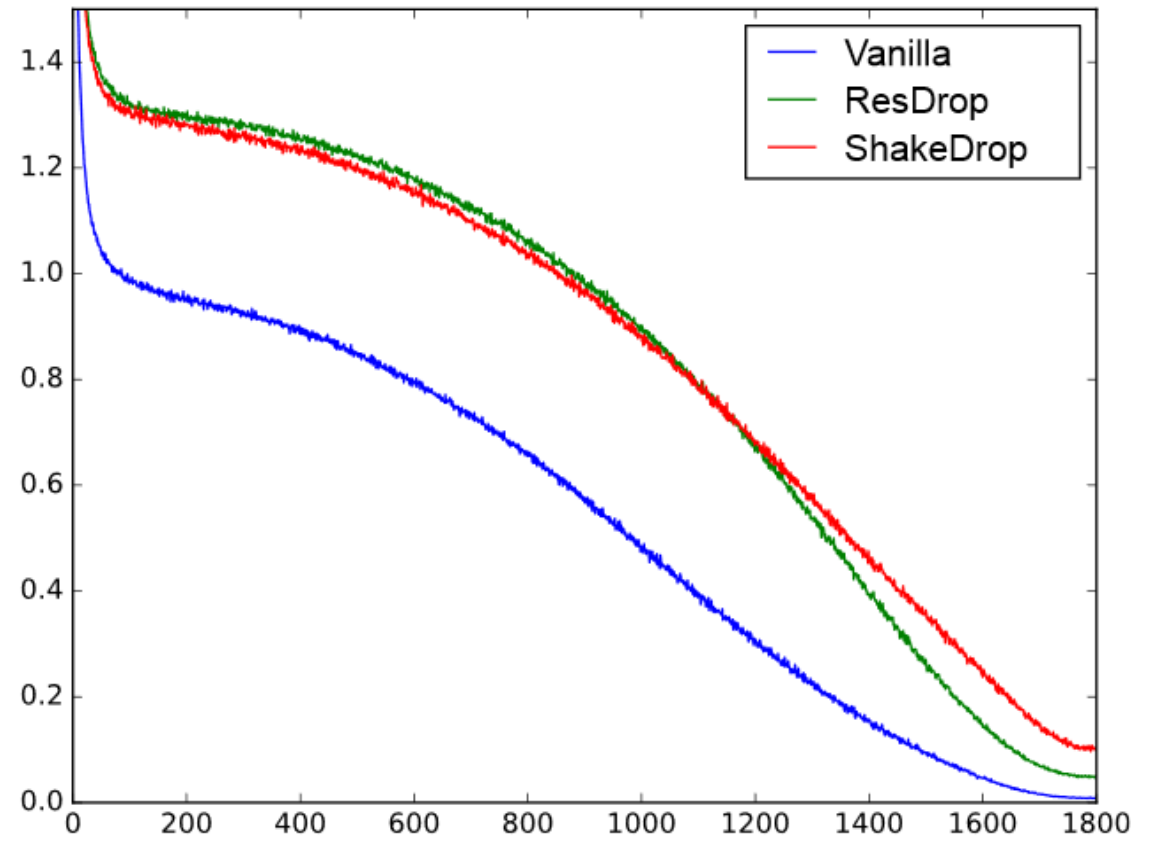
Method	Reg	Cos	Fil	Depth	#Param	CIFAR -10 (%)	CIFAR -100 (%)
Coupled Ensemble (Dutt et al., 2017)				118	25.7M	*2.99	*16.18
				106	25.1M	*2.99	*15.68
				76	24.6M	*2.92	*15.76
				64	24.9M	*3.13	*15.95
				-	50M	*2.72	*15.13
				-	75M	*2.68	*15.04
				-	100M	*2.73	*15.05
ResNeXt (Xie et al., 2017)		✓		26	26.2M	+3.58	-
				29	34.4M	-	+16.34
ResNeXt + Shake-Shake (Gastaldi, 2017)	SS	✓		26	26.2M	*2.86	-
				29	34.4M	-	*15.85
ResNeXt + Shake-Shake + Cutout (DeVries & Taylor, 2017b)	SS	✓	CO	26	26.2M	*2.56	-
				29	34.4M	-	*15.20
PyramidNet (Han et al., 2017b)				272	26.0M	*3.31	*16.35
		✓	RE	272	26.0M	3.42	16.66
PyramidDrop (Yamada et al., 2016)	RD			272	26.0M	3.83	15.94
	RD	✓	RE	272	26.0M	2.91	15.48
PyramidNet + ShakeDrop (Proposed)	SD			272	26.0M	3.41	<b>14.90</b>
	SD		RE	272	26.0M	2.89	<b>13.85</b>
	SD	✓		272	26.0M	2.67	<b>13.99</b>
	SD	✓	RE	272	26.0M	<b>2.31</b>	<b>12.19</b>



# Training Loss



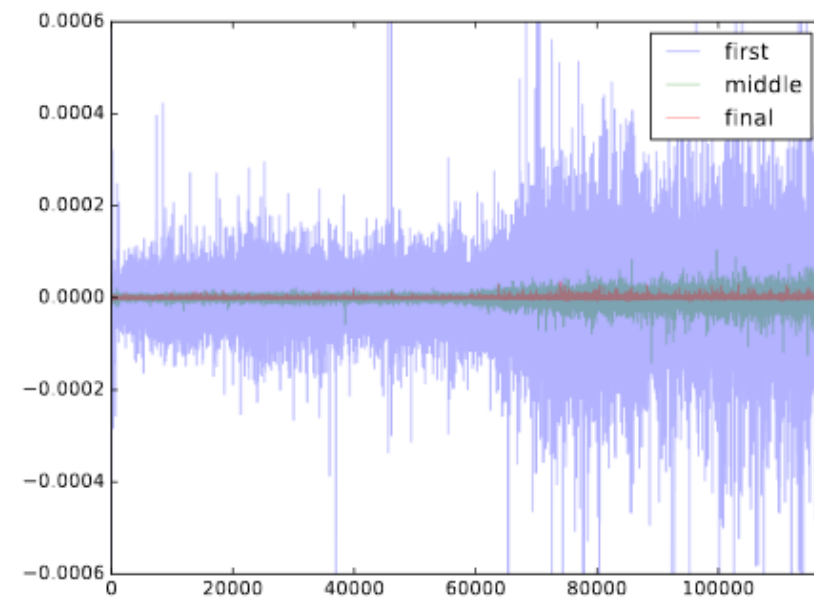
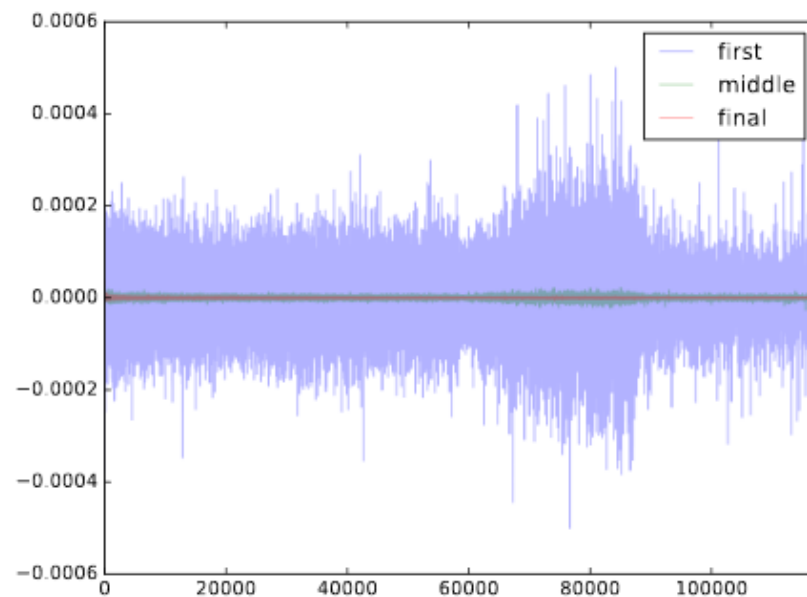
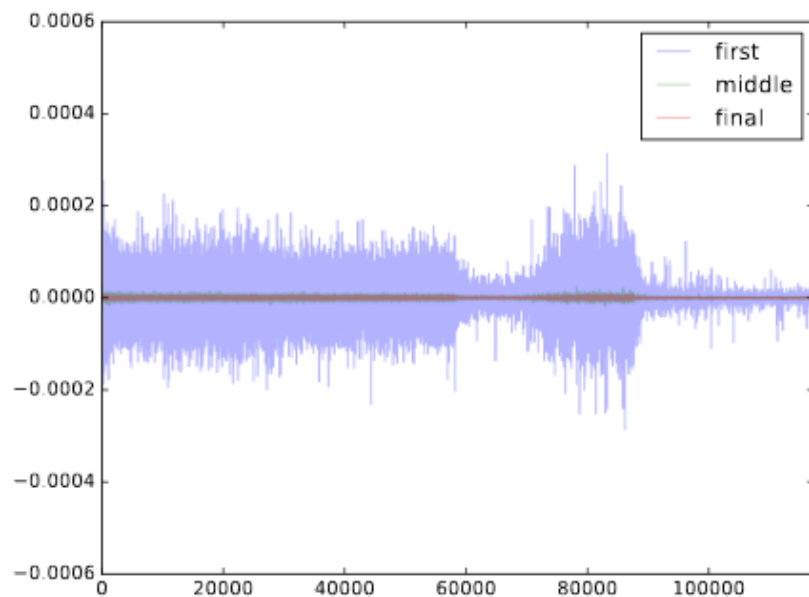
(a) 300-epoch training loss.



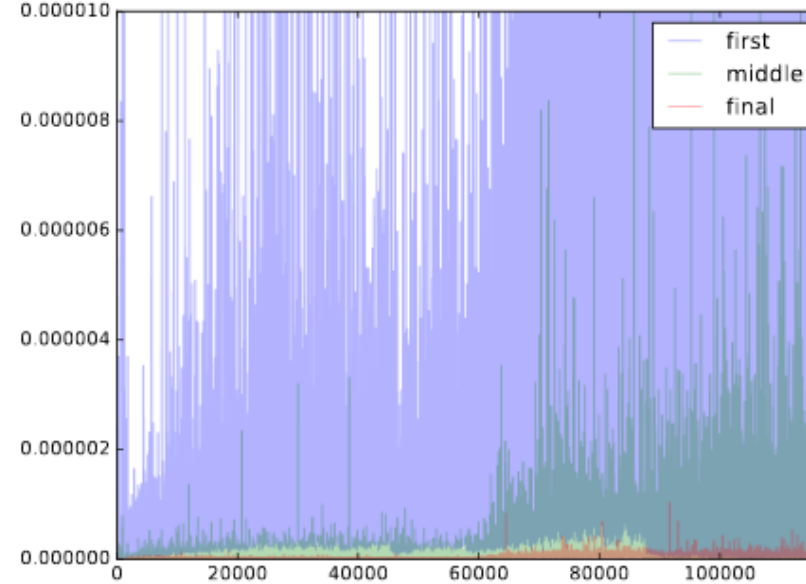
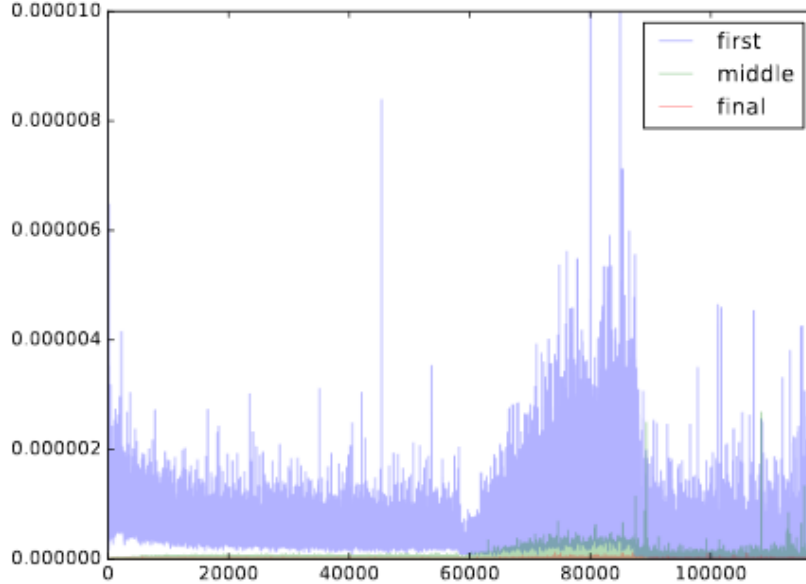
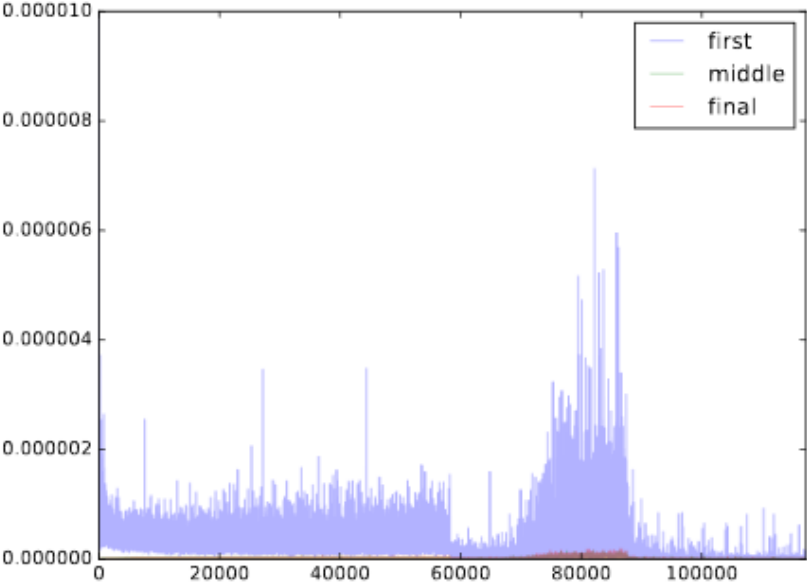
(b) 1800-epoch training loss.



# Gradient Averages During Training



# Gradient Variance During Training



# Conclusion & Critique

- ShakeDrop is a meaningful development for State-of-the-art image classification, improving classification accuracy across for all tested networks, without dramatically increasing the number of parameters used.
- Limited mathematical justification, relies heavily on intuition and empirical results



# References

- [Yamada et al., 2018] Yamada Y, Iwamura M, Kise K. ShakeDrop regularization. arXiv preprint arXiv:1802.02375. 2018 Feb 7.
- [He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proc. CVPR, 2016.
- [Zagoruyko & Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Proc. BMVC, 2016.
- [Han et al., 2017] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In Proc. CVPR, 2017a.
- [Xie et al., 2017] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proc. CVPR, 2017.
- [Huang et al., 2016] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth. arXiv preprint arXiv:1603.09382v3, 2016.
- [Gastaldi, 2017] Xavier Gastaldi. Shake-shake regularization. arXiv preprint arXiv:1705.07485v2, 2017.
- [Loshilov & Hutter, 2016] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.
- [DeVries & Taylor, 2017b] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017b.
- [Zhong et al., 2017] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. arXiv preprint arXiv:1708.04896, 2017.
- [Dutt et al., 2017] Anuvabh Dutt, Denis Pellerin, and Georges Qunot. Coupled ensembles of neural networks. arXiv preprint 1709.06053v1, 2017.
- [Veit et al., 2016] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. Advances in Neural Information Processing Systems 29, 2016.

