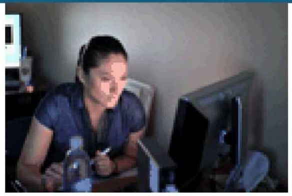
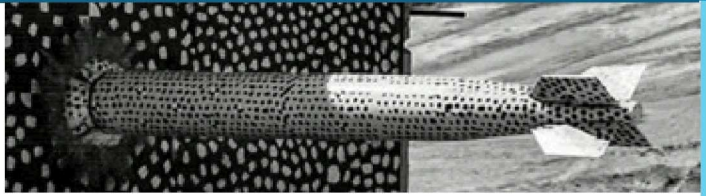
SAND2019-8957PE

Self-Normalizing Neural Networks with SELU Activation





PRESENTED BY

Tyler LaBonte, MARTIANS Intern (9365/9323)

July 25, 2019

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

It starts with a story…

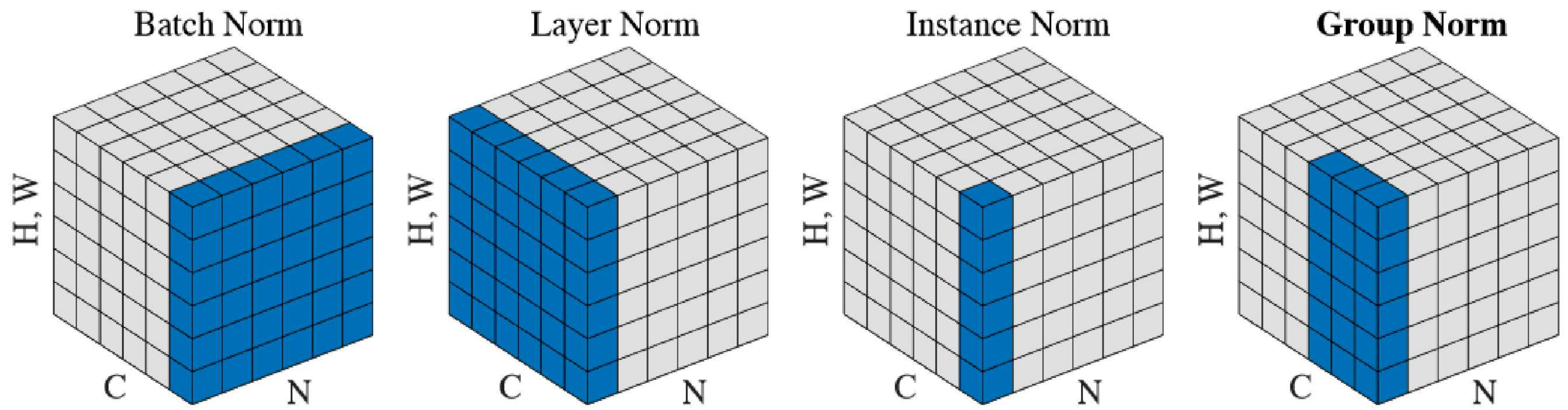
- Getting 98% train accuracy and NaN validation accuracy
  - Tried everything to diagnose, tracked it down to 1st batch normalization layer
- Batch norm normalizes activations based on mean, variance of batch
  - Increased training stability and resistance to vanishing gradients
- Batch norm uses Bessel's correction for the variance, where **N** is the batch size:

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

- NaN if **N = 1**…
- Or if **len(train) % batch_size = 1**

Other downfalls of batch normalization

- Needs batch size > 8 to obtain accurate batch statistics
- Can’t be used in RNNs because each time-step has different statistics
- Challenged by authors proposing Weight/Layer/Instance/Group/Spectral normalization
  - Nobody agrees on the “best” way to normalize



Ioffe & Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” 32nd International Conference on Machine Learning, 2015. https://arxiv.org/abs/1502.03167

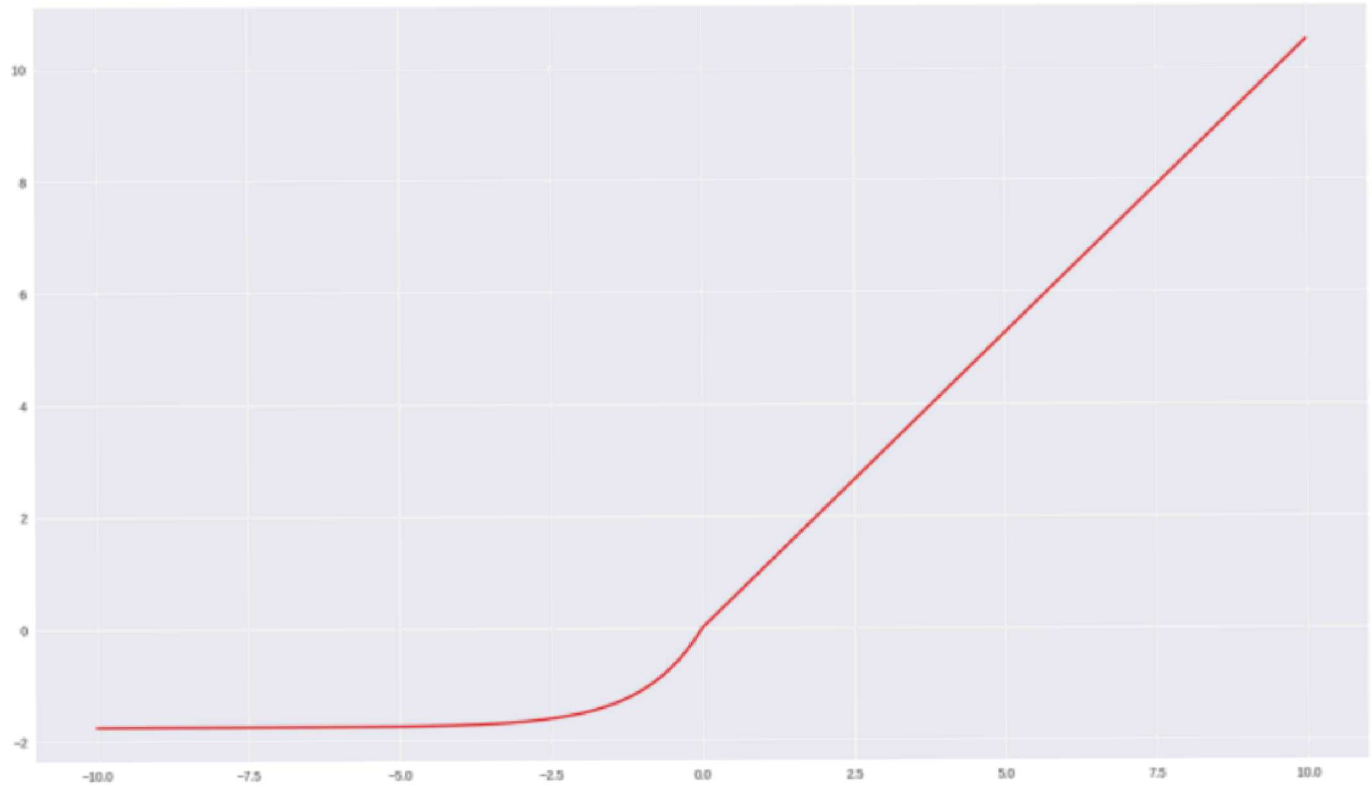
Wu & He. “Group Normalization.” 15th European Conference on Computer Vision, 2018. https://arxiv.org/abs/1803.08494

Wouldn't it be great if deep neural networks just *knew* how best to normalize?

Types of normalization

- Input normalization
  - Ex. Normalizing images between 0-1
- Training normalization
  - Ensures zero mean, unit variance between layers of the network
  - Ex. Batch normalization
- **Internal normalization**
  - Imposed on networks by virtue of their architecture
  - Normalization via activation

SELU Activation



$$\mathrm{selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leqslant 0 \end{cases}$$

Benefits of SELU

- Self-normalizing: automatically converges to zero mean, unit variance
- Allows training of very deep networks
- Allows strong regularization schemes
- Ensures learning robustness
- Theoretically, makes vanishing/exploding gradients impossible

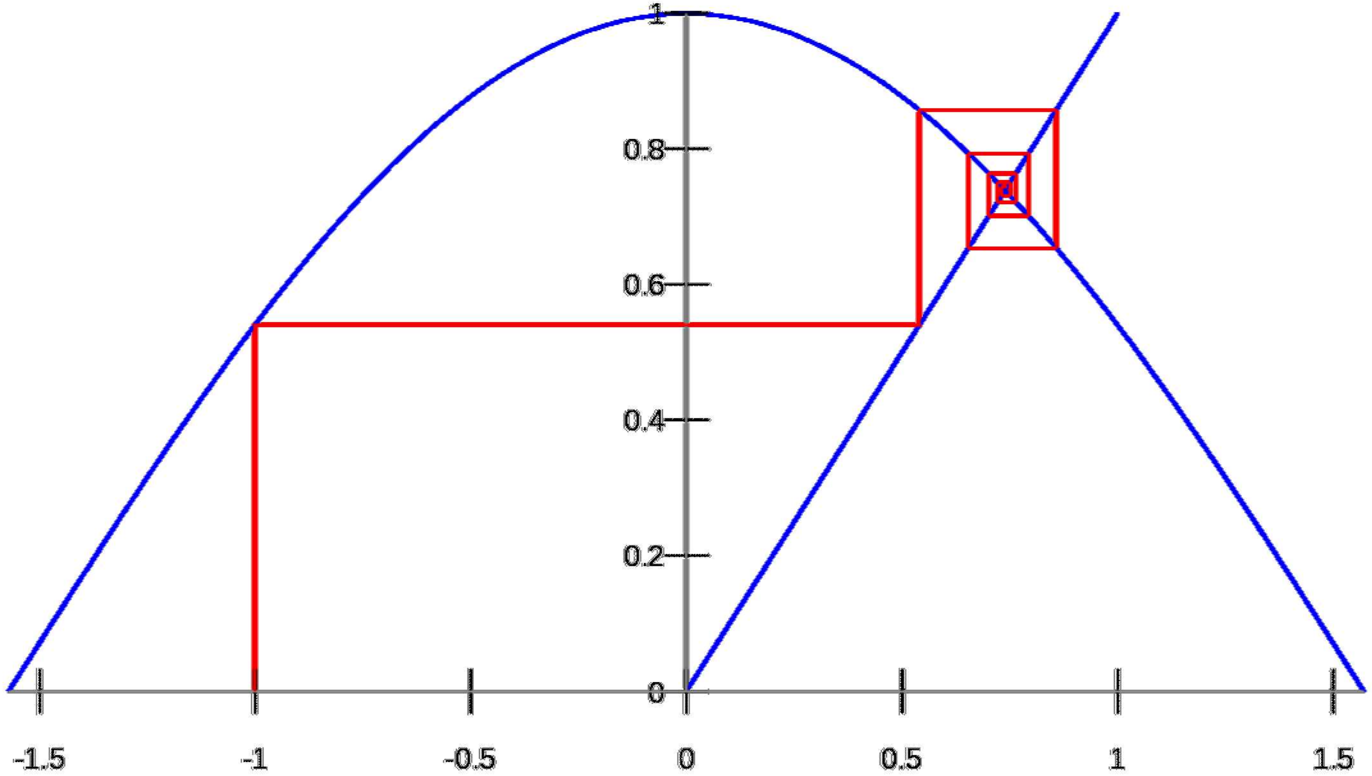
Klambauer et. al. "Self-Normalizing Neural Networks." 31st Conference on Neural Information Processing Systems, 2017.
https://arxiv.org/abs/1706.02515

How does SELU normalize?

- Fixed point of f: A point c with $f(c) = c$.
- Attractive fixed point of f: A fixed point c such that the iterative function

$$x, f(x), f\big(f(x)\big), \ldots \text{ converg}$$

- Ex: iterative cos converges to 0.739



How does SELU normalize?

- Deep neural network is an iterative function: $x, selu(x), selu(selu(x)), \ldots$
- Banach Fixed Point Theorem (1922): Proof of fixed point uniqueness and construction method
  - Applies to a complete metric space with a contraction mapping
  - 71 pages of proofs that BFPT applies to SELU networks
- Authors solve for fixed point $(\mu = 0, \upsilon = 1)$, obtaining $(\alpha = 1.67, \lambda = 1.05)$

$$\mathrm{selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leqslant 0 \end{cases}$$

SELU Success Stories

- Enables very deep, stable networks
- 13,654 input, 18-layer feedforward NN for patient mortality
- Improved convergence & reliability of actor-critic reinforcement learning
  - "SELU is unexpectedly good"
- 50+ layer CNNs only converge with SELU

https://github.com/bioinf-jku/SNNs/tree/master/SNN-successes

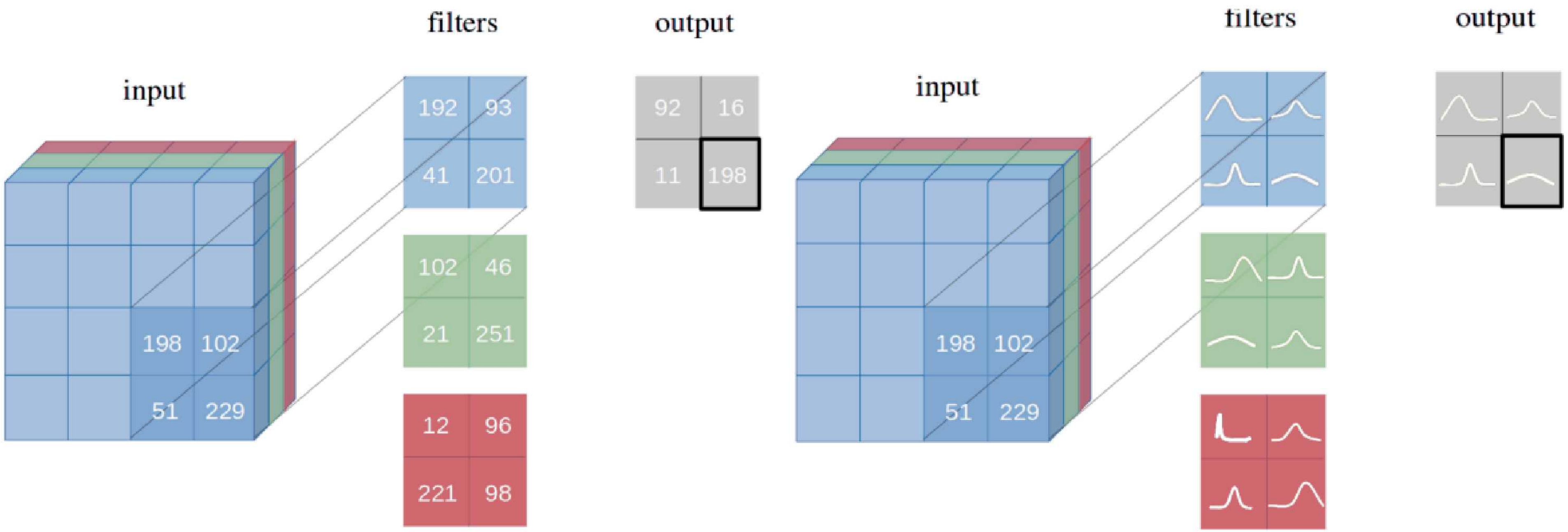
Huang et. al. "Learning to Run with Actor-Critic Ensemble." 31st Conference on Neural Information Processing Systems, 2017. https://arxiv.org/abs/1712.08987

Avati et. al. "Improving Palliative Care with Deep Learning." IEEE International Conference on Bioinformatics and Biomedicine, 2017. https://arxiv.org/abs/1711.06402

Molina & Vila. "Solving Internal Covariate Shift in Deep Learning with Linked Neurons." IEEE Conference on Computer Vision and Pattern Recognition, 2018. https://arxiv.org/abs/1712.02609

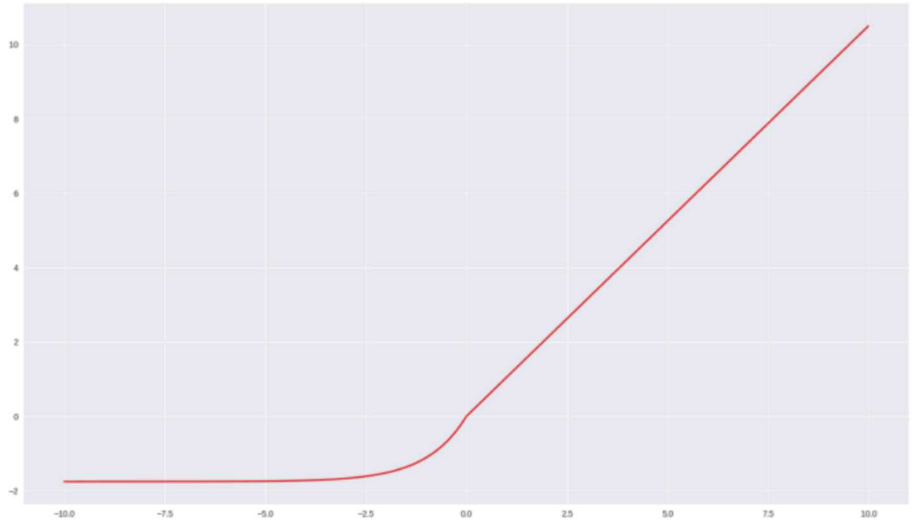
Bayesian Convolutional Neural Networks

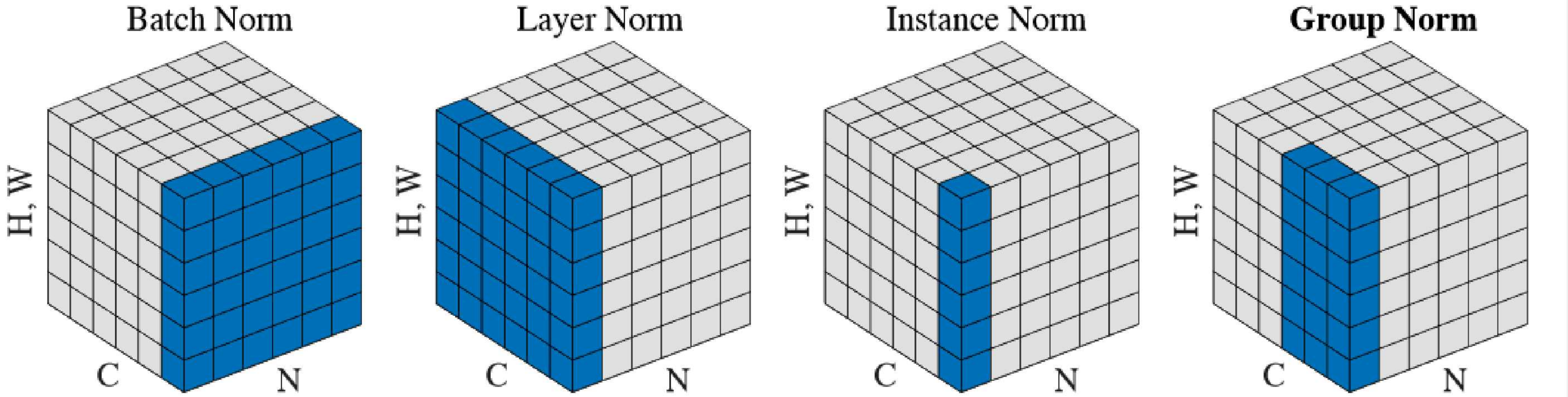
- Learn distributions over weights rather than point-estimates
- Great for uncertainty quantification
- Difficult to train
  - Double the parameters of a normal model
  - Susceptible to vanishing gradients



Normalization Alternatives for BCNN

- Batch normalization
  - Normalizes activations per batch
- Group normalization
  - Normalizes groups of channels independent of batch size
- SELU
  - Self-normalizes via activation function





Experimental Results

- Trained Bayesian U-Net on battery segmentation dataset of 1008x1008 slices, batch size 8

# of Conv Layers	Normalization	Best Validation Accuracy
14 (Param ~160k)	None	0.9639
	Batch Norm	0.9681
	Group Norm	**0.9758**
	SELU	0.9745*
19 (Param ~670k)	None	0.9699
	Batch Norm	0.9718
	Group Norm	**0.9780**
	SELU	0.9691*
24 (Param ~2.7m)	None	0.9692*
	Batch Norm	0.9685
	Group Norm	**0.9778**
	SELU	0.5751

Experimental Analysis

- SELU accuracy peaks and then rapidly decays
  - Hypothesis: Bayesian layers are initialized with a $N(0,1)$ prior, not the LeCun-normal initialization that the authors worked with (which has way smaller variance)
  - SELU may be sensitive to the weight initialization (possibly different α and λ)
  - Further research: implement Bayesian prior similar to LeCun-normal
- Group norm beats batch norm due to small batch size

Conclusion

- SELUs are a self-normalizing activation function for deep networks
  - They work by cleverly converging to the fixed point of zero mean unit variance
- Applications in feedforward networks, CNNs, RNNs, reinforcement learning
- In actuality: best normalization technique highly situational
  - Batch normalization in generic cases
  - Group normalization when batch size is low
  - SELU in deep cases where LeCun-normal initialization is possible
    - Not supported by my experiments, but proven body of literature