

One Initialization to Rule Them All?

Stefan Wijnja, Ellis Wierstra, Thomas Hamburger

January 27, 2020

Neural networks are very cool

- ▶ Images of cool applications.

But they can be very demanding

- ▶ Numbers for huge networks and how long it takes to train them.

One possible solution: pruning

- ▶ Explain pruning, iterative pruning, discuss known results so far.

Pruning can lead to Winning Tickets

- ▶ Explain what winning tickets are.

Question

Does weight initialization matter looking for Winning Tickets?

Method

Plan

1. Pick dataset.
2. Pick model.
3. Pick initialization methods.
4. **Write the code.**
5. **Do lots of experiments.**
6. Answer the question.

Method

Dataset: MNIST



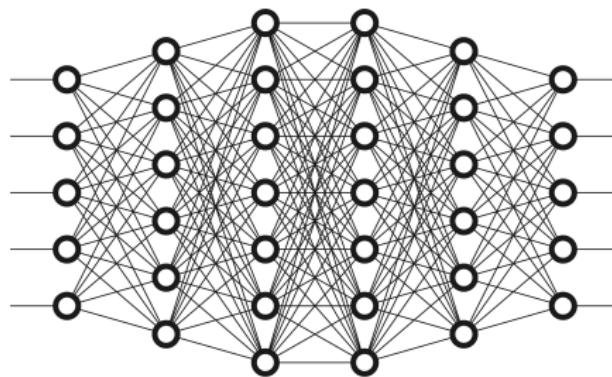
A 4x10 grid of handwritten digits. The digits are arranged in four rows. The first row contains ten '0's. The second row contains ten '1's. The third row contains ten '2's. The fourth row contains ten '3's. The fifth row contains ten '4's. The sixth row contains ten '5's. The seventh row contains ten '6's. The eighth row contains ten '7's. The ninth row contains ten '8's. The tenth row contains ten '9's. The digits are written in a cursive style and are slightly irregular.

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9

- ▶ Handwritten digits
- ▶ 28×28 pixels
- ▶ Training set: 60k
- ▶ Test set: 10k

Method

Model: LeNet (Lecun et al. 1998)



- ▶ Fully connected
- ▶ Two hidden layers: 300 & 100 neurons → 266k weights.
- ▶ Leaky ReLU (negative slope: 0.05)

Method

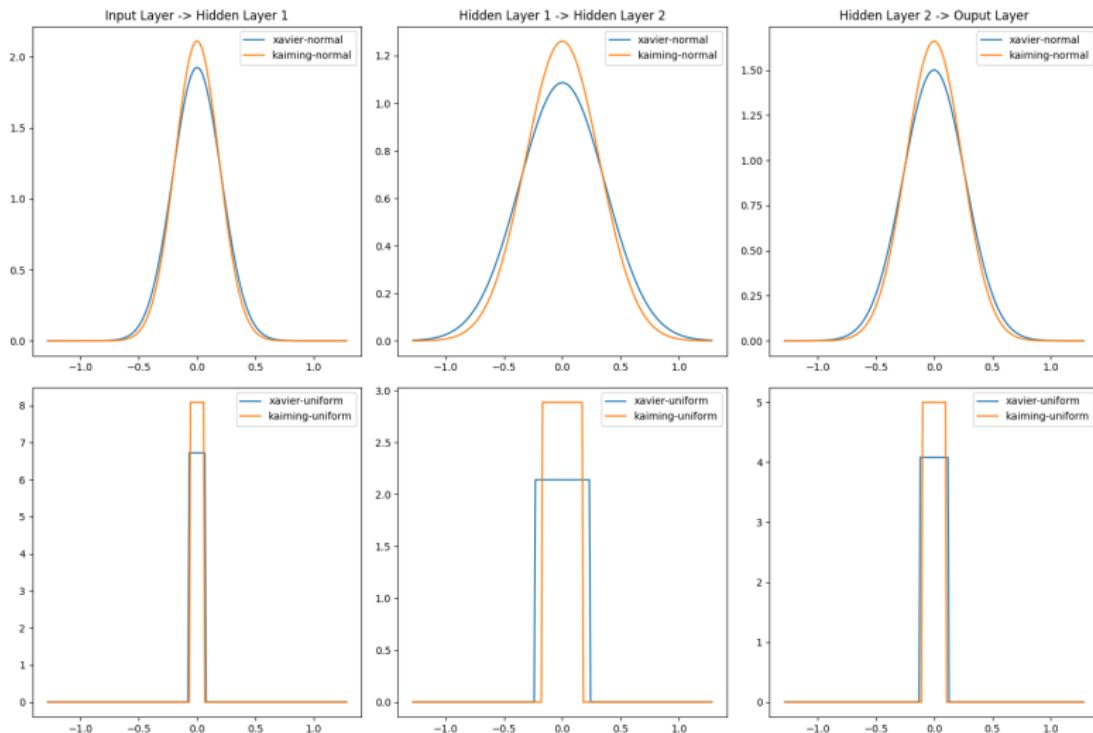
Initialization Methods

	Xavier	Kaiming
$\mathcal{U}(-a, a)$	$\sqrt{\frac{6}{\text{fan_in}+\text{fan_out}}}$	$\sqrt{\frac{3}{\text{fan_mode}}}$
$\mathcal{N}(0, \text{var})$	$\sqrt{\frac{2}{\text{fan_in}+\text{fan_out}}}$	$\frac{1}{\sqrt{\text{fan_mode}}}$

- ▶ $\text{fan_in}/\text{fan_out}$: number of inputs to/outputs from a neuron.
 fan_mode : fan_in or fan_out .

Method

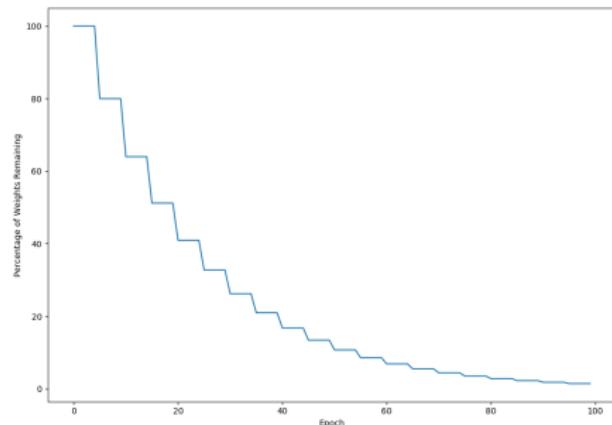
Initialization Methods' Probability Density Functions



Method

Training

1. Initialize network and save weights.
2. Train for 5 epochs.
3. Trim smallest 20% of the weights.
4. Reset the weights to the saved ones.
5. GO TO 2



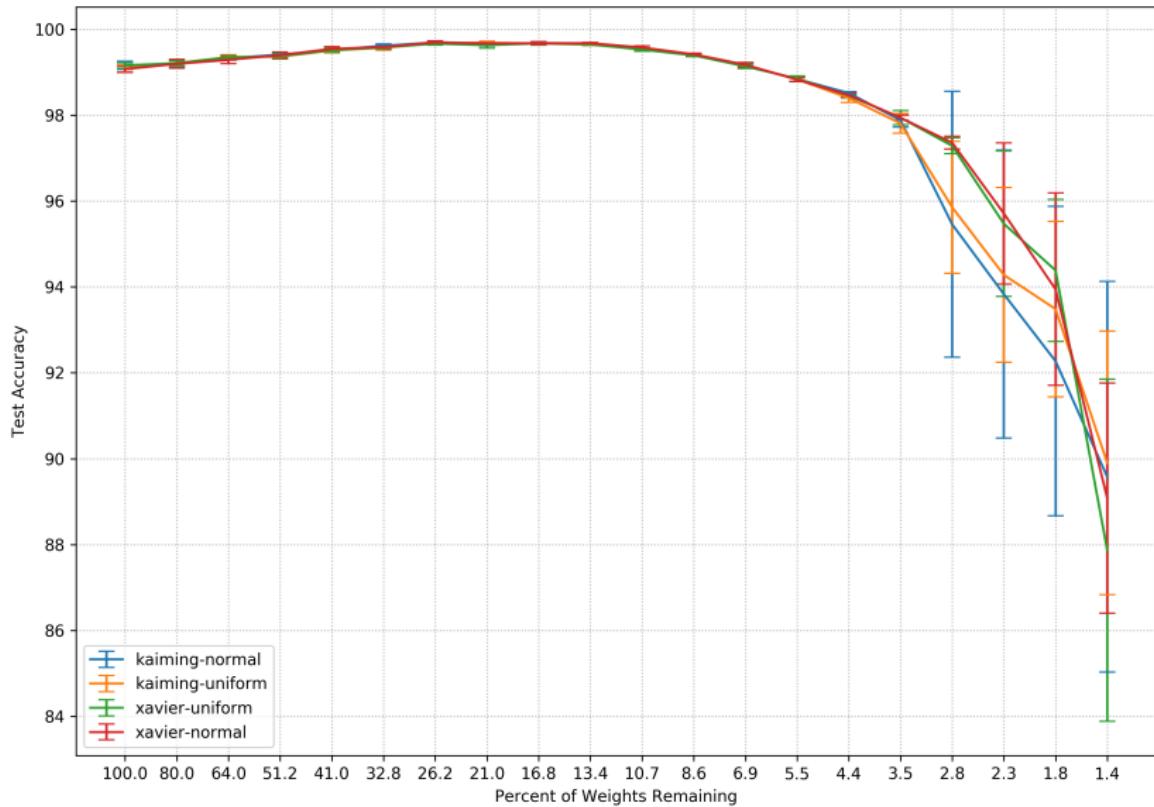
Method

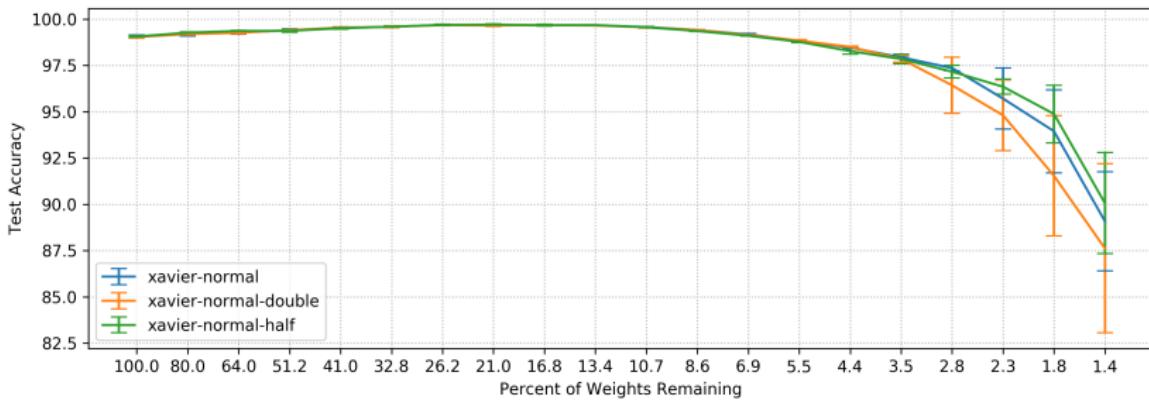
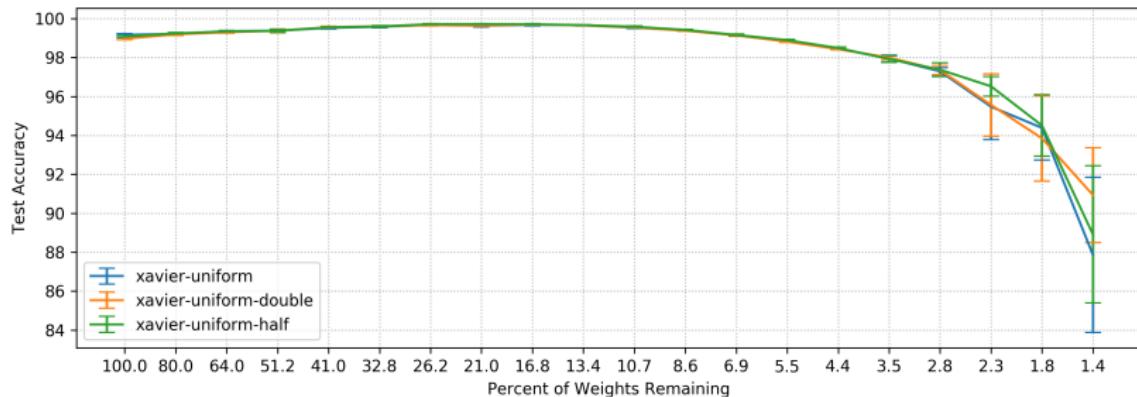
Testing

- ▶ Save loss and accuracy on the training and testing set after every epoch.



Results





Further Research

Get interns to try:

- ▶ More datasets.
- ▶ More models.
- ▶ More pruning rates and pruning intervals.
- ▶ More pruning methods.

Questions?

Lecun, Y, L Bottou, Y Bengio, and P Haffner. 1998.
“Gradient-Based Learning Applied to Document Recognition.”
Proceedings of the IEEE 86 (11): 2278, 2324.