EPFL

[Re]: Numerical Influence of ReLU'(0) on Backpropagation

Tommaso Martorella,
Héctor Ramírez

NeurIPS 2023
ReScience C

Scope of Reproducibility

First let's consider that:

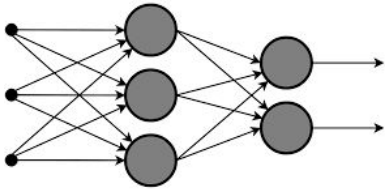
- ★ It is important to be aware of the implications that the **hardware** and **technical limitations** can have in the model training, architecture selection, memory and computational cost.
- ★ In the theoretical framework, there are no memory or bit-precision, nevertheless, **computational limitations must be taken into account** when training the model.
- ★ We also ran **additional experiments** to further extend the authors' idea of the value of the subgradient being a hyperparameter to tune during training.

Scope of Reproducibility

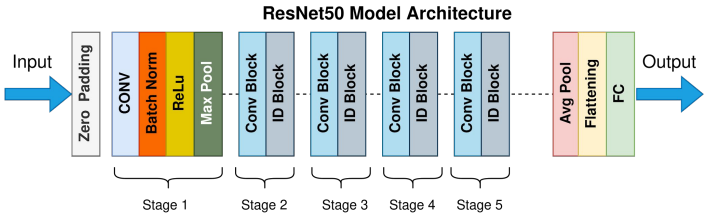
- ★ The choice of ReLU'(0) becomes **computationally meaningful** and influences the training and test accuracy.
- ★ There's the **trend** to lower the precision to make the model training more efficient in terms of **energy, memory** and resources
  → Which are the implications?
- ★ The main takeaway is that the arbitrary choice of mathematically negligible factors (such as the ReLU'(0)) **might not be computationally negligible.**

Models

- Fully connected NN



- ResNet 18



- MobileNet V3

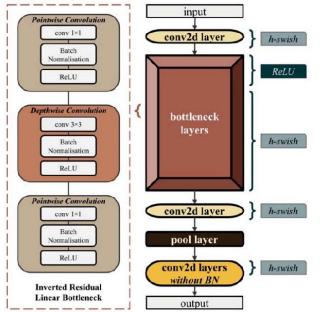


Figure 1 MobileNetV3 architecture. The general architectures are the same for both MobileNetV3-Large and MobileNetV3-Small.

Tommaso Martorella, Héctor Ramírez

Data

- MNIST
- Fashion MNIST
- 3-layer grayscale MNIST for MobileNetV3

Tommaso Martorella, Héctor Ramírez

Hyperparameters

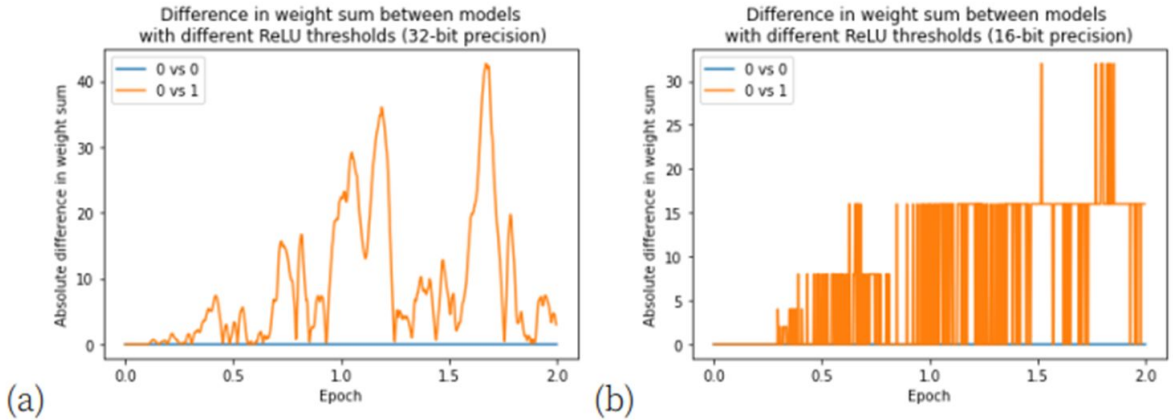
- Precision: 16 or 32-bit
- Model Architecture
- Activation function: ReLU vs ReLU6 (vs LeakyReLU)
- Value of subgradient(s)

For all experiments:

- Optimizer → ADAM with ɣ=0.001
- Batch Size: 128

Initial Experiment

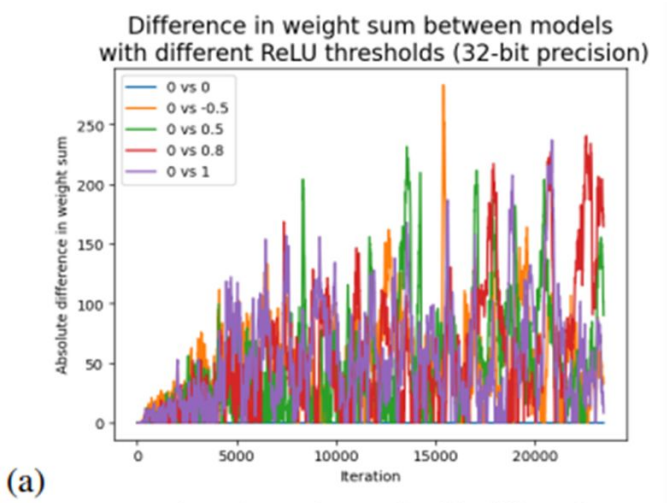
- Compare weight difference of the parameters with 16 and 32 bit precision models.
- Use of L1 norm between weight matrices.
- No performance comparison, just magnitude change.
- Tested in the interval [0,1] for subgradient value.
- At low precision the change is more unstable.

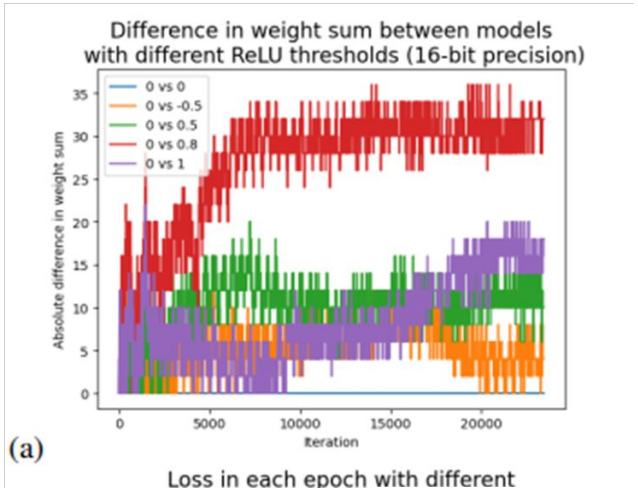


Tommaso Martorella, Héctor Ramírez

Fully Connected NN

- Made a volume comparison (L1 norm) between the models using 32 and 16 bit precision.
- Differences between the choice of RELU'(0) became evident.
- The difference between weight matrices is considerable.

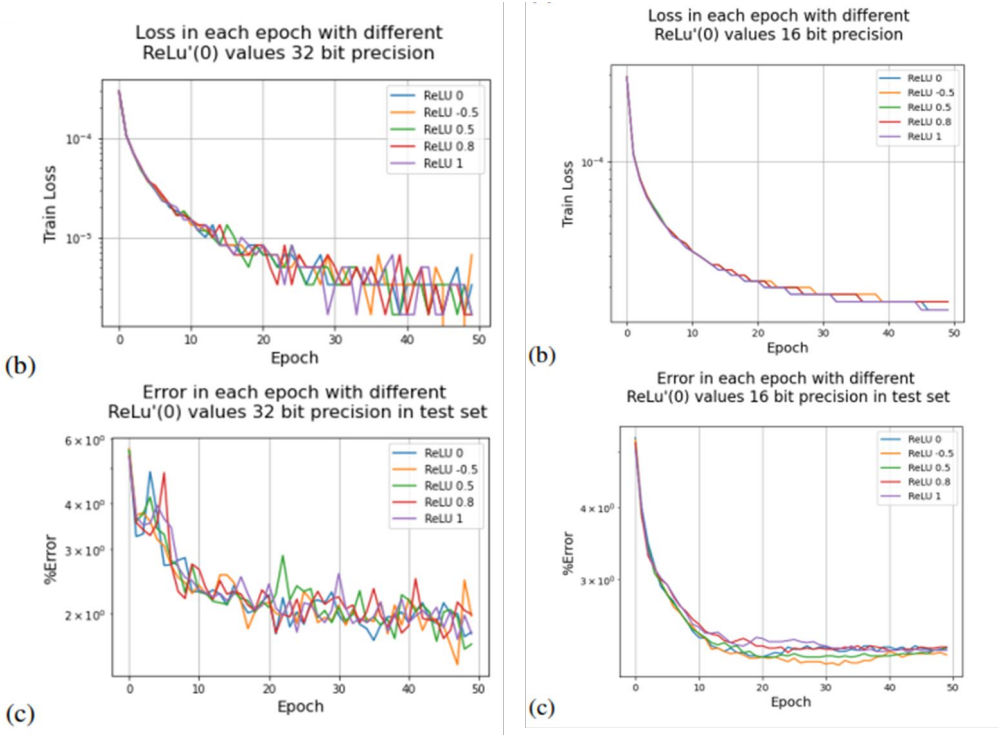




Tommaso Martorella, Héctor Ramírez

Fully Connected NN

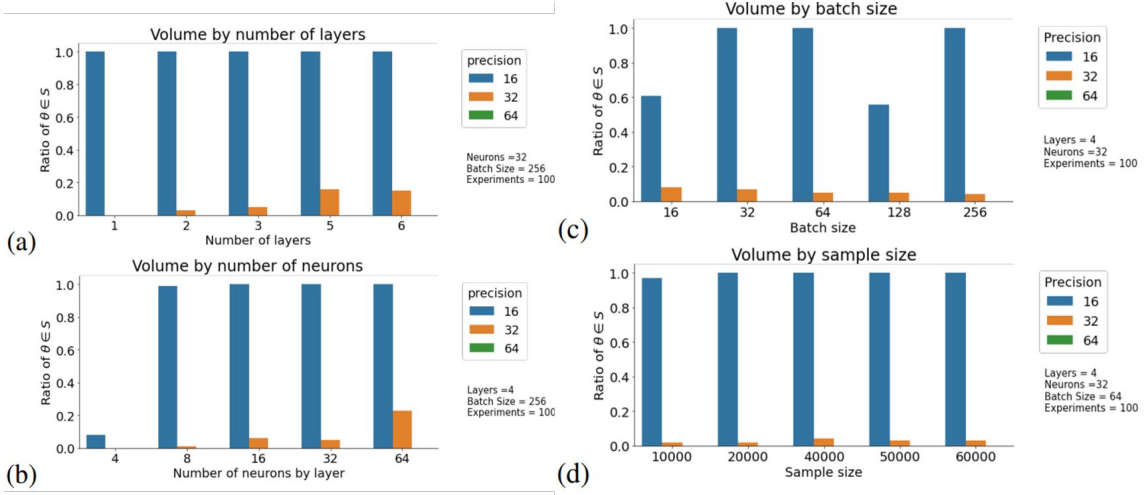
- Performance is also affected by the decision of subgradient value.
- At 32 bit train loss and error are unstable, but seem to oscillate the same mean value.
- At 16 bit precision there is less variation but there's an offset in both results (small for this model)



Tommaso Martorella, Héctor Ramírez

Fully Connected NN

- Understand how these hyperparameters are affected by the decision of taking a lower precision.
- Compared number of layers, batch size, number of neurons and sample size.

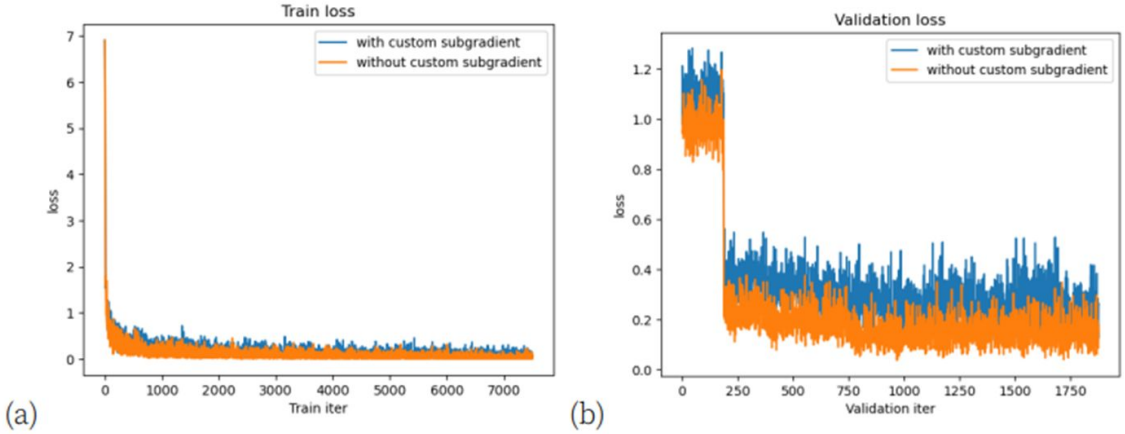


Tommaso Martorella, Héctor Ramírez

Tommaso Martorella, Héctor Ramírez

Subgradient as an Hyperparameter: MobileNet V3

- Hyperparameter tuning for MobileNet V3 with and without the subgradient as an hyperparameter.
- With the default subgradient, performance was more stable across different hyperparameter configurations and reached a better overall performance.



Summary

- In theory the value of the surrogate of the subgradient in a non-differentiable point should not impact the outcome.
- Even though in theory this is correct, when using numerical methods to perform a backpropagation, altogether with numerical bit-precision it becomes relevant as rounding errors can lead to different solutions.
- As the use of 32-bit precision is widely used as a standard in neural network training and as 16-bit is becoming a trend to speed up the training in GPUs and energy saving, the choice of subgradient becomes relevant.