DeepHochuli: Numerical Precision Analysis in Deep CNNs

Undergraduate Research Project (Iniciação Científica) - UFPR

Pietro Comin

December 3, 2025

Abstract

This collaborative research (conducted with colleague Thamiris Fischer) investigates the theoretical and practical implications of the non-differentiability of the Rectified Linear Unit (ReLU) activation function at exactly zero ($z = 0$). We analyze how this singularity interacts with network depth and floating-point precision (FP16, FP32, FP64). Specifically, we aim to understand if numerical rounding errors in lower precision formats artificially trigger "dead neurons" by collapsing small, non-zero activations to zero, thereby halting gradient flow.

1 Problem Statement

In modern Deep Learning, the ReLU activation function, defined as $f(x) = \max(0, x)$, is non-differentiable at the origin. While sub-gradient descent typically handles this by arbitrarily assigning a gradient of 0 or 1, the numerical representation of inputs near zero is critical. Our primary investigation focuses on whether hardware-level constraints (such as TensorFloat-32 on NVIDIA GPUs) introduce sufficient noise or rounding errors (swamping) to impact the training dynamics of deep architectures, potentially increasing the sparsity of the network unintendedly.

2 Methodology

We developed a custom PyTorch framework to:

- Track the distribution of pre-activations (z) magnitude across layers during training.
- Isolate and compare arithmetic behavior between CPU (IEEE 754 standard) and GPU (NVIDIA Ampere architecture using TF32).
- Simulate "Swamping" effects to determine the minimum representable signal in the presence of large accumulators.

3 Preliminary Observations

3.1 Numerical Absorption (Swamping)

We empirically verified that in single-precision (FP32), the addition of small perturbations to larger activation values results in information loss due to mantissa truncation. Our tests confirmed that for an accumulator magnitude of 1.0, signals smaller than $\approx 10^{-8}$ are mathematically erased (absorbed), effectively creating a "blind spot" for the network.

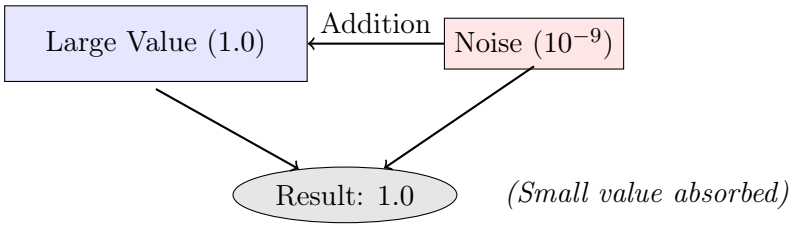
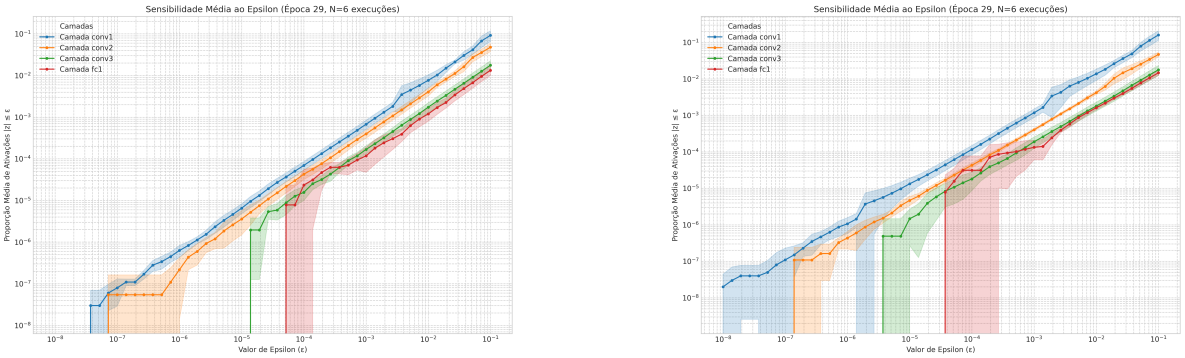


Figure 1: Conceptual visualization of the "Swamping" phenomenon observed in FP32. Small gradients or activations fail to update the accumulator due to precision limits.

3.2 Activation Magnitude Thresholds

Analysis of the pre-activation values (z) suggests a strong linear decay in the probability of observing small values. As shown in Figure 2a, the distribution of values cuts off around 10^{-8} for FP32.

Interestingly, FP64 (Double Precision) experiments exhibited a similar cutoff range. This leads to two main current hypotheses: first that hardware or software implementations may be the cause to this limit or second that it might be driven by the statistical rarity of such small sums in the forward pass (a "Probability Barrier") or bias anchoring, rather than purely hardware limitations. This hypothesis is currently under rigorous testing.



(a) Float32 (Standard)

(b) Float64 (Double Precision)

Figure 2: **Probability Barrier Comparison.** Log-Log plots of activation magnitudes for FP32 (left) and FP64 (right). Despite FP64's capability to represent values down to 10^{-308}, the distribution naturally cuts off around 10^{-9} in both cases.

3.3 Architecture Divergence

We identified a significant discrepancy between CPU and GPU calculations. Due to the usage of TensorFloat-32 (TF32) on modern GPUs (which uses a 10-bit mantissa compared to the 23-bit mantissa of standard FP32), arithmetic errors in the forward pass were observed to be up to **830x larger** on the GPU compared to the CPU baseline.

4 Next Steps

The research is now moving towards the **Backward Pass analysis**. We are implementing hooks to capture gradient flows (∇w) to determine if the numerical absorption proven in Section 3.1 is causing "vanishing gradients" at a microscopic scale, which would permanently kill neurons that attempt to recover from zero.

Note: This project contains confidential code and data. This document serves as a brief description of the current ongoing investigations, it does not include all the results and thoughts developed throughout the research process.