

Peer Review

Review of: "Neural Networks Use Distance Metrics"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper asks a simple question with a long history behind it: when a hidden unit fires, does a large value mean "the feature is strongly present" (what the author calls an intensity metric), or does a small value mean "the input is close to a learned prototype" (a distance metric)? Building on the author's earlier theoretical note linking linear units with absolute-value activation to the Mahalanobis distance (reference [5]), this preprint tries to settle the question empirically. A one-hidden-layer network, $x \rightarrow \text{Linear}(784) \rightarrow \text{Perturbation} \rightarrow \text{Activation (ReLU/Abs)} \rightarrow \text{Linear}(10) \rightarrow y$ (Eq. 1), is trained on MNIST twenty times per activation. After training, a frozen perturbation layer applies $y = \min(\text{scale} \cdot x + \text{offset}, \text{clip})$ to the pre-activations, and accuracy is measured as the offset, the scale, or the clip threshold is swept.

The headline result is an asymmetry. Small offsets (a few percent of each unit's range) damage accuracy quickly, while scaling over a wide range does not. From this, the paper concludes that both ReLU and Abs networks rely on distances to decision boundaries rather than on activation magnitude, and, in §7, that the experiments provide "empirical validation for the theoretical connection between neural networks and Mahalanobis distance". A secondary claim is that Abs networks are more sensitive to offsets than ReLU networks because the Abs fold selects a single feature, while ReLU selects a set (§6, Figures 2 and 3).

2. Whether the conclusion holds on the data

I like several things here, and I say so first because they are real. The design is minimal and easy to follow: the perturbation layer is present during training with identity parameters, so the trained weights are not changed by the probe. The author repeats each experiment twenty times, reports confidence bands (Figure 1 caption), and prints every number in the appendix tables instead of hiding them in plots.

The code is linked (footnote 1). The conceptual pictures in Figures 2 and 3, where ReLU accepts a set of features $\{a, b, c\}$ and Abs isolates one, give a concrete and testable story for why the two activations respond differently. And §6.3 and §6.4 openly admit that the intensity side of the comparison is not well defined. That honesty makes the paper easier to review.

The raw numbers support the descriptive part of the story. Table 4 shows Abs dropping from 99.99 (row Baseline) to 86.25 at an offset of -10% and to 81.40 at +10%, while ReLU goes from 98.33 to 96.55 and 96.31 at the same offsets. Table 2 shows almost no change for Abs between the 10% row (99.98) and the 1000% row (99.99). So the asymmetry exists. My concerns are about what the asymmetry can tell us.

2.1 The scaling test is close to a mathematical identity. Both ReLU and $|\cdot|$ are positively homogeneous: for $p > 0$, $\text{ReLU}(pz) = p\text{ReLU}(z)$ and $|pz| = p|z|$. Because the perturbation sits before the activation (Eq. 1) and is followed by a single linear layer, scaling every unit by p turns the logits into $pW_2h + b_2$. The argmax of that vector changes only through the output bias b_2 . This predicts exactly the pattern in Table 2: no effect at large p (row 1000%: 99.99 for Abs, 98.33 for ReLU) and a drop only when p is small enough for b_2 to dominate (row 1%: 94.76 for Abs, 75.33 for ReLU). Any network of this shape would show that invariance, whatever its units represent. I would therefore not count the scaling result as evidence against the intensity interpretation. The last sentence of §6.4's scaling paragraph ("this begins to sound suspiciously like a distance metric") comes close to this point; I think it deserves to be made explicitly and moved into §4.

2.2 The offset test would hurt an "intensity" network too. Adding pr_i to each pre-activation changes which ReLU units are on and shifts every surviving activation by a unit-specific constant. After W_2 , that becomes a class-dependent shift of the logits. A reader who holds the intensity view would also predict an accuracy loss here. The paper does not state what the intensity hypothesis predicts for offsets, so the experiment does not yet discriminate between the two hypotheses; it shows that the network is sensitive to bias shifts. A matched control would help a lot: for example, a post-activation offset (a pure magnitude shift that leaves the boundary alone), or random perturbations with the same L2 norm as the offset at each p .

2.3 The perturbations are not matched in size. A 50% scale moves an activation of size v by $0.5v$, which for large units approaches $0.5r$. A 10% offset moves every unit by $0.1r$. So the "large" intensity perturbation is often bigger in absolute terms than the "small" distance perturbation that breaks the model. The abstract's contrast between "large" intensity and "small" distance perturbations, repeated in

§5 ("relatively small distance offset perturbations"), rests on these words, and I would like to see them expressed in one common unit.

2.4 The statistics do not match the text. §5 states that fluctuations under scaling and clipping were "not statistically significant ($p > 0.05$), as detailed in Table 2 and Table 3". In Table 2, ReLU has $p = 2.9\text{e-}09$ at scale 10%, $1.7\text{e-}04$ at 25%, and $1.6\text{e-}03$ at 50%. In Table 3, every p -value for both models is below $1\text{e-}12$, including the 50% and 75% rows. Only four Abs rows in Table 2 (25%, 50%, Baseline, 1000%) have $p > 0.05$. There is a second puzzle. Each table has a Baseline row compared "to baseline performance," yet its t -statistic is not zero: Table 2, Baseline, ReLU, $t = -3.6$, $p = 1.7\text{e-}03$; Table 3, Baseline, Abs, $t = -26.5$; Table 4, Baseline, Abs, $t = -42.8$. A condition compared with itself should give $t = 0$. The sign of t also does not follow the accuracy: in Table 4, Abs, the -10% row (86.25, well below baseline) has $t = -31.5$, the same sign as the Baseline row, while the -25% row (49.03) has $t = +32.0$. I could not reconstruct from the text which reference value or which test produced these columns. Until that is fixed, I would not rely on any p -value in Appendix A.

2.5 Evaluation on the training set. §4.2 explains the choice, and I understand it. Still, Table 1 shows a generalization gap: Abs training accuracy 99.99 against test accuracy 95.29, a gap of $99.99 - 95.29 = 4.70$ points; for ReLU, $98.33 - 95.61 = 2.72$ points. A network that has nearly memorized its training set may be unusually fragile to any perturbation of its boundaries. Repeating Table 4 on the test set would show whether the offset sensitivity is a property of the learned representation or of memorization.

2.6 The conclusion goes further than the experiment. Nothing in §4 or §5 measures a Mahalanobis distance. The claim in §7 of "empirical validation" for the link proposed in [5] would need a direct test, for instance comparing each Abs unit's output with the Mahalanobis distance of the input along a principal component of the class-conditional pre-activations. The abstract's phrase "robust performance under large intensity-based perturbations" also covers clipping, but Table 3 shows Abs at 52.90 and ReLU at 75.93 for the 1% cutoff, and Abs at 88.55 for 20%. The abstract should say that clipping above roughly half the maximum is tolerated, not clipping in general.

Taken together: the phenomenon is real and interesting, but the current controls cannot tell the two readings apart, and the statistical columns need to be recomputed. These are fixable, and on MNIST they are cheap to fix.

3. Reproducibility

The good news first. The implementation is public (footnote 1), the learning rate (0.001), loss (cross-entropy), optimizer (SGD, full batch), number of epochs (5000), normalization constants ($\mu = 0.1307$, $\sigma = 0.3081$), and number of repeats (20) are all in §4. That is more than many papers of this size provide. I did not run the repository, so I cannot confirm that it reproduces the tables.

Several things are missing or unclear in the text itself.

- **Hidden width.** Eq. 1 writes `Linear(784)`. I cannot tell whether 784 is the input size, the hidden size, or both. The number of hidden units matters for every per-unit percentage in §4.2.
- **Seeds and initialization.** The twenty runs presumably differ by seed, but seeds, initialization scheme, and whether biases are used in both layers are not stated. Point 2.1 depends on b_2 , so this is not a formality.
- **Optimizer wording.** "Trained on the entire MNIST dataset (rather than using minibatches)" with "Stochastic Gradient Descent" describes full-batch gradient descent. It would help to say so and to confirm that "entire MNIST dataset" means the 60,000 training images only, since Table 1 also reports test accuracy.
- **How r is computed.** §4.2 says percentages are "applied to individual node ranges over the input set". Is $r = \max - \min$ of the pre-activation over the training set, per unit, per run?
- **The statistical test.** Test type, paired or unpaired, degrees of freedom, reference value, and any multiple-comparison correction are not given (see 2.4). With Abs training accuracy reported as 99.99 ± 0.00 in Table 1, a t-test against that row is degenerate, which may explain some of the extreme values.
- **Software and compute.** Framework and version, hardware, and run time are not stated. The mention of LogSoftmax in §6.2 suggests PyTorch, but I could not confirm it from the text.
- **Figure 1.** As rendered on Qeios, I can read only the caption of Figure 1; the curves are not available in the text I reviewed, so my reading relies on Tables 2 to 4.

4. What specifically to fix

1. §4.1, the offset equation $\{\text{scale} = (1 - p)r, \text{offset} = pr\}$. At $p = 0$, this gives $\text{scale} = r$, not the identity $\text{scale} = 1$. I suspect the intended form is $\text{scale} = 1 - p$ (or $\text{scale} = 1$) with $\text{offset} = pr$, but I could not tell. Please write the equation that the code actually applies, and say whether the offset sweep also rescales the units, because at $p = -200\%$ the printed form would multiply by $3r$.

2. Add the homogeneity argument from my point 2.1 to §4.1 and say plainly what the scaling test can and cannot show. An easy check: rerun the Table 2 sweep with b_2 set to zero. If accuracy becomes flat for all $p > 0$, the scaling result is fully explained by the architecture.
3. **Controls for Table 4.** I would add at least one of these: a post-activation offset, a norm-matched random perturbation at each level, or an offset applied to a random half of the units. Each gives the reader a baseline for "how much does any perturbation of this size hurt?".
4. §5, second paragraph: the sentence citing $p > 0.05$ for Tables 2 and 3 contradicts the tables (point 2.4). Recompute every T-stat and P-value column in Appendix A, state the test, and check that the Baseline rows give $t = 0$.
5. §5 says baseline accuracy is kept for scaling "from 10% to 200%". Table 2 has no 200% row; its rows are 1%, 5%, 10%, 25%, 50%, Baseline, and 1000%. Either add the row or change the text.
6. **One definition of the clip threshold.** §4.1 defines $\text{clip} = \text{pr}$ (a fraction of the range), the Table 3 caption says "percentages of the maximum activation", and §6.3 speaks of "the 50th percentile". These are three different quantities.
7. Because the perturbation precedes the activation in Eq. 1, $\min(\cdot, \text{clip})$ caps only the positive side of the pre-activation. For Abs, large negative pre-activations still become large positive activations after folding, so clipping does not cap Abs outputs at all on that side. Please discuss this in §6.3, or clip after the activation for a fair comparison between the two models.
8. Repeat Table 4 on the MNIST test set (point 2.5). One extra column per model would do.
9. §5, the ReLU claim of maintaining baseline "from -3% to +2%": Table 4 gives 98.13 and 98.14 at those offsets against 98.33. State the tolerance you used (for example, within 0.2 points) so the Abs and ReLU windows are defined the same way.
10. In §7, soften "empirical validation for the theoretical connection between neural networks and Mahalanobis distance" to "evidence consistent with", unless a direct Mahalanobis test is added (point 2.6). The abstract's claim about large intensity perturbations should exclude low cutoffs, given Table 3, rows 1% to 20%.
11. §6.4 would be stronger with a working definition of an intensity feature, even a provisional one, that makes a prediction opposite to the distance view. The passage about the "principal disponent" and Bigfoot is memorable, but it reads as rhetoric rather than argument. The sudden switch to "I don't think" in a paper otherwise written as "we" could also be smoothed.

12. §5 links the results to adversarial examples ([7], [8]). Those works perturb inputs, not hidden pre-activations; I would either explain the connection or drop the sentence.
13. A deeper network, even two hidden layers, and a second dataset such as Fashion-MNIST would show whether the offset/scale asymmetry survives beyond one linear readout. §7 already names deeper networks as future work; a small version of both here would help the claim.

5. Recommendation

Major revisions. The question is worth asking, the experiment is transparent and cheap to repeat, and the Abs versus ReLU comparison in §6 gives a concrete mechanism that others can test. My recommendation reflects the size of the fixes, not doubt about the effort. The scaling result follows from positive homogeneity plus a linear readout, the offset result lacks a control that the intensity view would fail, and the p-values in Appendix A do not agree with the text or with themselves. All three can be addressed with experiments that fit on a laptop. With those added and the conclusion in §7 matched to what is measured, I would expect the paper to make a clear and useful contribution to the interpretability discussion.

Declarations

Potential competing interests: No potential competing interests to declare.