Relationship to Adversarial Robustness

Although we're most interested in the implications of superposition for interpretability, there appears to be a connection to adversarial examples. If one gives it a little thought, this connection can actually be quite intuitive.

In a model without superposition, the end-to-end weights for the first feature are:

$$(W^TW)_0 \quad = \quad (1,\ 0,\ 0,\ 0,\ ...)$$

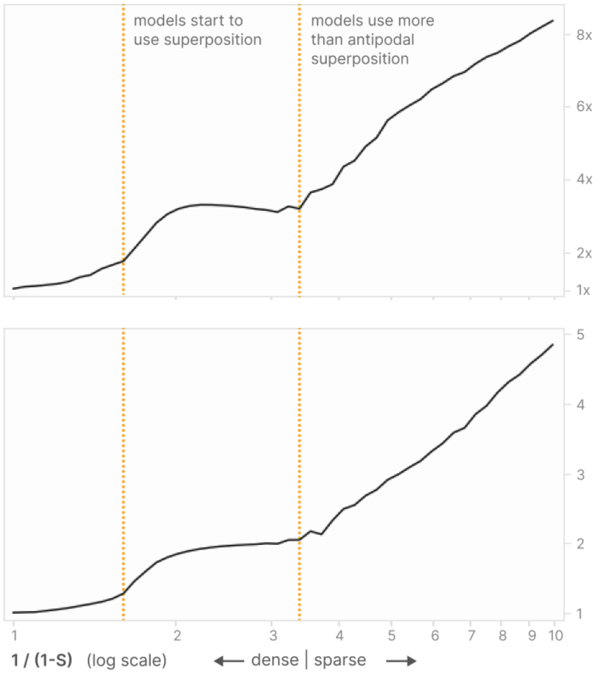
But in a model with superposition, it's something like:

$$(W^TW)_0 \quad = \quad (1,\ \epsilon,\ -\epsilon,\ \epsilon,\ ...)$$

The ϵ entries (which are solely an artifact of superposition "interference") create an obvious way for an adversary to attack the most important feature. Note that this may remain true even in the infinite data limit: the optimal behavior of the model fit to sparse infinite data is to use superposition to represent more features, leaving it vulnerable to attack.

To test this, we generated L2 adversarial examples (allowing a max L2 attack norm of 0.1 of the average input norm). We originally generated attacks with gradient descent, but found that for extremely sparse examples where ReLU neurons are in the zero regime 99% of the time, attacks were difficult, effectively due to gradient masking [32]. Instead, we found it worked better to analytically derive adversarial attacks by considering the optimal L2 attacks for each feature ($\lambda(W^TW)_i/||(W^TW)_i||_2$) and taking the one of these attacks which most harms model performance.

We find that vulnerability to adversarial examples sharply increases as superposition forms (increasing by >3x), and that the level of vulnerability closely tracks the number of features per dimension (the reciprocal of feature dimensionality).



Vulnerability to Adversarial Examples (Relative to Non-Superposition Model)

$(L' - L)/(L'_0 - L_0)$

The model's vulnerability to adversarial examples increases with feature sparsity, tracking the "features per dimension" measure of superposition plotted below.

Features Per Dimension

$1/D^* = ||W||_F^2/m$

Here we plot the number of features per dimension, the reciprocal of the "dimensionality" we defined in the pervious section. We see that the curve closely parallels vulnerability to adversarial examples.