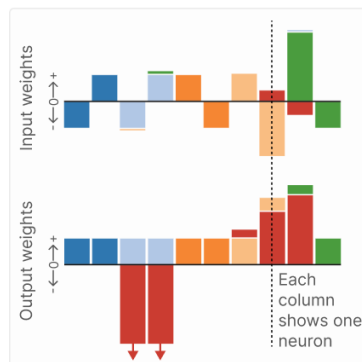
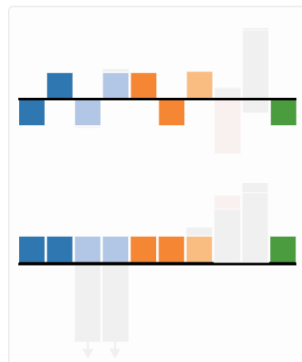


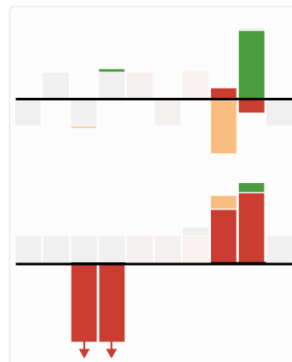
At first glance, this model is quite complicated and tricky to understand. However, we can (mostly) decompose it into two pieces...



Many weights are simply implementing absolute value, or a single side of absolute value, in the expected way.

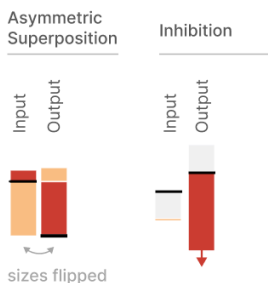


The main other thing is **asymmetric superposition with inhibition**. The model has two instances of this motif.

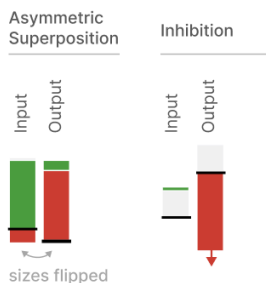


These other neurons implement two instances of asymmetric superposition and inhibition. Each instance consists of two neurons:

#### Asymmetric Superposition with Inhibition Instance 1



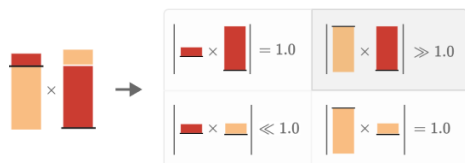
#### Asymmetric Superposition with Inhibition Instance 2



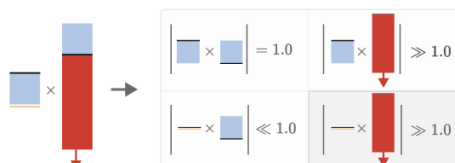
One neuron does *asymmetric superposition*. In normal superposition, one might store features with equal weights (eg.  $W = [1, -1]$ ) and then have equal output weights ( $W = [1, 1]$ ). In asymmetric superposition, one stores the features with different magnitudes (eg.  $W = [2, -\frac{1}{2}]$ ) and then has reciprocal output weights (eg.  $W = [\frac{1}{2}, 2]$ ). This causes one feature to heavily interfere with the other, but avoid the other interfering with the first!

To avoid the consequences of that interference, the model has another neuron heavily inhibit the feature in the case where there would have been positive interference. This essentially converts positive interference (which could greatly increase the loss) into negative interference (which has limited consequences due to the output ReLU).

One neuron represents two features (orange and red) with *asymmetric superposition*. This causes orange to heavily interfere with red, but not the reverse.



Large amounts of positive interference are bad, so the model then puts a small amount of orange into a neuron and uses it to massively inhibit red. This also forces the main feature the neuron is operating on (blue) to inhibit red.



There are a few other weights this doesn't explain. (We believe they're effectively small conditional biases.) But this asymmetric superposition and inhibition pattern appears to be the primary story.