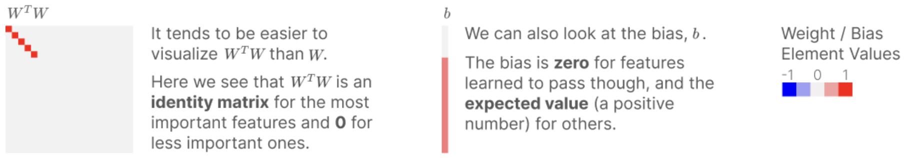
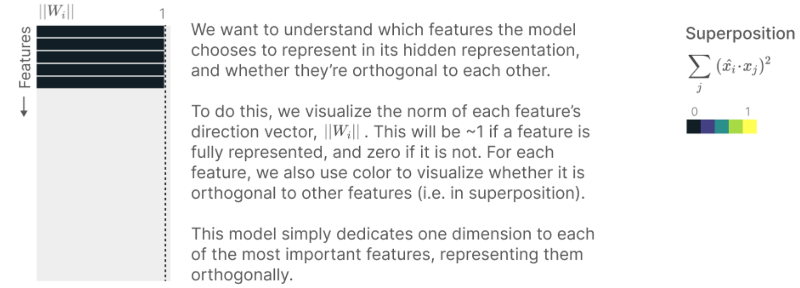


But the thing we really care about is this hypothesized phenomenon of superposition – does the model represent "extra features" by storing them non-orthogonally? Is there a way to get at it more explicitly? Well, one question is just how many features the model learns to represent. For any feature, whether or not it is represented is determined by $||W_i||$, the norm of its embedding vector.

We'd also like to understand whether a given feature shares its dimension with other features. For this, we calculate $\sum_{j\neq i}(\hat{W}_i \cdot W_j)^2$, projecting all other features onto the direction vector of W_i. It will be 0 if the feature is orthogonal to other features (dark blue below). On the other hand, values ≥ 1 mean that there is some group of other features which can activate W_i as strongly as feature i itself!

We can visualize the model we looked at previously this way:



Now that we have a way to visualize models, we can start to actually do experiments. We'll start by considering models with only a few features ($n = 20;\ m = 5;\ I_i = 0.7^i$). This will make it easy to visually see what happens. We consider a linear model, and several ReLU-output models trained on data with different feature sparsity levels: