1 Predictability of features

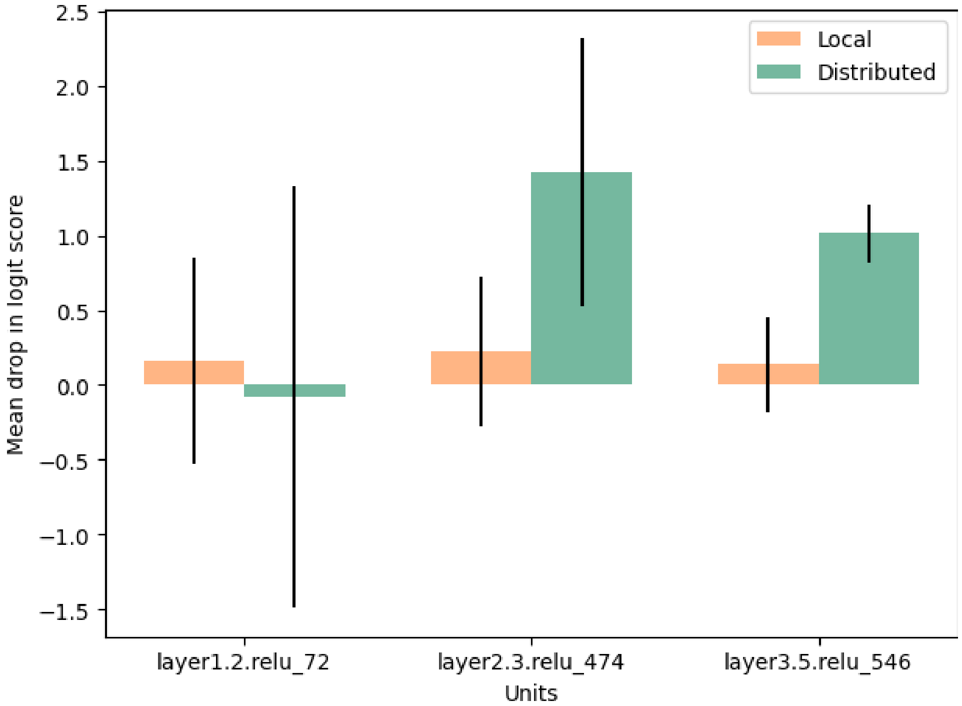

Figure 1: **Predictability of features.** Given an image **X**, the original prediction of the model on this image **Y**, and the feature vector **V** in the basis considered, we measure the drop in logit score for **Y** when we ablate **V**. For each feature, we measure the average drop in logit score across the top100 most activating images. Those early results suggest that features from distributed representations are at least as predictive as features from local representations.