Diffused Redundancy in Pre-trained Representations

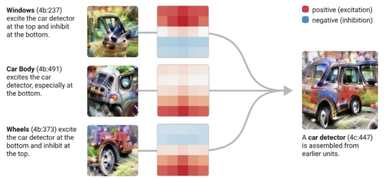
Vedant Nanda, Till Speicher, John P. Dickerson, Soheil Feizi, Krishna P. Gummadi, Adrian Weller

University of Maryland, College Park · Max Planck Institute for Software Systems · University of Cambridge

Understanding The Nature Of Learned Representations

Pre-trained representations are everywhere!

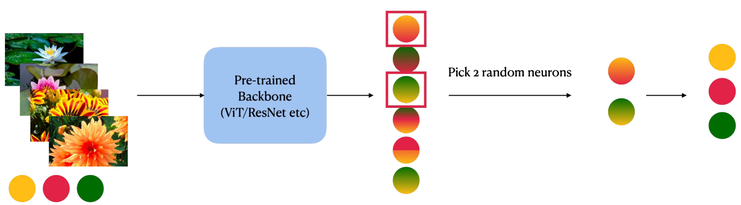
Interpretability: Compositionality between parts of a network



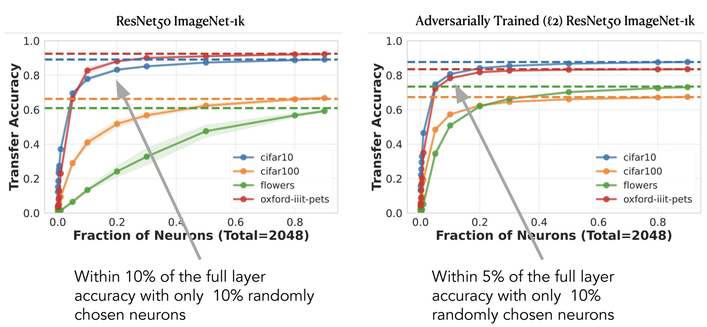
Zoom In: An Introduction to Circuits. Olah et al. 2020

DL Theory: Compression [Shwartz-Ziv and Tishby, 2017] & Neural Collapse [Papyan et al. 2020]
→ Representations need not store all information about the input

Diffused Redundancy



Learned features are spread throughout the layer, and thus a random subset of (of sufficient size) neurons suffices for most downstream tasks



Within 10% of the full layer accuracy with only 10% randomly chosen neurons

Within 5% of the full layer accuracy with only 10% randomly chosen neurons

We show that a randomly selected subset of neurons can perform (almost) as well as the full layer for downstream tasks

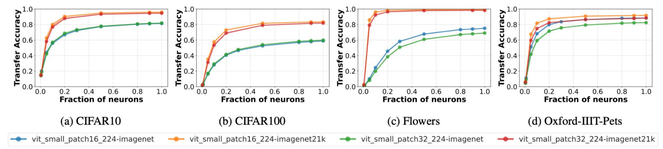
Our results have Implications for nature of learned representations: Do distinct parts of a network learn distinct features? Or are features diffused all over the neurons?

→ We find evidence for diffused redundancy!

arXiv tinyurl.com/diffused-redundancy

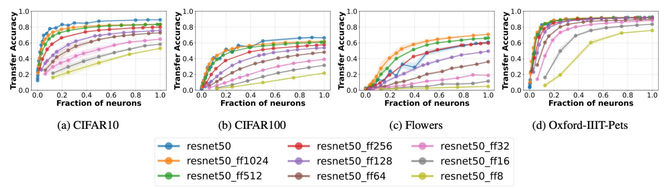
github.com/nvedant07/diffused-redundancy

Upstream Datasets



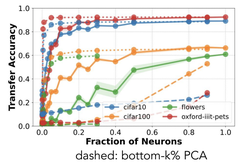
Bigger pretraining datasets lead to higher redundancy

Layer Width

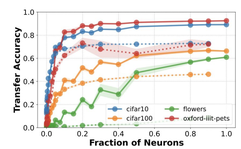


Redundancy disappears as we shrink the layer

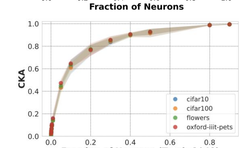
Intuitions for Diffused Redundancy



After a sufficiently large pick of random neurons (solid), performance closely follows projection on PCA (dotted)

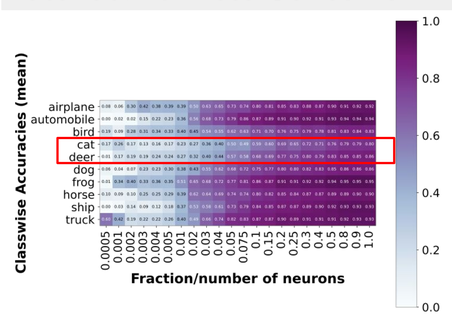


Random pick of neurons (solid) performs much better than random projections of the layer (dotted)



High Similarity between random subsets and full layer

Potential Fairness Concerns



Some classes are affected more than others