# Diffused Redundancy in Pre-trained Representations

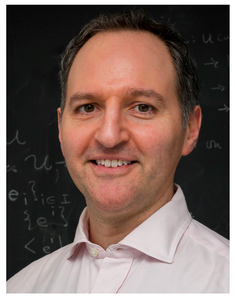
**NeurIPS 2023**

**Vedant Nanda**

Till Speicher

John P. Dickerson

Krishna P. Gummadi



Adrian Weller

Soheil Feizi

# TL;DR of Diffused Redundancy

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

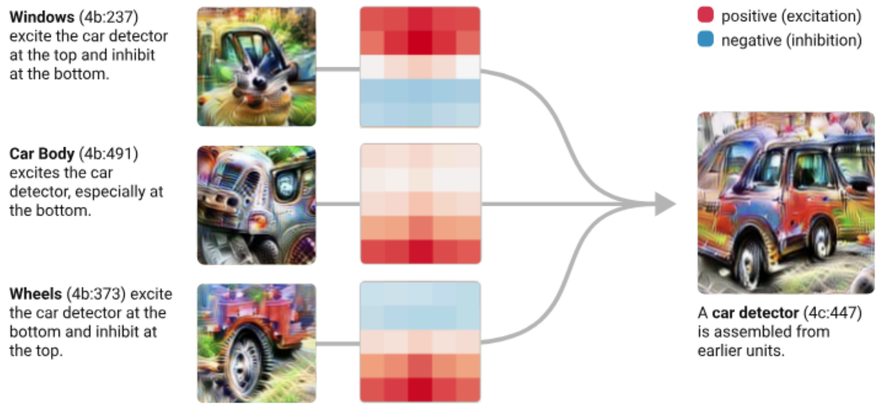
- Implications for nature of learned representations: do distinct parts of a network learn distinct features? Or are features diffused all over the neurons?
- Our results suggest diffused redundancy of features!
- While primarily an “understanding” paper, it also opens new directions for efficient finetuning / inference
- We highlight possible fairness tradeoffs when using random subsets of neurons

# Pretrained Representations Are Everywhere

- For any NLP/Vision task:
  - Pick a pre-trained backbone
  - Solve the downstream task using features extracted from this backbone
  - Eg: Image Classification
- Ground-breaking performance!
- However, understanding the nature of learned features is an ongoing research effort

# Understanding The Nature of Learned Representations

- **Explainability:** Compositionality between parts of a network

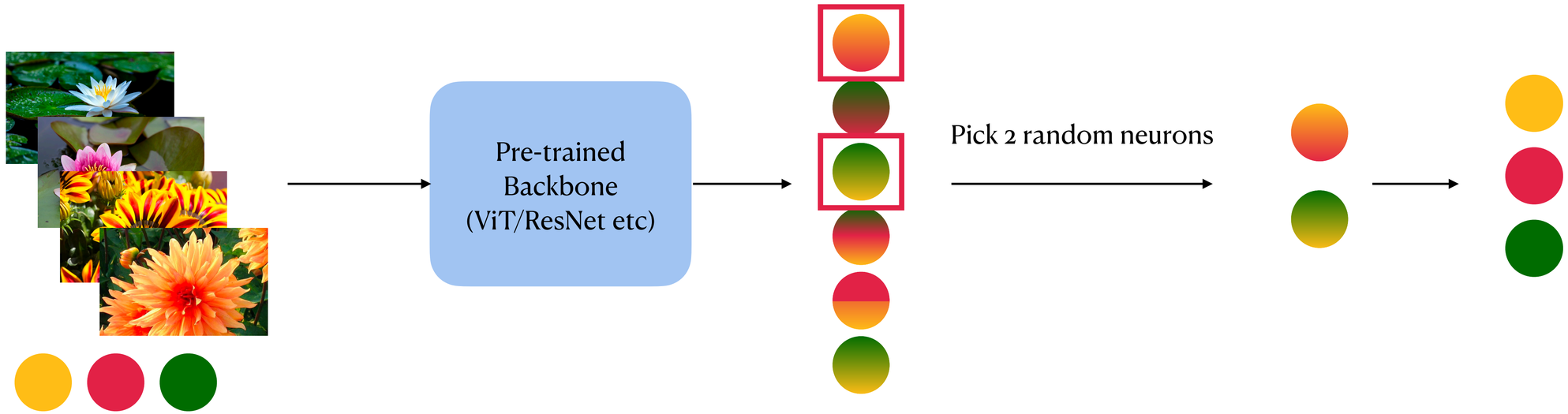



**Zoom In: An Introduction to Circuits.** Olah et al. 2020

- **Deep Learning Theory:** Compression Phase [Shwartz-Ziv and Tishby, 2017] & Neural Collapse [Papyan et al. 2020]
  - Representations need not store all information about the input
  - Do we need all neurons?

# Diffused Redundancy

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

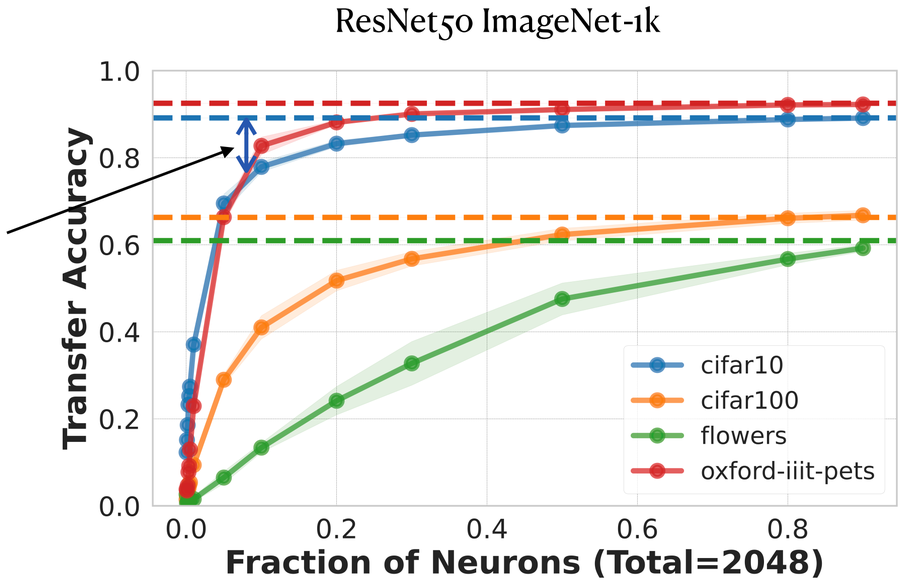



**DL Theory:** Few neurons contain enough information to efficiently transfer

**Explainability:** Information is redundantly spread out over many neurons, thus still allowing compositionality

# Evidence of Diffused Redundancy

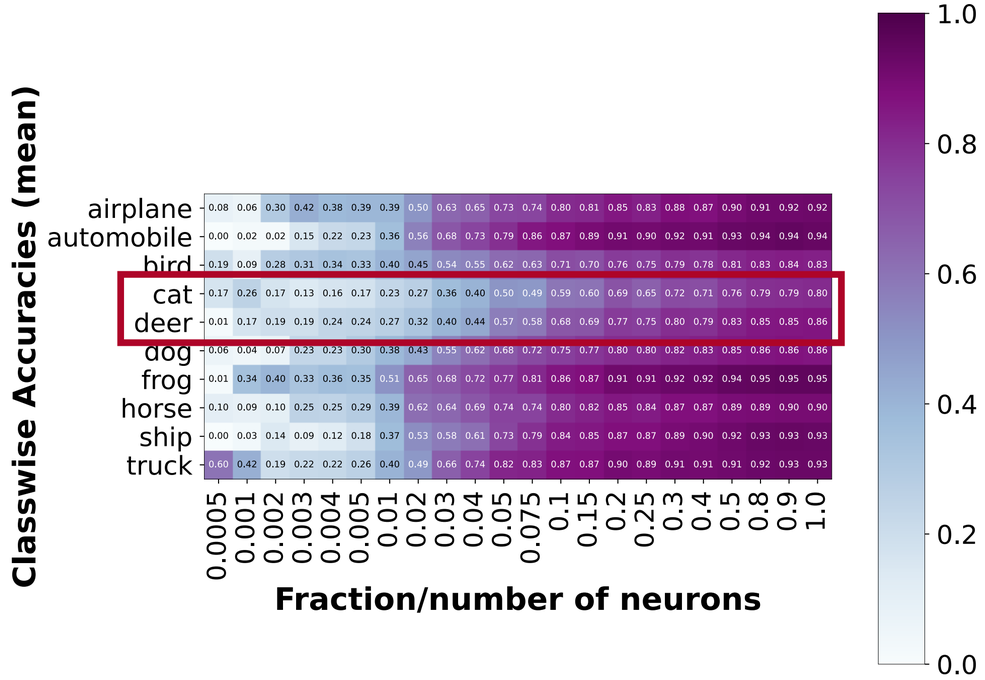
Within 10% of the full layer with only 10% randomly chosen neurons




- Degree of diffused redundancy depends on downstream task and pretraining loss
- More evidence in the paper!

# Possible Fairness Considerations

- A natural application — use the “compact” representation for efficient transfer
- However, this can lead to potential biases




**Some classes are affected more than others!**

# More Results in The Paper!

- How is diffused redundancy affected by:
  - last layer size,
  - different pretraining losses,
  - different pretraining datasets
- Why this happens — comparison to PCA and random projections
- Come chat with me at my NeurIPS poster! #634 Wed 13 Dec, Poster Session #3

Please reach out if you'd like to chat! **vnanda@mpi-sws.org**

arxiv.org/abs/2306.00183

github.com/nvedant07/diffused-redundancy