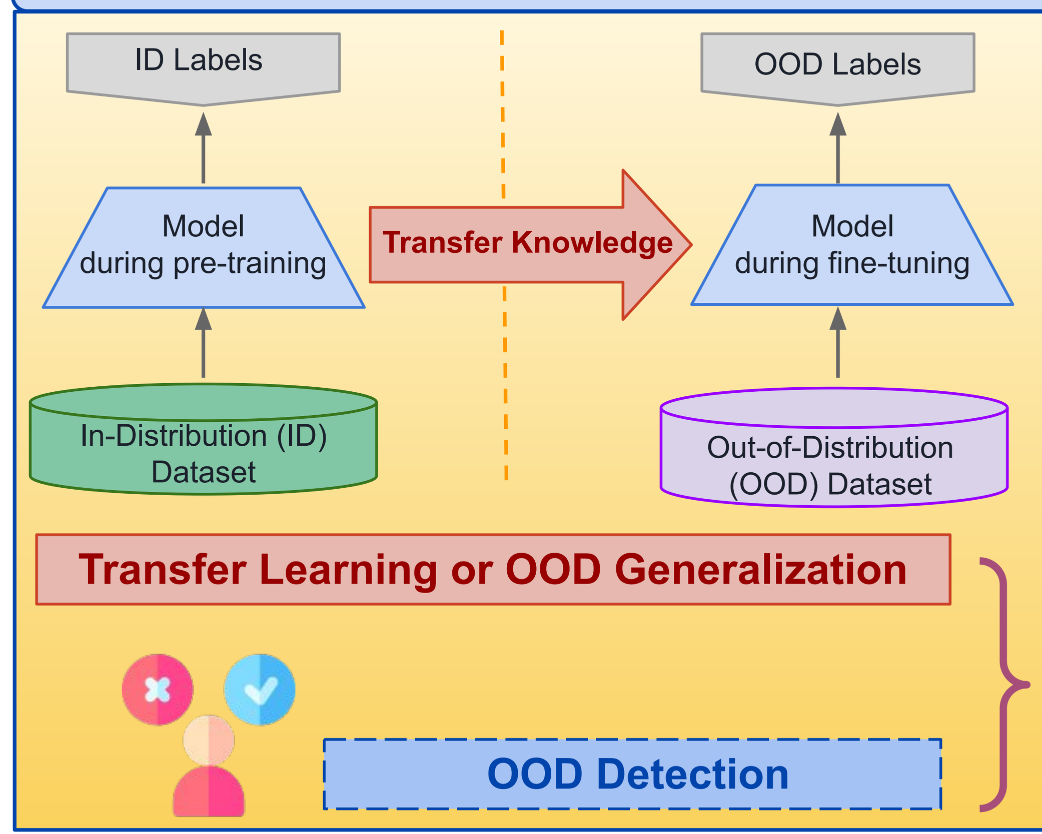
Controlling Neural Collapse Enhances Out-of-Distribution Detection and Transfer Learning

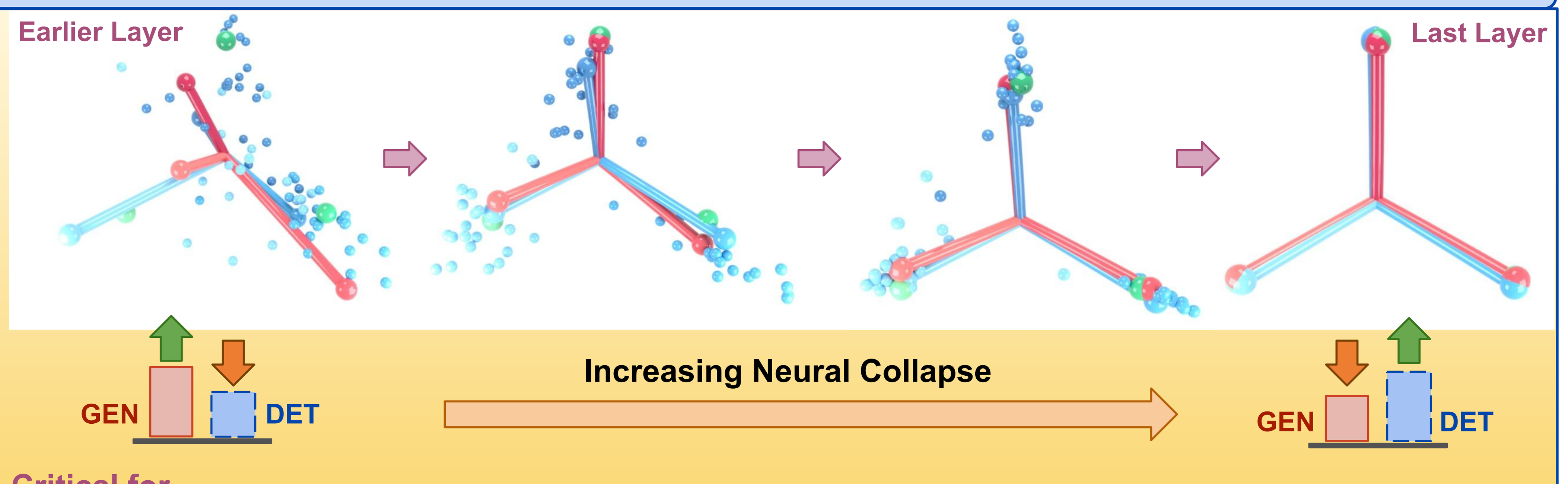
Md Yousuf Harun[1*], Jhair Gallardo[1], Christopher Kanan[2]

[1]Rochester Institute of Technology, [2]University of Rochester, *Correspondence: mh1023@rit.edu

Project Website

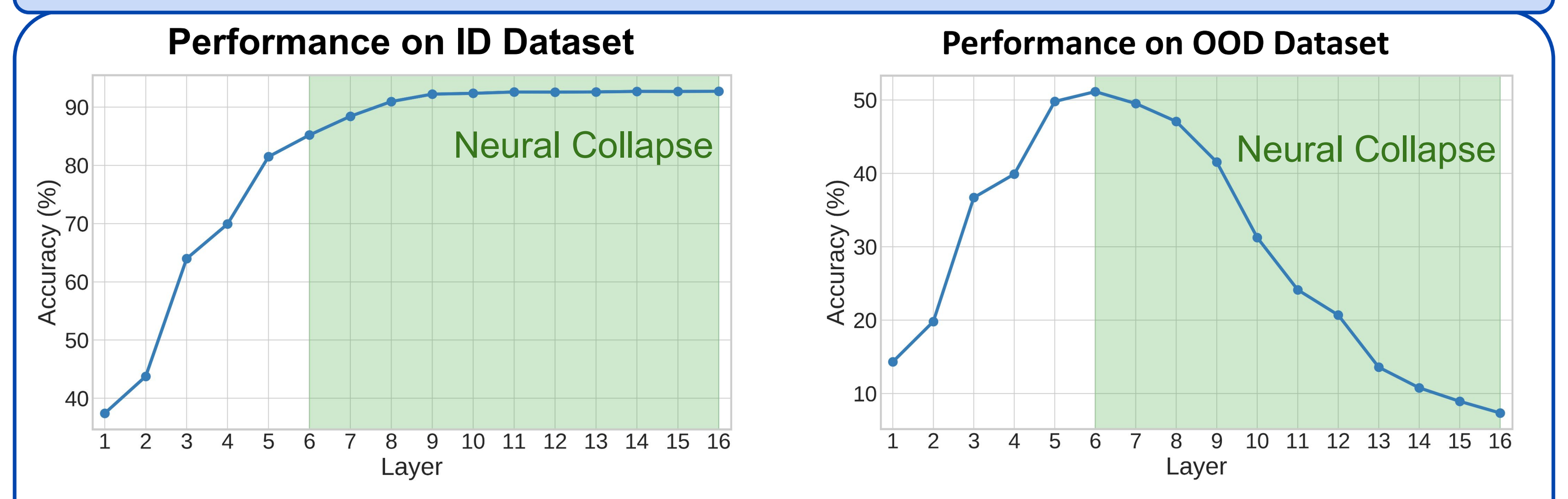
Motivation





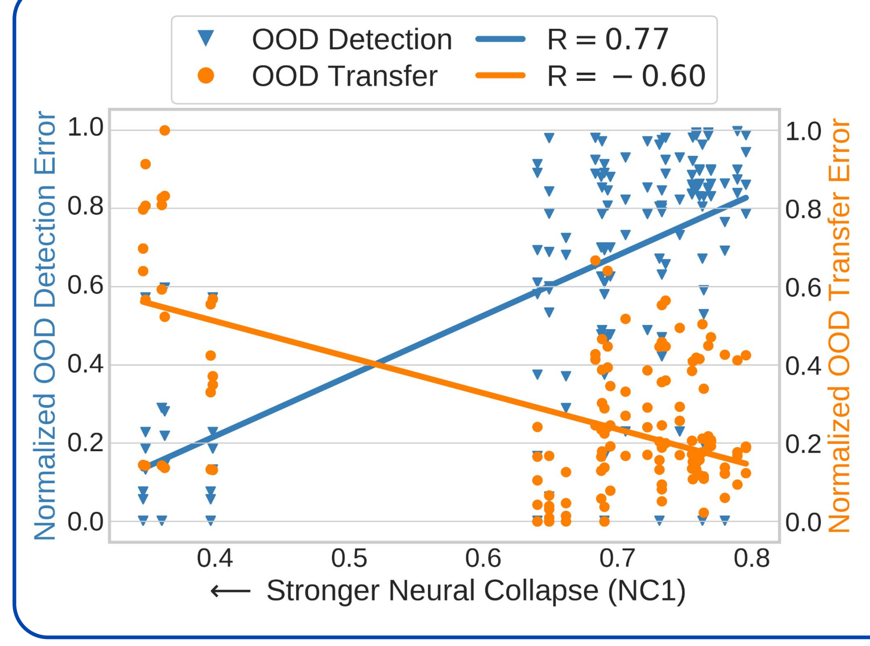
TL;DR: We developed a method to jointly optimize OOD detection and generalization by leveraging the Neural Collapse phenomenon

Neural Collapse Degrades OOD Generalization



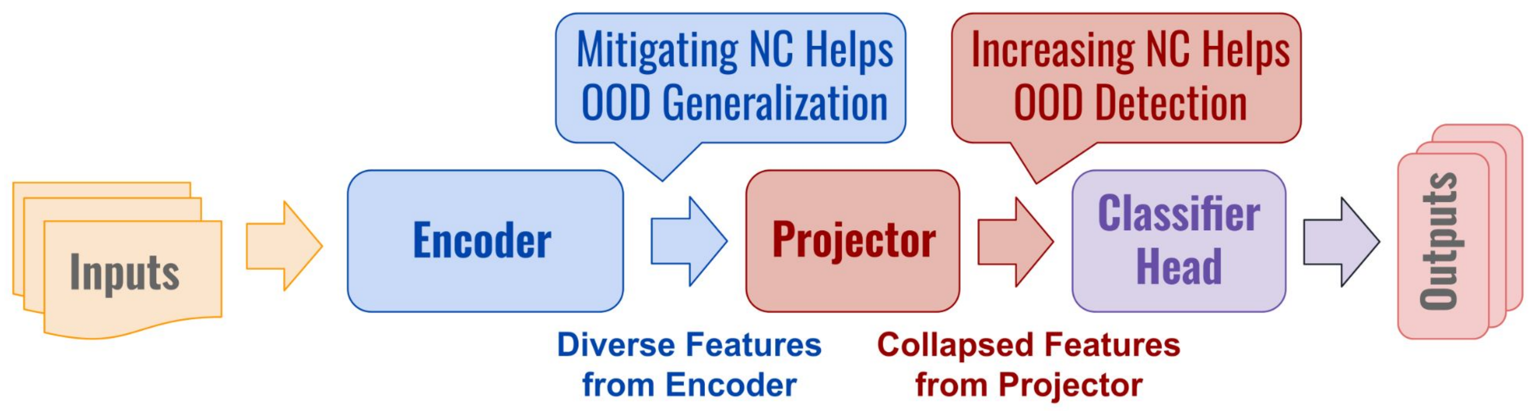
- **Neural Collapse (NC):** NC is a phenomenon where DNNs develop compact and structured class representations. While typically seen in the final layer, it can also occur to varying degrees in the last K layers – known as *intermediate* NC.
- In our NeurIPS 2024 paper "*What Variables Affect Out-of-Distribution Generalization in Pretrained Models?*", we show that intermediate NC degrades OOD generalization.
- As shown above, linear probe ID accuracy monotonically increases as a function of layers, but OOD accuracy only increases until the *Neural Collapse* is reached and then decreases.
- In this follow-up work, we demonstrate that the degree of NC plays a major role in both OOD detection and generalization.

Core Insight: Neural Collapse Relates to OOD Detection & Generalization



- Neural Collapse exhibits a strong positive correlation with OOD detection. And, it shows negative correlation with OOD generalization.
- Stronger NC improves OOD detection but degrades OOD generalization – and vice-versa.
- This inverse relationship reveals that a single feature space cannot effectively optimize both tasks.
- We measure NC1, OOD detection, and OOD generalization (via linear probing) across DNN layers.

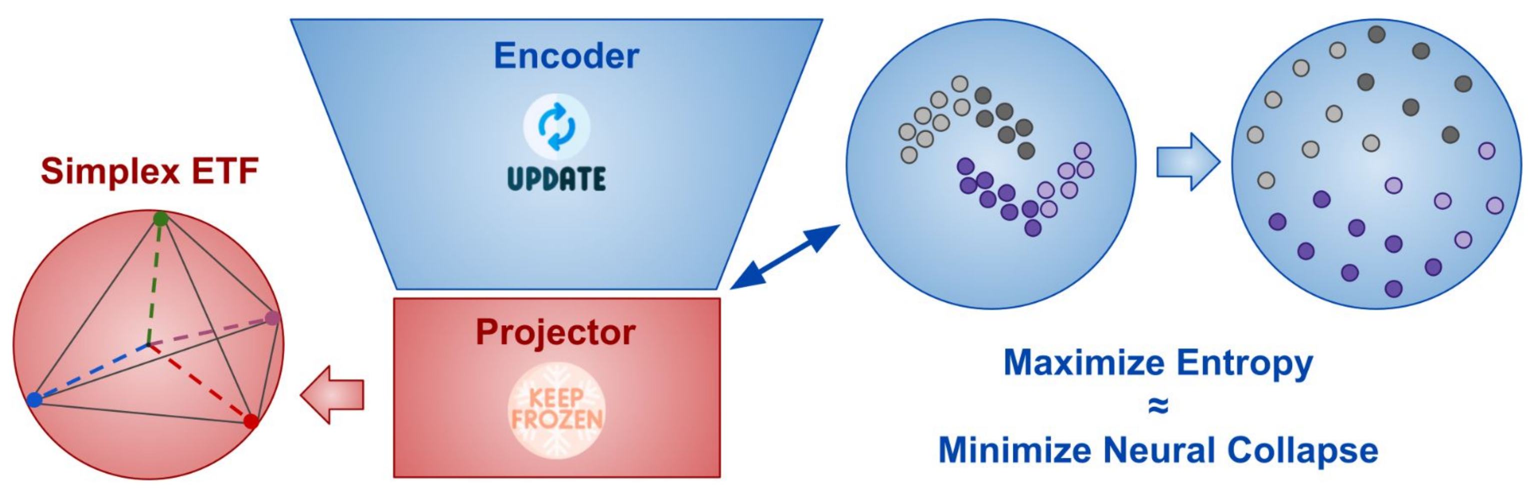
Method Overview: Controlling Neural Collapse



Implication of Neural Collapse:

- ➢ A single feature space cannot effectively achieve both OOD detection and generalization.
- ➢ To address this, we control NC at different DNN layers, using an encoder optimized for generalization and a projector tailored for detection.

How Do We Control Neural Collapse?



➔ **Layer for OOD generalization:** we introduce **entropy regularization** to mitigate NC in the encoder, improving feature diversity for OOD generalization.

- ◆ We develop a theoretical framework that explains how entropy regularization mitigates NC. In particular, we show that collapsing implies entropy diverges to negative infinity.
- ◆ For the entropy regularization, we leverage nearest-neighbor-based density estimation.

➔ **Layer for OOD detection:** we leverage a **fixed simplex Equiangular Tight Frame (ETF) projector** to induce NC in the final layer, improving feature compactness for detection.

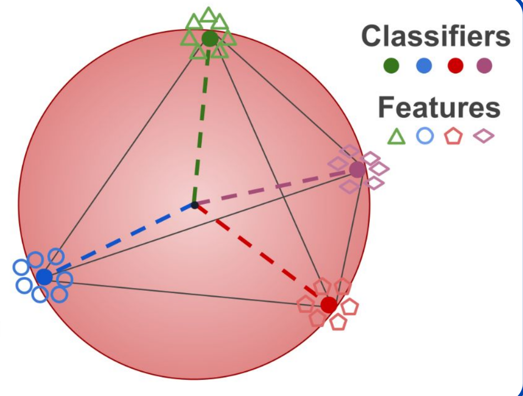
- ◆ For the ETF projector, we configure a two-layer MLP as a simplex ETF (equinorm and maximum equiangularity) and keep it frozen during training.

Experimental Setup

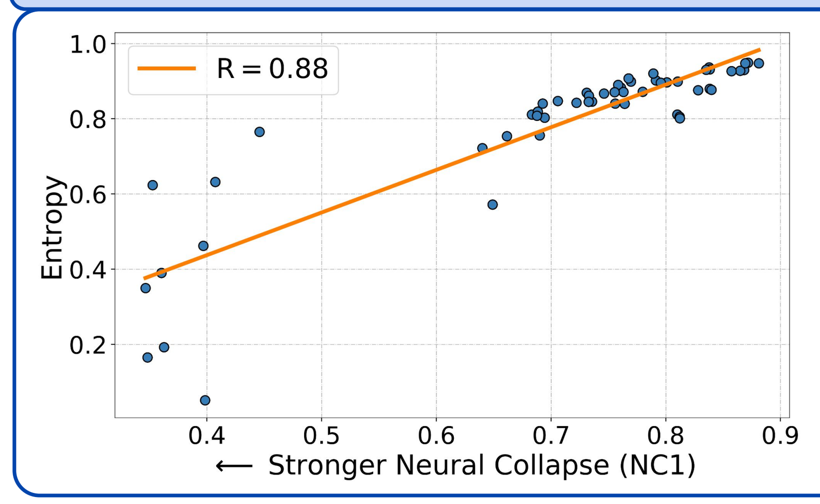
- **Datasets:** Used ImageNet-100 as ID dataset and eight OOD datasets: NINCO, CUB-200, CIFAR-100, ImageNet-R, Flowers-102, Aircrafts, Oxford Pets, and STL-10
- **Architecture:** VGG17, ResNet18, ResNet34, ViT-Tiny, and ViT-Small
- **NC Evaluation:** Four NC metrics NC1, NC2, NC3, and NC4 characterized by NC criteria. A lower NC indicates stronger Neural Collapse and vice-versa.
- **Performance Metrics:** ID generalization error, OOD generalization error, and OOD detection error (FPR95).

Neural Collapse Criteria

1. **Feature Collapse (NC1):** Intra-class features collapse to a single mean with low variability.
2. **Simplex ETF (NC2):** Class means, centered at the global mean, form a maximally spaced simplex on a hypersphere.
3. **Self-Duality (NC3):** Classifiers align tightly with class means, creating a nearly self-dual configuration.
4. **Nearest Class Mean (NCM) Decision (NC4):** Classification resembles a NCM scheme, based on class mean proximity.



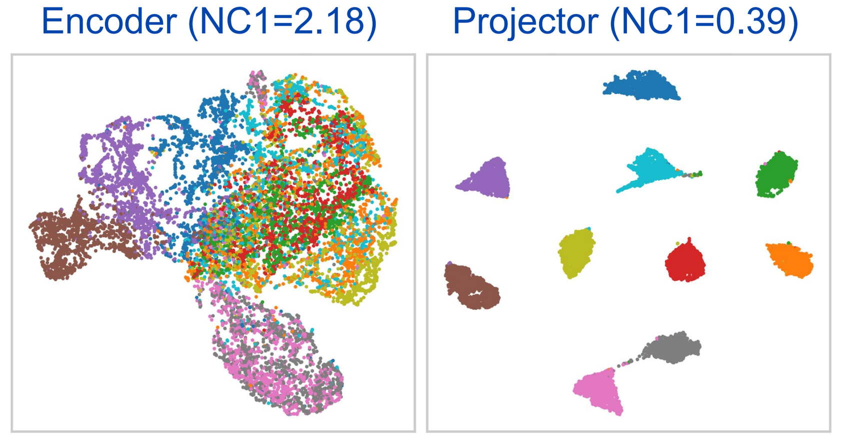
Correlation between Neural Collapse and Entropy



- Neural collapse (NC1) correlates with entropy. **The stronger the neural collapse, the lower the entropy** and vice-versa.
- Increasing the entropy of the encoder's embeddings may mitigate NC and enhance OOD generalization.
- We analyze different layers of VGG17 networks that are pre-trained on ImageNet-100 dataset (ID).

Qualitative Results: Encoder Vs. Projector

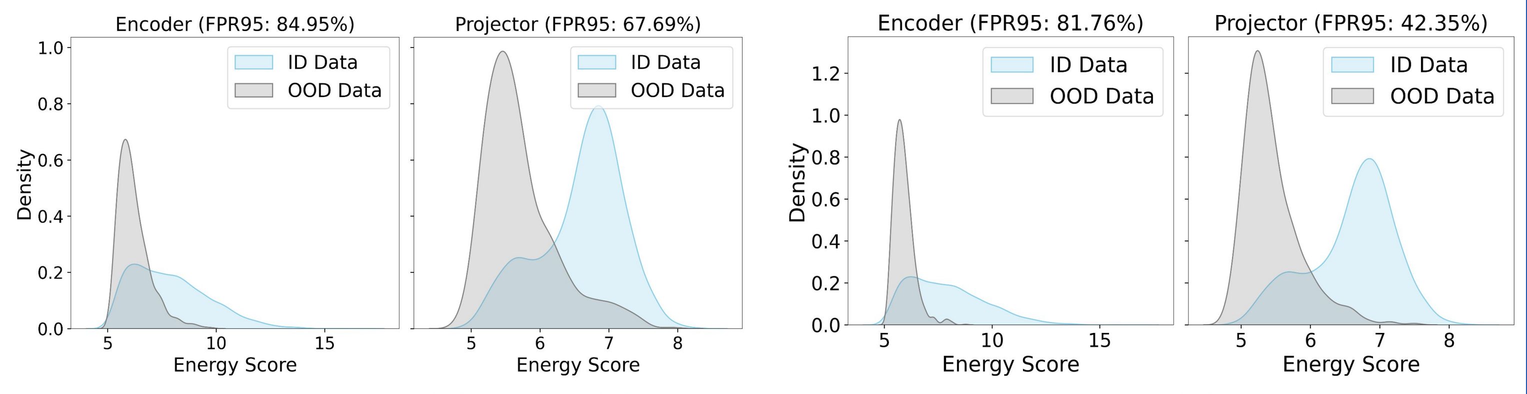
The following results are based on VGG17 models pre-trained on ImageNet-100 dataset (ID).



UMAP Visualization of Embeddings:

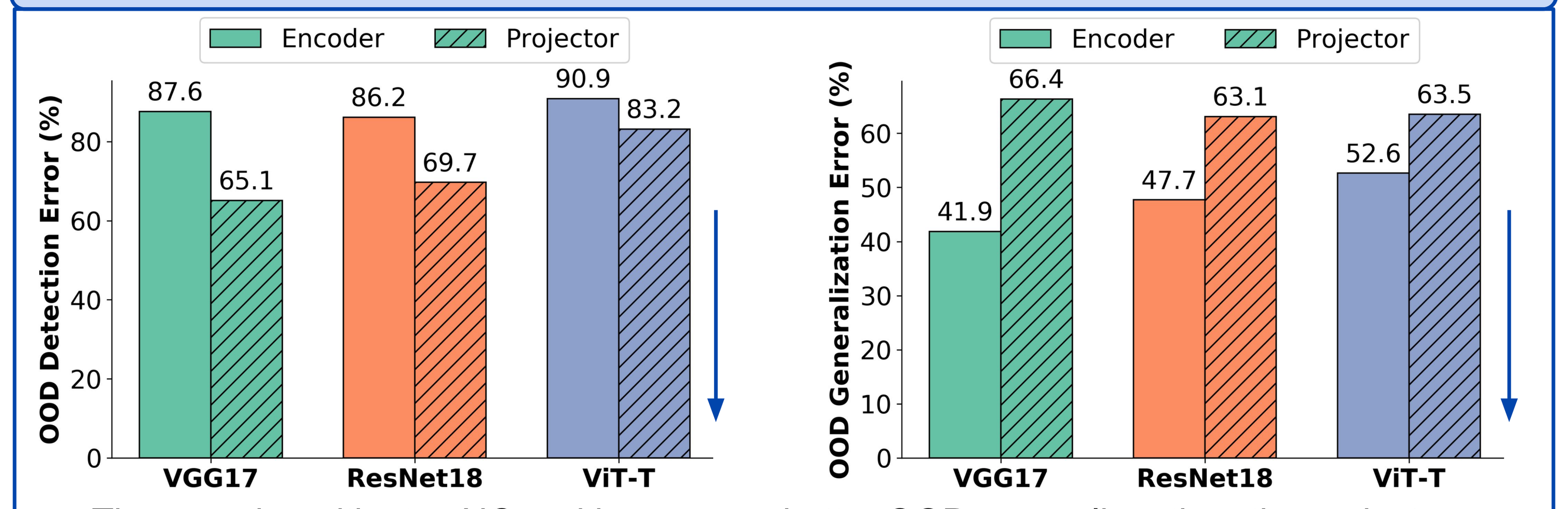
- Projector embeddings exhibit significantly stronger neural collapse–evidenced by 5.6x lower NC1 and tighter clustering around class means–compared to encoder embeddings.
- We show 10 ImageNet classes by distinct colors.

OOD Detection: As shown in energy score distribution, the projector creates a greater separation between ID & OOD data and achieves a lower FPR95 than the encoder.



ID: ImageNet-100, OOD: NINCO-64 ID: ImageNet-100, OOD: Flowers-102

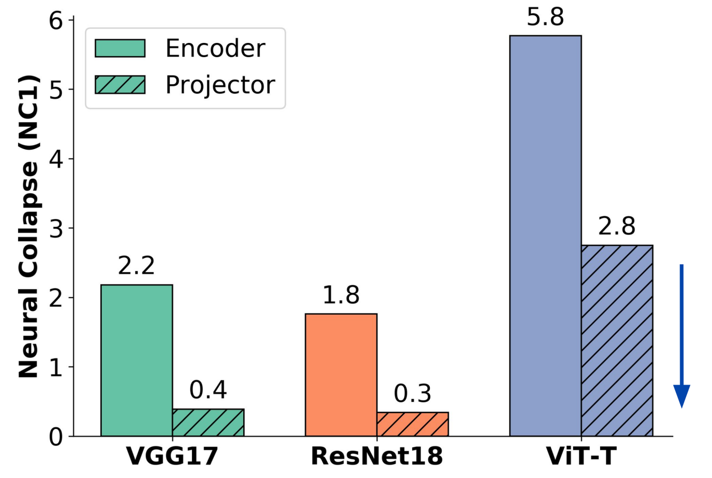
Quantitative Results: Encoder Vs. Projector



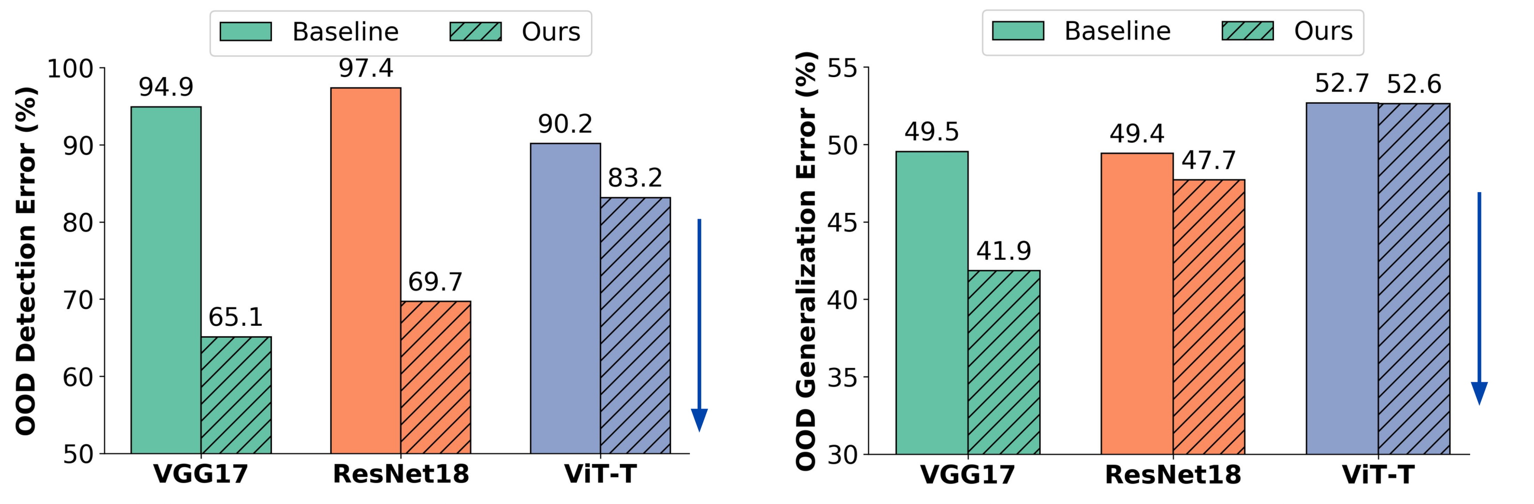
- The encoder mitigates NC and becomes a better OOD generalizer than the projector.
- The projector intensifies NC and becomes a better OOD detector than the encoder.

Neural Collapse Evaluation:

- The projector exhibits lower NC1 values (i.e., stronger neural collapse) than the encoder across DNN architectures.
- We report NC1 (feature collapse), the most dominant indicator of neural collapse.
- All above results are averaged across eight OOD datasets



Comparison with Baseline

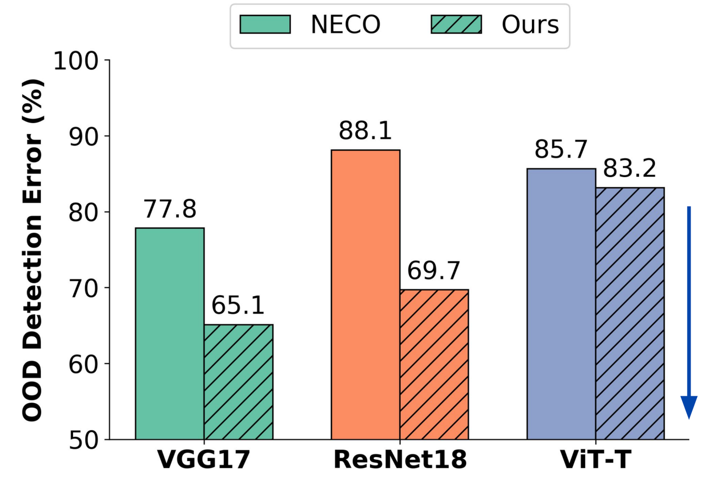


OOD Detection Performance OOD Generalization Performance

- ➢ Baseline DNNs lack mechanisms to control NC, resulting in poor performance.
- ➢ Our method controls NC and achieves significant improvements over these baselines.
- ➢ Results are based on DNNs pre-trained on ImageNet-100 dataset (ID), with performance averaged across eight OOD datasets.

Comparison with SOTA OOD Detector:

- We compare our method with NECO, a state-of-the-art OOD detection method based on NC properties.
- Since NECO does not address OOD generalization, we restrict this comparison to OOD detection only.
- Our method consistently outperforms NECO across all settings.



Acknowledgments

We thank US National Science Foundation for supporting our research.