International Conference On Machine Learning

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

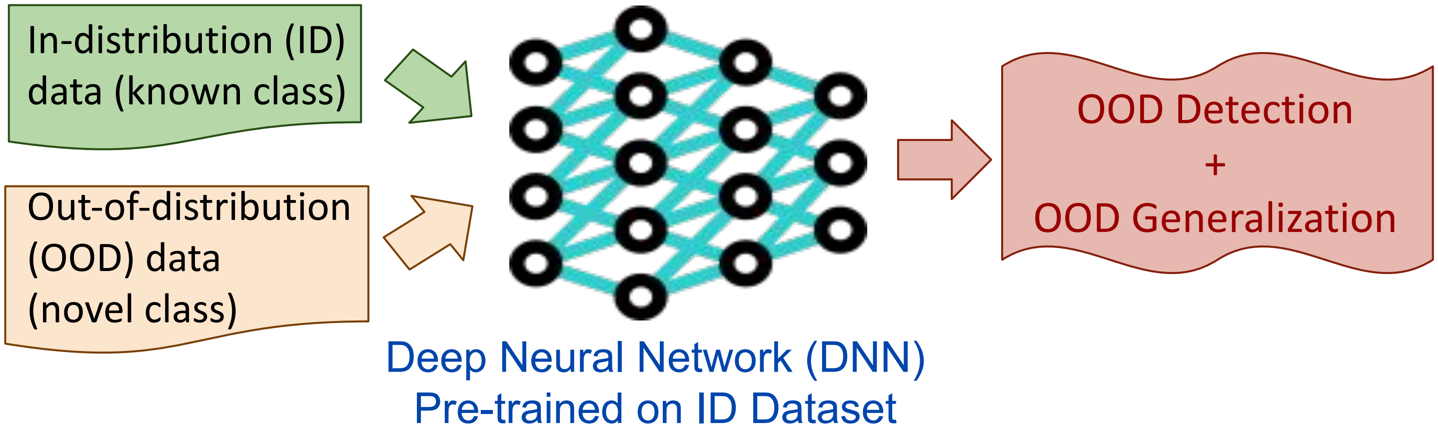
Md Yousuf Harun[1], Jhair Gallardo[1], Christopher Kanan[2]

[1]Rochester Institute of Technology,[2]University of Rochester

Project Website

Motivation & Problem Statement

- **Goal:** In open-world settings, DNNs must detect novel concepts and maximize forward transfer to facilitate efficient learning.
- **Research Question:** How can we build representations in a DNN to simultaneously achieve both OOD detection and generalization?
- **Challenge:** Optimizing OOD detection hurts OOD generalization and vice-versa.
- **TL;DR:** We developed a method for jointly optimizing the OOD detection and forward transfer (OOD generalization) based on the **Neural Collapse** phenomenon.



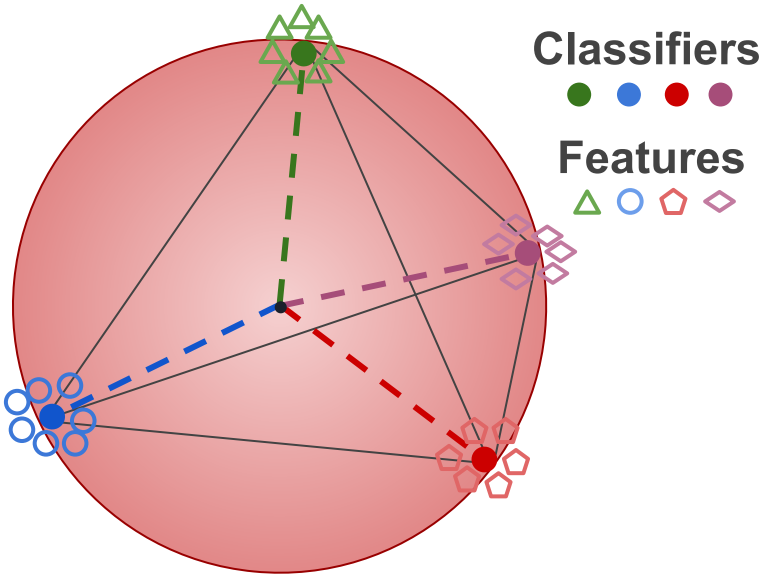
Constraints:

1. No Semantic overlap between ID & OOD
2. No access to additional OOD training data

What is Neural Collapse?

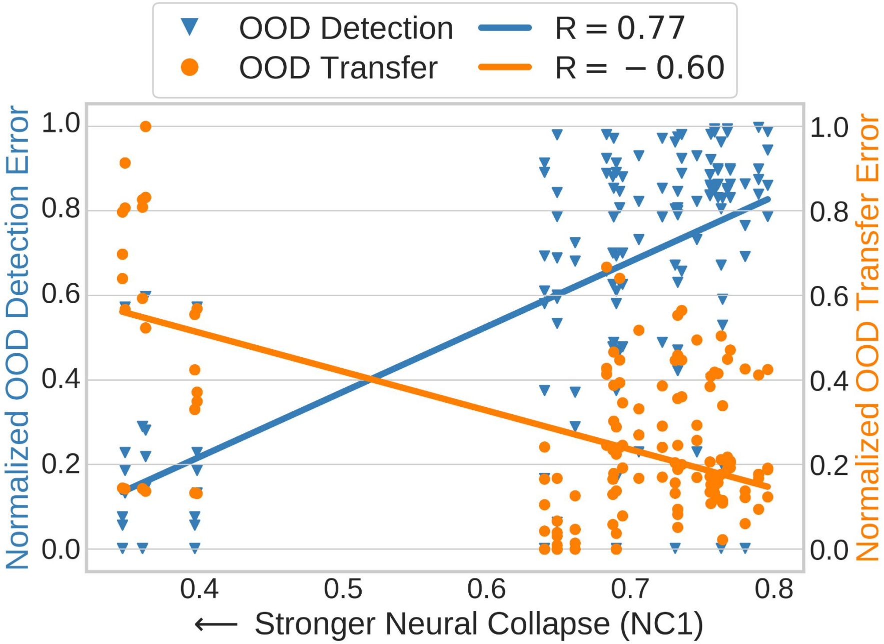
- **Neural Collapse (NC)** arises when class features become tightly clustered, often converging toward a Simplex *Equiangular Tight Frame* (ETF)

- Neural collapse is characterized by following **four** criteria:
  - ***NC*1**: feature collapse
  - ***NC*2**: simplex ETF structure
  - ***NC*3**: self-duality
  - ***NC*4**: nearest class mean decision



Neural Collapse Insights

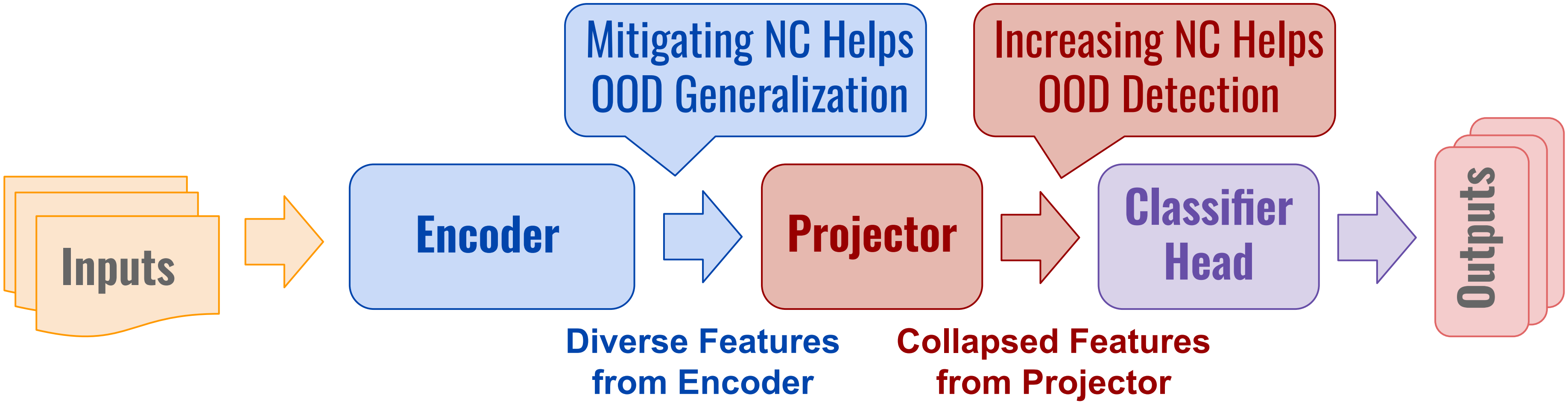
- **Observation:** Increasing neural collapse improves OOD detection but hinders OOD generalization and vice-versa.
- **Takeaway:** A single feature space cannot simultaneously achieve both tasks.



Correlation between NC & OOD detection/ generalization

Method Overview: Controlling NC

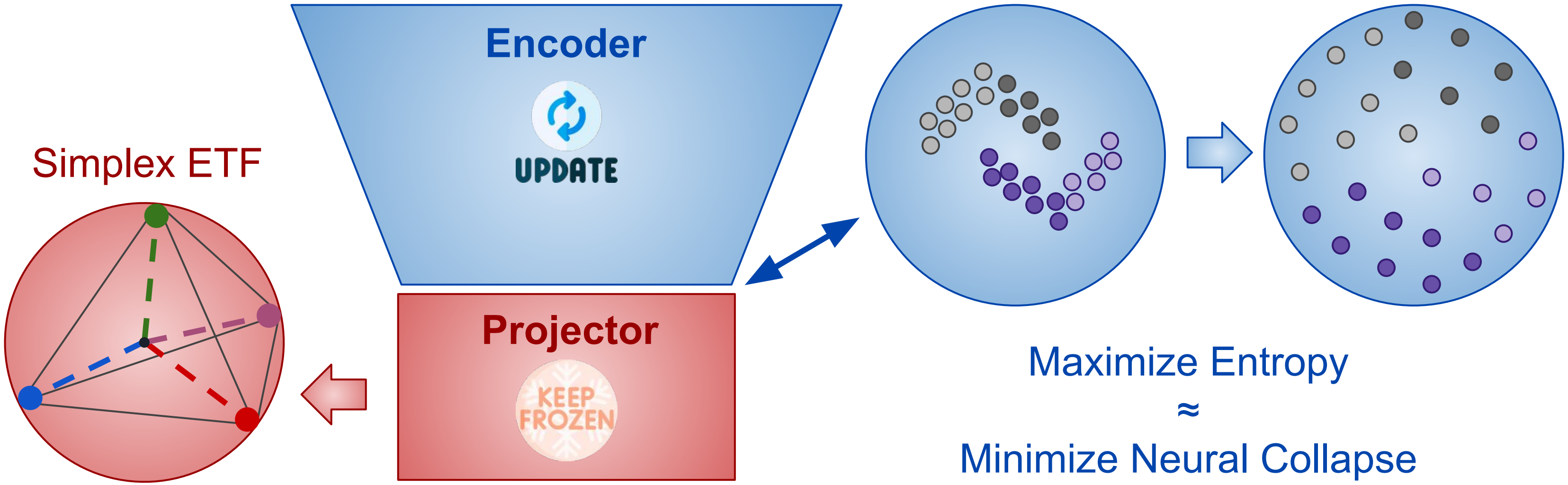
- A single feature space cannot effectively achieve both OOD detection and generalization.
- To address this, we control NC at different DNN layers, using an encoder optimized for generalization and a projector tailored for detection.



Method Overview: Controlling NC

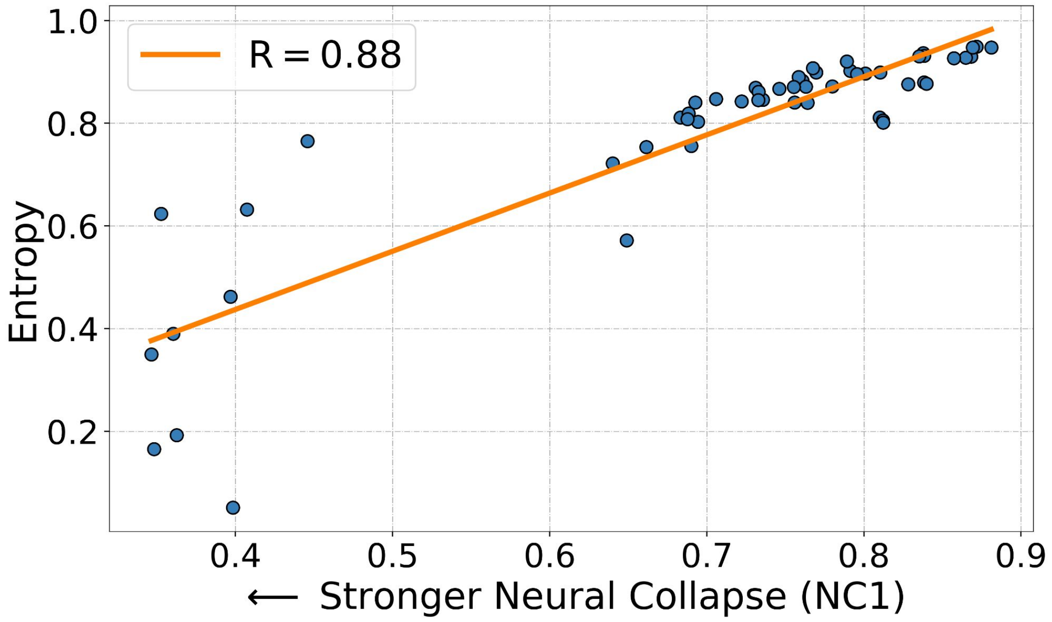
- **Layer for OOD generalization:** entropy regularization mitigates NC in the encoder → improves feature diversity for OOD generalization.

- **Layer for OOD detection:** a fixed simplex ETF projector increases NC in the final layer → improves feature compactness for OOD detection.



Entropy Vs. Neural Collapse

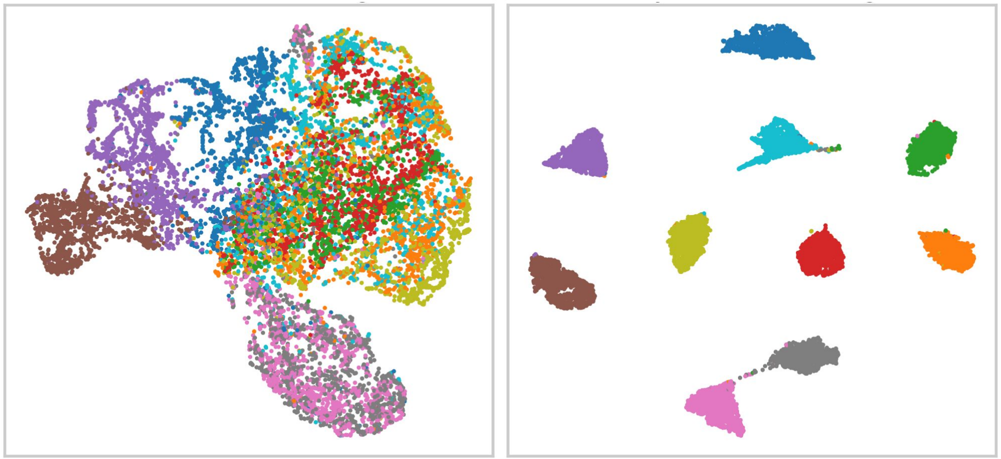
- Neural Collapse (NC1) correlates with entropy. **The stronger the neural collapse, the lower the entropy** and vice-versa.
- It suggests that increasing entropy of encoder embeddings may decrease NC and increase OOD generalization



Qualitative Results: Encoder Vs. Projector

- 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.

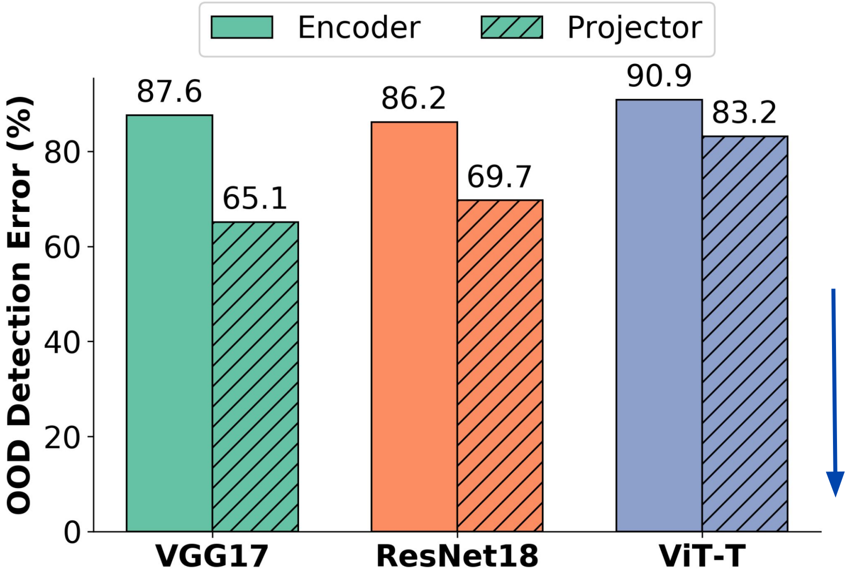
Encoder Embeddings (NC1 = 2.18) | **Projector Embeddings (NC1 = 0.39)**



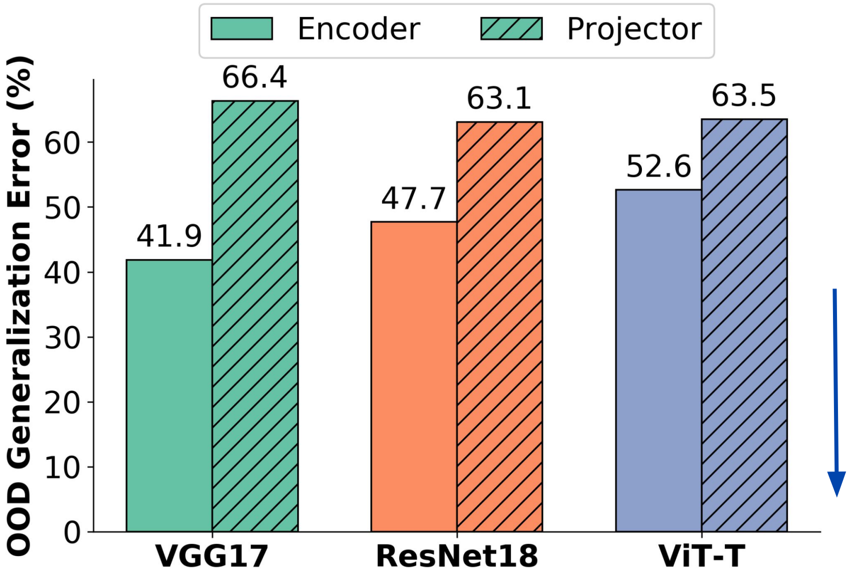
UMAP visualization of embedding

Results: Encoder Vs. Projector

- We train various DNNs on ImageNet-100 (ID) and use eight OOD datasets for evaluations. Reported results are averaged across eight OOD datasets.
- 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.



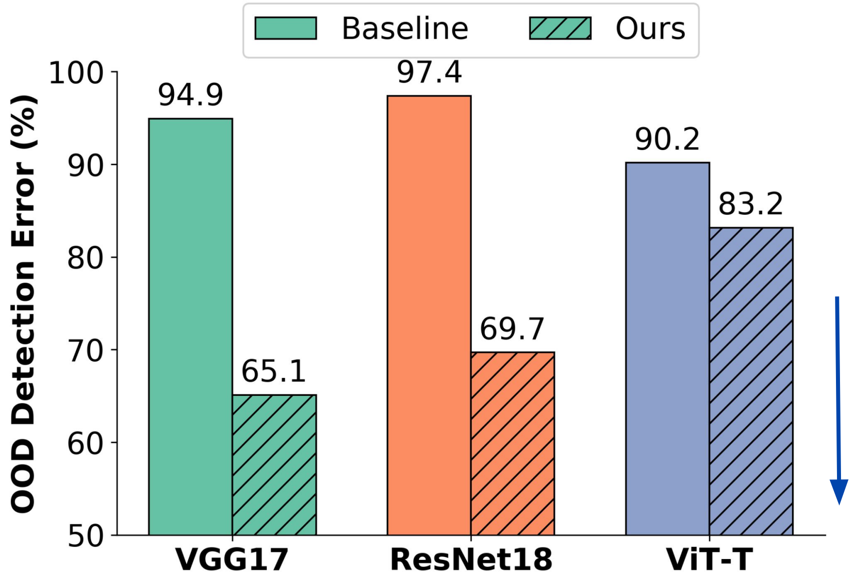
OOD Detection



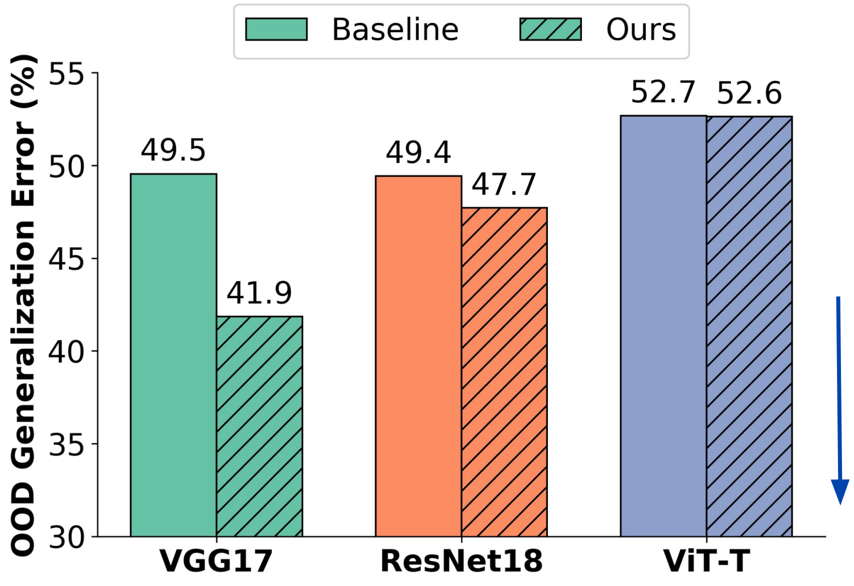
OOD Generalization

Results: Comparison with Baseline

- We train various DNNs on ImageNet-100 (ID) and use eight OOD datasets for evaluations. Reported results are averaged across eight OOD datasets.
- Baseline DNNs lack mechanisms to control NC, resulting in poor performance.
- Our method controls NC and achieves significant improvements over the baselines.



OOD Detection



OOD Generalization

Summary

- ❖ We demonstrated that Neural Collapse has an inverse relationship with OOD detection and generalization
- ❖ Motivated by this inverse relationship, our method enhances OOD detection by enforcing NC while promoting OOD generalization by mitigating NC.
- ❖ Our method excels at both OOD detection & generalization tasks without any additional OOD training data.
- ❖ This work has implications for open-world problems where both OOD detection and generalization are critical.

Thank You

Paper Link:

https://arxiv.org/abs/2502.10691