École Nationale Supérieure de Techniques Avancées - ENSTA Paris

Class : Deep Learning for Computer Vision

Practical Work Report

Membres : DE QUEIROZ GARCIA Luana, NOGUEIRA FALABELLA Leonardo

Palaiseau
20 février 2026

Table des matières

Introduction

Out-of-Distribution (OOD) detection is a fundamental challenge for the safe deployment of deep learning models in real-world applications. Classification models trained on a closed set of classes tend to make overconfident predictions for unknown samples, creating significant risks in critical systems such as autonomous vehicles, medical diagnosis, and security systems.

This practical work aims to explore the intersection between two important phenomena in deep learning : OOD detection and Neural Collapse.

In this project, we will :

1. Train a ResNet-18 classifier on the CIFAR-100 dataset
2. Implement and compare different OOD detection methods : Max Softmax Probability (MSP), Maximum Logit Score, Mahalanobis, Energy Score, and ViM
3. Study the Neural Collapse phenomenon (NC1 to NC4) at the end of training
4. Investigate the NC5 property (ID/OOD orthogonality)
5. Implement the NECO method, which leverages the geometric properties of Neural Collapse for OOD detection

The goal is to understand how OOD detection metrics behave and how Neural Collapse can be leveraged for analysis and detection of out-of-distribution samples. Through visualizations and quantitative analysis, we will explore the geometric structure of learned representations and their impact on the ability to detect unknown samples.

Training a ResNet18 classifier

For the image classification task, the CIFAR-100 dataset was used. Image preprocessing included data augmentation techniques such as random horizontal flips and random crops, along with pixel normalization and conversion to tensors. The data was organized into batches of 256 images to feed the model during training and evaluation.

The chosen model was a ResNet-18 architecture, leveraging weights pre-trained on ImageNet for a better starting point. Its classification layer was adapted to recognize the 100 distinct classes of CIFAR-100. The training was conducted for 400 epochs, using cross-entropy loss and the SGD optimizer with an initial learning rate of 0.01 and momentum of 0.9. The learning rate was adjusted during the process using a cosine annealing scheduler. This will be referenced as model 1.

The same model was trained a second time, where we tried to reproduce the regime done in [1]. The learning rate started at 0.1 and followed a multi-step decay of rate 0.1 in two milestones equally distributed along the 400 epochs. This will be referenced as model 2.

Throughout training, the loss and accuracy were monitored, demonstrating the evolution of the model's learning. Upon completion, the final state of the model, optimizer, and scheduler, along with the training metrics, were saved in a checkpoint file (resnet18_cifar100.pth and resnet18_cifar100_reproduc.pth) . A generated plot (Figure 2 illustrates the trajectory of loss reduction and accuracy increase over the epochs, highlighting the learning progress.

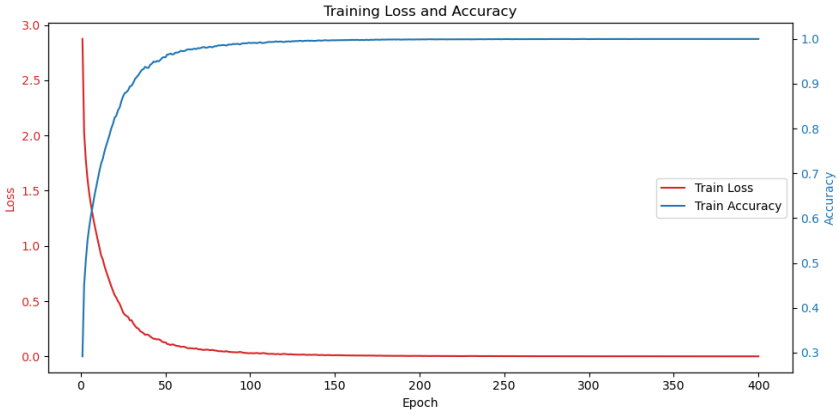

FIGURE 1 – Training Loss and Accuracy of model 1

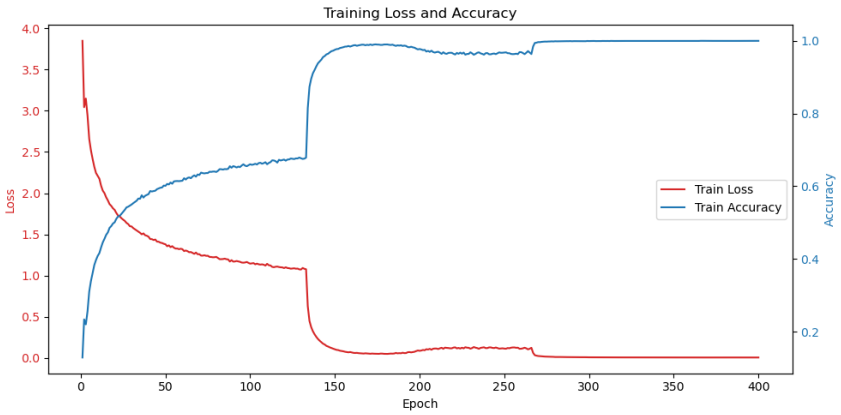

FIGURE 2 – Training Loss and Accuracy of model 1

1 OOD Scores

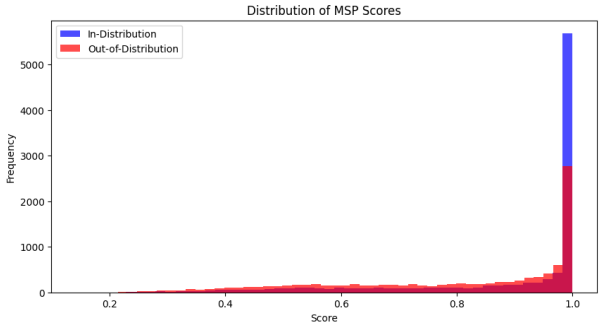

FIGURE 3 – Distribution of MSP Score

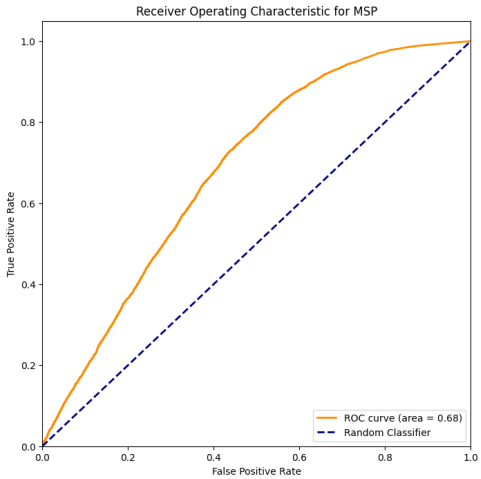

FIGURE 4 – ROC for MSP

Next, it is presented an analysis of the OOD detection capabilities of a ResNet-18 model on CIFAR-100 (In-Distribution) against CIFAR-10 (Out-of-Distribution) using Max Softmax Probability (MSP), Maximum Logit Score, Mahalanobis distance, Energy Score, and Virtual-logit Matching (ViM) methods. Both models performed similarly in all metrics presented in this section, as such we will only discuss the results of model 1 on this report.

The Maximum Logit method demonstrated the best performance, achieving an AUROC of 0.6850 and an AUPR of 0.6338

1.1 Analysis of Each Method

1. Maximum Logit Score Calculation : ID scores ranged from 0.1 to 1.0, while OOD scores ranged from 0.2 to 1.0. Achieved the highest performance with an AUROC of 0.6850 and an AUPR of 0.6338. As seen in Figures 3 and 4.
2. Max Softmax Probability (MSP) : ID scores ranged from 6.4 to 50.0, while OOD scores ranged from 6.4 to 43.6 (really similar ranges of score). Showed comparable

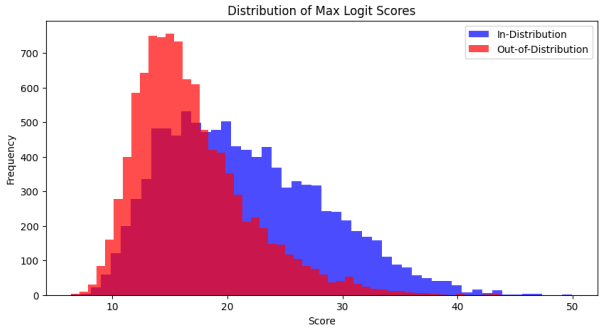

FIGURE 5 – Distribution of Max Logit Score

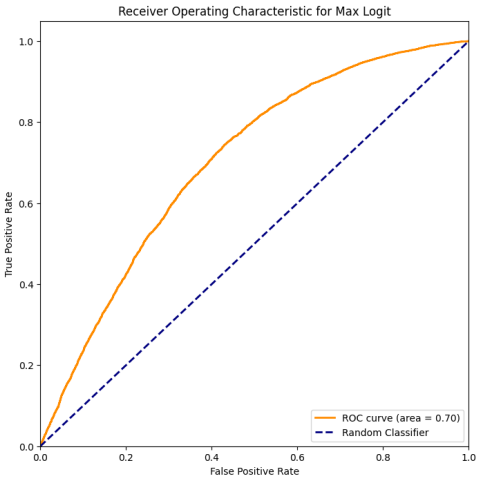

FIGURE 6 – ROC for Max Logit

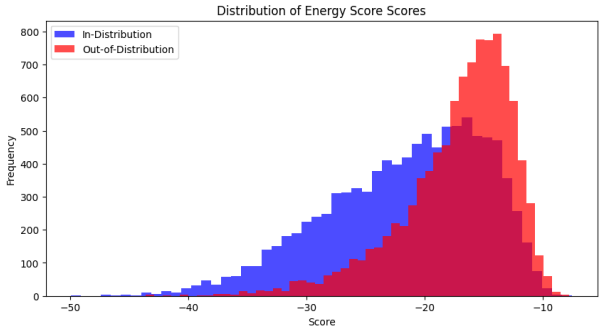

FIGURE 7 – Distribution of Energy Score

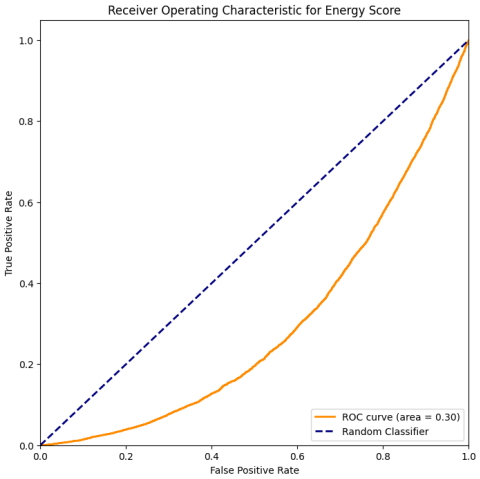

FIGURE 8 – ROC for Energy

performance (to Maximum Logit) with an AUROC of 0.7028 and an AUPR of 0.6588. As seen in Figures 5 and 6.

3. Energy Score Calculation : ID scores ranged from -50.0 to -7.5, and OOD scores ranged from -43.6 to -7.8. Showed significantly lower performance, with AUROC values of 0.2971 and an AUPR of 0.3734. As seen in Figures 7 and 8.
4. Mahalanobis Distance Score : ID scores ranged from 10.1 to 36.0, and OOD scores from 15.9 to 36.0. Performed close to random, yielding an AUROC of 0.5779 and an AUPR of 0.5469. Also showed significantly lower performance, with AUROC values of 0.3759 and an AUPR of 0.4067. As seen in Figures 9 and 10.
5. ViM Score : ID scores ranged from -41.2 to -19.6, and OOD scores ranged from -40.9 to -19.3. Performed close to random, yielding an AUROC of 0.4886 and an AUPR of 0.4836. As seen in Figures 11 and 12.

We can conclude that the raw score-based methods (Maximum Logit and MSP) generally performed better than feature-based methods (Mahalanobis, Energy, ViM) for OOD detection on this specific CIFAR-100 vs. CIFAR-10 task, indicating that the final logits or softmax probabilities are strong indicators of distribution shift. This may be explained by the caracteristics of both OOD and ID datasets studied, as the features of the classes

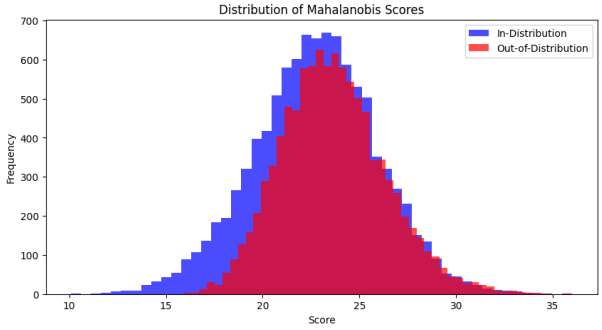

FIGURE 9 – Distribution of Mahalanobis Score

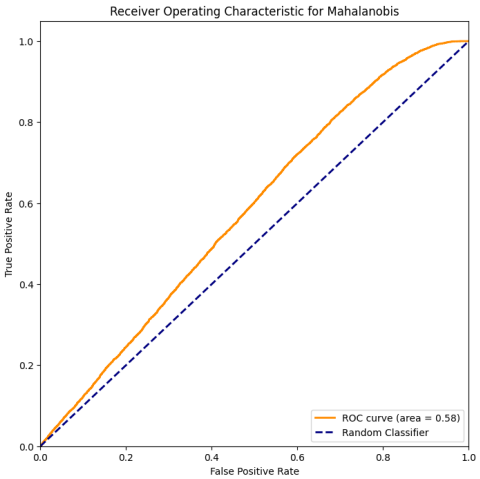

FIGURE 10 – ROC for Mahalanobis

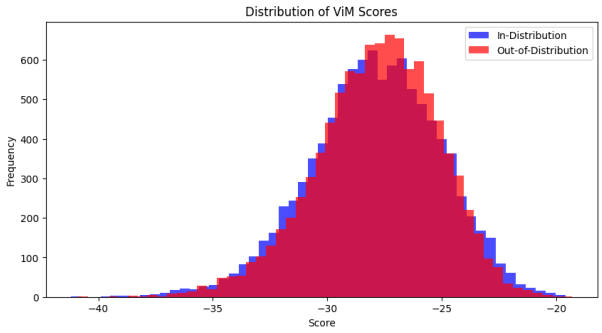

FIGURE 11 – Distribution of ViM Score

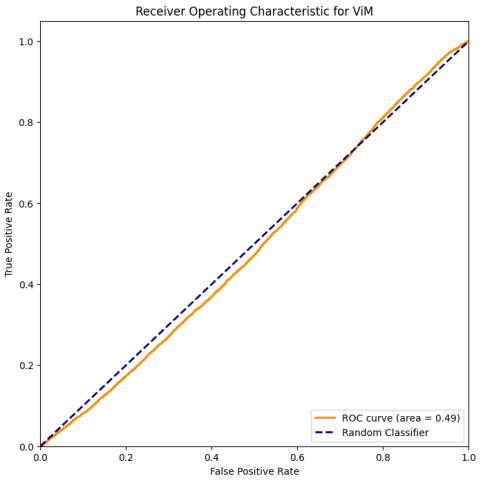

FIGURE 12 – ROC for ViM

of CIFAR-10 may sit in between features of classes of CIFAR-100 that are closely related with the latter.

2 Neural Collapse (NC)

Neural Collapse is a geometric structure that emerges in the final layers of deep neural networks trained to near-zero error. It is defined by four properties :

— NC1 : Features of samples from the same class collapse to their class mean.
— NC2 : The classifier weight vectors align with these class means.
— NC3 : The feature means and classifier weights become perfectly aligned.
— NC4 : The network achieves perfect classification on training data.

Together, these properties show that features and classifiers converge to the vertices of a maximally separated geometric object called a Simplex Equiangular Tight Frame (ETF).

2.1 Why it Matters ?

It helps explain why deep networks generalize well, as the structured representations create clear, robust decision boundaries. It also reveals an inductive bias in neural networks trained with cross-entropy loss toward learning this optimal configuration.

2.2 Why it Happens Late in Training ?

Neural Collapse typically occurs toward the end of training because :

— Optimization Convergence : As loss approaches zero, the cross-entropy loss pushes features closer to their correct classifier weights.
— Implicit Regularization : In overparameterized networks, the optimization process guides the model toward this maximally separated, symmetric solution.

2.3 NC 1 to NC 4

2.3.1 Neural Collapse Stage 1 (NC1) : Within-Class Variability Collapse

NC1 is the first property of Neural Collapse, describing the phenomenon where feature vectors of samples from the same class become nearly identical in the network's penultimate layer by the end of training.

Some characteristics of NC1 :

1. Within-Class Variability Collapses to Zero : All data points from a specific class converge to a single point in the feature space, meaning their variance effectively disappears.
2. Formation of Class Prototypes : Each class forms a unique, fixed representation (its class mean) in the latent space.
3. Perfect Clustering : The training data becomes perfectly clustered, with each cluster being a single, tight point corresponding to a class.

2.3.2 Neural Collapse Stage 2 (NC2) : Class Means Collapse to a Simplex Equiangular Tight Frame (ETF)

NC2 describes the geometric arrangement of class means (prototypes) after they have formed in NC1. Specifically, these class means collapse to the vertices of a Simplex Equiangular Tight Frame (ETF).

What is a Simplex ETF ?

An ETF is a mathematically optimal configuration of vectors where :
— All vectors have identical norms (same length).
— All pairs of vectors have equal pairwise angles (same distance from each other).
— They are centered around the origin, forming a regular simplex (the most symmetric arrangement of points, like an equilateral triangle in 2D or a tetrahedron in 3D).

Some characteristics of NC2 :

1. Optimal Separation : This structure maximizes the angular distance between class means, ensuring classes are as distinct as possible.
2. Robustness : The symmetric, balanced arrangement creates stable decision boundaries, improving generalization.
3. Simplified Classification : It allows the final classifier to function like a simple nearest-neighbor model.
4. Fair Representation : It ensures all classes are treated equally in the feature space.

2.3.3 Neural Collapse Stage 3 (NC3)

NC3 describes the relationship between the last-layer classifier's parameters and the geometric structure established in NC2. It states that the classifier's weights and biases become directly proportional to the class means (the ETF vertices).

Some characteristics of NC3 :
— Weight Alignment : The weight vector for each class becomes proportional to its corresponding class mean.
— Bias Adjustment : The bias term for each class becomes proportional to the negative squared norm of that class mean.

2.3.4 Neural Collapse Stage 4 (NC4)

NC4 describes the implicit optimization of the feature extractor (the network layers before the classifier). It explains how the feature extractor learns to produce the structured representations observed in the earlier stages of Neural Collapse.

Some characteristics of NC4 :

— Projection onto ETF Directions : The feature extractor learns to project input data onto the directions defined by the Simplex ETF formed by the class means (from NC2).

— Alignment with Class Means : For any input, the feature extractor is optimized to align its output closely with the specific mean vector of its corresponding class.

— Discarding Irrelevant Information : The network learns to retain only the discriminative features needed for classification while discarding within-class variations.

2.4 Summary

Neural Collapse (NC) reveals that deep neural networks, when trained to near-zero error, converge toward a geometrically optimal and highly structured configuration in their final layers. The four stages form an interconnected, sequential framework that collectively explains how networks achieve exceptional performance and generalization :

Stage	Description	Function
NC1	Features of the same class collapse to their class mean	Eliminates within-class variability, creating distinct class prototypes
NC2	Class means arrange as vertices of a Simplex ETF	Maximizes inter-class separation through symmetric, equiangular geometry
NC3	Last-layer classifiers align with these class means	Creates optimal decision boundaries determined by the feature geometry
NC4	Feature extractor projects data onto ETF directions	Ensures the entire network produces these structured representations

Table 1 – Summary of each Step

In summary, the stages build upon each other—NC1 creates the prototypes, NC2 arranges them optimally, NC3 aligns the classifier with this arrangement, and NC4 ensures the feature extractor generates these structures from raw data. NC demonstrates that overparameterized networks implicitly discover maximally separable and symmetric representations (Simplex ETFs) as a natural outcome of optimization, explaining their strong generalization despite theoretical capacity for overfitting. Understanding NC opens pathways for designing more robust, interpretable models by deliberately inducing or controlling this geometric structure, potentially improving training stability and transfer learning performance.

In essence, Neural Collapse reveals that deep networks learn to organize their internal representations according to fundamental geometric principles of optimal separation and symmetry.

NC level	Metric in notebook	Results
NC1	$\mathrm{tr}(\Sigma_W)/\mathrm{tr}(\Sigma_B)$	0.542
NC1	$\|\Sigma_W\|_F/\|\Sigma_B\|_F$	0.326
NC2	$\mathrm{std}(\|\mu_c - \mu_G\|)/\mathrm{mean}(\|\mu_c - \mu_G\|)$	0.057
NC2	$\|G - G_{\mathrm{etf}}\|_F/\|G_{\mathrm{etf}}\|_F$	0.840
NC2 (diag.)	Off-diagonal cosine mean/std	-0.01 / 0.085
NC3	$\left\|\frac{W^T}{\|W\|_F} - \frac{\dot{M}}{\|\dot{M}\|_F}\right\|_F$	0.311
NC3 (diag.)	$\cos(w_c, \mu_c - \mu_G)$ (mean/min)	0.956 / 0.941
NC4	$\mathbb{P}$(network = NCC)	0.9998
NC4 (context)	Train acc (network / NCC)	0.99968 / 0.099972

TABLE 2 – Summary of Neural Collapse (NC) metrics.

2.5 Quantitative Results

In our experiments, we measured a number of different metrics related to neural collapse. We observed that model 2 presents much stronger neural collapse than model 1, thus we only present results of the former in this report. Following the same notation as [1], table 2 summarize the results. All but the last 3 metrics should converge to 0, the others should converge to 1, in total NC. We can observe strong NC, particularly in level 2 and 4.

3 Neural Collapse Inspired OOD Detection (NECO)

Based on the article [2], we implemented the NECO method and evaluated its performance on the CIFAR-100 dataset (ID) and CIFAR-10 dataset (OOD). NECO leverages the geometric properties of Neural Collapse, specifically the NC5 property (ID/OOD Orthogonality), to distinguish between in-distribution and out-of-distribution samples by measuring the relative norm of features within the Simplex ETF subspace.

In NC5, OOD representations align with subspaces orthogonal to the ID feature space (Simplex ETF), occupying distinct and complementary dimensions. The orthogonality intensifies with continued training beyond convergence, improving ID/OOD discriminability. In addition, since ID data lies within the ETF subspace while OOD lies in its orthogonal complement, the relative norm within the ETF subspace serves as an effective OOD score — the foundation of the NECO method.

This novel observation extends the neural collapse framework, providing both theoretical understanding and practical utility for OOD detection.

3.1 Quantitative Results

Table 3 presents the quantitative performance of the NECO detector on our test set for model 2. The metrics indicate better than random performance, suggesting that the model reached Neural Collapse regime and that NECO can be used to separate OOD samples, althought it is far from being a perfect OOD identifier.

Metric	**Value**
FPR@95	0.7907
AUROC	0.6693
AUPR	0.6226

TABLE 3 – NECO detection metrics on CIFAR-100

3.2 Qualitative Analysis

To better understand the detector's behavior, we visualize the results through several complementary plots.

3.3 Analysis of Results

The NECO results show **partial** ID/OOD separation rather than complete overlap : AUROC = 0.6693, AUPR = 0.6226, and FPR@95 = 0.7907 (threshold-accuracy : 0.6106). As seen in Figure 14, the ID and OOD score distributions are shifted but still substantially overlapping, which explains why detection remains weak at high recall (high FPR@95).

This suggests that the NECO signal is present but not yet strong enough for reliable OOD rejection. In practice, this can indicate that Neural Collapse geometry is only partially realized on this model/dataset setup, and that NC5-like orthogonality between ID and OOD features is incomplete.

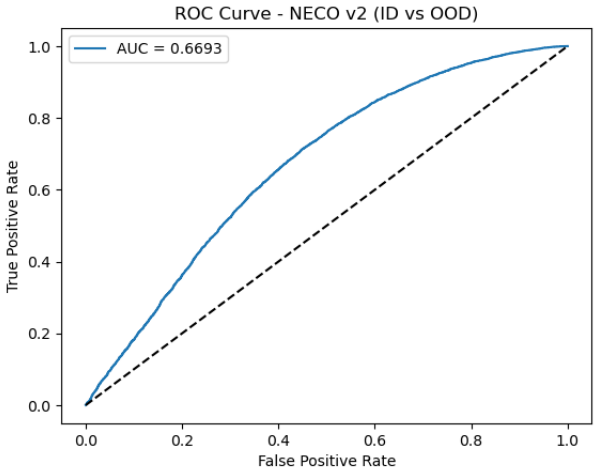

FIGURE 13 – ROC Curve - NECO Detector

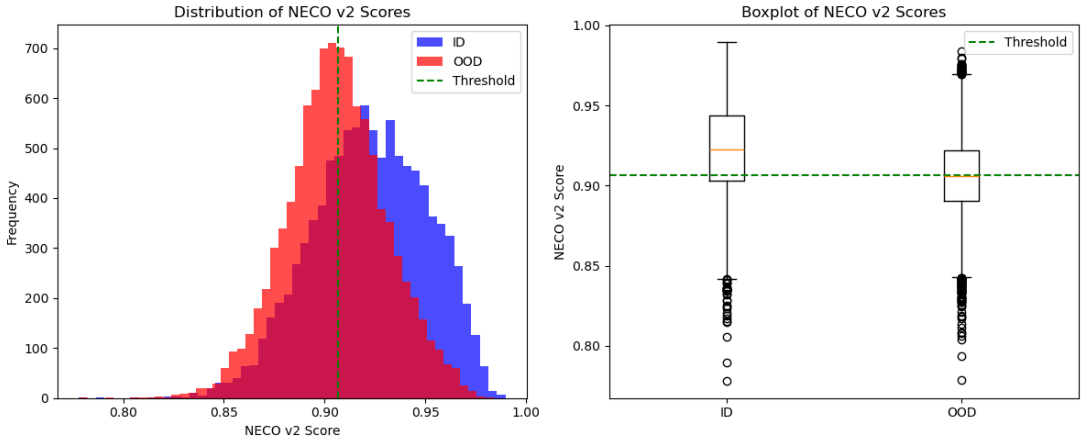

FIGURE 14 – Distribution and Boxplot of NECO Scores for ID and OOD Samples

Conclusion

This work shows that the trained classifier exhibits several Neural Collapse trends (reduced within-class variability, more structured class-mean geometry, and improved classifier/feature alignment), but not full collapse in the strict ideal sense. The NECO-based OOD analysis is consistent with that : performance is above random (AUROC 0.6693, AUPR 0.6226) yet still limited, especially at high recall (FPR@95 0.7907), indicating substantial overlap between ID and OOD feature behavior. Overall, the results suggest partial emergence of NC geometry with incomplete ID/OOD separation. A possible next step is to track NC and NECO metrics across training checkpoints and calibration settings to identify when (and if) stronger collapse and better OOD discrimination appear.

Bibliographie et références

Références

[1] Vardan Papyan, X. Y. Han et David L. Donoho. "Prevalence of Neural Collapse during the Terminal Phase of Deep Learning Training". In : *Proceedings of the National Academy of Sciences* 117.40 (2020), p. 24652-24663. DOI : 10.1073/pnas.2015509117.

[2] Mouïn Ben Ammar et al. *NECO : NEural Collapse Based Out-of-distribution detection*. 2024. arXiv : 2310.06823 [stat.ML]. URL : https://arxiv.org/abs/2310.06823.