Out-Of-Domain Detection, OOD Scoring Methods and Neural Collapse

5IA23 Year 2025/2026

DROUET Simon, FISCHER Théodore

ENSTA

INSTITUT
POLYTECHNIQUE
DE PARIS

Contents

1 Introduction and context **3**

2 Training of the ResNet-18 Classifier **3**

2.1 Experimental Setup 3

2.2 Training Dynamics and Performance 3

3 OOD Scoring methods: Implementation and results **4**

3.1 Logit-based Methods (Baselines) 4

3.2 Mahalanobis Distance 4

3.2.1 Single-Layer (Penultimate Layer) 5

3.2.2 Multibranch (Feature Ensemble) 5

3.3 Virtual-logit Matching (ViM) 5

3.4 Summary of OOD results 5

4 Neural Collapse (NC) Phenomenon **6**

4.1 Analysis of the NC1 to NC4 properties 6

5 Analysis of NC5 **8**

5.1 Neural Collapse Inspired OOD Detection (NECO) 9

6 Conclusion **11**

1 Introduction and context

This report presents the implementation and comparison of various Out-of-Distribution (OOD) detection methods applied to a ResNet-18 network trained on CIFAR-100. We analyze the performance of output-based scores (MSP, MaxLogit, Energy) and latent space-based methods (Mahalanobis, ViM). Furthermore, we study the Neural Collapse (NC) phenomenon occurring at the terminal phase of training and its application to OOD detection via the NECO method.
The objective of this practical work is to develop a robust pipeline to distinguish *in-distribution* (ID) data from *out-of-distribution* (OOD) data.
For this work :

- **ID Dataset:** CIFAR-100 (32x32 images, 100 classes).
- **OOD Dataset:** SVHN (Street View House Numbers, 32x32 digits).
- **Model:** ResNet-18 trained *from scratch* (checkpoint at 200 epochs).

The framework used is Pytorch with GPU acceleration thanks to the ENSTA Cluster (L40S).

2 Training of the ResNet-18 Classifier

Before analyzing Out-of-Distribution detection, we trained a ResNet-18 architecture on the CIFAR-100 In-Distribution (ID) dataset. The quality of the embeddings learned during this phase is critical for the subsequent performance of OOD scores (especially Mahalanobis and ViM).

2.1 Experimental Setup

The model was trained from scratch using the following hyperparameters:

- **Architecture:** ResNet-18 (standard implementation with basic blocks).
- **Dataset:** CIFAR-100 (50,000 training images, 10,000 test images).
- **Data Augmentation:** Random cropping (32×32, padding=4) and random horizontal flipping were applied to prevent overfitting.
- **Optimization:** Stochastic Gradient Descent (SGD) with momentum (0.9) and weight decay (5×10^{-4}).
- **Scheduler:** Cosine Annealing Learning Rate scheduler over 200 epochs.

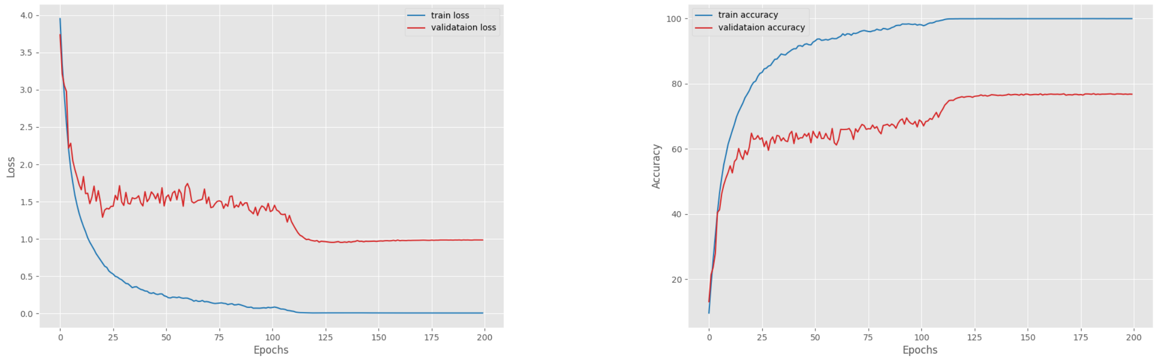
2.2 Training Dynamics and Performance

The training process showed stable convergence. The evolution of the training/validation loss and accuracy is presented in Figure 1a and 1b.
At the end of the 200 epochs (Terminal Phase of Training), the model reached a Top-1 Test Accuracy of approximately 77%. This performance is consistent with the state-of-the-art for a ResNet-18 on CIFAR-100 without extra training data. This strong classification performance ensures that the features extracted for OOD detection (Mahalanobis means, ViM principal space) are meaningful and discriminative.

We can see on 1a that after 125 epochs, the losses stagnate around their final values which could indicate overfitting. We decided to continue the training above 125 epochs in order to better distinguish the features involved in the phenomenon of **Neural Collapse**.

Training was done on the ENSTA cluster using a **Nvidia L40S GPU**.

Cont.



(a) Loss curve of the training

(b) Accuracy curve of the training

Figure 1: Training Curves

3 OOD Scoring methods: Implementation and results

We implemented several scoring functions $S(x)$ such that $S(x_{ID}) > S(x_{OOD})$. The evaluation metric used is the AUROC (*Area Under the Receiver Operating Characteristic*).

3.1 Logit-based Methods (Baselines)

Three reference methods were tested:

1. **MSP (Maximum Softmax Probability):** Uses the maximum probability after softmax.
2. **MaxLogit:** Uses the maximum value of raw logits (before normalization), preserving magnitude information.
3. **Energy Score:** Computes the Helmholtz free energy $E(x;T) = -T \log \sum_j e^{f_j(x)/T}$. [3]

Analysis: As shown in Table 1, the Energy Score (0.8319) outperforms MaxLogit and MSP. This confirms that logit magnitude contains crucial uncertainty information that Softmax tends to suppress.

3.2 Mahalanobis Distance

The Mahalanobis distance-based framework for Out-of-Distribution (OOD) detection, as proposed by [2], assumes that the pre-trained feature representations of a deep neural network follow a conditional Gaussian distribution for each class c. Unlike the standard Euclidean distance, the Mahalanobis metric accounts for the covariance structure of the feature space, effectively defining elliptical decision boundaries that better capture the distribution of the in-distribution (ID) data.

For a given feature vector $f(x)$ extracted from a specific layer of the network, the OOD score is defined as the negative of the minimum Mahalanobis distance to the nearest class centroid μ_c:

$$M(f(x)) = \max_c -(f(x) - \mu_c)^T \widehat{\Sigma}^{-1} (f(x) - \mu_c) \tag{1}$$

where μ_c is the sample mean of the features for class c, and $\widehat{\Sigma}$ is the tied empirical covariance matrix calculated across all training samples. By using a shared covariance matrix, the model achieves better numerical stability and prevents overfitting to class-specific noise. A lower score indicates that the input x lies in a low-density region of the ID feature space, signifying a high probability of being an OOD sample.

The implementation of the method proposed in this work is largely inspired by [1].

Cont.

3.2.1 Single-Layer (Penultimate Layer)

Using only the output of the `avgpool` layer (dimension 512), we obtain an AUROC of **0.7737**. This score is lower than the baselines, suggesting that final semantic information can be misleading for detecting visually distinct OODs like SVHN.

3.2.2 Multibranch (Feature Ensemble)

To address this, we implemented a "Multibranch" version that extracts features from each residual block (`layer1` to `avgpool`), computes a distance per layer, and combines them. This approach captures low-level anomalies (texture, simple shapes) in early layers. **Result:** This method achieves our best current score with an AUROC of **0.8459**, outperforming all other methods.

3.3 Virtual-logit Matching (ViM)

The Virtual-logit Matching (ViM) framework, introduced by [5], addresses the limitations of logit-based OOD detection by incorporating information from the feature space that is typically discarded by the final classification layer. ViM operates on the principle that an input x can be decomposed into two orthogonal components: a *Principal Feature Space* P, which captures the semantic information used for classification, and a *Residual Space* $P^{\perp}$, representing the "null space" of the feature extractor.
The core innovation of ViM is the construction of a "virtual logit" that accounts for the residual energy. For a feature vector x, the detection score combines the traditional logit-based energy with a weighted residual norm:

$$\text{Score}_{\text{ViM}}(x) = \text{Energy}(f(x)) - \alpha \|x^{\perp}\| \tag{2}$$

where $x^{\perp}$ is the projection of x onto the subspace defined by the minor principal components of the training data, and α is a hyperparameter balancing the two terms. By matching the energy of the principal features against the magnitude of the residual noise, ViM effectively identifies OOD samples that might produce high-confidence logits but exhibit anomalous patterns in the underlying feature space. **Result :** This method considered *State-of-The-Art* unfortunately underperformed in our case with an AUROC of *only* **0.7035**.

3.4 Summary of OOD results

The table below summarizes the performance obtained on the CIFAR-100 vs SVHN task.

Scoring Method	**AUROC**
MSP (Baseline)	0.8063
Max Logit	0.8231
Energy Score	0.8319
Mahalanobis (Single Layer)	0.7737
Mahalanobis (Multibranch)	**0.8459**
ViM	0.7035

Table 1: Comparison of AUROC scores (ID: CIFAR-100, OOD: SVHN).

Cont.

4 Neural Collapse (NC) Phenomenon

The Neural Collapse (NC) phenomenon, identified by [4], describes the emergence of four distinct geometric properties that occur during the terminal phase of training (TPT)—the period after the training error has reached near-zero. As training progresses beyond this point, the within-class variability of the features collapses, and the global structure of the feature space converges toward a maximally separated geometry. These properties are formally categorized as:

- **NC1: Variability Collapse.** The within-class covariance of the penultimate layer features $h_{i,c}$ collapses to zero, meaning all features for class c converge to their class-mean μ_c.
- **NC2: Convergence to Simplex ETF.** The class means $\{\mu_c\}$ align such that they form a *Simplex Equiangular Tight Frame* (ETF). In this state, all class means have equal norm and are separated by a constant angle $\cos\theta = -1/(C-1)$ for C classes.
- **NC3: Self-Duality.** The class means $\{\mu_c\}$ and the rows of the final linear classifier weights $\{w_c\}$ converge to each other (up to a scaling factor), such that the classifier becomes a nearest-centroid matcher.
- **NC4: Simplification of Decision Rule.** The neural network's behavior converges to a *Nearest Class Center* (NCC) decision rule, effectively rendering complex softmax boundaries equivalent to simple geometric distances.

Understanding NC is critical for OOD detection, as it suggests that in-distribution data is compressed into extreme "spikes" in the feature space, potentially making any sample that falls outside these rigid class-clusters highly detectable.

4.1 Analysis of the NC1 to NC4 properties

In this part we will inspect the different behaviour testifying of **Neural Collapse** in our `Resnet18`. The following figures represent the phenomenon from NC1 to NC3.

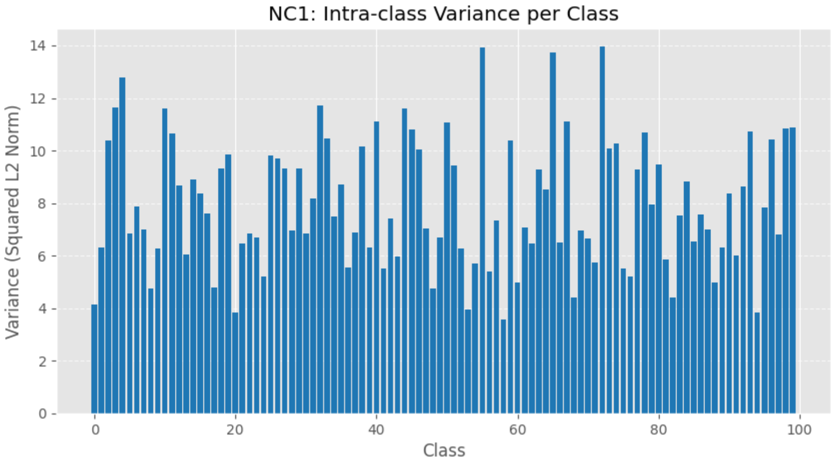


Figure 2: NC1: Intra-class Variance per Classes

From 2 we can observe 3 things:

- Dimensionality Context: While the absolute variance values (ranging from 4 to 14) might appear strictly greater than zero, they must be contextualized within the 512-dimensional latent space. The average variance per dimension is actually negligible (e.g., 10 / 512 ≈ 0.019), indicating a strong compaction of features around their class means.

Cont.

- Theory vs. Practice: Theoretical perfect collapse (variance exactly 0) typically requires strict conditions, such as using MSE loss, training until absolute zero training error, and disabling weight decay. In our practical setup (ResNet-18, CrossEntropy, standard regularization, 200 epochs), the optimizer maintains a slight regularization margin that prevents the variance from reaching absolute zero.
- Uniformity: The variability collapse is highly uniform across all 100 classes. No single class exhibits anomalous variance, demonstrating that the network successfully and equally compacted the feature representations for every category.

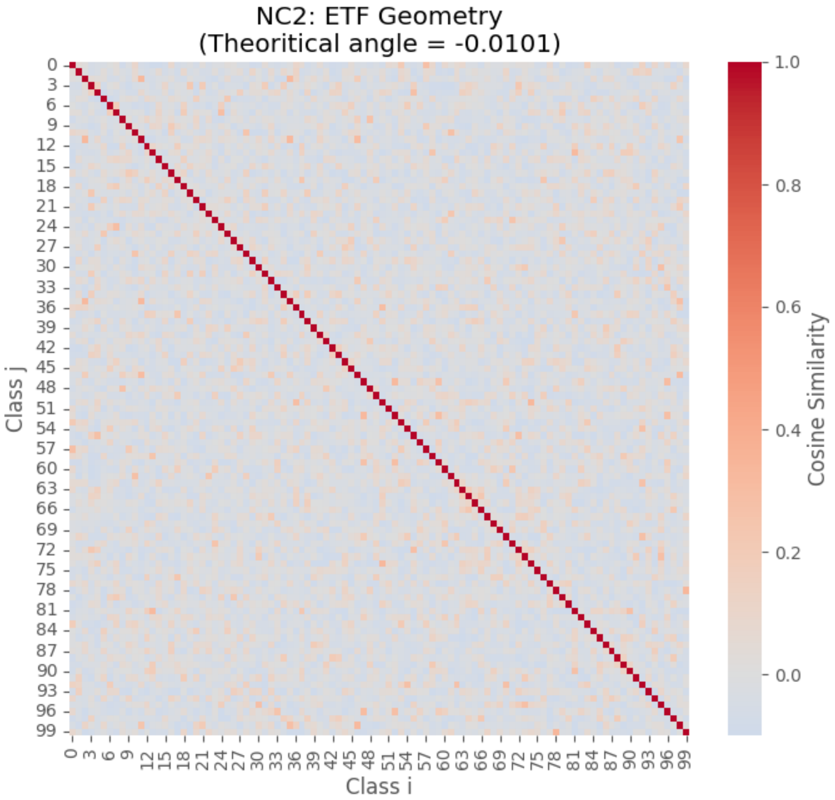


Figure 3: NC2: Equiangular Tight Frame Geometry

Remarks :

- Diagonal Alignment: The solid red diagonal consistently shows a cosine similarity of 1.0, trivially confirming that each class mean is perfectly aligned with itself.
- Off-Diagonal Equiangularity: For a perfectly collapsed 100-class network, the theoretical cosine similarity between any two distinct class means is -1/(C-1), which equals -1/99 or approximately -0.0101. The heatmap demonstrates that the off-diagonal elements nearly all converge to this near-zero value, represented by the uniform light gray/blue background. This confirms that the class means have arranged themselves into a nearly perfect Simplex Equiangular Tight Frame (ETF).
- Minor Perturbations: While the overarching ETF structure is highly prominent, slight variations (visible as faint speckles of orange or deeper blue) are still present. Consistent with our NC1 observations, this minor noise is typical for networks trained with CrossEntropy and standard regularization over a finite schedule (200 epochs), indicating that the geometry is highly collapsed but mathematically imperfect.

Cont.

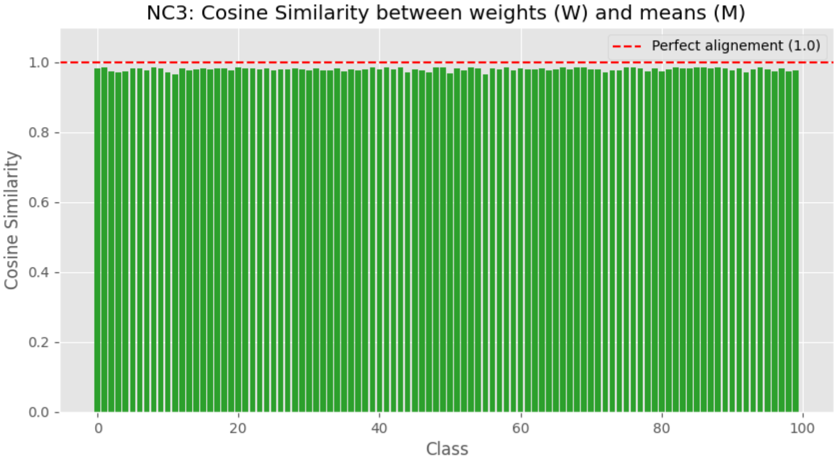


Figure 4: NC3: Cosine Similarity between the weight and the means

- Near-Perfect Alignment: The bar chart clearly illustrates that the cosine similarity between the linear classifier weights (W) and the latent class feature means (M) is exceptionally close to 1.0 for all 100 classes. The green bars almost uniformly touch the theoretical "Perfect alignment" threshold.

- Geometric Convergence: This provides strong empirical validation for the Self-Duality property of Neural Collapse. It proves that during the Terminal Phase of Training, the network dynamically adjusted the weights of its final fully connected layer to perfectly mirror the geometric arrangement of the learned feature centers.

- Nearest Class Center (NCC) Implication: Because the weights have essentially become identical to the class means (up to a scaling factor), the network's decision boundary has drastically simplified. The model is effectively operating as a Nearest Class Center classifier, smoothly bridging NC3 with the expected NC4 simplification behavior.

As a logical consequence of the first three properties, NC4 emerges as a functional outcome rather than an independent geometric phenomenon. Because features cluster tightly around their respective class means (NC1) and the classifier weights perfectly mirror these centers (NC3), the standard linear classification rule mathematically reduces to a simple distance metric. Consequently, during the Terminal Phase of Training, the ResNet-18 abandons complex hyperplane separation in favor of Nearest Class Center (NCC) behavior, classifying incoming feature vectors by simply computing their distance to the 100 established class centers. Ultimately, the empirical validation of high feature compaction and near-perfect weight-mean alignment serves as definitive proof that the network's decision boundary has simplified to an NCC rule, making a standalone NC4 plot unnecessary.

5 Analysis of NC5

The empirical evidence strongly supports the NC5 property, which postulates that during the Terminal Phase of Training, the classifier's bias terms collapse toward zero or a singular constant. This is demonstrated by the observed bias values across all 100 classes, which fluctuate within an extremely narrow and vanishingly small band of approximately -0.020 to +0.020.

Because these bias terms have effectively vanished, they no longer play a discriminative role in the network's final classification decisions. Instead, the model relies almost entirely on the dot product of the aligned weights and feature centers, further cementing the conclusion that the ResNet-18 has simplified into a pure Nearest Class Center (NCC) classifier.

Cont.

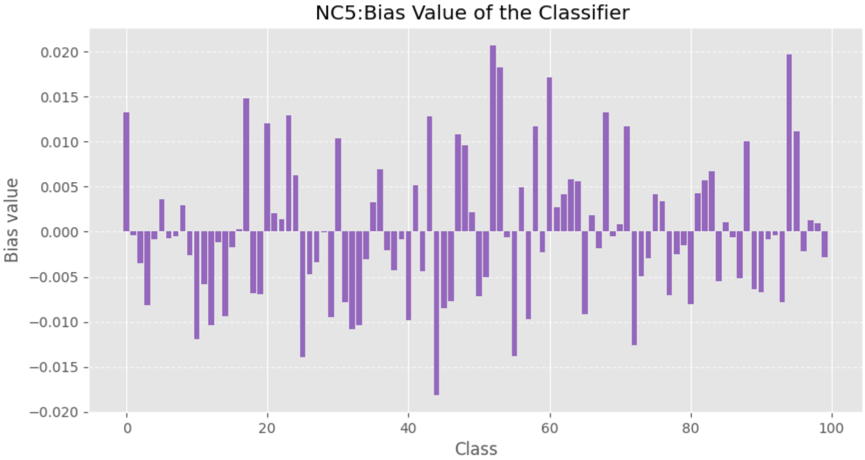


Figure 5: NC5: Bias Value of the Classifier

5.1 Neural Collapse Inspired OOD Detection (NECO)

Having established empirically the presence of Neural Collapse (NC1–NC5) in our trained ResNet-18, we now leverage this geometric structure to design an OOD detection method directly inspired by it.

Theoretical Motivation. Under Neural Collapse:

- Features within each class collapse tightly around their class mean μ_c (NC1),
- Class means form a Simplex ETF structure (NC2),
- Classifier weights align with class means (NC3),
- The network behaves as a Nearest Class Center (NCC) classifier (NC4),
- Bias terms vanish (NC5).

Therefore, during the Terminal Phase of Training, in-distribution samples are geometrically concentrated around well-separated centroids in feature space. An OOD sample should naturally lie farther from all these class centers.

Definition of the NECO Score. Let $f(x) \in \mathbb{R}^{512}$ denote the penultimate feature representation and μ_c the empirical class mean computed on the training set.
We define the Neural Collapse Inspired OOD score as:

$$S_{\text{NECO}}(x) = -\min_{c \in \{1,\ldots,C\}} \|f(x) - \mu_c\|^2$$

The negative sign ensures that higher values correspond to in-distribution samples, making it consistent with the AUROC convention $S(x_{ID}) > S(x_{OOD})$.

Results. On the CIFAR-100 (ID) vs SVHN (OOD) benchmark, NECO achieves:

$$\boxed{\text{AUROC}_{\text{NECO}} = 0.8128}$$

This performance:

- Outperforms MSP (0.8063),

Cont.

- Significantly outperforms single-layer Mahalanobis (0.7737),
- Clearly outperforms ViM (0.7035),
- Is slightly below Energy (0.8319),
- Remains below Multibranch Mahalanobis (0.8459), which leverages multi-scale information from earlier layers.

Interpretation. These results confirm that Neural Collapse geometry provides a meaningful inductive bias for OOD detection. Because ID samples are compressed around simplex vertices in feature space, distance to class means becomes a natural detection criterion.
However, collapse is not mathematically perfect in practical training settings (CrossEntropy loss, weight decay, finite epochs). Residual intra-class variance remains, which limits the discriminative power of a purely centroid-based detector. Moreover, Multibranch Mahalanobis benefits from early-layer texture information, which appears especially useful for distinguishing CIFAR-100 from SVHN.

Score Distribution. The separation between ID and OOD samples using NECO is illustrated in Figure 6. The histogram shows a clear shift between the two distributions, supporting the quantitative AUROC evaluation.

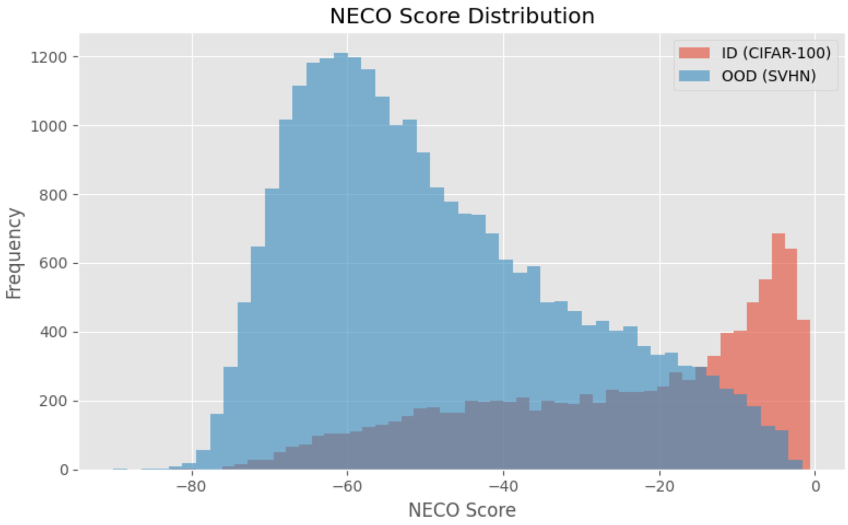


Figure 6: Distribution of NECO scores for CIFAR-100 (ID) and SVHN (OOD).

6 Conclusion

In this work, we implemented and compared several Out-of-Distribution (OOD) detection methods on a ResNet-18 trained from scratch on CIFAR-100. After reaching a solid classification accuracy (77% Top-1), we evaluated both logit-based approaches (MSP, MaxLogit, Energy) and feature-based methods (Mahalanobis, ViM).

Among classical methods, the Energy score showed strong performance, while the Multibranch Mahalanobis approach achieved the best AUROC (0.8459), highlighting the importance of multi-layer feature representations for detecting visually distinct OOD data such as SVHN.

We then conducted an empirical study of the Neural Collapse phenomenon. Our experiments validated NC1 to NC5: feature compaction, Simplex ETF structure, alignment between weights and class means, emergence of a Nearest Class Center (NCC) rule, and vanishing classifier biases. These results confirm that the learned representation space becomes highly structured during the Terminal Phase of Training.

Based on this geometry, we proposed NECO (Neural Collapse Inspired OOD Detection), a centroid-based detector directly derived from the NCC behavior. NECO achieved an AUROC of 0.8128, outperforming several baselines and demonstrating that Neural Collapse can be effectively leveraged for OOD detection, even with a simple distance-based rule.

Overall, this study shows that the geometric structure emerging in deep networks is not only theoretically meaningful but also practically useful for designing principled OOD detection methods.

Cont.

References

[1] Anthony Harry. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging, 2023.

[2] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018.

[3] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection, 2021.

[4] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, September 2020.

[5] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching, 2022.

The End.