Out-of-Distribution Detection and Neural Collapse

Fares Boudeaa, Tiena Soro

Abstract

This report studies Out-of-Distribution (OOD) detection methods and the Neural Collapse phenomenon in deep networks. We train a ResNet-18 classifier on CIFAR-100 and compare multiple OOD scoring methods, including MSP, Energy, Mahalanobis, and ViM. We then analyze Neural Collapse (NC1–NC5) at the end of training and implement the NECO method for OOD detection.

Introduction

Motivation

Deep neural networks often produce overconfident predictions on out-of-distribution samples. Reliable uncertainty estimation is therefore crucial in safety-critical applications such as autonomous driving, medical imaging, and fraud detection. In these settings, a model that silently assigns high confidence to inputs it has never seen can cause serious harm. OOD detection addresses this by equipping a classifier with the ability to recognize when an input falls outside its training distribution, and to flag it accordingly rather than producing a potentially misleading prediction. In this report, we study several OOD scoring methods and explore how the geometric structure that emerges at the end of training, known as Neural Collapse, can be leveraged for this purpose.

Setup

Residual Networks (ResNets) were originally designed for large-scale image classification on ImageNet, which consists of high-resolution 224×224 images. We adapt ResNet-18 for CIFAR-100 by replacing the first convolutional layer with a smaller kernel and stride, and removing the initial max-pooling layer, so that the spatial resolution of the small 32×32 inputs is preserved through the early layers of the network.

We train a ResNet-18 adapted for CIFAR-100 from scratch and analyze the feature representations of dimension $H = 50000 \times 512$ at the penultimate layer before the classification head.

The CIFAR-100 dataset consists of 60,000 color images of size 32×32 divided into 100 classes, further grouped into 20 superclasses. Every class has 600 examples, 500 for training and 100 for testing. It mainly contains images of everyday objects, animals, vehicles, and vegetation. We perform data augmentation for a more robust model: random crop with padding, random horizontal flip, and normalization with CIFAR-100 statistics. We then train the ResNet model until the training loss reaches 0, even though the training accuracy already reached 99.9%, to ensure that the model reaches the terminal training phase.

Training Procedure

Optimization We train using Stochastic Gradient Descent (SGD) with momentum combined with hyperparameter tuning, as it outperforms Adam for this setting. The chosen parameters are the following:

- Optimizer: SGD with momentum 0.9
- Weight decay: 5×10^{-4}
- Initial learning rate: 0.1
- LR schedule: cosine decay / step decay
- Loss: Cross Entropy Loss

Training Duration

- Batch size: 128
- Number of epochs: 360
- Training Loss: 0.001
- Training Accuracy: 99.9%

In order to evaluate the out-of-distribution metrics as well as evaluate NECO for OOD detection, a different dataset containing classes not present in CIFAR-100 is used. Here we chose the Street View House Numbers (SVHN) dataset, which contains labeled images of house numbers extracted from Google Street View. We compare both datasets' means and standard deviations to confirm that the two datasets are not very similar:

- CIFAR-100 : mean $= (0.5071, 0.4867, 0.4408)$
- SVHN : mean $= (0.4377, 0.4438, 0.4728)$
- CIFAR-100 : std $= (0.2675, 0.2565, 0.2761)$
- SVHN : std $= (0.1980, 0.2010, 0.1970)$

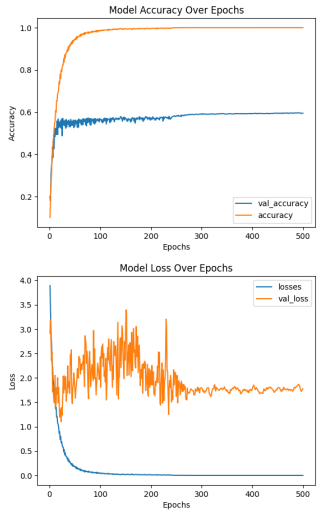

Figure 1: Training and Test, Loss and Accuracy

OOD Scoring Methods

Problem Setup

Let $(x, y) \sim \mathcal{D}_{\text{ID}}$ denote in-distribution data. At test time, inputs may instead come from $\mathcal{D}_{\text{OOD}}$.

The goal of OOD detection is to define a score function $S(x)$ such that:

$$S(x_{\text{ID}}) > S(x_{\text{OOD}})$$

Evaluation metrics: *AUROC*, *AUPR*, *FPR@95*
Let $f(x) \in R^K$ be the logits of the network.

Max Softmax Probability (MSP)

$$S_{\text{MSP}}(x) = \max_k \text{softmax}(f(x))_k$$

Maximum Logit Score

$$S_{\text{Logit}}(x) = \max_k f_k(x)$$

Energy Score

$$S_{\text{Energy}}(x) = -T \log \sum_{k=1}^{K} e^{f_k(x)/T}$$

Mahalanobis Score

Let μ_c be class means and Σ the shared covariance matrix.

$$S_{\text{Mahalanobis}}(x) = -\min_c (h(x) - \mu_c)^T \Sigma^{-1} (h(x) - \mu_c)$$

ViM

ViM decomposes features into principal and residual subspaces:

$$h(x) = h_{\parallel}(x) + h_{\perp}(x)$$

The score combines logit magnitude and residual norm.

NECO: Neural Collapse Inspired OOD Detection

Motivation

OOD samples violate the geometric structure induced by Neural Collapse. In-distribution samples, at the end of training, have features that cluster tightly around their respective class means, which are arranged in a Simplex ETF configuration. OOD samples do not belong to any of the trained classes and therefore do not align with this structure. NECO exploits this by measuring how much of a feature vector lies in the subspace spanned by the class means: an ID sample should project strongly onto this subspace, while an OOD sample should not.

Method

Let P denote the orthogonal projector onto the subspace spanned by the centered class mean vectors. The NECO score is defined as the relative projection of the feature onto this subspace:

$$\text{NECO}(x) = \frac{\|Ph_\omega(x)\|}{\|h_\omega(x)\|} = \frac{\sqrt{h_\omega(x)^\top PP^\top h_\omega(x)}}{\sqrt{h_\omega(x)^\top h_\omega(x)}}$$

This score is large when the feature vector aligns with the class-mean subspace, which happens for in-distribution inputs. For OOD inputs, the feature vector has a significant component orthogonal to the class-mean subspace, making the score smaller. The score is scale-invariant by construction, which makes it robust to the norm growth described in NC1.

Neural Collapse

Neural Collapse emerges at the terminal phase of training when cross-entropy loss approaches zero.

Let $h_{i,c}$ denote features of sample i from class c.

NC1: Variability Collapse

Within-class covariance:

$$\Sigma_W = \frac{1}{N} \sum_{i,c} (h_{i,c} - \mu_c)(h_{i,c} - \mu_c)^T \to 0$$

In practice we use:

$$\text{Tr}\left[\frac{\Sigma_W \Sigma_B^\dagger}{C}\right]$$

We use this metric instead of computing Σ_W directly because as training progresses the norms of $h_{i,c}$ increase. This is due to gradient descent: $w = w - \eta \nabla L$. As L decreases, the gradient is negative, so $-\eta \nabla L$ is positive and the weights keep growing in magnitude. This is seen in the following figure. In classification, Σ_W captures within-class noise while Σ_B represents the between-class covariance:

$$h_{i,c} = \mu_c + \epsilon_{i,c}$$

$$\Sigma_B = \frac{1}{C} \sum_c (\mu_c - \mu_G)(\mu_c - \mu_G)^T$$

$$\Sigma_W = \frac{1}{N} \sum_{i,c} \epsilon_{i,c} \epsilon_{i,c}^T$$

By using $\mathrm{Tr}[\Sigma_W \Sigma_B^\dagger / C]$, we obtain a scale-invariant measure of within-class variability relative to between-class spread. As this ratio converges to zero, the within-class noise becomes negligible compared to the separation between classes, which is precisely the condition of variability collapse.

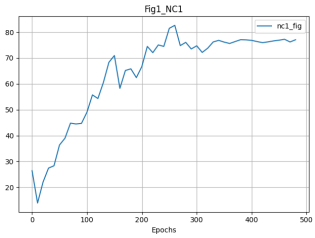

Figure 2: Computation of $\|\Sigma_W\|$ through the epochs. The norm keeps growing, which shows its unreliability as a direct measure of collapse.

NC2: Simplex ETF Structure

Class means become equiangular and equal norm:

$$|\|\mu_c - \mu_G\|_2 - \|\mu_{c'} - \mu_G\|_2| \to 0 \quad \forall c, c'$$

$$\langle \tilde{\mu}_c, \tilde{\mu}_{c'} \rangle \to \frac{C}{C-1} \delta_{c,c'} - \frac{1}{C-1} \quad \forall c, c'$$

In practice we use:

$$\mathrm{EN}_{\text{class-means}} = \frac{\mathrm{std}_c\{\|\mu_c - \mu_G\|_2\}}{\mathrm{avg}_c\{\|\mu_c - \mu_G\|_2\}}$$

and equi-angularity

$$\mathrm{Equi\_ang}_{\text{c-m}} = \mathrm{Avg}_{c,c'} \left| \frac{\langle \mu_c - \mu_G, \mu_{c'} - \mu_G \rangle + \frac{1}{C-1}}{\|\mu_c - \mu_G\|_2 \|\mu_{c'} - \mu_G\|_2} \right|$$

The practical metrics are self-evident. For the first, we compute the standard deviation of the norms of the class mean vectors centered around the global mean, divided by the average norm to account for the increasing magnitude mentioned earlier. The geometric interpretation of this metric approaching 0 is that all class mean vectors are at equal distance from the global mean in the high-dimensional feature space. The second metric computes the average pairwise deviation from the ideal cosine similarity $-\frac{1}{C-1}$. Geometrically, as this metric approaches 0, the angle between any two centered class means converges to the maximum possible value, which corresponds to the Simplex ETF structure.

NC3: Self-Duality

Classifier weights w_c align with class means:

$$\left\| \frac{W^T}{\|W\|_F} - \frac{\dot{M}}{\|\dot{M}\|_F} \right\|_F \to 0$$

NC3 states that the classifier weight matrix and the class mean matrix become proportional at the end of training. The normalized weight matrix $\frac{W^T}{\|W\|_F}$ converges to the normalized class mean matrix $\frac{\dot{M}}{\|\dot{M}\|_F}$, meaning the classifier weights are the dual basis of the class means. In other words, the linear classification head no longer learns a separate geometric structure from the features; it mirrors the arrangement of the class mean vectors exactly. This is a direct consequence of the simplex ETF structure, since both the class means and the classifier weights converge to the same equiangular tight frame.

NC4: Nearest Class Center Rule

Classification becomes equivalent to:

$$\arg\max_{c'} \langle w_{c'}, h \rangle + b_{c'} \to \arg\min_{c'} \|h - \mu_{c'}\|_2$$

NC4 is a direct consequence of NC2 and NC3. Since the weight vectors align with the class means and all class means have equal norm, the dot product between a feature vector and a class weight vector becomes a monotone function of the Euclidean distance to the corresponding class mean. The linear classifier therefore degenerates to a simple nearest-class-center decision rule: classifying a sample amounts to finding the class whose mean is closest to the feature vector in Euclidean distance. This behavioral simplification is what makes the classifier's geometry interpretable and is also exploited by OOD methods such as Mahalanobis and NECO.

NC5

NC5 characterizes the relationship between the in-distribution class mean vectors and the global mean of OOD features.

$$\forall c, \ \frac{\langle \mu_c, \mu_G^{\text{OOD}} \rangle}{\|\mu_c\|_2 \|\mu_G^{\text{OOD}}\|_2} \to 0$$

In practice, we use the following metric:

$$\mathrm{OrthoDev}_{\text{classes-OOD}} = \mathrm{Avg}_c \left| \frac{\langle \mu_c, \mu_G^{\text{OOD}} \rangle}{\|\mu_c\|_2 \|\mu_G^{\text{OOD}}\|_2} \right|$$

NC5 states that as training converges, the in-distribution class mean vectors become approximately orthogonal to the global mean of OOD features. Intuitively, the training objective pushes the class means to spread out in a symmetric ETF configuration centered near the origin, so their average direction in feature space becomes diffuse. An OOD sample, whose feature vector does not align with any of the trained class means, will tend to have a non-negligible projection onto the OOD global mean. This orthogonality property provides a geometric basis for separating ID and OOD samples and is directly exploited by the NECO score described in Section .

Results

We compute the same metrics used in (1), and find very similar results:

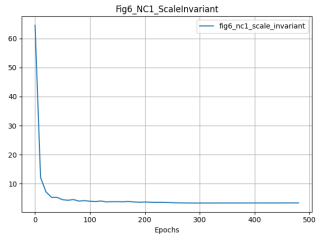

Figure 3: Scale-invariant within-class covariance metric converging to zero over training epochs.

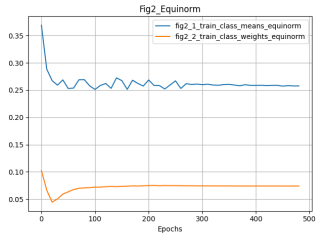

Figure 4: Standard deviation of class mean norms and weight vector norms, each scaled by their average, converging to 0. This confirms that all class mean vectors and weight vectors converge to the same norm.

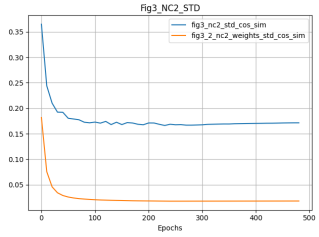

Figure 5: Average pairwise cosine similarity deviation for class mean vectors and weight vectors, converging to 0. This confirms that all pairs of class means and weight vectors form the same maximum angle, consistent with the Simplex ETF structure.

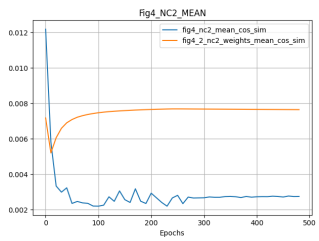

Figure 6: Mean cosine similarity between class mean vector pairs and weight vector pairs over training.

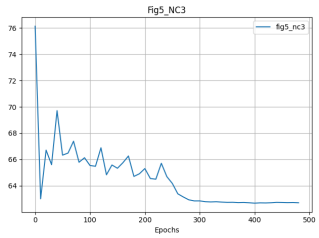

Figure 7: The Frobenius norm $\left\| \frac{W^T}{\|W\|_F} - \frac{\dot{M}}{\|\dot{M}\|_F} \right\|_F$ approaches 0 over training, confirming that the normalized weight matrix converges to the normalized class mean matrix. The classifier weights become the dual basis of the class means.

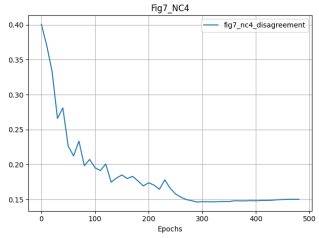

Figure 8: The disagreement between the linear classifier and the nearest-class-center rule tends to zero, confirming that classification simplifies to the nearest class-mean decision rule at the end of training.

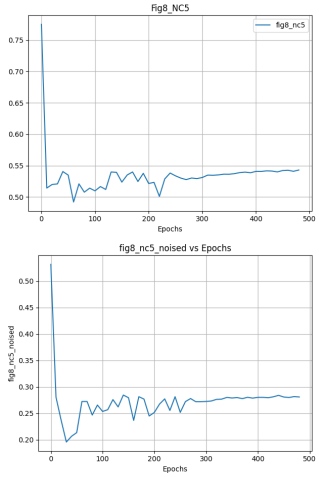

Figure 9: OrthoDev metric measuring the average cosine similarity between in-distribution class means and the OOD global mean for the CIFAR100 and SVHN, and a random noised mean vector.

Discussion

Neural Collapse

Our experiments on CIFAR-100 reproduce the Neural Collapse properties reported in (1) with high fidelity. The scale-invariant NC1 metric converges to zero, the equinorm and equiangularity metrics for both class means and classifier weights collapse as predicted, and the nearest-class-center disagreement vanishes at the terminal phase of training. The one result that differs slightly from the idealized theory is NC3: the Frobenius norm $\left\| \frac{W^T}{\|W\|_F} - \frac{\dot{M}}{\|\dot{M}\|_F} \right\|_F$ decreases over training but plateaus around 40 rather than reaching zero, which is not consistent with the values observed in (1) and reflects the fact that perfect self-duality is an asymptotic property that is approached but not fully attained in finite training. Overall, these results confirm that ResNet-18 trained on CIFAR-100 undergoes Neural Collapse and reaches the terminal training phase.

Average absolute cosine similarity between in-distribution class means and the global mean of OOD features, computed for SVHN and a randomly generated noise vector as a baseline. The metric stabilizes around 0.5 for SVHN and around 0.2 for the random noise vector.

The difference between the two values is informative. The random noise vector, being unstructured, produces features that are spread approximately uniformly across the feature space, so its global mean has very little systematic alignment with any of the in-distribution class means, giving a value close to 0.

SVHN, on the other hand, consists of real natural images that share low-level visual statistics with CIFAR-100, such as edges, textures, and color distributions. The ResNet feature extractor, having been trained on natural images, maps SVHN inputs to a region of feature space that partially overlaps with the in-distribution subspace, producing a noticeably higher cosine similarity of 0.5. This shows that the orthogonality predicted by NC5 is better approximated by truly unstructured inputs than by semantically unrelated but visually similar datasets, which has practical implications for the reliability of NECO when the OOD dataset shares low-level statistics with the training distribution.

Comparison of OOD Methods

Method	AUROC	AUPR	FPR@95
MSP			
Energy			
Mahalanobis			
ViM			
NECO			

Table 1: OOD detection results on CIFAR-100 vs. SVHN.

Conclusion

References

[1] V. Papyan, H. Xue, and D. L. Donoho, *Prevalence of neural collapse during the terminal phase of deep learning training*, Proceedings of the National Academy of Sciences (PNAS), vol. 117, no. 40, pp. 24652–24663, 2020.