

Out-of-Distribution Work

Practical Work - Deep Learning

Gabriel BAPTISTA TRELLESSE
Rodrigo BOTELHO ZUIANI

2026

ENSTA



IP PARIS

Contents

1	Practical Work	2
1.1	ResNet-18 Training on CIFAR-100	2
1.2	Out-of-Distribution Detection Scores	3
1.2.1	Max Softmax Probability (MSP)	3
1.2.2	Maximum Logit Score	4
1.2.3	Mahalanobis	4
1.2.4	Energy Score	4
1.2.5	ViM	5
1.2.6	NECO	5
1.3	Neural Collapse Analysis	6
1.3.1	Neural Collapse Visualizations	7
1.4	Bonus: Neural Collapse Across Layers	10

1 Practical Work

All the code used for this project can be seen in [GitHub](#).

1.1 ResNet-18 Training on CIFAR-100

We trained a ResNet-18 classifier on the CIFAR-100 dataset using standard data augmentation and stochastic gradient descent. The objective was to reach the Terminal Phase of Training in order to evaluate Out-of-Distribution detection methods and later analyze the Neural Collapse phenomena.

Training Setup: The model was trained with the following configuration:

- Optimizer: SGD with momentum
- Initial learning rate: 0.1
- Weight decay: 5×10^{-4}
- Batch size: 128
- Number of epochs: 450
- Scheduler: MultiStepLR with decay factor $\gamma = 0.1$ at epochs corresponding to 1/3 and 2/3 of the initial total training duration of 350 epochs (later increased to 450 epochs)

Standard CIFAR-100 data augmentation was used, applying random cropping and horizontal flipping for the training set.

Training Dynamics. Figure 1 shows the evolution of the training loss, training accuracy, and test accuracy across epochs.

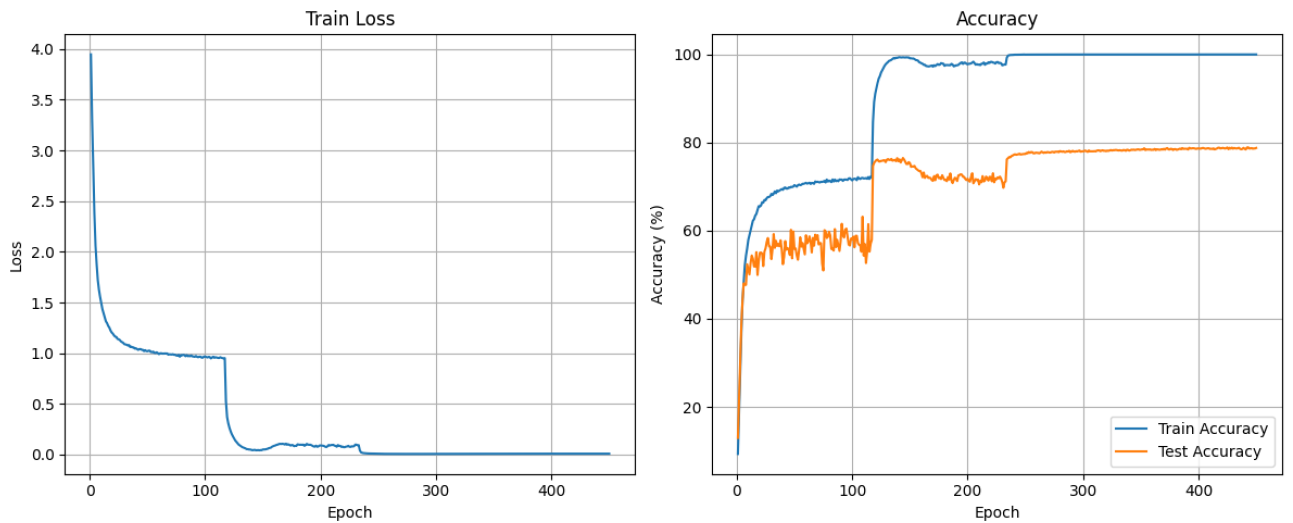


Figure 1: Training curves for ResNet-18 on CIFAR-100. Left: training loss across epochs. Right: training and test accuracy.

We observe a sharp decrease in training loss during the first phase of training, followed by additional improvements after each learning rate decay. The learning rate reductions at epochs 117 and 233 produce noticeable jumps in both training and test accuracy.

The final model achieves nearly 100% training accuracy and approximately 78% test accuracy on CIFAR-100. The near-zero training loss and saturated training accuracy indicate that the network has entered the Terminal Phase of Training, which is a necessary condition for observing Neural Collapse behavior.

In conclusion, the final test performance is consistent with standard ResNet-18 benchmarks on CIFAR-100. The gap between training and test accuracy is expected for CIFAR-100 due to its high number of classes and high variability between elements of the same class.

1.2 Out-of-Distribution Detection Scores

After training the ResNet-18 model on CIFAR-100 (in-distribution data), we evaluated several OOD detection methods using SVHN as the out-of-distribution dataset. All methods operate on features extracted from the penultimate layer of the trained network.

Performance is evaluated using AUROC (Area Under the Receiver Operating Characteristic curve) and FPR95 (False Positive Rate at 95% True Positive Rate). Higher AUROC and lower FPR95 indicate better OOD detection performance.

1.2.1 Max Softmax Probability (MSP)

The MSP score is defined as:

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

It measures the maximum predicted class probability. Lower confidence values indicate potential OOD samples.

Results:

- AUROC: 0.8496
- FPR95: 0.6644

1.2.2 Maximum Logit Score

The Maximum Logit Score uses the largest raw logit value:

$$\max_k f_k(x).$$

Results:

- AUROC: 0.8603
- FPR95: 0.6578

This method slightly improves over MSP by avoiding softmax normalization effects.

1.2.3 Mahalanobis

The Mahalanobis method models class-conditional Gaussian distributions in feature space and computes the minimum Mahalanobis distance to class means.

Results:

- AUROC: 0.8290
- FPR95: 0.7522

Unlike we expected, the performance is lower than other energy based methods in this setup. This may have happened due to the challenge of performing estimations in a high-dimensional feature space with 100 classes.

1.2.4 Energy Score

The Energy Score is defined as:

$$E(x) = -\log \sum_k e^{f_k(x)}.$$

It considers the full logit distribution rather than only the maximum component.

Results:

- AUROC: 0.8638
- FPR95: 0.6393

The Energy Score method achieves the second best performance among the other methods, which confirms that taking the full logit distribution into account improves OOD detection.

1.2.5 ViM

ViM (Virtual Logit Matching) considers the classifier weight structure and feature geometry.

Results:

- AUROC: 0.8194
- FPR95: 0.7057

ViM performs competitively but does not outperform the Energy Score method in our setting.

1.2.6 NECO

NECO (Neural Collapse-based OOD detection) exploits the geometric structure emerging during Neural Collapse. It measures the ratio of projected feature norm onto the principal subspace of in-distribution data. Using an optimal subspace dimension (250 principal components), we obtain:

Results:

- AUROC: 0.8892
- FPR95: 0.5318

NECO achieves the best performance among all evaluated methods. This result supports the hypothesis that exploiting Neural Collapse geometry improves OOD detection capability.

OOD Method	AUROC	FPR95
Max Softmax Probability	0.8496	0.6644
Maximum Logit Score	0.8603	0.6578
Mahalanobis	0.8290	0.7522
Energy Score	0.8638	0.6393
ViM	0.8194	0.7057
NECO (best neco_dim)	0.8892	0.5318

Table 1: Comparison of Out-of-Distribution detection performance on SVHN (OOD) vs CIFAR-100 (ID).

Comparison Summary. Overall ranking:

$$\text{NECO} > \text{Energy Score} > \text{Max Logit} > \text{MSP} > \text{Mahalanobis} > \text{ViM}$$

This demonstrates that the geometric approaches of Neural Collapse properties have the potential to provide better OOD separation than purely confidence based methods.

1.3 Neural Collapse Analysis

Neural Collapse is a phenomenon that occurs in the late stage of training deep neural networks, known as the Terminal Phase of Training (TPT). During this phase, as the model approaches an optimal solution, it becomes increasingly capable of identifying distinctive characteristics for each class. For example, in CIFAR-100, classes such as bird and automobile develop clearly separated feature representations. At the same time, samples belonging to the same class become more similar in the feature space, meaning that their representations concentrate around their corresponding class mean.

Due to this convergence of class means and the reduction of within-class variance, the features in the representation space progressively organize into a geometric structure known as a Simplex Equiangular Tight Frame (Simplex ETF). These can be summarized into the following 4 inter-related characteristics of the final and penultimate layer, whereas NC5 was added by NECO to increasing OOD detection using OOD and ID properties during NC:

1. NC1 – Within-Class Collapse

Features of samples from the same class collapse toward their class mean.

2. NC2 – Simplex ETF Structure

Class means form a simplex equiangular tight frame configuration.

3. NC3 – Self-Duality

Classifier weights align with class means.

4. NC4 – Nearest Class Center Behavior

The classifier behaves as a nearest-class-mean classifier in feature space.

5. NC5 – ID/OOD Orthogonality

OOD data and In-distributions(ID) data starts becoming more and more orthogonal to each other.

To illustrate these properties in practice, we trained our ResNet-18 until reaching the Terminal Phase of Training (TPT). It is important to emphasize that a sufficiently small learning rate is required to ensure stable convergence and precise geometric alignment between features and classifier weights. If the learning rate remains too large, weight updates become unstable and the Neural Collapse structure may not properly emerge. Therefore, an appropriate learning rate scheduler is essential.

In our experiments, we used the following configuration:

- Batch size: 128

- Number of workers: 2
- Initial learning rate: 0.1
- Weight decay: 5×10^{-4}
- Optimizer: SGD
- Number of epochs: 450
- Scheduler: MultiStepLR with $\gamma = 0.1$ at epochs 117 and 233

The learning rate milestones were chosen proportionally (at one-third and two-thirds of training) following standard training protocols.

1.3.1 Neural Collapse Visualizations

The following figures illustrate key aspects of Neural Collapse for in-distribution (ID) data. They help visualize how feature representations and classifier weights evolve during training.

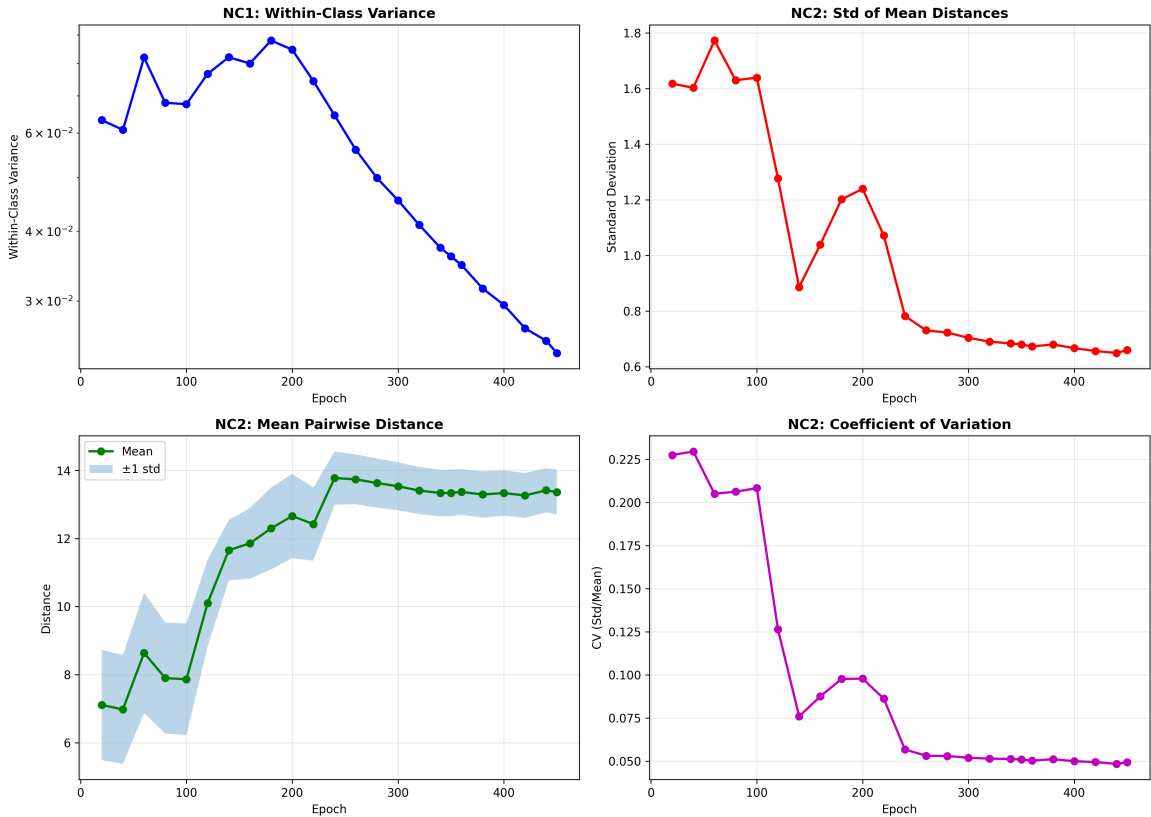


Figure 2: Neural Collapse metrics for ID. This plot shows the evolution of NC1 (within-class variance) and NC2 (mean pairwise distance and coefficient of variation) across training epochs, highlighting the convergence of features towards their class means.

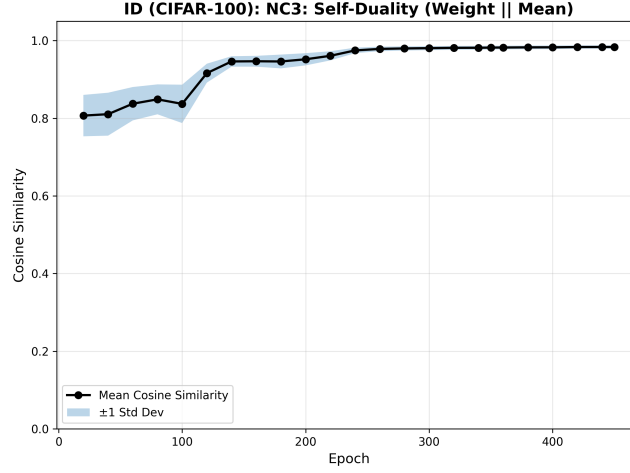


Figure 3: Cosine similarity between classifier weights and class means for ID. This illustrates NC3 (self-duality), where the classifier weights progressively align with the corresponding class feature means as training progresses.

These visualizations demonstrate that, as training approaches the Terminal Phase, the within-class variance decreases (converging to zero), the class means form a more symmetric structure (pairwise distances maximize while the coefficient of variation converges to zero), and the classifier weights align with the class centers (cosine similarity converges to 1), all of which are characteristic behaviors of Neural Collapse.

When plotting these metrics for OOD data, in this case we used the Street View House Number(SVHN), we obtain the following:

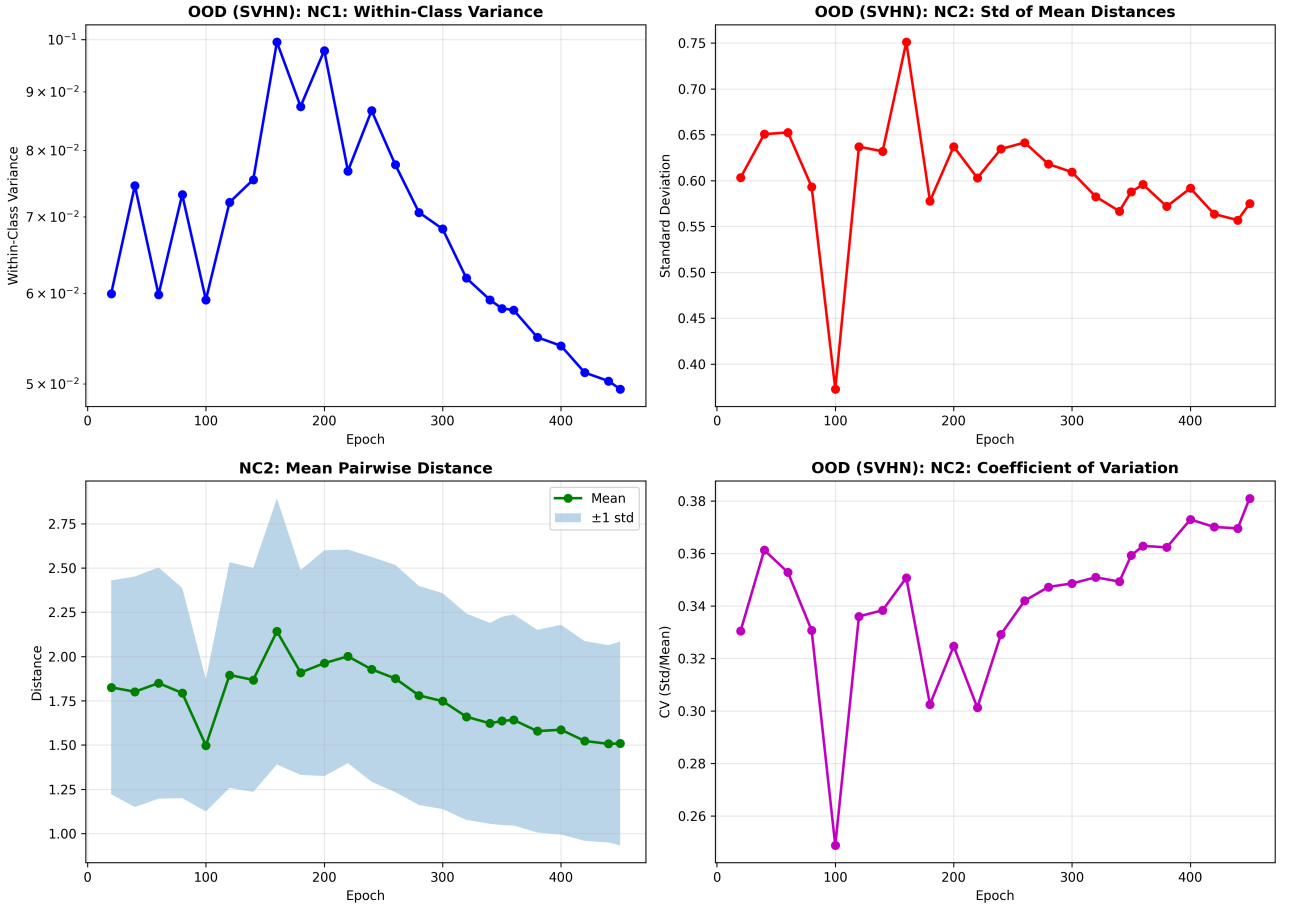


Figure 4: Neural Collapse metrics for OOD.

This shows that the within-class variance still decreases, but the pairwise distances between class means do not form the structured Simplex ETF pattern. Instead, the features resemble an unstructured cluster without a clear geometric arrangement.

To visualize both metrics at the same time, see Figure 5.

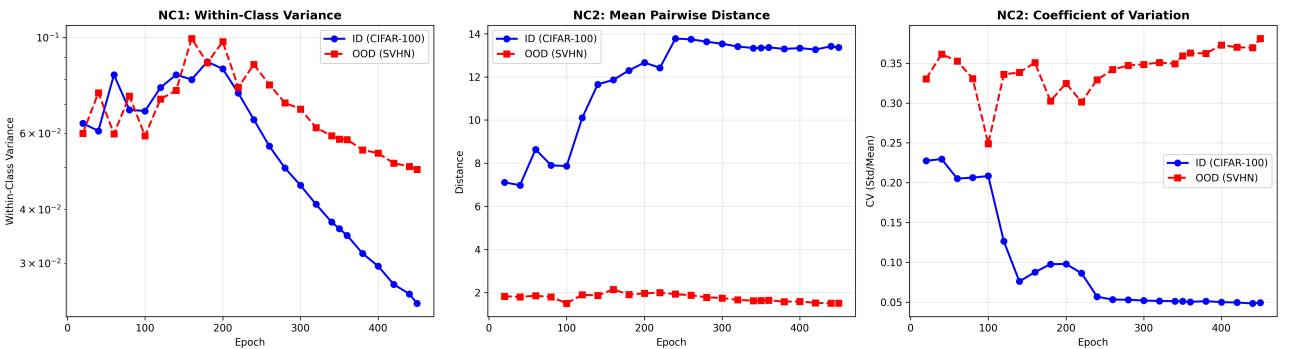


Figure 5: Neural Collapse metrics comparing ID and OOD data.

1.4 Bonus: Neural Collapse Across Layers

The last experiment analyzed the collapsing behavior across layers to determine whether Neural Collapse occurs predominantly in early, middle, or final layers. For this analysis, we extracted features from the initial convolution ('conv1'), the three groups of residual blocks ('layer1', 'layer2', 'layer3'), and the final fully connected layer ('layer4'). The resulting within-class variance and mean pairwise distance across layers are shown in Figure 6.

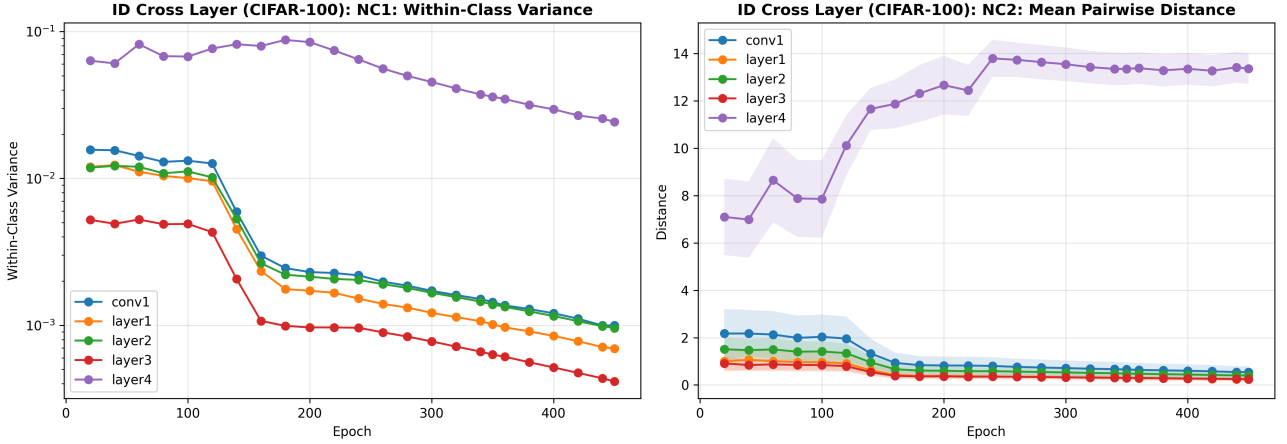


Figure 6: Neural Collapse metrics across layers. The plots show how within-class variance and mean pairwise distance evolve across epochs for each layer.

We observe that the within-class variance collapses progressively in each successive layer, with a noticeable gap at the fully connected layer, likely due to the larger feature space. A similar trend is observed for the mean pairwise distance, where distances between class means increase more prominently in later layers.

References

- AMMAR, Mouin Ben *et al.* *NECO: NEural Collapse Based Out-of-distribution detection*. 2024. arXiv: 2310.06823 [stat.ML]. Available from: <https://arxiv.org/abs/2310.06823>.
- KOTHAPALLI, Vignesh. *Neural Collapse: A Review on Modelling Principles and Generalization*. 2023. arXiv: 2206.04041 [cs.LG]. Available from: <https://arxiv.org/abs/2206.04041>.