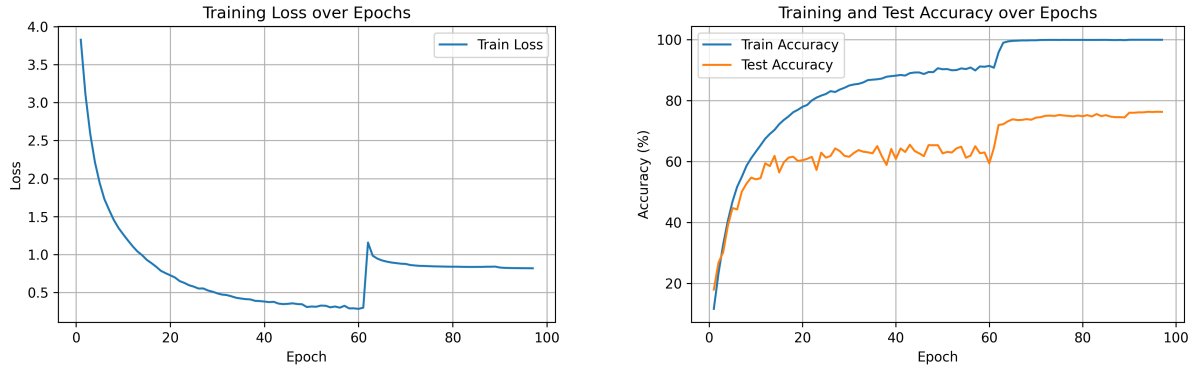
TP3 FDL

Adèle Dejoie

December 2025

1 Question 2



(a) Training loss over epochs

(b) Training and test accuracy over epochs

Figure 1: Training dynamics of the CIFAR-100 model

Figure 1 highlights a clear change of training regime around epoch 60. Before this point, the model is trained with a learning rate of 0.005 and without label smoothing. During this phase, the training loss decreases monotonically while training accuracy steadily increases. However, test accuracy saturates around 63–65%, indicating limited generalization despite continued optimization on the training set.

After epoch 60, label smoothing with coefficient 0.1 is introduced and the learning rate is reduced to 0.003. This induces a sharp increase in the training loss, which is expected since label smoothing prevents the loss from converging to zero even for correct predictions. At the same time, training accuracy rapidly saturates near 100%, while test accuracy significantly improves and stabilizes around 76%.

This transition marks the entry into the terminal phase of training, where representations become more structured and better calibrated. This phase is critical for the emergence of Neural Collapse properties and explains the strong performance of geometry-based OOD detectors such as NECO.

2 Question 3

Comparison of OOD Detection Methods. We compare five out-of-distribution (OOD) detection methods: Maximum Softmax Probability (MSP), MaxLogit, Energy-based score, Mahalanobis distance, and ViM. These methods differ both in their underlying assumptions and in the type of information they exploit, ranging from output logits to deep feature representations.

Confidence-based methods. MSP is defined as the maximum predicted class probability:

$$\mathrm{MSP}(x) = \max_k \ \mathrm{Softmax}(f_k(x)),$$

where $f_k(x)$ denotes the logit associated with class k. Samples with low MSP values are considered OOD. However, MSP is known to suffer from overconfidence, as neural networks often assign high softmax probabilities to OOD inputs.

MaxLogit directly uses the maximum logit value:

$$\mathrm{MaxLogit}(x) = \max_k f_k(x),$$

thereby avoiding the normalization effect of the softmax function. This often leads to slightly improved robustness compared to MSP.

The Energy-based score is defined as:

$$E(x) = -\log \sum_k \exp(f_k(x)).$$

Energy provides a smoother confidence measure by aggregating information across all logits, and has been shown to improve calibration and OOD detection compared to MSP and MaxLogit.

In the experiments, these three confidence-based methods exhibit similar performance. On SVHN, their AUROC values are around 0.82, while on CIFAR-10 they achieve the highest AUROC values among the considered methods, around 0.78. This indicates that confidence-based scores remain relatively robust in near-OOD settings.

Feature-based methods. The Mahalanobis method assumes that deep features follow class-conditional Gaussian distributions. For a feature representation $z(x)$, the OOD score is defined as:

$$D_{\mathrm{Mahalanobis}}(x) = \min_c (z(x) - \mu_c)^\top \Sigma^{-1} (z(x) - \mu_c),$$

where μ_c is the mean feature vector of class c and Σ is the shared covariance matrix estimated on training data. Samples far from all class centers are considered OOD.

ViM (Virtual Logit Matching) builds upon the observation that in-distribution features lie in a low-dimensional subspace. A PCA is first applied to the training features to estimate this subspace. This OOD score combines the Energy score with the norm of the projection onto the orthogonal complement:

$$\mathrm{ViM}(x) = E(x) + \alpha \left\| (I - UU^\top) z(x) \right\|,$$

where U contains the principal components of the in-distribution feature space and α is a scaling factor.

Empirical comparison. On the distant OOD dataset SVHN, ViM significantly outperforms all other methods, achieving an AUROC of 0.91 compared to approximately 0.82 for confidence-based scores. This confirms that SVHN samples exhibit strong components outside the in-distribution feature subspace, which ViM is able to exploit effectively. Mahalanobis does not yield a comparable improvement, likely due to the high number of classes and the limitations of the Gaussian assumption.

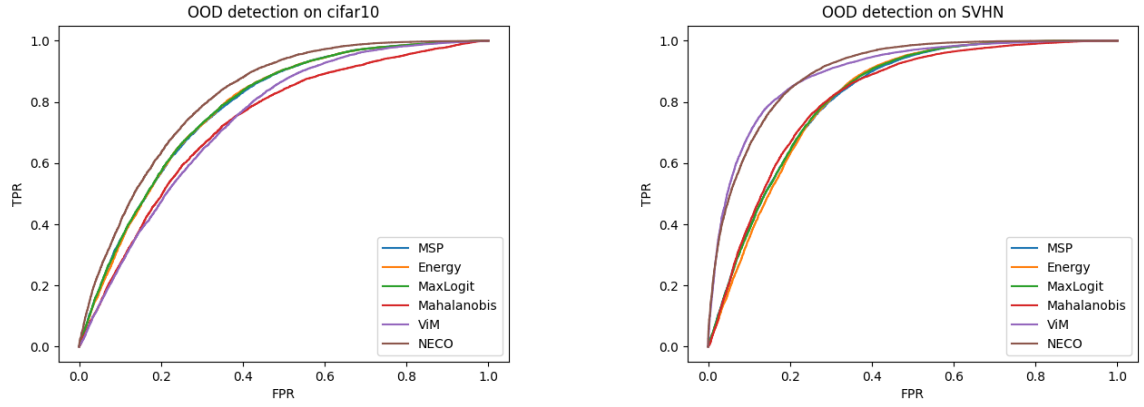
On the near-OOD dataset CIFAR-10, all methods struggle. Confidence-based scores achieve the best performance, with AUROC values around 0.78, while feature-based methods perform worse. In particular, both Mahalanobis and ViM degrade noticeably, indicating that CIFAR-10 features largely overlap with the CIFAR-100 feature space. In this regime, the orthogonal subspace exploited by ViM does not provide a reliable separation signal.

ROC curve analysis (TPR vs FPR). To further analyze the behavior of the different OOD detection methods, we also consider the ROC curves, which plot the true positive rate (TPR) as a function of the false positive rate (FPR) for varying decision thresholds.

On the distant OOD dataset SVHN, the ROC curves of MSP and Energy are almost perfectly superimposed, indicating very similar trade-offs between OOD detection and false rejections of in-distribution samples. In contrast, the ViM curve is clearly shifted towards the top-left corner of the plot. Its shape is close to a right angle, which indicates that a high true positive rate can be achieved while keeping the false positive rate relatively low. For instance, it is possible to reach a TPR of approximately 0.8 with an FPR around 0.2, which illustrates the strong separation capability of ViM in this setting.

On the near-OOD dataset CIFAR-10, the situation is markedly different. The ROC curves of MSP, MaxLogit, and Energy are again almost superimposed, confirming that these confidence-based methods behave very similarly. However, ViM is clearly below these curves and closer to the diagonal, indicating a weaker separation between in-distribution and OOD samples. The Mahalanobis method performs slightly better than ViM, but the difference remains limited. Overall, the ROC curves highlight the difficulty of near-OOD detection and visually confirm that feature-based methods do not provide a clear advantage over confidence-based scores in this scenario.

Conclusion. Overall, confidence-based methods are simple and robust, especially for near-OOD detection, while feature-based methods such as ViM are highly effective for distant OOD detection. These results highlight the importance of matching the OOD detection method to the nature of the OOD data and underline the intrinsic difficulty of near-OOD detection.



(a) Near OOD detection

(b) Far OOD detection

Figure 2: OOD detection

	Near-OOD (CIFAR-10)			Far-OOD (SVHN)		
Method	AUROC	AUPR	FPR@95	AUROC	AUPR	FPR@95
MSP	0.7809	0.7367	0.6125	0.8189	0.8928	0.4944
MaxLogit	0.7824	0.7378	0.6121	0.8206	0.8928	0.4832
Energy	0.7810	0.7336	0.6114	0.8162	0.8872	0.4778
Mahalanobis	0.7484	0.7024	0.7298	0.8190	0.8907	0.5268
ViM	0.7353	0.6807	0.6755	**0.9102**	**0.9546**	0.3652
NECO	**0.8160**	**0.7780**	**0.5243**	0.9001	0.9500	**0.3538**

Table 1: OOD detection performance on Near-OOD (CIFAR-10) and Far-OOD (SVHN) datasets in terms of AUROC, AUPR, and FPR@95. Best results for each metric are highlighted in bold.

3 Question 4 Neural Collapse Analysis (NC1–NC4)

We study the Neural Collapse phenomenon at the end of training by analyzing the geometry of the learned representations on the CIFAR-100 training set. Features are extracted from the penultimate layer of the network, and all measurements are performed without data augmentation to ensure stable statistics. We report results across multiple training checkpoints as well as at the checkpoint achieving the best test accuracy.

NC1: Within-Class Variability Collapse. Neural Collapse predicts that the within-class variance of features progressively vanishes during the terminal phase of training. Figure 3a shows that the ratio $\mathrm{tr}(\Sigma_W)/\mathrm{tr}(\Sigma_T)$ decreases monotonically throughout training, from 0.58 at early epochs to approximately 0.37 at late epochs. While the variance does not fully collapse to zero, the consistent downward trend indicates a clear contraction of intra-class features, suggesting a partial realization of NC1.

NC2: Simplex Equiangularity of Class Means. NC2 states that the centered class means converge to a simplex equiangular configuration. This behavior is clearly observed in Figure 3b, where the mean cosine similarity between centered class means converges toward the theoretical value $-1/(C-1)$, which equals -0.0101 for CIFAR-100. In parallel, the standard deviation of cosine similarities steadily decreases, indicating increasing concentration around the simplex configuration. This result is further confirmed by the cosine similarity heatmap in Figure 4a, where off-diagonal entries form a nearly uniform pattern, providing strong evidence of NC2.

NC3: Alignment Between Classifier Weights and Class Means. NC3 predicts that the classifier weights align with their corresponding class means. As shown in Figure 3a, the average cosine similarity between classifier weights and class means increases steadily during training, reaching values around 0.69 at late epochs, with low variance. This indicates a strong, although not perfectly saturated, alignment consistent with a partial realization of NC3.

NC4: Nearest-Class-Mean Equivalence. Finally, NC4 asserts that classification using the nearest class mean becomes equivalent to the linear classifier. Figure 3a shows that nearest-class-mean accuracy rapidly increases and reaches more than 97% on the training set at late epochs, closely matching the classifier accuracy. This demonstrates that the learned representation is effectively explained by class means alone, strongly validating NC4.

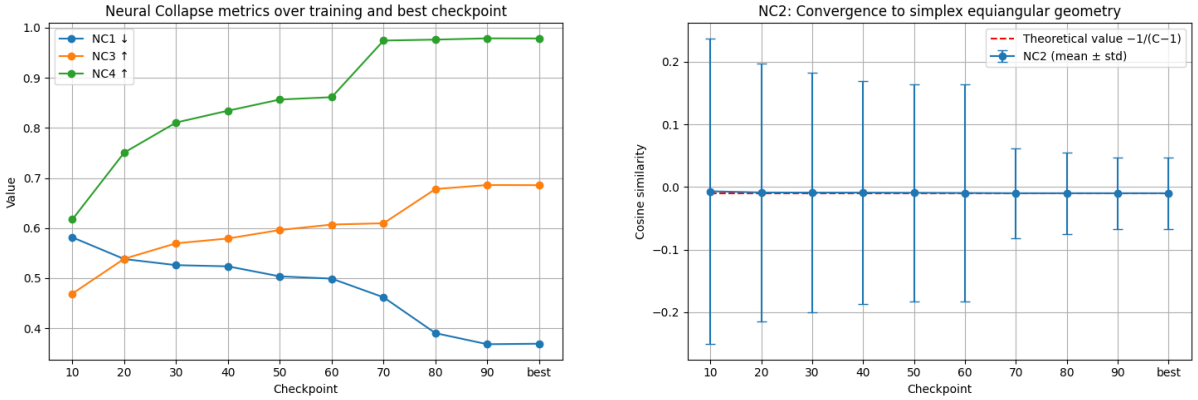
Stability at the Best Checkpoint. We additionally report Neural Collapse metrics at the checkpoint achieving the best test accuracy. As summarized in Table 2, all NC indicators at this checkpoint closely match those observed at late training epochs. This confirms that Neural Collapse properties remain stable once the network enters the terminal phase, and are not tied to a specific epoch.

Conclusion Overall, the network exhibits clear Neural Collapse behavior, particularly in its inter-class geometry (NC2) and nearest-class-mean equivalence (NC4), while within-class variance collapse (NC1) remains partial. These observations are consistent with Neural Collapse theory and highlight that inter-class geometric structures stabilize earlier than full intra-class contraction during training.

am

Metric	Ep10	Ep20	Ep30	Ep40	Ep50	Ep60	Ep70	Ep80	Ep90	Best
NC1 ↓	0.582	0.538	0.526	0.524	0.504	0.499	0.462	0.390	0.368	0.369
NC2 mean	-0.0069	-0.0088	-0.0090	-0.0090	-0.0093	-0.0096	-0.0100	-0.0100	-0.0100	-0.0100
NC2 std ↓	0.244	0.206	0.191	0.179	0.173	0.174	0.071	0.065	0.057	0.057
NC3 mean ↑	0.469	0.539	0.570	0.579	0.596	0.607	0.610	0.678	0.686	0.686
NC3 std ↓	0.044	0.049	0.056	0.058	0.061	0.054	0.052	0.051	0.047	0.047
NC4 ↑	0.617	0.751	0.811	0.834	0.856	0.861	0.974	0.976	0.979	0.978

Table 2: Evolution of Neural Collapse metrics (NC1–NC4) during training on CIFAR-100. Results are reported at multiple epochs and at the checkpoint achieving the best test accuracy. Arrows indicate the expected direction of improvement.



(a) NC1, NC3 and NC4 over training

(b) NC2 over training

Figure 3: Neural Collapse metrics over training

4 Question 5 : Neural Collapse Analysis (NC5)

Following the course definition, we study NC5 as the orthogonality between the in-distribution (ID) class-mean configuration and out-of-distribution (OOD) data in the representation space. Let $f(x) \in \mathbb{R}^d$ denote the penultimate-layer feature, μ_c the mean feature of class c over the ID training set, and μ_G the global ID mean. We center class means and the OOD mean by μ_G:

$$\tilde{\mu}_c = \mu_c - \mu_G, \qquad \tilde{\mu}_{\text{OOD}} = \mu_{\text{OOD}} - \mu_G,$$

and quantify orthogonality by the average absolute cosine similarity:

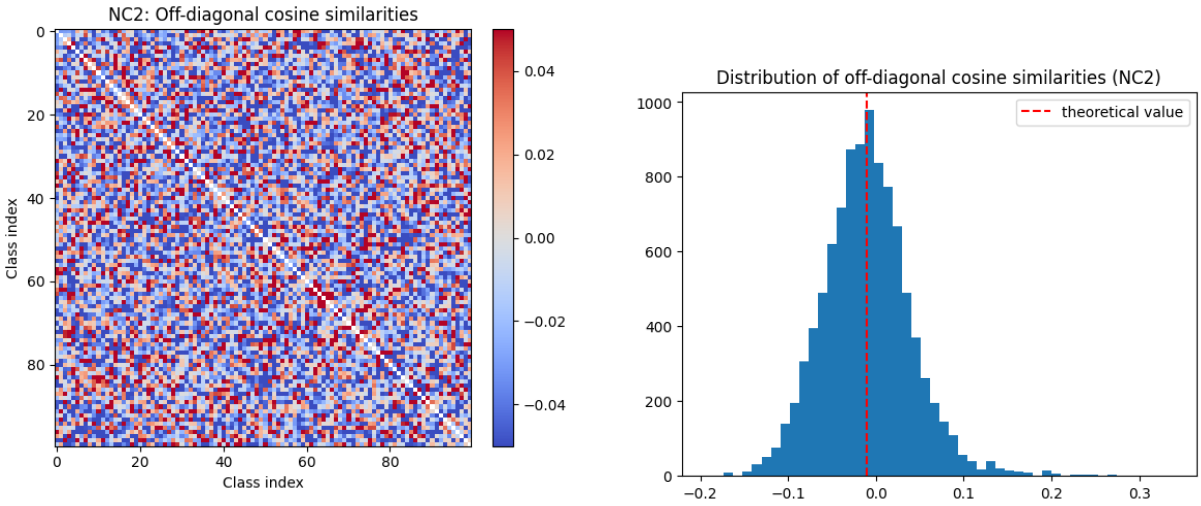
$$\text{NC5} = \frac{1}{C} \sum_{c=1}^{C} \left| \cos\left(\tilde{\mu}_c, \tilde{\mu}_{\text{OOD}}\right) \right|.$$

Neural Collapse predicts that this quantity becomes small in the terminal phase, reflecting that OOD representations are nearly orthogonal to the ID simplex structure.

In the experiments, we obtain NC5 ≈ 0.052 for SVHN and NC5 ≈ 0.054 for CIFAR-10 (with comparable dispersion across classes). These low values provide clear evidence that OOD feature means are close to orthogonal to the centered ID class means, in accordance with NC5. Moreover, SVHN is slightly more orthogonal than CIFAR-10, consistent with SVHN being semantically further from CIFAR-100.

5 Question 6 : NECO

We finally implement the NECO method, a Neural Collapse–inspired OOD detector that leverages the geometric structure emerging in the representation space at the end of training. NECO



(a) Cosine similarities heatmap

(b) Cosine similarities distribution

Figure 4: Cosine similarities

measures the relative energy of features projected onto the principal subspace learned from in-distribution (ID) data, and rescales this quantity using the maximum classifier logit to incorporate class-discriminative information.

Principal components are estimated via PCA on the penultimate-layer features of the CIFAR-100 training set. Given a test sample x with feature representation $h(x)$ and global ID mean μ_G, the NECO score is computed as the ratio between the norm of the projection of $h(x) - \mu_G$ onto the leading PCA subspace and the norm of $h(x) - \mu_G$, further multiplied by the maximum logit.

We evaluate NECO on SVHN and CIFAR-10 as OOD datasets. Results show that NECO achieves strong performance, with an AUROC of 0.900 and FPR@95 of 0.354 on SVHN, and an AUROC of 0.816 and FPR@95 of 0.524 on CIFAR-10. NECO consistently outperforms classical score-based methods such as MSP, Energy, and Mahalanobis, and achieves performance comparable to or better than ViM.

An analysis of the ROC curves further supports these quantitative results. On far OOD data (SVHN), NECO exhibits performance similar to the strongest baseline (ViM), maintaining high true positive rates across a wide range of false positive rates. On near OOD data (CIFAR-10), NECO clearly outperforms all other methods, achieving the best trade-off between TPR and FPR. This behavior highlights the robustness of NECO across different OOD regimes, and indicates that it provides the most reliable criterion among the evaluated methods.

These results are consistent with our previous Neural Collapse analysis. In particular, the strong presence of NC2–NC4 and the observed ID/OOD orthogonality (NC5) provide a geometric explanation for the effectiveness of NECO. This confirms that Neural Collapse is not only a theoretical phenomenon but also a practical foundation for robust OOD detection.