Practical Work: Out-of-Distribution Detection, OOD Scoring Methods, and Neural Collapse

Fondements Théoriques de l'Apprentissage Profond (MVA)

Aziz Bacha

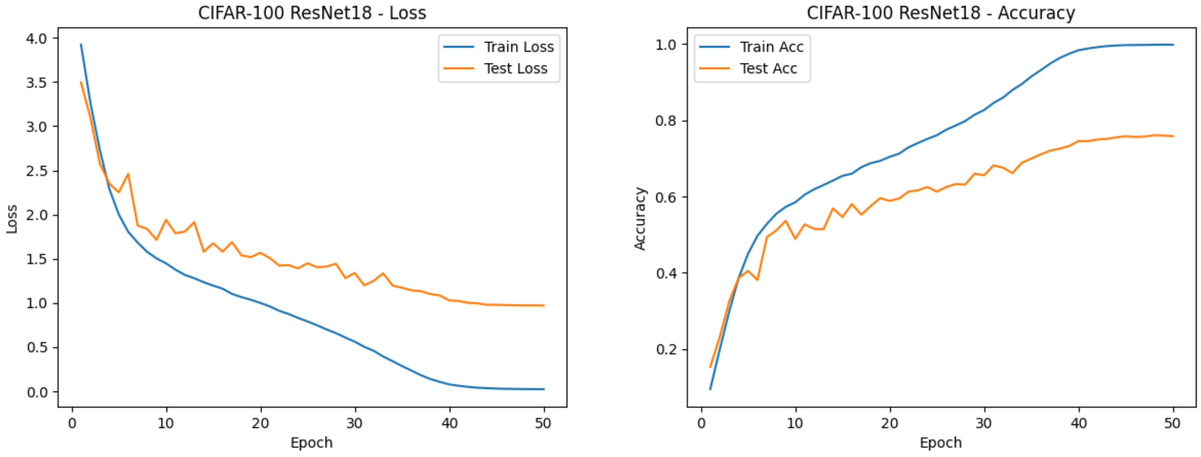
1 Introduction

Out-of-Distribution (OOD) detection aims to identify inputs that do not belong to the training distribution, even when the classifier outputs confident predictions. This practical work explores how different OOD scoring methods behave for a ResNet-18 trained on CIFAR-100, and how the Neural Collapse phenomenon can be leveraged for OOD analysis.

2 Model training

We train a ResNet-18 classifier from scratch on CIFAR-100 using PyTorch. The network is adapted to 32×32 images by replacing the first convolution with a 3×3 layer (stride 1, padding 1) and removing the max-pooling layer. Training is performed for 50 epochs with SGD (learning rate 0.1, momentum 0.9, weight decay $5 \cdot 10^{-4}$), a cosine annealing learning rate schedule, and mini-batches of size 128.

For data augmentation, we apply random cropping with padding 4 and random horizontal flips on the training set, followed by normalization with the standard CIFAR-100 mean and standard deviation; test images are only normalized. We track loss and accuracy on both the training and test sets at each epoch and save the best checkpoint according to test accuracy. After 50 epochs, the model reaches about 99.8% training accuracy and 76.1% best test accuracy, Figure 1 shows the evolution of loss and accuracy during training.



(a) Training and test loss over epochs.

(b) Training and test accuracy over epochs.

Figure 1: Training curves for ResNet-18 on CIFAR-100.

3 OOD scoring methods

3.1 Implementation

We evaluate several standard OOD scoring methods on the trained ResNet-18. For the logit-based methods (MSP, Max Logit, and Energy), scores are computed directly from the network outputs

$f(x) \in \mathbb{R}^{100}$. For feature-based methods (Mahalanobis and ViM), we additionally extract the penultimate-layer representations $z(x)$ using a modified forward pass. All scores are oriented so that higher values indicate in-distribution samples.

Given logits $f(x)$ and softmax probabilities $p(x) = \mathrm{softmax}(f(x))$, we use:

- **MSP**: $s_{\mathrm{MSP}}(x) = \max_k p_k(x)$.
- **Max Logit**: $s_{\mathrm{MaxLogit}}(x) = \max_k f_k(x)$.
- **Energy**: $s_{\mathrm{Energy}}(x) = T \log \sum_k \exp\left(f_k(x)/T\right)$, with $T = 1$ and the sign chosen so that larger scores indicate ID.

For Mahalanobis and ViM, we rely on the penultimate features $z(x)$:

- **Mahalanobis**: we estimate class means μ_c and a shared covariance Σ on CIFAR-100 test features, using shrinkage for numerical stability. The score is

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

  i.e., the negative squared Mahalanobis distance to the nearest class mean.
- **ViM**: we center features, perform an SVD to split the space into a principal subspace and a residual subspace, and project $z(x)$ onto the residual subspace V_n. Let $r(x)$ be the norm of this residual and $m(x)$ the maximum logit. The ViM-style score is

  $$s_{\mathrm{ViM}}(x) = -\alpha\, r(x) + m(x),$$

  where α is chosen to roughly match the scales of both terms (here based on their empirical standard deviations on ID data).

3.2 Results

We report OOD detection results with CIFAR-100 as the in-distribution dataset, and SVHN or CIFAR-10 as OOD datasets. SVHN represents a *large distribution shift* (digit images vs. natural images), while CIFAR-10 constitutes a *near-OOD* setting, since it shares similar visual statistics with CIFAR-100 but contains different semantic classes. Evaluating on both allows us to assess how OOD scores behave under easy and difficult distribution shifts.

For each method, we compute three standard OOD metrics using `scikit-learn`: AUROC, AUPR (treating in-distribution samples as the positive class), and FPR@95.

AUROC measures the overall separability between ID and OOD score distributions, AUPR emphasizes performance when operating at high precision, and FPR@95 quantifies how many OOD samples are incorrectly accepted as ID when the true positive rate on ID data is fixed at 95%. Higher AUROC/AUPR and lower FPR@95 indicate better OOD detection.

Table 1: OOD detection performance on SVHN (ID: CIFAR-100).

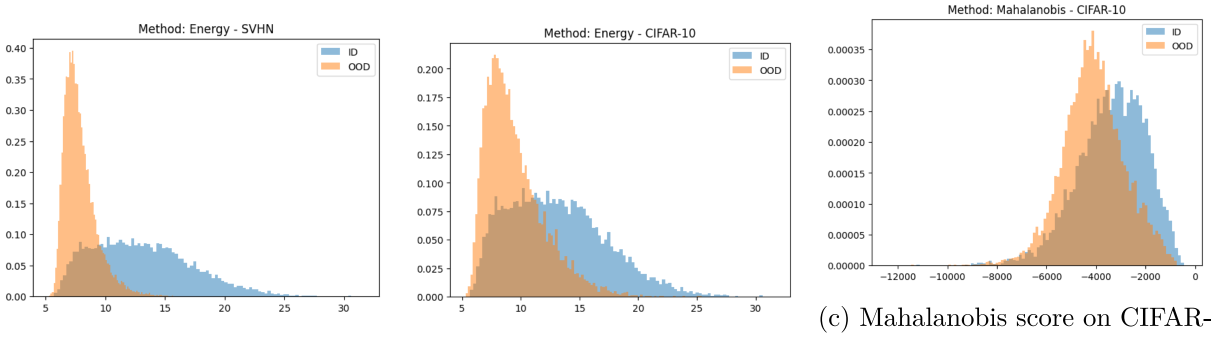
Method	AUROC	AUPR	FPR@95
MSP	0.8413	0.7706	0.7180
Max Logit	0.8910	0.8418	0.6318
Energy	0.8952	0.8467	0.6174
Mahalanobis	0.7315	0.6480	0.9330
ViM	0.8611	0.8151	0.7578

Overall, Energy and Max Logit perform best on the easier OOD shift (SVHN), while all methods degrade on CIFAR-10, which is more similar to CIFAR-100. The Mahalanobis score performs notably worse, especially in terms of FPR@95, and ViM lies between Mahalanobis and the logit-based scores.

Figure 2 shows score distributions for representative OOD methods under large and near distribution shifts. For SVHN (left), the Energy score yields a clear separation between ID and

Table 2: OOD detection performance on CIFAR-10 (ID: CIFAR-100).

Method	AUROC	AUPR	FPR@95
MSP	0.7644	0.7921	0.8235
Max Logit	0.7713	0.7907	0.8225
Energy	0.7688	0.7891	0.8244
Mahalanobis	0.6639	0.6626	0.9214
ViM	0.7271	0.7413	0.8701



(a) Energy score on SVHN.

(b) Energy score on CIFAR-10.

(c) Mahalanobis score on CIFAR-10.

Figure 2: Score distributions for representative OOD detection methods. Higher scores indicate in-distribution samples.

OOD samples, consistent with its high AUROC and low FPR@95. For CIFAR-10 (middle), the Energy distributions overlap more, illustrating the difficulty of near-OOD detection and explaining the degradation of all metrics. The Mahalanobis score on CIFAR-10 (right) exhibits poor separation, accounting for its particularly high FPR@95. This behavior is consistent with known results showing that, after sufficiently long training (50 epochs here), the model enters an over-parameterized regime in which Mahalanobis-based OOD detection degrades.

4 Neural Collapse at the end of training (NC1–NC4)

We study Neural Collapse on the penultimate features $z(x)$ and final linear classifier $W \in \mathbb{R}^{C\times D}$ at the end of training.

NC1: within-class variance collapse. NC1 states that features within each class concentrate around their class mean μ_c, so that the within-class variance becomes small compared to the between-class variance:

$$\frac{\mathbb{E}[\|z-\mu_c\|^2 \mid y=c]}{\mathbb{E}[\|\mu_c-\mu\|^2]} \to 0,$$

where μ is the global mean. In our case, the average within-class variance is 47.4, the average between-class variance is 71.8, and the ratio is about 0.66. This shows that between-class separation dominates within-class spread (features have clearly clustered by class), but NC1 is not “perfectly” collapsed (the within-class variability remains non-negligible for a 100-class problem).

NC2: simplex ETF structure of class means. NC2 predicts that the centered class means $\mu_c - \mu$ have almost equal norms and form a simplex equiangular tight frame (ETF), i.e. all

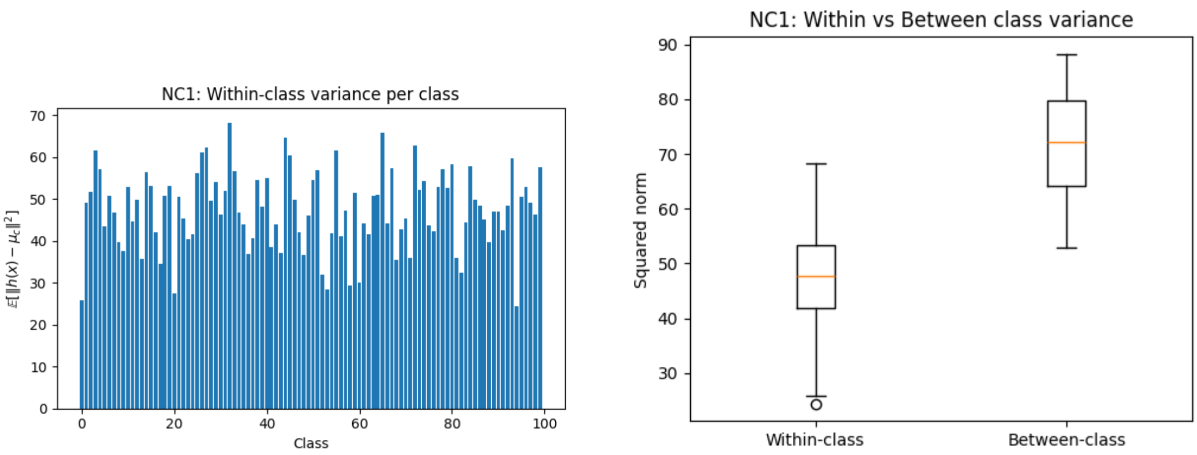


Figure 3: NC1 diagnostics on CIFAR-100. Left: within-class variance for each class. Right: comparison between within-class and between-class variances.

pairwise angles are equal and

$$\langle \tilde{\mu}_c, \tilde{\mu}_{c'} \rangle \approx -\frac{1}{C-1}, \quad c \neq c',$$

where $\tilde{\mu}_c$ are the normalized centered means. We observe that the norms of $\mu_c - \mu$ are tightly concentrated (mean 8.46, std 0.54), and the cosine Gram matrix has diagonal entries exactly 1 and off-diagonal mean ≈ -0.010, very close to the theoretical $-1/(C-1) \approx -0.0101$, with a moderate spread (std ≈ 0.12). This indicates that the class means are well arranged in an approximate simplex ETF.

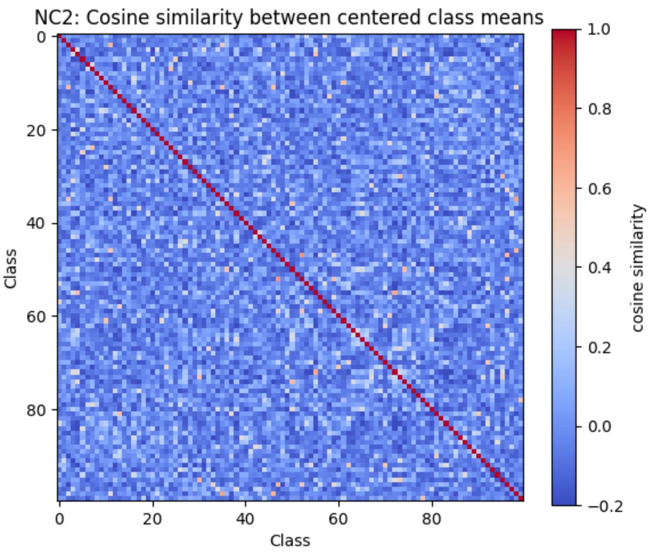


Figure 4: NC2: Cosine similarity matrix between centered class means. Diagonal entries are equal to 1, while off-diagonal entries concentrate around $-1/(C-1)$, consistent with a simplex ETF geometry.

NC3: alignment between classifier weights and class means. NC3 states that each classifier weight vector w_c becomes aligned with its corresponding class mean μ_c, i.e. $\cos(w_c, \mu_c) \to 1$. In our experiment, the cosine similarities have mean ≈ 0.94, with a small standard deviation

(≈ 0.016), and all values lie in $[0.89, 0.97]$. Thus, the learned classifier weights are strongly aligned with the feature class means, in line with NC3.

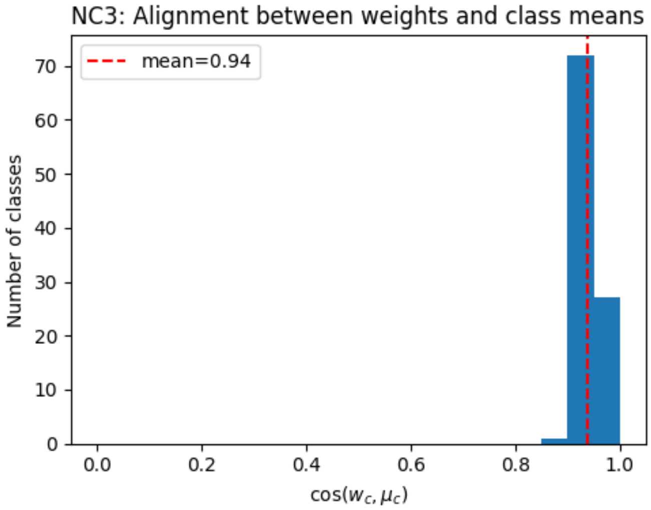


Figure 5: NC3: Alignment between classifier weights and class means. Histogram of cosine similarities $\cos(w_c, \mu_c)$, showing strong alignment with values concentrated close to 1.

NC4: simplification to nearest-class center classification. NC4 states that, at the end of training, the network decision rule becomes equivalent to a nearest-class center (NCC) classifier in the penultimate feature space: predictions are obtained by assigning each sample to the closest class mean μ_c. To evaluate this, we compare the standard linear classifier with an NCC rule using class means computed on the training set. On the test set, the network achieves 76.07% accuracy, while the NCC classifier reaches 75.56%, with a decision agreement of 93.81% between both rules. This strong agreement indicates that the learned classifier largely reduces to a nearest-class center decision, in line with the NC4 prediction.

NC5: ID/OOD orthogonality. NC5 extends Neural Collapse to out-of-distribution data and predicts that, as training progresses, the global mean of OOD features μ_G^{OOD} becomes increasingly orthogonal to the subspace spanned by the in-distribution class means. This can be quantified by the average cosine similarity $\langle \mu_c, \mu_G^{\text{OOD}} \rangle / (\|\mu_c\| \|\mu_G^{\text{OOD}}\|)$ over classes.

For SVHN, we obtain a mean cosine similarity of 0.52 (std 0.07), while for CIFAR-10 the mean is higher at 0.65 (std 0.06). These values indicate partial, but not complete, orthogonality, with stronger alignment for the near-OOD dataset CIFAR-10. This behavior is consistent with the intuition that OOD features lie increasingly outside the ID simplex subspace, while remaining closer to it for semantically similar distributions.

5 NECO: Neural Collapse based OOD detection

Method. NECO leverages the geometric structure induced by Neural Collapse. Under NC1 and NC2, in-distribution features concentrate around a low-dimensional simplex ETF subspace, while NC5 predicts that OOD features become increasingly orthogonal to this subspace. Given the penultimate representation $h(x)$, NECO measures the proportion of energy of $h(x)$ lying in the Neural Collapse subspace:

$$\text{NECO}(x) = \frac{\|Ph(x)\|}{\|h(x)\|},$$

where P is the projection onto the top principal components obtained by PCA on in-distribution training features. Following the original formulation, this geometric score is rescaled by the maximum logit in order to inject class-dependent confidence information.

Results and analysis. On SVHN, NECO achieves an AUROC of 0.892, an AUPR of 0.844, and an FPR@95 of 0.626, performing on par with the strongest logit-based methods. On the more challenging near-OOD setting CIFAR-10, NECO reaches an AUROC of 0.772 and an FPR@95 of 0.821, slightly improving over MSP while remaining comparable to Energy and MaxLogit. The score distributions in Figure 6 show that NECO induces a separation between ID and OOD samples for SVHN, while exhibiting bigger overlap for CIFAR-10, consistent with the partial ID/OOD orthogonality observed in NC5. Overall, these results confirm that NECO effectively exploits Neural Collapse geometry, but that near-OOD detection remains fundamentally difficult when OOD data lie close to the in-distribution feature subspace.

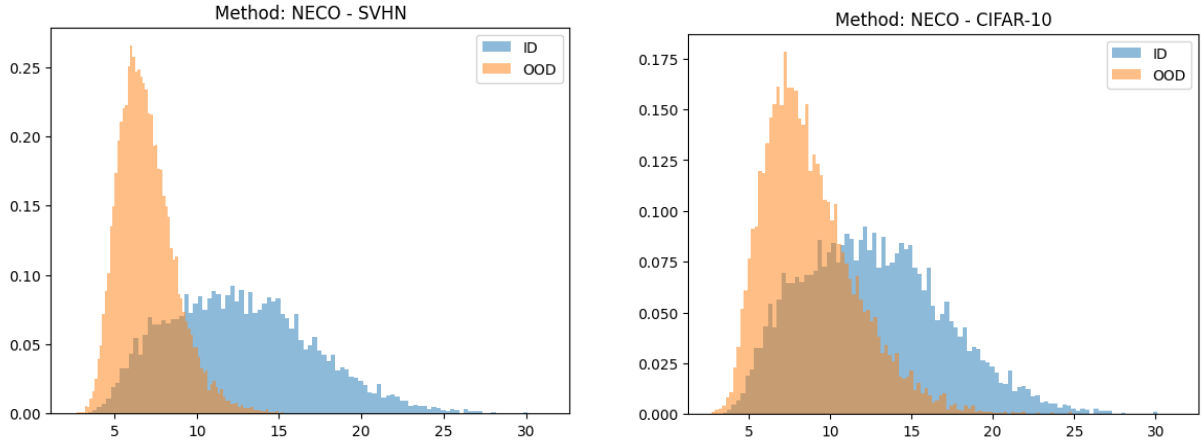


Figure 6: NECO score distributions for ID and OOD samples. Higher scores indicate in-distribution data.

Table 3: OOD detection performance summary (ID: CIFAR-100). Best results for each metric and dataset are shown in bold.

OOD Dataset	Method	AUROC ↑	AUPR ↑	FPR@95 ↓
SVHN	MSP	0.8413	0.7706	0.7180
	MaxLogit	0.8910	0.8418	0.6318
	Energy	**0.8952**	**0.8467**	**0.6174**
	Mahalanobis	0.7315	0.6480	0.9330
	ViM	0.8611	0.8151	0.7578
	NECO	0.8922	0.8438	0.6262
CIFAR-10	MSP	0.7644	**0.7921**	0.8235
	MaxLogit	0.7713	0.7907	0.8225
	Energy	0.7688	0.7891	0.8244
	Mahalanobis	0.6639	0.6626	0.9214
	ViM	0.7271	0.7413	0.8701
	NECO	**0.7715**	0.7908	**0.8212**