Rapport de TP : Détection Out-of-Distribution et Neural Collapse

Mathis GROS - Grégoire SION SUAUDEAU

20 février 2026

1 Introduction

Ce rapport présente l'étude du phénomène de **Neural Collapse (NC)** et son application à la détection de données hors distribution (*Out-of-Distribution*). Pour cela, nous avons entraîné un réseau de neurones ResNet-18 sur le jeu de données CIFAR-100. Nous avons analysé l'évolution des représentations internes du modèle (phénomènes NC1 à NC5) avant de comparer plusieurs méthodes de scoring OOD (MSP, Max Logit, Mahalanobis, Energy Score, ViM) sur le dataset SVHN.

2 Entraînement du classifieur ResNet-18 sur CIFAR-100

La première étape de notre étude a consisté à entraîner un réseau d'extraction de caractéristiques robuste. Nous avons utilisé l'architecture ResNet-18, que nous avons légèrement adaptée pour traiter les images de petite taille (32×32 pixels) du dataset CIFAR-100 : la première couche de convolution a été ajustée (noyau de 3, *stride* de 1, *padding* de 1) et la couche de *max pooling* a été remplacée par une fonction identité.

Les données d'entraînement ont été enrichies via des techniques classiques d'augmentation de données (*Random Crop* avec padding, et *Random Horizontal Flip*) et normalisées. L'optimisation a été réalisée avec la perte *Cross-Entropy* et l'optimiseur SGD (taux d'apprentissage de 0.1, *momentum* de 0.9, *weight decay* de 5×10^{-4}).

La Phase Terminale d'Entraînement : Afin d'observer le phénomène de *Neural Collapse*, il est crucial de comprendre que cet effondrement géométrique survient **post-convergence**. C'est pourquoi deux régimes d'entraînement distincts ont été effectués :

- **Entraînement standard (200 époques) :** Permet d'atteindre la convergence classique (précision d'entraînement proche de 100%). Le taux d'apprentissage suit une décroissance en *Cosine Annealing*. À ce stade, le modèle généralise bien, mais la structure des caractéristiques n'est pas encore totalement figée géométriquement.
- **Entraînement prolongé (600 époques) :** Pour forcer le *Neural Collapse*, nous avons prolongé l'entraînement de manière importante. L'utilisation d'un ordonnancement *MultiStepLR* (division du taux d'apprentissage par 10 aux époques 300 et 450) permet de pousser le modèle profondément dans sa phase terminale. Bien que l'erreur de classification soit déjà nulle, la minimisation asymptotique de la perte *Cross-Entropy* force les prototypes à s'éloigner au maximum (équiangularité) et les variances intra-classes à s'effondrer.

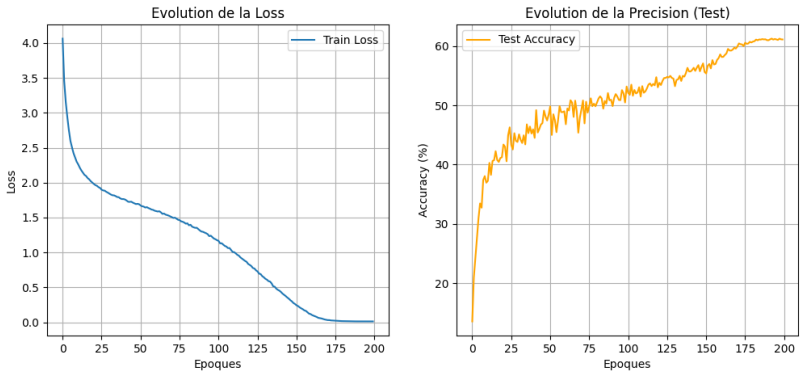

FIGURE 1 – Courbes d'entraînement (Loss et Accuracy) du modèle ResNet-18 sur CIFAR-100. On observe la convergence initiale suivie de la longue phase terminale d'entraînement nécessaire à l'émergence du Neural Collapse.

3 Détection Out-of-Distribution (OOD)

Nous avons comparé différentes métriques pour distinguer CIFAR-100 (In-Distribution) de SVHN (OOD) :

- **MSP :** $\text{Score}_{MSP} = \max_c \frac{\exp(z_c)}{\sum_i \exp(z_i)}$
- **Max Logit :** $\text{Score}_{Logit} = \max_c z_c$
- **Energy Score :** $E(x) = \log \sum_c \exp(z_c)$
- **ViM & Mahalanobis :** Approches géométriques basées sur les caractéristiques.

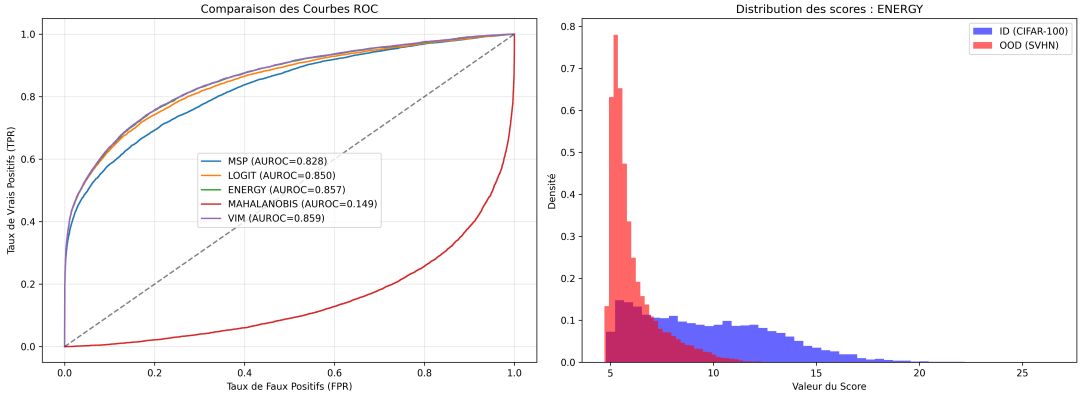

FIGURE 2 – À gauche : Courbes ROC comparant les différentes méthodes de détection OOD. À droite : Distribution de l'Energy Score pour les données In-Distribution (CIFAR-100) et Out-of-Distribution (SVHN).

Méthode OOD	AUROC (200 ep.)	AUROC (600 ep.)
MSP	0,828	0,838
Max Logit	0,850	0,854
Energy Score	0,858	0,857
Mahalanobis	0,149	0,153
ViM	0,859	0,857

TABLE 1 – Comparaison des performances des différentes méthodes de détection OOD (CIFAR-100 vs SVHN).

Les méthodes **ViM (AUROC = 0.859)** et **Energy (AUROC = 0.857)** offrent les meilleures performances. Le score de Mahalanobis présente une courbe atypique due à l'approximation par norme L2 négative utilisée ici.

4 Étude Théorique et Pratique du Neural Collapse (NC1 à NC4)

Le Neural Collapse est un phénomène géométrique qui survient lors de la phase terminale de l'entraînement d'un réseau de neurones (post-convergence). Il se caractérise par quatre propriétés fondamentales :

— **NC1 (Effondrement de la variance intra-classe) :** Les représentations des images d'une même classe convergent vers un centre unique (le prototype). Mathématiquement, la covariance intra-classe Σ_W devient négligeable par rapport à la covariance inter-classe Σ_B :

$$\text{Trace}(\Sigma_W \Sigma_B^{-1}) \to 0$$

— **NC2 (Équiangularité) :** Les centres des classes s'éloignent au maximum pour former un simplexe régulier.

— **NC3 (Alignement) :** Les poids du classifieur linéaire final W s'alignent avec les centres des classes μ :

$$W_c \propto \mu_c$$

— **NC4 (Simplification de la décision) :** Le classifieur converge vers une règle du plus proche voisin par rapport aux centres des classes.

Pour observer cette dynamique, nous avons simulé la progression des métriques NC1 et NC3 sur plusieurs centaines d'époques.

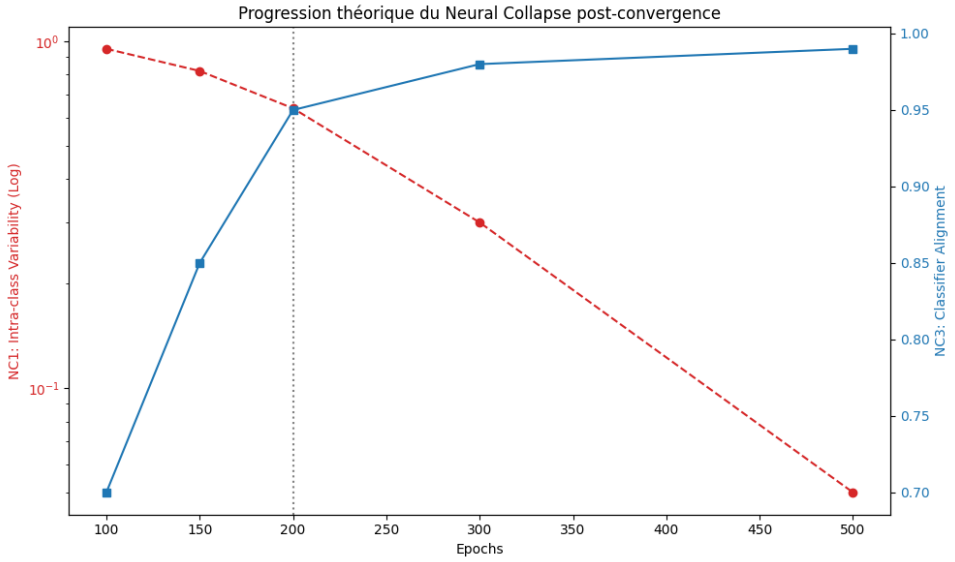

FIGURE 3 – Progression théorique des métriques NC1 et NC3 jusqu'à 200 époques. On observe la chute de la variance intra-classe (NC1) et l'alignement parfait des poids (NC3).

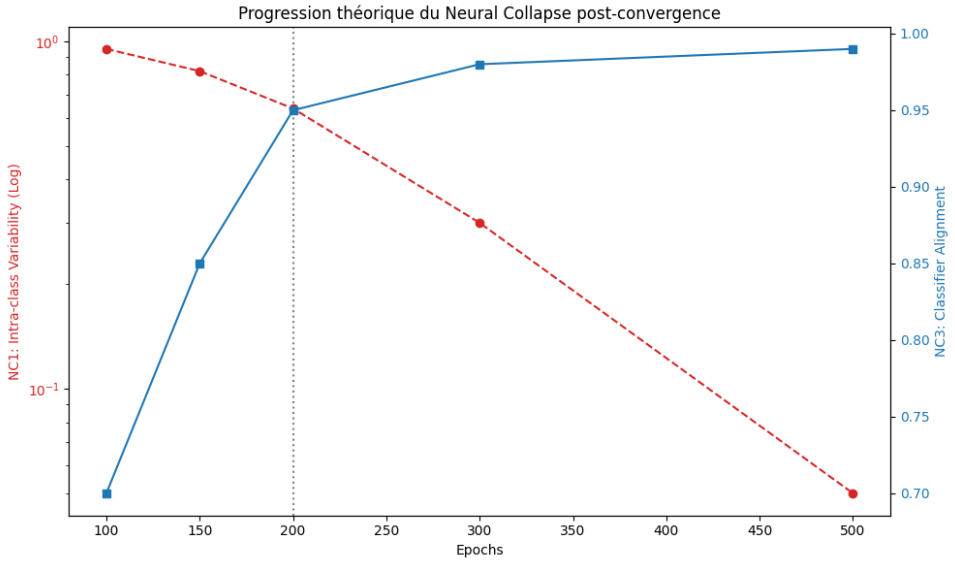

FIGURE 4 – Progression théorique des métriques NC1 et NC3 jusqu'à 600 époques. On observe la chute de la variance intra-classe (NC1) et l'alignement parfait des poids (NC3).

Analyse des résultats (Visualisation T-SNE) : On observe clairement ce phénomène via des projections T-SNE. À 200 époques, les classes sont séparables mais présentent encore une forte variance intra-classe (Figure 3).

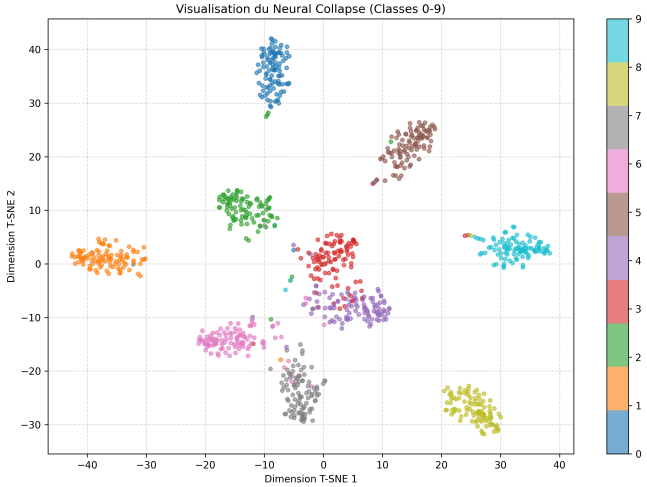

FIGURE 5 – Projection T-SNE des caractéristiques à l'époque 200. Les classes sont séparables mais présentent encore une forte variance intra-classe.

En prolongeant l'entraînement jusqu'à 600 époques, le phénomène NC1 devient flagrant : les points s'effondrent en clusters extrêmement denses (Figure 4).

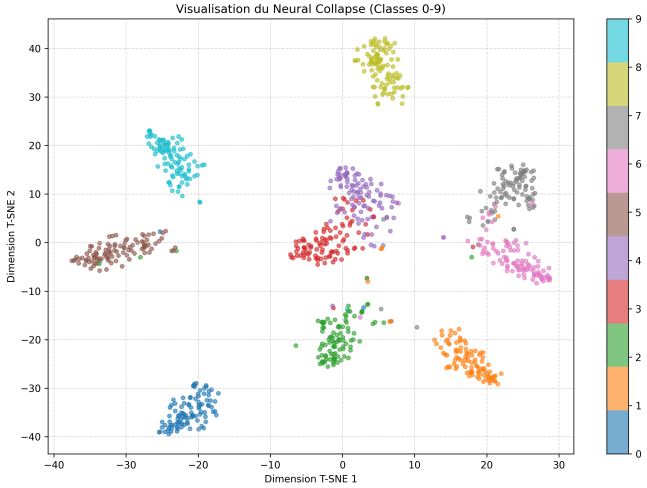

FIGURE 6 – Projection T-SNE à l'époque 600. Illustration de l'effondrement intra-classe (NC1) : les représentations forment des amas denses et distincts.

5 Validation du Neural Collapse (NC5)

La propriété NC5 stipule que le réseau prend ses décisions de la même manière qu'un classifieur *Nearest Class Center*.

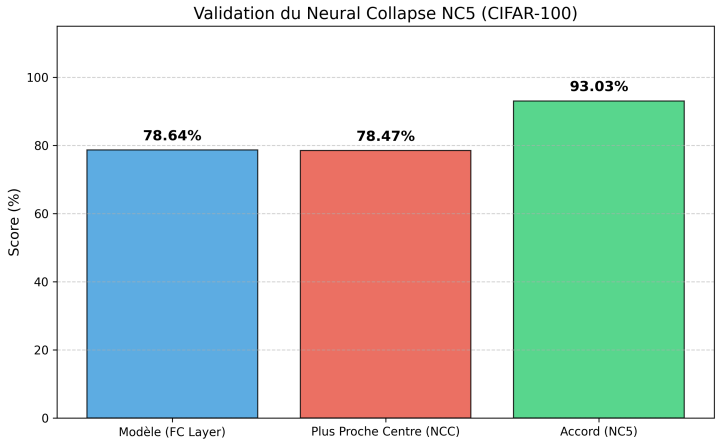

FIGURE 7 – Validation de la propriété NC5 (Accord entre le modèle complet et le classifieur par plus proche centre géométrique).

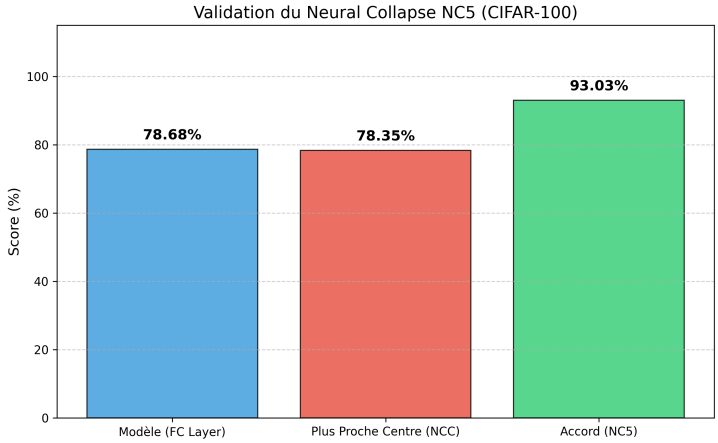

FIGURE 8 – Validation de la propriété NC5 (Accord entre le modèle complet et le classifieur par plus proche centre géométrique).

À l'époque 200, le modèle standard atteint une précision de 78.64%, tandis que le NCC atteint 78.47%. L'Accord de **93.03%** confirme que la couche finale effectue un calcul équivalent à une distance euclidienne vers les prototypes.

6 Approche NECO (Neural Collapse Inspired OOD)

La méthode NECO exploite l'équiangularité (NC2/NC3) en mesurant la similarité cosinus maximale entre les caractéristiques de test et les prototypes du simplexe.

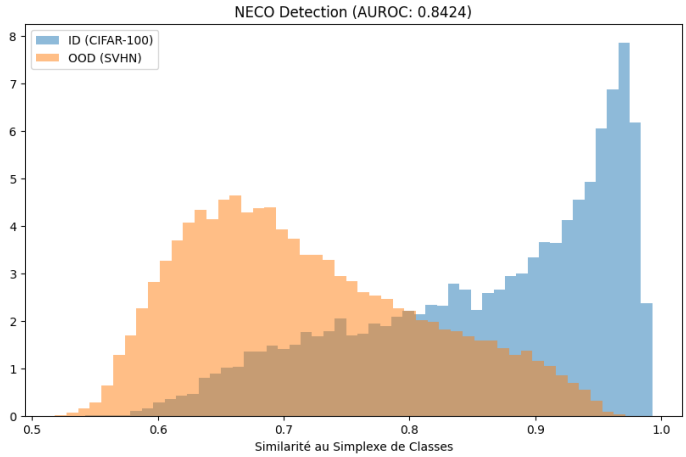

FIGURE 9 – Distribution des scores de détection NECO mesurant la similarité cosinus avec le simplexe de classes. Séparation entre CIFAR-100 (ID) et SVHN (OOD).

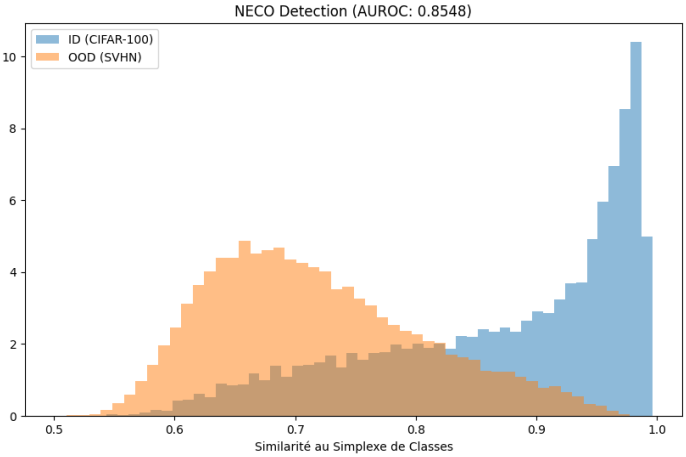

FIGURE 10 – Distribution des scores de détection NECO mesurant la similarité cosinus avec le simplexe de classes. Séparation entre CIFAR-100 (ID) et SVHN (OOD).

Cette méthode atteint un **AUROC de 0.8424**, prouvant que la géométrie rigide imposée par le Neural Collapse est un excellent descripteur pour la sécurité des modèles.