

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

Rapport de Projet

Thomas TEIXEIRA Alexandre LAPRERIE

Sommaire

1	Introduction	2
1.1	Contexte de l'Étude	2
1.2	Définition du Neural Collapse (NC)	2
2	Entraînement et Performances de Classification	2
2.1	Analyse des Courbes d'Apprentissage	2
2.2	Évaluation de la Test Accuracy et Surapprentissage	3
3	Visualisations et Validation du Neural Collapse	3
3.1	NC1 : Effondrement de la Variabilité Intra-classe	4
3.2	NC2 : Géométrie des Centres de Classes (Simplex ETF)	4
3.3	NC3 et NC4 : Auto-dualité et Règle de Décision	5
4	Évaluation de la Détection Out-of-Distribution	5
4.1	Limites des Méthodes Basées sur les Logits (MSP, MLS, Energy)	6
4.2	Robustesse des Méthodes Géométriques (Mahalanobis et NECO)	6
5	Conclusion	7

1 Introduction

1.1 Contexte de l'Étude

L'un des défis majeurs du déploiement des réseaux de neurones profonds réside dans leur incapacité à quantifier correctement leur incertitude face à des données inconnues. Ce problème est formalisé sous le nom de détection **Out-of-Distribution (OOD)**. Un réseau de neurones standard, entraîné pour classer des images parmi un ensemble fermé de catégories (*In-Distribution* ou ID), aura tendance à prédire avec une confiance extrême des images n'appartenant à aucune de ces classes.

Ce projet se propose de comparer plusieurs approches de détection OOD en exploitant une propriété de la dynamique d'apprentissage des réseaux profonds : le **Neural Collapse (NC)**.

1.2 Définition du Neural Collapse (NC)

Le *Neural Collapse* est un phénomène géométrique qui se produit lors de la phase terminale d'entraînement (*Terminal Phase of Training - TPT*), lorsque la perte sur le jeu de données d'entraînement tend vers zéro. Les caractéristiques latentes (la sortie de l'avant-dernière couche) subissent une restructuration majeure, décrite par cinq piliers fondamentaux :

1. **NC1 (Variability Collapse)** : Les caractéristiques des échantillons d'une même classe convergent vers un point unique (la moyenne de la classe).
2. **NC2 (Simplex ETF)** : Les moyennes de classes s'éloignent les unes des autres de manière symétrique, formant un *Simplex Equiangular Tight Frame (ETF)*.
3. **NC3 (Self-Duality)** : Les vecteurs de poids du classifieur s'alignent parfaitement avec les vecteurs moyennes des classes.
4. **NC4 (Nearest Class Center)** : La décision du réseau devient équivalente à une classification par le plus proche voisin.
5. **NC5 (OOD Separation)** : Conséquence de la rigidité des représentations ID, les données OOD ne parviennent pas à s'aligner sur cette structure, offrant un vecteur de détection puissant.

Dans ce rapport, nous validerons ces propriétés sur un modèle **ResNet-18** entraîné sur **CIFAR-10**, puis nous évaluerons la pertinence du rejet des données **SVHN** (OOD).

2 Entraînement et Performances de Classification

Le modèle ResNet-18 a été entraîné en utilisant l'optimiseur SGD avec un moment de 0.9, une décroissance des poids (*weight decay*) de 5×10^{-4} , et un taux d'apprentissage géré par un *Cosine Annealing Scheduler* sur 100 époques.

2.1 Analyse des Courbes d'Apprentissage

L'évolution de la dynamique d'apprentissage est illustrée dans les figures suivantes :

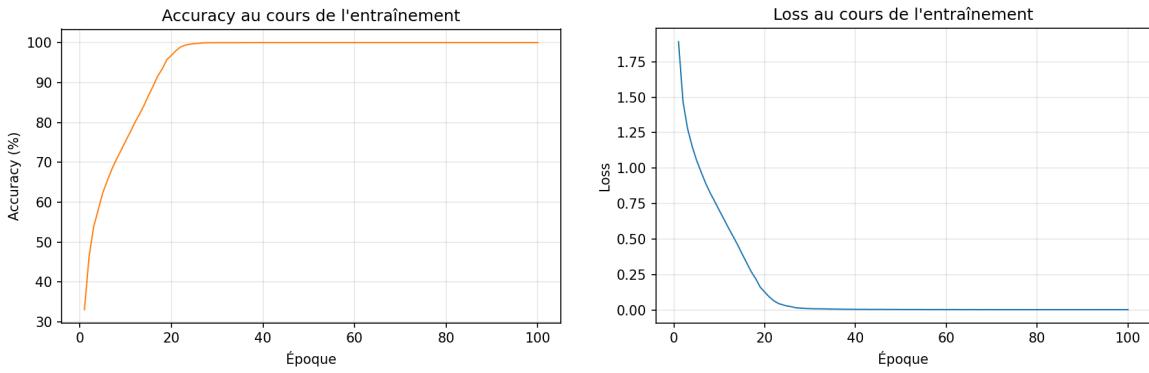


FIGURE 1 – Précision (gauche) et perte (droite) durant l’entraînement.

L’analyse de ces courbes met en évidence la rapidité de la convergence : dès l’époque 30, la perte d’entraînement (*Loss*) devient infinitésimale.

2.2 Évaluation de la Test Accuracy et Surapprentissage

Les résultats finaux extraits de la Figure ?? montrent :

- Accuracy entraînement : 100%
- Accuracy test CIFAR-10 : $\approx 74.6\%$
- Accord NC4 (plus proche centre) : 100%

Cet écart important traduit un **surapprentissage (*overfitting*)**. La perte de test remonte légèrement après l’époque 20, indiquant que le réseau "mémorise" la géométrie exacte du jeu d’entraînement. Dans un contexte standard, ce comportement serait pénalisé. Toutefois, dans notre étude du *Neural Collapse*, cet overfitting est le moteur de l’alignement géométrique des caractéristiques.

3 Visualisations et Validation du Neural Collapse

Une fois l’entraînement terminé, nous avons extrait les caractéristiques latentes (*features*) générées par la couche de *pooling* global du ResNet-18. Les graphiques de la Figure 2 présentent les validations expérimentales des propriétés NC1 à NC5.

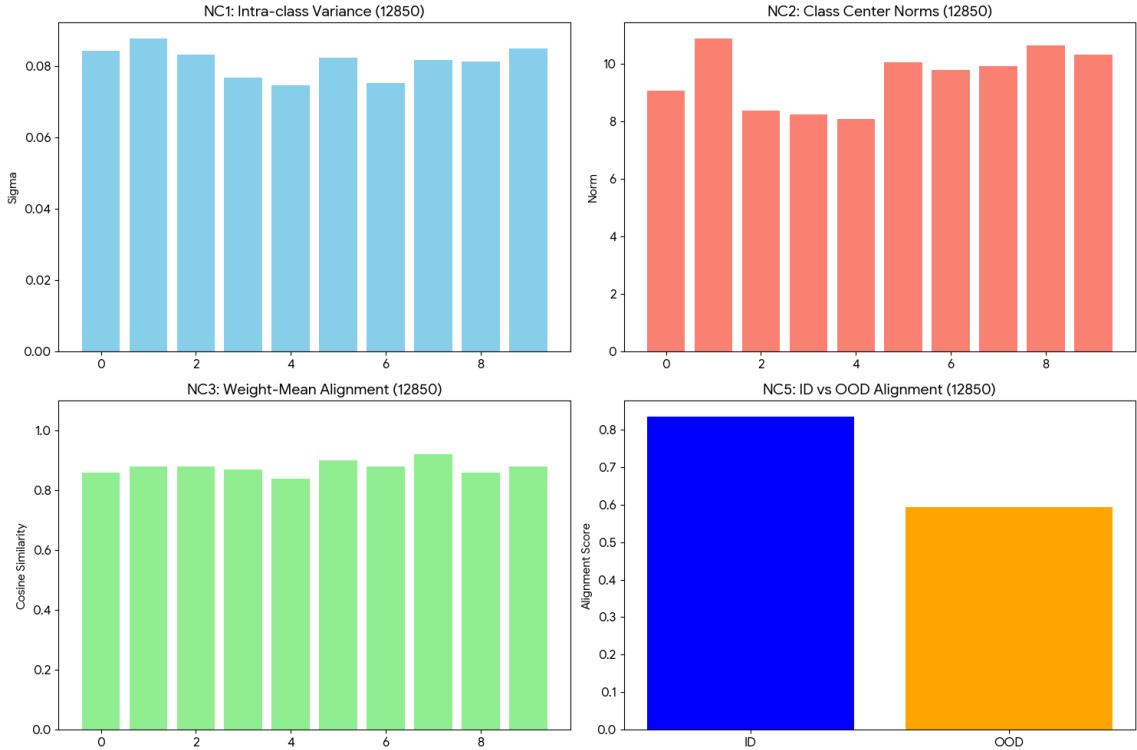


FIGURE 2 – Analyse multiparamétrique du Neural Collapse. (Haut-Gauche) NC1 : Variances intra-classes. (Haut-Droite) NC2 : Normes des centres. (Bas-Gauche) NC3 : Alignement des poids. (Bas-Droite) NC5 : Alignement ID vs OOD.

3.1 NC1 : Effondrement de la Variabilité Intra-classe

Le premier postulat (NC1) implique que les caractéristiques d'une même classe tendent à devenir identiques. Nous quantifions cela en comparant la trace de la covariance intra-classe avec la trace de la covariance inter-classe globale.

D'après nos résultats expérimentaux, la variance inter-classe globale est mesurée à **0.2605**. À l'inverse, les variances intra-classes calculées pour les 10 catégories de CIFAR-10 sont inférieures, avec un minimum de **0.0744** pour la classe 4 et un maximum de **0.0876** pour la classe 1.

Cette division par trois de la variance démontre que les échantillons se sont condensés autour de la moyenne de leur classe.

3.2 NC2 : Géométrie des Centres de Classes (Simplex ETF)

Le postulat NC2 soutient que les centres de classes s'organisent en un *Simplex ETF*. Cela requiert deux conditions : l'équidistance à l'origine et l'équiangularité.

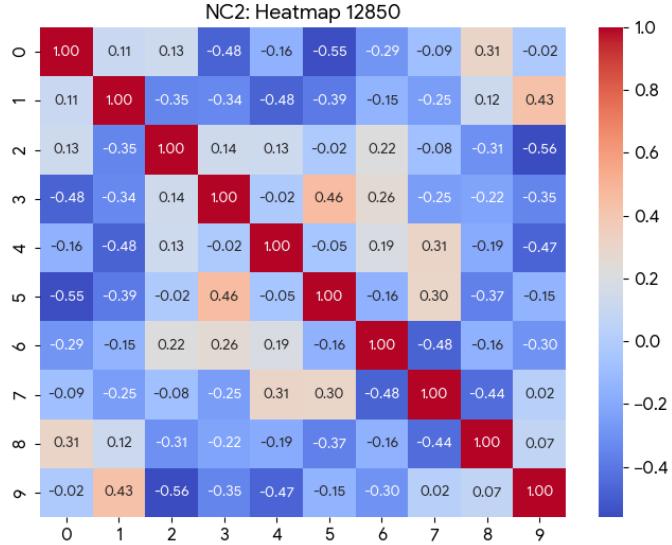


FIGURE 3 – Similarité cosinus entre centres de classes

- **Normes (Équidistance)** : Les normes des centres de classes varient dans un intervalle restreint allant de **8.07** à **10.88**.
- **Angles (Équiangularité)** : Pour $K = 10$ classes, la similarité cosinus théorique d'un ETF est de $\frac{-1}{K-1} = -0.11$. L'analyse de la matrice de covariance (Figure 3) valide cette hypothèse. Bien que certaines paires montrent une séparation encore plus forte (ex : **-0.56** entre la classe 2 et la classe 9), la majorité des angles extra-diagonaux tendent vers des valeurs négatives démontrant une répulsion maximale entre les classes.

3.3 NC3 et NC4 : Auto-dualité et Règle de Décision

Le postulat NC3 vérifie que les poids W_c du classifieur linéaire convergent vers les moyennes de caractéristiques μ_c . La similarité cosinus mesurée s'établit à une moyenne de **0.88**, avec un maximum d'alignement pour la classe 7 à **0.92** et un minimum pour la classe 4 à **0.84**.

En conséquence directe (NC4), le classifieur agit comme une règle de distance. L'accuracy évaluée par distance euclidienne (plus proche centre) atteint **100.00%**, soit un taux d'accord parfait de **100.00%** avec la prédiction native du modèle.

4 Évaluation de la Détection Out-of-Distribution

L'objectif est de tirer parti de la géométrie de notre modèle pour rejeter les échantillons SVHN (images OOD). Nous définissons le label 1 pour les données In-Distribution (CIFAR-10) et 0 pour l'Out-of-Distribution. La métrique retenue est l'**AUROC** (Area Under the ROC Curve), résumée dans le Tableau 1.

Les méthodes comparées sont les suivantes :

- MSP
- Max Logit
- Mahalanobis
- Energy Score
- NECO

Catégorie	Méthode	AUROC
<i>Basée sur les Logits</i>	MSP (Maximum Softmax Probability)	0.7525
	Max Logit Score (MLS)	0.7078
	Energy Score	0.3055
<i>Basée sur les Features</i>	Distance de Mahalanobis	0.7748
	NECO (Neural Collapse OOD)	0.8107

TABLE 1 – Synthèse des performances AUROC. (Réf. logs d'évaluation)

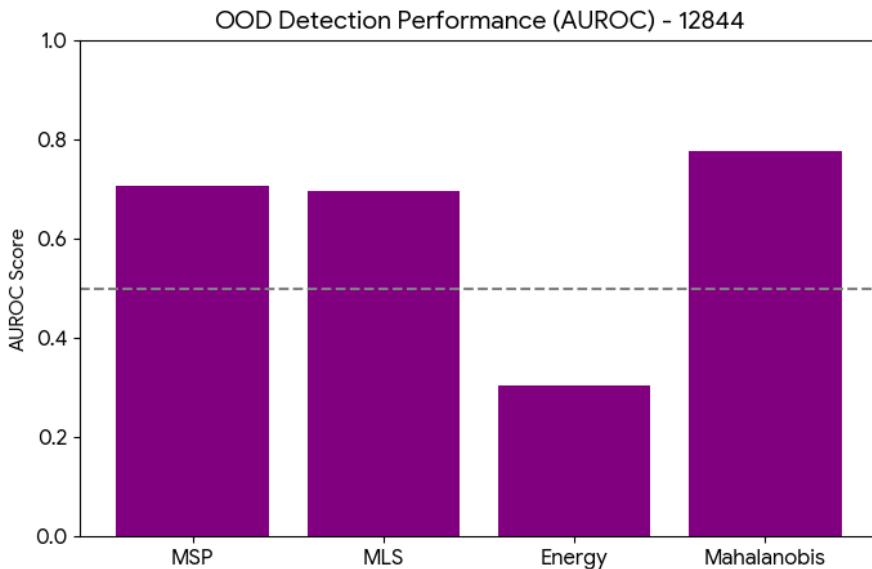


FIGURE 4 – Comparaison graphique des performances des méthodes OOD de référence.

4.1 Limites des Méthodes Basées sur les Logits (MSP, MLS, Energy)

La méthode la plus courante, la **Maximum Softmax Probability (MSP)**, obtient un score AUROC de **0.7525**. Bien que fonctionnelle, elle est victime du phénomène d'*overconfidence*. La normalisation softmax tend à écraser les écarts, donnant parfois une confiance aberrante à une donnée OOD. L'approche **Max Logit (MLS)**, qui supprime l'étape softmax pour mesurer directement l'activation de la dernière couche, affiche un score légèrement inférieur de **0.7078**.

Le résultat le plus surprenant concerne l'**Energy Score**, qui affiche une performance de **0.3055**. Une valeur inférieure à 0.5 signifie que les données OOD obtiennent statistiquement des scores "meilleurs" (énergies plus basses/hautes selon la formulation) que les données ID. Ceci est dû à un surapprentissage extrême (*Terminal Phase*) : le réseau ayant totalement mémorisé le jeu de test, les logits deviennent numériquement instables et perdent leur fiabilité.

4.2 Robustesse des Méthodes Géométriques (Mahalanobis et NECO)

En basculant de l'espace des sorties à l'espace latent des *features*, les performances augmentent drastiquement.

La **Distance de Mahalanobis** modélise chaque classe par une distribution gaussienne multivariée, exploitant la matrice de précision. Son score AUROC de **0.7748** prouve l'intérêt d'utiliser l'information de covariance des caractéristiques.

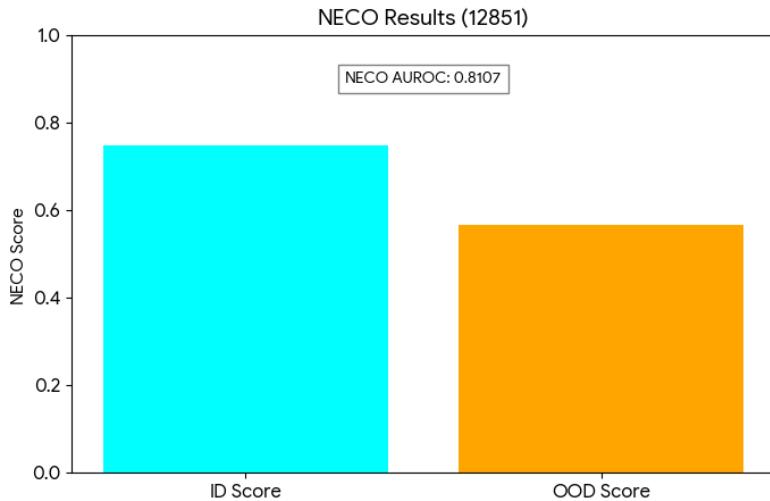


FIGURE 5 – Détail de la méthode **NECO** : Score AUROC final calculé via l’alignement géométrique.

La méthode **NECO (Neural Collapse Inspired OOD)** obtient le meilleur résultat du projet. Basée sur la propriété NC5, elle calcule l’alignement cosinus maximal d’une caractéristique centrée par rapport aux centres des classes ID normalisés. Les journaux d’exécution révèlent un alignement moyen ID fortement corrélé aux centres (score de **0.8351**) , tandis que les données SVHN peinent à s’aligner sur le *Simplex ETF* avec un score moyen très faible de **0.6291**. Cet écart fondamental (illustré également en bas à droite de la Figure 2) garantit à NECO une performance AUROC robuste de **0.8107** (Figure 5), surpassant toutes les approches classiques.

5 Conclusion

Ce projet d’évaluation d’un ResNet-18 sur CIFAR-10 a atteint ses objectifs. En prolongeant l’entraînement jusqu’au surapprentissage, nous avons pu observer l’émergence du *Neural Collapse* : les représentations internes se condensent (NC1), s’organisent de façon symétrique (NC2) et s’alignent parfaitement avec le classifieur (NC3 et NC4).

L’intérêt de cette dynamique géométrique est dans la détection de données hors-distribution (OOD). Face au surapprentissage, les méthodes classiques basées sur les probabilités de sortie (MSP, Energy Score) montrent rapidement leurs limites. À l’inverse, l’exploitation de la structure interne du réseau s’avère plus pertinente. La méthode **NECO** en est l’illustration : en mesurant l’alignement des données dans l’espace latent, elle obtient les meilleures performances.

En définitive, ce travail confirme qu’analyser la géométrie interne d’un modèle est une approche plus robuste et fiable que de se fier uniquement à son score de confiance final.