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

Theory of Deep Learning (MVA Program + ENSTA)

Objectives

In this practical work, you will:

1. Train a ResNet-18 classifier on CIFAR-100 using PyTorch.
2. Implement and compare the following OOD scores:
   - Max Softmax Probability (MSP)
   - Maximum Logit Score
   - Mahalanobis
   - Energy Score
   - ViM
3. Study the **Neural Collapse** phenomenon at the end of training NC 1 to NC4. (Please explain them a bit)
4. Study the **Neural Collapse** phenomenon at the end of training NC5.
5. Implement the **NECO** method (Neural Collapse Inspired OOD Detection).
6. (**Bonus**) Analyse Neural Collapse on earlier layers.
7. (**Bonus (IMPOSSIBLE)**) Link it to mechanistic Interpretability.

This TP is designed to be practical and exploratory. The goal is to understand how OOD detection metrics behave and how neural collapse can be leveraged for OOD analysis.

Deliverables

- Training curves and test accuracy.
- Comparison of OOD for:
  - MSP
  - Max Logit
  - Mahalanobis
  - Energy Score
  - Energy Score
  - NECO
- Visualizations of Neural Collapse:
  - class mean distances
  - within-class variance
  - cosine similarity of classifier weights vs means
- Bonus: Neural Collapse across layers.

Good luck, and have fun exploring OOD detection!