Article choisi: EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

DÉPOTS :

- Git :
  - Code source
  - Readme
  - Notebook (demo execution des experiences de l'article)
- Rapport 6 à 8 pages en ANGLAIS
- Vidéo 5 min de présentation du projet

BAREME :

- Rapport /40
- Code /30
- Soutenance /30

Problème étudié	Les réseaux de neurones se trompent sur des entrées légèrement perturbées.
Ce qu'on pensait avant	Que c'était à cause de leur non-linéarité ou de l'overfitting.
Nouvelle explication	En fait, c'est leur **nature trop linéaire** (à grande échelle) qui cause ces erreurs.
Conséquence	On peut générer facilement des exemples adversariaux en exploitant cette linéarité.
Application utile	En utilisant ces exemples pour l'entraînement, on améliore la robustesse du modèle.

Pas compliqué, ressources internet

Objectifs : Produire des exemples qui amènent à une mauvaise classification

➔ Gros focus sur **évaluation et résultats**

2 équipes :

- Réimplémenter l'article
  - Générer des exemples adversarios
- Se concentrer sur l'évaluation
  - Entraîner le modèle
  - Récup un modèle et voir comment il marche

2sem :

- récup base de donnée et présenter, cmbn de classes, ce qu'on retient, résultats qualitatifs (brouillage qui marche et qui marche pas)
- récup ttes infos sur internet à ce sujet

Ds 1 mois :

- majorités des entrainements commencés/finis
- Majorité de l'architecture faites

On peut utiliser code ou IA génératives mais à indiquer

Pas de courbe époque = f(loss) on s'en balec (jsp ce que ça veut dire mais oké), reproduire les résultats de l'article

Pour aller plus loin : Regarder comment les modèles se comportent avec d'autres bases de données